



HAL
open science

Exploitation de Connaissances Formelles dans des Processus Décisionnels Automatisés et Apprentissage de Modèles Prédicatifs

Sébastien Harispe

► **To cite this version:**

Sébastien Harispe. Exploitation de Connaissances Formelles dans des Processus Décisionnels Automatisés et Apprentissage de Modèles Prédicatifs. Intelligence artificielle [cs.AI]. Université de Montpellier, 2024. tel-04947879

HAL Id: tel-04947879

<https://hal.science/tel-04947879v1>

Submitted on 14 Feb 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MEMOIRE POUR OBTENIR L'HABILITATION à DIRIGER des RECHERCHES DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale n°166 Information, Structures, Systèmes - I2S

Unité de recherche Euromov Digital Health in Motion

Exploitation de Connaissances Formelles dans des Processus Décisionnels Automatisés et Apprentissage de Modèles Prédicatifs

Présenté par Sébastien HARISPE
le 5 juin 2024

Devant le jury composé de

Mme Sylvie DESPRES, Professeur émérite, LIMICS, Université Sorbonne Paris Nord
Mme Catherine FARON, Professeur des universités, I3S, Université Côte d'Azur
M. Christophe MARSALA, Professeur des universités, LIP6, Sorbonne Université
M. Eric GAUSSIER, Professeur des universités, LIG, Université Grenoble Alpes
M. Guillaume GRAVIER, Directeur de recherche, IRISA, CNRS
M. Pascal PONCELET, Professeur des universités, LIRMM Université de Montpellier
M. Jacky MONTMAIN, Professeur de l'IMT, Euromov DHM, IMT Mines Alès

Présidente du jury
Rapporteur
Rapporteur
Rapporteur
Examineur
Examineur



UNIVERSITÉ
DE MONTPELLIER

« Je déclare avoir respecté, dans la conception et la rédaction de ce mémoire d'HDR, les valeurs et principes d'intégrité scientifique destinés à garantir le caractère honnête et scientifiquement rigoureux de tout travail de recherche, visés à l'article L.211-2 du Code de la recherche et énoncés par la Charte nationale de déontologie des métiers de la recherche et la Charte d'intégrité scientifique de l'Université de Montpellier. Je m'engage à les promouvoir dans le cadre de mes activités futures d'encadrement de recherche. » source : École doctorale « Information, Structures et Systèmes » (I2S, ED 166).

A mes très chers parents

Préambule

Ce manuscrit expose une synthèse de notre activité de recherche en vue de prétendre au diplôme d'Habilitation à Diriger des Recherches en Informatique délivré par l'Université de Montpellier.¹

La première partie du manuscrit - non partagée dans la version publique - se présente sous la forme d'un curriculum vitae détaillé qui fait état du parcours recherche présenté. Cette partie évoque par ailleurs différentes activités en lien étroit avec la recherche que nous menons depuis plusieurs années au sein d'IMT Mines Alès - pour l'auteur en qualité de Maître-assistant de classe normale de l'Institut Mines Télécom.

La seconde partie du manuscrit débute à proprement dit le développement de l'exposé de nos activités de chercheur, et de recherche. Un propos introductif nous permet de discuter des éléments de contexte institutionnels et disciplinaires, et de détailler quelques responsabilités en lien avec la recherche seulement évoquées dans le curriculum vitae détaillé. Cette partie permet en particulier de préciser certaines de nos activités menées en qualité de chercheur à la fois membre et au service d'un collectif, e.g. implications et responsabilités dans l'animation de la recherche.

Le corps du manuscrit expose par la suite une sélection de travaux de recherche en Informatique, et plus particulièrement en Intelligence Artificielle. Ceux-ci ont pour l'essentiel été réalisés avec différents collègues chercheurs et étudiants-chercheurs depuis l'obtention par l'auteur du titre de docteur (2014). Après un développement sur la notion de mesure sémantique - qui sous-tend plusieurs de nos travaux -, nous structurons notre exposé au travers des axes de recherche suivants :

- **L'exploitation de connaissances formelles dans des processus décisionnels automatisés.** Nous discutons comment des représentations de connaissances de type ontologies légères, peuvent servir la production de résumés sémantiques, l'évaluation de la véracité de sources de données, ou encore contribuer à la mise en place de modèles prédictifs lorsqu'une connaissance métier est disponible *a priori*.
- **L'apprentissage de modèles prédictifs à partir de données.** Nous présentons nos travaux qui portent pour l'essentiel sur l'étude de techniques d'apprentissage automatique pour la résolution de tâches spécifiques, e.g. apprentissage supervisé appliqué à des tâches de traitement automatique du langage naturel, ou apprentissage par renforcement appliqué à la résolution de problèmes d'optimisation combinatoire. Des travaux disciplinaires, détachés d'un applicatif spécifique, seront aussi exposés dans une section dédiée, e.g. optimisation du mécanisme d'attention populaire dans les réseaux de neurones, interprétabilité de prédicteurs.

Les contributions apportées dans ces différents axes ont été appliquées à des domaines multiples : analyse psychosensorielle, domaine juridique, domaines des sciences du mouvement humain, du sport et de la santé. . . Bien que nous insistions sur le contexte appliqué de certaines contributions, nous tâcherons aussi, tant que possible, de détacher le caractère générique de nos propositions.

Le mémoire se termine par la présentation de notre programme de recherche futur axé sur (1) la poursuite de l'étude du **transfert de connaissances de l'Homme vers la Machine** dans la continuité des travaux déjà réalisés, mais aussi sur (2) l'étude du **transfert de connaissances de la Machine vers l'Homme** dans le contexte de l'apprentissage machine.

1. Nous faisons le choix de préférer l'utilisation du *nous* dit de modestie car l'essentiel du travail que nous exposerons résulte selon nous d'une entreprise collective. Nous ferons aussi dans certains cas référence à l'auteur pour préciser notre propos ou évoquer une prise de position discutable qu'il souhaite assumer seul.

Table des matières

Préambule	i
I Curriculum Vitae détaillé	3
II Mémoire	7
1 Introduction	13
1.1 Contexte professionnel	13
1.2 Synthèse	26
2 Synthèse des contributions recherche	27
2.1 Étude des mesures sémantiques	28
2.2 Exploitation de connaissances formelles dans des processus d'aide à la décision	44
2.3 Apprentissage de modèles prédictifs à partir de données	85
3 Projet de recherche	123
3.1 Propos généraux	123
3.2 Apprentissage Machine appliqué	125
3.3 Transfert de connaissances de l'Homme vers la Machine	128
3.4 Transfert de connaissances de la Machine vers l'Homme	130
Annexes	135
A Sélection d'articles	135
B Notes	227
Bibliographie	230

Première partie

Curriculum Vitae détaillé

Non partagé dans cette version publique (environ 25 pages accessibles sur demande).

Deuxième partie

Mémoire

Résumé

Ce manuscrit expose une synthèse de notre activité de recherche en vue de prétendre au diplôme d'Habilitation à Diriger des Recherches en Informatique délivré par l'Université de Montpellier.

Nous présentons dans un premier temps dans la version examinée, les contextes institutionnel et disciplinaire dans lesquels la recherche que nous menons depuis plusieurs années au sein d'IMT Mines Alès - pour l'auteur en qualité de Maître-assistant de classe normale de l'Institut Mines Télécom - est réalisée. Cette introduction est exclue de la version rendue publique afin de concentrer le document sur les aspects scientifiques du manuscrit.

Le corps du manuscrit expose par la suite une sélection de travaux de recherche en Informatique, et plus particulièrement en Intelligence Artificielle. Ceux-ci ont pour l'essentiel été réalisés avec différents collègues chercheurs et étudiants-chercheurs depuis l'obtention par l'auteur du titre de docteur (2014). Après un développement sur la notion de mesure sémantique - qui sous-tend plusieurs de nos travaux -, nous structurerons notre exposé au travers des axes de recherche suivants :

- **L'exploitation de connaissances formelles dans des processus décisionnels automatisés.** Nous discutons notamment comment des représentations de connaissances de type ontologies légères, peuvent servir la production de résumés sémantiques, l'évaluation de la véracité de sources de données, ou encore contribuer à la mise en place de modèles prédictifs lorsqu'une connaissance métier est disponible *a priori*.
- **L'apprentissage de modèles prédictifs à partir de données.** Nous présentons nos travaux qui portent pour l'essentiel sur l'étude de techniques d'apprentissage automatique pour la résolution de tâches spécifiques, e.g. apprentissage supervisé appliqué à des tâches de traitement automatique du langage naturel, ou apprentissage par renforcement appliqué à la résolution de problèmes d'optimisation combinatoire. Des travaux disciplinaires, détachés d'un applicatif spécifique, seront aussi exposés dans une section dédiée, e.g. optimisation du mécanisme d'attention populaire dans les réseaux de neurones, interprétabilité de prédicteurs.

Les contributions apportées dans ces différents axes ont été appliquées à des domaines multiples : analyse psychosensorielle, domaine juridique, domaines des sciences du mouvement humain, du sport et de la santé. . . Bien que nous insistions sur le contexte appliqué de certaines contributions, nous tâcherons aussi tant que possible de détacher le caractère générique de nos propositions.

Le mémoire se termine par la présentation de notre programme de recherche futur axé sur (1) la poursuite de l'étude du **transfert de connaissances de l'Homme vers la Machine** dans la continuité des travaux déjà réalisés, mais aussi sur (2) l'étude du **transfert de connaissances de la Machine vers l'Homme** dans le contexte de l'apprentissage machine.

Summary

This manuscript presents a summary of our research activity with a view to applying for the French Habilitation à Diriger des Recherches (accreditation to supervise research) in Computer Science awarded by the University of Montpellier.

We begin by presenting the institutional and disciplinary contexts in which the research that we have been carrying out for several years at IMT Mines Alès - for the author as an associate professor of the Institut Mines Télécom.

The body of the manuscript is then devoted to the presentation of a selection of research work in Computer Science, and more particularly in Artificial Intelligence. Most of this work has been carried out with various research colleagues and research students since the author obtained his PhD in Computer Science (2014). After a discussion of the notion of semantic measures - a subject that underpins several of the works we have carried out - we structure the presentation of our work through the following lines of research :

- **Exploiting formal knowledge in decision-making processes** : The use of formal knowledge in automated decision-making processes. We discuss how knowledge representations, such as lightweight ontologies, can be used to produce semantic summaries, assess the veracity of data sources, or contribute to the implementation of predictive models when business knowledge is available a priori.
- **Learning predictive models from data**. This work mainly concerns the study of machine learning techniques for solving specific tasks, e.g. supervised learning applied to automatic natural language processing tasks, or reinforcement learning applied to solving combinatorial optimization problems. Disciplinary work on this theme, detached from a specific application, will also be presented in a dedicated section, e.g. optimization of the attention mechanism popular in numerous neural networks, interpretability of predictors.

The contributions made in these different areas have been applied to a wide range of fields : psychosensory analysis, the legal domain, the sciences of human movement, sport and health, etc. Although we emphasize the applied context of certain contributions, we will also try, as far as possible, to detach the generic nature of our proposals.

The dissertation concludes with a presentation of our future research programme, focusing on (1) the study of **knowledge transfer from Human to Machine** in the continuity of work already carried out, but also on (2) the study of **knowledge transfer from Machine to Human** in the context of machine learning.

Table des matières

1	Introduction	13
1.1	Contexte professionnel	13
1.1.1	Contexte institutionnel	14
1.1.2	Contexte disciplinaire	16
1.2	Synthèse	26
2	Synthèse des contributions recherche	27
2.1	Étude des mesures sémantiques	28
2.1.1	La notion de mesure sémantique	29
2.1.2	Synthèse des travaux de doctorat	34
2.1.3	Synthèse des travaux post-doctoraux sur les mesures sémantiques	35
2.1.4	Regard critique sur nos travaux et évolution du domaine	39
2.2	Exploitation de connaissances formelles dans des processus d'aide à la décision	44
2.2.1	Résumé d'annotations terminologiques et conceptuelles	45
2.2.2	Estimation de la qualité de sources d'information et vérification d'assertions	59
2.2.3	Enrichissement de modèles prédictifs et d'inférence en TALN	67
2.2.4	Découverte de règles d'association à partir de données imparfaites	74
2.3	Apprentissage de modèles prédictifs à partir de données	85
2.3.1	Contributions au Traitement Automatique du Langage Naturel	85
2.3.2	Contributions à l'étude de problèmes d'optimisation combinatoire	102
2.3.3	Contributions disciplinaires en Apprentissage Machine	109
3	Projet de recherche	123
3.1	Propos généraux	123
3.2	Apprentissage Machine appliqué	125
3.3	Transfert de connaissances de l'Homme vers la Machine	128
3.4	Transfert de connaissances de la Machine vers l'Homme	130
	Annexes	135
A	Sélection d'articles	135

Table des matières

B Notes 227

Bibliographie 230

Introduction

Sommaire

1.1	Contexte professionnel	13
1.1.1	Contexte institutionnel	14
1.1.2	Contexte disciplinaire	16
1.2	Synthèse	26

1.1 Contexte professionnel

Le travail du chercheur est largement dépendant du contexte institutionnel dans lequel il évolue, e.g. missions annexes à ses recherches qui lui sont confiées, mais aussi de l'évolution des champs disciplinaires qui l'intéressent. Nous vous proposons dans cette section deux présentations succinctes des contextes institutionnel et disciplinaire qui permettront à la fois de situer les travaux de recherche que nous exposerons et de comprendre le parcours recherche de l'auteur. Nous nous permettons en préambule de section une synthèse de son parcours personnel de l'enseignement supérieur à son poste actuel d'enseignant-chercheur en Informatique. Cette note plus personnelle permettra d'apprécier des éléments de contexte complémentaires à ceux qui seront évoqués par la suite.

1.1.1 Contexte institutionnel

Précision sur le statut de Maître-assistant

Le statut de Maître-assistant est adopté dans différents établissements d'enseignement supérieur sous tutelle d'un ministère autre que le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI). Les corps des Maîtres-assistants ont été intégrés dans le nouveau corps des Maîtres de Conférences en 1984. L'auteur appartient au corps des Maîtres-assistants de l'Institut Mines-Télécom et relève des ministres chargés de l'Industrie et des Communications Électroniques. Nous assurons statutairement des missions d'enseignement et de recherche classiques chez les enseignants-chercheurs (sans obligation de service, e.g., pas d'obligation statutaire d'enseignement). Nous participons aussi à une mission de valorisation des travaux de la recherche auprès des acteurs industriels et économiques français. ¹

Nous réalisons dans ce contexte des activités pour l'essentiel en lien avec une recherche appliquée. En effet, la mission de valorisation des travaux de la recherche qui nous est confiée a des incidences importantes sur le positionnement de notre recherche, les partenariats que nous établissons, les guichets de financement que nous sollicitons, mais aussi les niveaux de TRL (*Technology Readiness Levels*) de nos travaux. La recherche de l'école est par ailleurs labellisée par l'Institut Carnot qui reconnaît les recherches orientées vers les besoins des entreprises et de la société.

Organisation interne de la recherche à IMT Mines Alès

IMT Mines Alès mène des activités de recherche en Informatique depuis les années 90 et a connu plusieurs réorganisations de ses centres de recherche au fil des ans. Nos activités de recherche ont été menées dans les structures détaillées ci-après :

- **LGI2P**, Laboratoire de Génie informatique et d'Ingénierie de production, dirigé par Yannick Vimont ² à mon arrivée en 2011, puis par Jacky Montmain à partir de 2017. Le LGI2P a pour l'essentiel existé sur le site nîmois d'IMT Mines Alès fermé avant notre rapatriement sur Alès en 2017. Il regroupait une vingtaine d'ingénieurs et d'enseignants-chercheurs permanents structurés en deux équipes de recherche parmi lesquelles l'équipe KID - *Knowledge representation and Image analysis for Decision* ³ - dirigée par Jacky Montmain, dans laquelle l'auteur a effectué sa thèse puis ses premières années de recherche post-doctorales. L'équipe KID se concentrait sur la notion d'*automatisation cognitive* en vue de proposer des outils et techniques pour améliorer les capacités cognitives de l'opérateur humain en situations décisionnelles complexes. Trois axes de recherche étaient distingués : la représentation et l'analyse de connaissances, l'analyse d'images, et l'aide à la décision. L'auteur a évolué dans cette entité de 2011, début de nos travaux de thèse, à 2021, date officielle de fin du laboratoire LGI2P. ⁴ Les travaux sur l'apprentissage machine, très minoritaires à la date d'arrivée de l'auteur, se sont naturellement largement développés au fil des ans.

1. Dans le texte, les enseignants des écoles des mines oeuvrent à l'accomplissement des missions des écoles des mines et participent à « *Une mission de valorisation des résultats de la recherche, au titre de laquelle ils participent au développement des relations avec le monde industriel et économique, notamment en matière de recherche partenariale, et contribuent ainsi à l'amélioration de la compétitivité des entreprises.* » (source décret n°2007-468 [70]).

2. A ce jour directeur de la recherche aux Mines de Paris.

3. <https://kidknowledge.wp.imt.fr/>.

4. Le dernier rapport d'évaluation de l'Hcéres est disponible à l'adresse suivante : lien.

- **CERIS / I3A.** L'organisation de la recherche et de l'enseignement d'IMT Mines Alès est aujourd'hui structurée en interne par des Centres d'Enseignement et de Recherche. L'auteur appartient de fait au Centre d'Enseignement et de Recherche en Informatique et Système (CERIS) dirigé par Jacky Montmain. Ce centre, qui regroupe une cinquantaine d'ingénieurs et chercheurs, distingue deux équipes dont l'équipe Informatique, Image et Intelligence Artificielle (I3A) dirigée par Gérard Dray dans laquelle nous évoluons. A noter que le CERIS et son équipe I3A n'ont pas pour ambition de porter un projet scientifique au sens de l'évaluation Hcéres. Les membres d'I3A inscrivent leurs travaux de recherche dans l'Unité de Recherche Euromov Digital Health in Motion détaillée ci-après.

L'Unité de Recherche Euromov Digital Health in Motion

Euromov *Digital Health in Motion*⁵ est une unité de recherche créée en janvier 2021 entre IMT Mines Alès, l'Université de Montpellier, et les CHU de Montpellier et de Nîmes. Dirigée par Stéphane Perrey (Université de Montpellier) et codirigée par Jacky Montmain (IMT Mines Alès), l'unité regroupe une cinquantaine d'ingénieurs et d'enseignants-chercheurs permanents. Les recherches de l'unité, qui se veulent interdisciplinaires, se positionnent à l'interface entre l'Informatique, les Sciences du Mouvement Humain et le domaine de la Santé. Le manifeste du projet de l'unité précise la volonté d'amener une fertilisation croisée entre ces différents champs d'étude.

L'objectif sur le volet Sciences du Mouvement est d'améliorer notre compréhension de la plasticité comportementale chez l'Homme afin d'améliorer sa performance sensorimotrice et d'envisager de nouvelles approches thérapeutiques. Sur les aspects cliniques, nombre de nos travaux ont trait à l'amélioration de la récupération sensorimotrice ou à la réadaptation à la suite d'accidents traumatiques, e.g. AVC, ou dans des conditions cliniques de type lombalgies chroniques. Plusieurs travaux sont aussi menés dans le domaine sportif, e.g. prédiction de blessures ou du niveau de forme d'athlètes élités.

Les Sciences du Mouvement et de la Santé reposent très largement sur l'analyse de données hétérogènes et sur la prise en compte de connaissances métier, médicales en particulier. Elles constituent donc un domaine d'étude idéal pour tout chercheur en Sciences des données et en IA, entre autres intéressé par le développement de modèles d'IA hybrides couplant des techniques issues de l'IA symbolique et de l'Apprentissage Automatique. Les collaborations menées au sein d'Euromov DHM vont cependant au-delà de cet aspect *bac à sable de choix* dans lequel nous éprouvons et perfectionnons les techniques de l'IA. Nous envisageons une recherche interdisciplinaire en vue d'amener une fertilisation croisée entre les différents champs disciplinaires. Nous postulons notamment que la compréhension des mécanismes d'apprentissage et de contrôle moteur chez l'être humain peut être une source d'inspiration pour la formulation d'approches novatrices en IA, e.g. en lien avec les mécanismes d'apprentissages adaptatif, par renforcement, par imitation. . .

L'organisation scientifique de l'unité de recherche est structurée autour de trois thèmes scientifiques (équivalent à des équipes de recherche) :

- *Monitoring & Improving Behaviors* (MIB) qui regroupe les travaux en lien avec l'étude de comportements humains mais aussi logiciels dans des environnements complexes, e.g. médicaux ou liés au sport.
- *Perception in Action & Synchronisation* qui fédère les travaux sur la perception et la synchronisation dans des contextes de mouvements humains.

5. «La Santé Numérique en Mouvement» : <https://dhm.euromov.eu/>.

- *Learning & Complexity* (LAC) qui regroupe les travaux menés sur les notions d'Apprentissage et de Complexité qui sous-tendent toute forme d'agent intelligent (humain ou machine).

L'auteur dirige avec deux collègues le thème LAC depuis sa création ; celui-ci sera par conséquent détaillé en section ??.

Deux axes transversaux enrichissent aussi cette structuration en thème : (i) l'axe SemTaxM qui porte sur l'étude de la sémantique et de la taxonomie du mouvement, et (ii) l'axe Factory qui vise à structurer les initiatives portées par les membres de l'unité en lien avec les développements méthodologiques et logiciels relatifs entre autres à l'acquisition et au traitement de données.

Sur les aspects disciplinaires de l'Informatique, les travaux de recherche menés au sein de l'unité Euromov DHM portent aujourd'hui en grande partie sur l'étude de techniques d'analyse adaptées au traitement de données physiologiques - e.g., électroencéphalographie (EEG), électromyographie (EMG), électroencéphalographie (ECG) - et de données liées au mouvement humain, e.g. données de systèmes Mocap (*Motion capture*). Les techniques à base d'IA, et notamment celles à base d'apprentissage machine, ont une place toute particulière dans les travaux de l'unité. Des exemples de contributions seront proposés dans ce manuscrit.

Département d'enseignement Informatique et IA (2IA)

IMT Mines Alès dispose de 6 départements d'enseignement qui structurent son offre de formation dans les spécialités proposées aux formations d'ingénieurs généralistes et par apprentissage. Les activités d'enseignement que nous menons, dont certaines sont naturellement en lien avec nos activités de chercheur, se concentrent pour l'essentiel dans le département Informatique et Intelligence Artificielle (2IA) dirigé par Sylvie Ranwez. Ce département est structuré en deux options :

- IAIL : Intelligence Artificielle et Ingénierie Logicielle (resp. Thomas Lambolais).
- IASD : Intelligence Artificielle et Sciences des Données (resp. Sébastien Harispe).

Nous détaillerons par la suite nos activités d'enseignement et de responsable de l'option IASD.

1.1.2 Contexte disciplinaire

L'Informatique, source de révolutions technologiques L'Informatique a d'ores et déjà transformé l'Humanité. Ce champ de recherche permet aujourd'hui à l'Homme d'être augmenté chaque jour un peu plus, véritablement enrichi de capacités décuplées : celle de pouvoir discuter de vive voix en visioconférence avec son semblable à l'autre bout du monde, celle de pouvoir accéder en un simple clic à des quantités pharamineuses d'informations, celle de pouvoir traduire dans presque toutes les langues, celle de pouvoir croiser des données hétérogènes, massives et complexes pour aider à la formulation de nouveaux médicaments, explorer notre Univers, comprendre notre incidence sur la planète à l'heure des défis écologiques qui nous sont proposés. . . L'Informatique, qui est par essence à la fois au service et nourrie de nombreux domaines de recherche et appliqués, ne cessera très probablement plus jamais de transformer l'Humanité.

Ces quelques lignes sont rédigées par un Informaticien qui a nécessairement un regard biaisé sur sa discipline. Par souci d'objectivité nous devons aussi souligner que l'Informatique est très souvent synonyme d'automatisation, de supervision, de productivité, de substitution de l'opérateur humain. Certains diront aussi - à juste titre - que ce champ de recherche donne aussi à l'Homme la capacité de coder sa perte en nous faisant croire que l'instantané doit s'établir règle, en faisant en sorte que nous rapprocher nous éloigne et que le virtuel et le paraître deviennent parfois plus importants que le réel.

Il contribue aussi largement à ce que l'information devienne plus que jamais source potentielle de manipulation à des échelles sans précédent, et à ce que les systèmes sociaux-économiques, médicaux et industriels qui sous-tendent nos sociétés se retrouvent à la merci d'éventuelles cyberattaques déstabilisatrices. . . Pour finir et recentrer la discussion sur le thème de l'IA abordé par la suite, certains diront aussi que, aujourd'hui, l'Informatique interroge plus que jamais en nous proposant d'implémenter de la manière des plus stupides l'Intelligence qui nous mènera à notre perte.

Nous considérerons quant à nous que les transformations bénéfiques que peut proposer l'Informatique sont tout aussi profondes que les craintes légitimes que nous venons d'exprimer, qui nous interrogent et accompagnent par ailleurs dans la pratique de notre recherche au quotidien.

La naissance d'une ère perpétuelle de la donnée et de son exploitation. Les développements et la démocratisation des technologies de type capteurs (e.g. smartphones), ou plus généralement d'acquisition (e.g. réseaux sociaux) et de stockage de données offrent des perspectives nouvelles et multiples à de nombreux champs disciplinaires. Difficile effectivement aujourd'hui de distinguer des disciplines qui ne sont pas bousculées par les technologies de numérisation ouvrant la porte à de nouvelles analyses, plus riches, plus larges. Nous vivons aujourd'hui grâce à (ou à cause de) cela dans une ère résolument caractérisée par la « Donnée ». Tantôt substrat de nouvelles ruptures, tantôt caractérisée comme toxique ou source de manipulation [49], la donnée, ou plutôt les données sont aujourd'hui au carrefour d'enjeux stratégiques multiples et complexes, e.g. économiques, géopolitiques. Multifformes, de plus en plus souvent massives, hétérogènes dans leur nature, de qualité variable, ces données représentent un véritable *or gris*. Elles forment en effet très souvent la pierre angulaire de nos systèmes informatiques qu'ils soient à base d'Intelligence Artificielle ou nourris par des techniques issues des Sciences des données. Ces données sont par ailleurs intimement liées aux questions relatives à la protection de la personne (e.g. RGPD), à la confiance que l'on peut leur accorder et à la cybersécurité pour ne citer que quelques exemples de sujets importants sur cette thématique.

Ces données, analysées et étudiées de manière contextuelle, permettent l'émergence de connaissances. Les Sciences des données et l'Intelligence Artificielle proposent par exemple différentes techniques qui permettent d'analyser des données variées et notamment d'extraire des tendances, des motifs caractéristiques, ou même des modèles utiles dans différents contextes, e.g. aide à la décision. Ces techniques permettent ainsi plus généralement d'implémenter l'extraction et la découverte (semi-)automatisées de connaissances ; elles font l'objet de nombre de nos travaux de recherche.

Comme nous le verrons par la suite, la recherche en Informatique que nous avons menée ces dernières années est étroitement liée à ces notions de données et de connaissances. Elle a initialement été axée sur l'utilisation de représentations des connaissances - qui rendent la connaissance exploitable formellement par l'outil informatique - pour des traitements variés, e.g. résumé d'information, recherche d'information. Une progression s'est opérée par la suite vers des problématiques d'apprentissage machine et plus spécifiquement d'apprentissage profond. Cette évolution thématique de la résolution de problèmes à l'aide d'objets issus de l'IA symbolique (e.g. ontologies), vers une approche à base d'apprentissage machine gouvernée par les données, n'est pas décorrélée des développements vécus ces dernières années par l'Informatique en tant que champ disciplinaire. Cette évolution n'est pas détachée de l'évolution de la place de l'IA en Informatique, puis de la position particulière prise par l'apprentissage machine dans le domaine de l'IA.

Cette section détaille quelques éléments de contexte qui permettront une meilleure lecture de la progression de notre positionnement recherche. Nous tâcherons en particulier d'introduire ce que nous entendons par IA, et de souligner le rôle particulier que joue aujourd'hui l'IA et l'apprentissage machine en Informatique (et plus largement de manière transversale dans tout domaine disposant de données). Nous évoquerons aussi plusieurs aspects qui expliquent cette situation, ainsi que les transformations importantes que cela a amenées, plus spécifiquement dans le contexte de notre recherche pour l'essentiel à finalité appliquée.

Intelligence Artificielle : éléments de définition L'IA fait référence à la fois à une technologie et à l'objet d'étude d'un domaine de recherche interdisciplinaire du même nom étroitement lié à l'Informatique. Différentes appréciations de la notion d'IA étant retrouvées dans la littérature, nous précisons ce que nous entendons par IA dans ce manuscrit.

Nous considérerons l'IA comme un domaine de recherche de l'Informatique centré sur l'étude d'un ensemble de techniques qui visent à adresser des problèmes complexes qui nécessitent l'expression de capacités cognitives pour être traités de manière satisfaisante par un opérateur humain. A ce titre, L'IA en tant que domaine de recherche disciplinaire regroupe un spectre large de travaux sur la représentation des connaissances et le raisonnement, l'apprentissage machine, le traitement du langage naturel, la vision par ordinateur, la planification. . .

La vision de l'IA que nous adoptons souscrit naturellement à celle exposée par le programme initial de l'étude de l'IA (Conférence de Dartmouth de 1956) [1]. Ce programme visait le développement de machines intelligentes capables entre autres de résoudre des problèmes à l'époque (et encore aujourd'hui pour la plupart) réservés à l'humain, i.e. dans le texte '*The study [of Artificial Intelligence] is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves*' [1]. Soulignons que la formulation de ce programme implique une définition évolutive de l'IA : la référence aux « problèmes aujourd'hui réservés à l'humain » conditionne même temporellement, dans sa genèse, ce qui intéresse l'IA. ⁶

Nos travaux reposent très souvent sur l'utilisation de représentations des connaissances et l'exploitation des paradigmes de l'apprentissage machine, avec des applicatifs multiples. Ils s'inscrivent donc, de fait, naturellement dans le domaine de l'IA. Il convient cependant de préciser que nos travaux n'aborderont pas l'étude de l'IA dans l'objectif de développer un agent capable d'exprimer une intelligence d'ordre général, au sens d'une intelligence artificielle générale (AGI, *Artificial General Intelligence*) : un agent superhumain qui serait autonome, capable même selon certains de développer une forme de conscience. ⁷ Certains travaux dissocient l'étude de l'IA faible, cantonnée à la résolution d'un problème précis, de l'étude de l'IA forte, à portée plus générale, e.g. AGI. Le périmètre dans lequel s'inscriront nos travaux est clairement celui de l'IA faible. Notons tout de même que les techniques fondamentales employées dans les deux cas ne diffèrent pas nécessairement

6. En effet les sujets qui intéressent l'IA évolueront au fil du temps ; à titre d'exemple, la recherche de chemins dans un graphe était même considérée comme un problème central de l'IA à certaines périodes, e.g. algorithmes A*.

7. L'AGI est l'objectif avoué de nombreux acteurs du domaine de l'IA, e.g. OpenAI, DeepMind, Anthropic.

(e.g. utilisation de réseaux de neurones); les travaux sur l'IA faible font cependant plus directement référence à des problématiques liées au traitement et à l'analyse de données dans le contexte d'une tâche précise. C'est en effet le positionnement que nous considérons dans tous les travaux que nous menons et qui seront exposés dans ce manuscrit.

Une période faste pour l'IA et l'Apprentissage Machine L'informatique a depuis toujours été un domaine générateur de ruptures qui ont sensiblement marqué nos sociétés (e.g., Internet, Web, ordinateurs personnels, smartphones, réseaux sociaux). L'IA est aujourd'hui annoncée comme la prochaine rupture. Les résultats qu'elle permet d'ores et déjà d'obtenir dans de nombreux domaines ne cessent par ailleurs d'asseoir la crédibilité de l'hypothèse, e.g. développements récents liés aux modèles génératifs utilisés pour la définition d'agents conversationnels de type ChatGPT [203], ou pour la génération d'images, e.g. StableDiffusion [230].

Au-delà de toute spéculation, l'omniprésence de la numérisation dans nos sociétés fait que la très grande majorité des champs disciplinaires et domaines sont aujourd'hui concernés par des problématiques liées au traitement et à l'analyse de données - et donc aux développements visés par l'IA faible. Les acteurs industriels et économiques avec qui nous collaborons sont naturellement aux premières loges des transformations amenées par cette numérisation. Nos missions institutionnelles (Recherche, Enseignement, Transfert vers l'Industrie) nous invitent donc naturellement à nous intéresser aux techniques adaptées aux traitements de données ; les techniques de l'IA et de l'apprentissage machine sont dans ce contexte des techniques de choix.

L'engouement dont bénéficie aujourd'hui l'IA est tout à fait particulier. Il s'explique en grande partie par l'essor des techniques d'apprentissage machine et en particulier des techniques d'apprentissage profond à partir de 2012.⁸ Les développements récents et rapides en apprentissage profond ont en effet graduellement permis l'obtention de gains de performance importants dans l'automatisation de nombreuses tâches complexes liées entre autres au langage et à la vision [166]. Les approches à base d'apprentissage profond ont bousculé nos communautés en permettant d'aborder la résolution de problèmes que l'on ne se pensait pas capable d'adresser si rapidement, et en déclassant littéralement les approches jusque-là établies dans l'état de l'art dans certains contextes. Il est par ailleurs intéressant de noter que ces améliorations ont touché de larges pans de la recherche en termes de types de données ou de domaines d'application. On retrouve ainsi des améliorations notables en classification d'images, de textes, de sons, ou encore dans la résolution de jeux (Atari, jeu de Go. . .) avec des applications dans des champs pour le moins variés de notre société.

Les développements permis par l'IA sont tels que nous commençons par ailleurs à nous habituer à la mise en place de systèmes à base d'IA capables d'exprimer des performances suprahumaines pour des tâches que l'on estimait difficiles - d'un point de vue automatisation - il y a encore quelques années. La notion même de *tâche difficile* n'a jamais été aussi instable. Les développements récents en termes de modèles génératifs d'images ou de textes arrivent même très souvent aujourd'hui à surprendre les chercheurs, plus ou moins jeunes, pourtant spécialistes du domaine de l'IA voire de l'apprentissage machine [197]. Le contexte est tel que des chercheurs du domaine en appellent même aujourd'hui à une pause des expériences massives dédiées au développement d'IA dans la crainte d'un développement incontrôlé de l'AGI [211].

8. Nous considérerons, comme c'est communément le cas, que l'essor de l'apprentissage profond a eu lieu à partir des années 2012, suite à la victoire du réseau AlexNet [149] à la compétition de classification d'images Imagenet [71].

Au-delà des performances parfois surprenantes des systèmes à base d'IA, le fort engouement pour le domaine s'explique aussi par la rapidité du déploiement des technologies de l'IA auprès du grand public et de leur forte adoption, e.g. ChatGPT est l'application grand public connaissant la croissance la plus rapide de l'histoire [47]. La Figure 1 illustre l'engouement marqué pour le domaine de l'IA à travers l'évolution de l'usage du terme *Artificial Intelligence* ces dernières années. Bien que le terme IA soit aujourd'hui dans toutes les bouches, sa popularité a largement fluctué depuis les années 60 (cf. succession de phases estivales et hivernales de l'IA au cours de son histoire). Le terme n'était que peu utilisé en 2011 dans le laboratoire LGI2P dans lequel l'auteur évoluait, qui traitait pourtant d'automatisation cognitive.⁹ L'IA regroupe en effet des sujets couverts dans de nombreux domaines, e.g. Informatique, Neurosciences, Mathématiques (logique, apprentissage statistique. . .), Robotique. Des connexions étroites peuvent par ailleurs être établies avec d'autres domaines qui ont su jouir d'une certaine popularité en Informatique par le passé, e.g. cybernétique, systèmes complexes, théorie de l'information, théorie des automates, reconnaissance des formes, aide à la décision. Beaucoup de chercheurs en Informatique s'intéressaient il y a quelques années à des techniques que l'on inscrit aujourd'hui dans l'IA sans que ce thème de recherche n'ait pour autant été systématiquement affiché. En ce qui concerne nos travaux, le choix d'une stratégie de communication très axée sur l'IA ou qui l'omettrait, n'aurait au final que peu d'incidence sur le fond car les recherches que nous menons sont orientées sur la résolution de problèmes (IA faible) ; les fondamentaux techniques que nous manipulons ne s'en verraient aucunement changés.

L'Apprentissage Machine, un domaine de recherche rayonnant et phagocytant L'IA ne s'est pas graduellement imposée seulement au grand public ces dernières années. Nos laboratoires de recherche français ainsi que nos acteurs industriels et économiques ont aussi été largement impactés par cette nouvelle *vague*, voire déferlante, de l'IA. L'impact ne s'est pas simplement traduit en termes d'affichage. Nombre d'équipes ont en effet largement modifié leur positionnement recherche pour y intégrer des composantes faisant référence à l'apprentissage machine. Ainsi, en une dizaine d'années, l'apprentissage machine et en particulier l'apprentissage profond, se sont graduellement et naturellement imposés à de nombreux laboratoires d'Informatique. Cette transition s'est opérée non pas du fait d'une injonction dictée par une mode supportée seulement par une forme de *hype* technologique ou des enjeux économiques - comme vu parfois au temps du *Big Data* -, mais plutôt par un enthousiasme fort des chercheurs pour ces systèmes à base d'apprentissage machine qui ont su, au fil des ans, asseoir des performances écrasantes dans de nombreux domaines.

Le fort degré de pénétration de l'Apprentissage Machine et de l'apprentissage profond dans de nombreux domaines s'explique aussi par la généricité de leurs paradigmes. En effet, comme nous l'avons souligné, un grand nombre de laboratoires en Informatique intègrent dans leurs équipes des chercheurs qui s'intéressent à des composantes liées à l'analyse de données au sens large, quelle que soit la modalité étudiée (i.e. type de données) : son, texte, image. . . Il était commun il y a encore quelques années d'entendre les experts de l'étude de chacune des spécialités de l'analyse de données exposer des techniques d'analyse propres à la modalité étudiée, e.g. Okapi BM25 pour la recherche d'information, N-gram pour l'étude de textes et la construction de modèles de langue, techniques d'analyse à base de filtres pour l'étude d'images. . . Le LGI2P répondait naturellement à ce canon. La situation a nettement changé aujourd'hui car nombre de problématiques traitées auparavant par des approches *ad hoc* peuvent aujourd'hui être abordées au travers de l'apprentissage machine - ce

9. Le rapport de l'Hcéres de l'IRIT - Institut de Recherche en Informatique de Toulouse, acteur national majeur en IA et central dans l'Institut 3IA Aniti -, ne contient qu'une unique mention à l'IA lors de la campagne d'évaluation 2014-2015 (au détour d'une mention d'un laboratoire dont il est issu : CERFIA, Cybernétique des Entreprises, Reconnaissances des Formes et IA) [82]; le rapport d'évaluation 2019-2020 mentionne quant à lui le terme 6 fois [83].

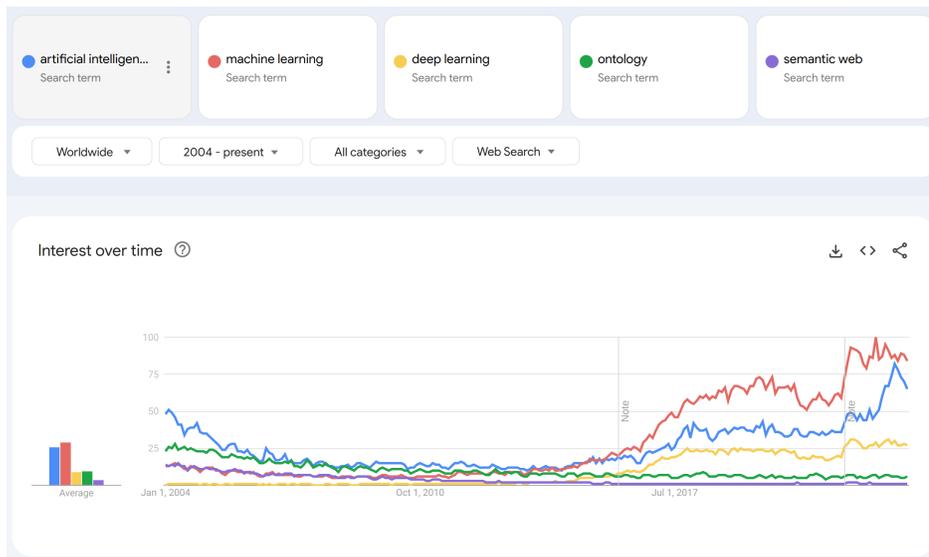


FIGURE 1 Aperçu de l'intérêt exprimé pour les différents sujets en lien avec nos thématiques de recherche. Source Google Trends <https://trends.google.com/>. On observe une forte augmentation de l'intérêt porté à l'IA, et plus spécifiquement à l'apprentissage machine.

qui ne décline pas cependant nécessairement les approches traditionnelles dans tous les contextes. Le constat est tel qu'il serait toutefois difficile aujourd'hui ne serait-ce que d'imaginer une équipe de recherche intéressée par le traitement de données qui n'intègre pas de composante liée, de près ou de loin, à l'Apprentissage Machine. Une forme de crainte légitime face à un risque de déclin des approches traditionnelles non basées sur de l'apprentissage machine explique aussi probablement la transition qui s'opère depuis plusieurs années. Cette transition ne s'effectuera pas par ailleurs dans certains cas sans un questionnement regrettable de la pertinence de certains programmes de recherche qui n'intègrent pas une forte composante en apprentissage machine.

Un accès sans précédent à la connaissance formelle Nous avons jusqu'à présent centré une grande partie de notre propos sur les développements importants, et rendus populaires ces dernières années, qui sont attribués à l'Apprentissage Automatique. Il convient cependant de préciser, s'il est nécessaire de le souligner à nouveau, que l'IA en tant que discipline ne se résume bien entendu pas à l'Apprentissage Automatique. Beaucoup de développements importants proposés par d'autres champs d'étude de l'IA mériteraient d'être soulignés. Compte tenu du fait que nous ferons dans nos travaux un large usage de modèles de connaissances, nous nous concentrerons par la suite sur quelques développements remarquables proposés par des domaines directement intéressés par la Représentation et l'Ingénierie des Connaissances

Nous soulignons plus spécifiquement les nombreux travaux proposés pour la définition de formalismes capables de représenter des connaissances complexes et de les engager dans des procédures de raisonnement, notamment dans des cadres déductifs et décidables. Les ontologies, en particulier basées sur des logiques de description plus ou moins expressives, permettent ainsi d'exprimer

formellement de la connaissance qui pourra être partagée de manière non ambiguë, et être traitée automatiquement selon un cadre précis et totalement maîtrisé [111]. Elles permettent ainsi de définir précisément des bases de connaissances décrivant des concepts, leurs relations (e.g. taxonomique), mais aussi des instances de ces concepts et leurs relations.

Les travaux amenés par la communauté du Web Sémantique ont aussi permis la définition de langages standard, aujourd'hui largement adoptés pour représenter des connaissances et les rendre accessibles sur le Web, e.g., RDF(S), OWL, SPARQL [124, 122]. Ces langages facilitent très largement le partage et l'utilisation de bases de connaissances. Le paradigme des données liées permet de plus une interconnexion de ces bases, parfois dans une vision ouverte. Point important, nous soulignons aussi qu'un grand nombre d'outils, souvent proposés et supportés par la communauté, permettent de faciliter le développement d'ontologies (e.g. Protégé [198]), de leurs appliquer des procédures de raisonnement via des raisonneurs (e.g. ELK [144]), de les manipuler via des bibliothèques de programmation (e.g. Jena [44]), et de stocker de larges bases de connaissances (e.g. Virtuoso [81]).

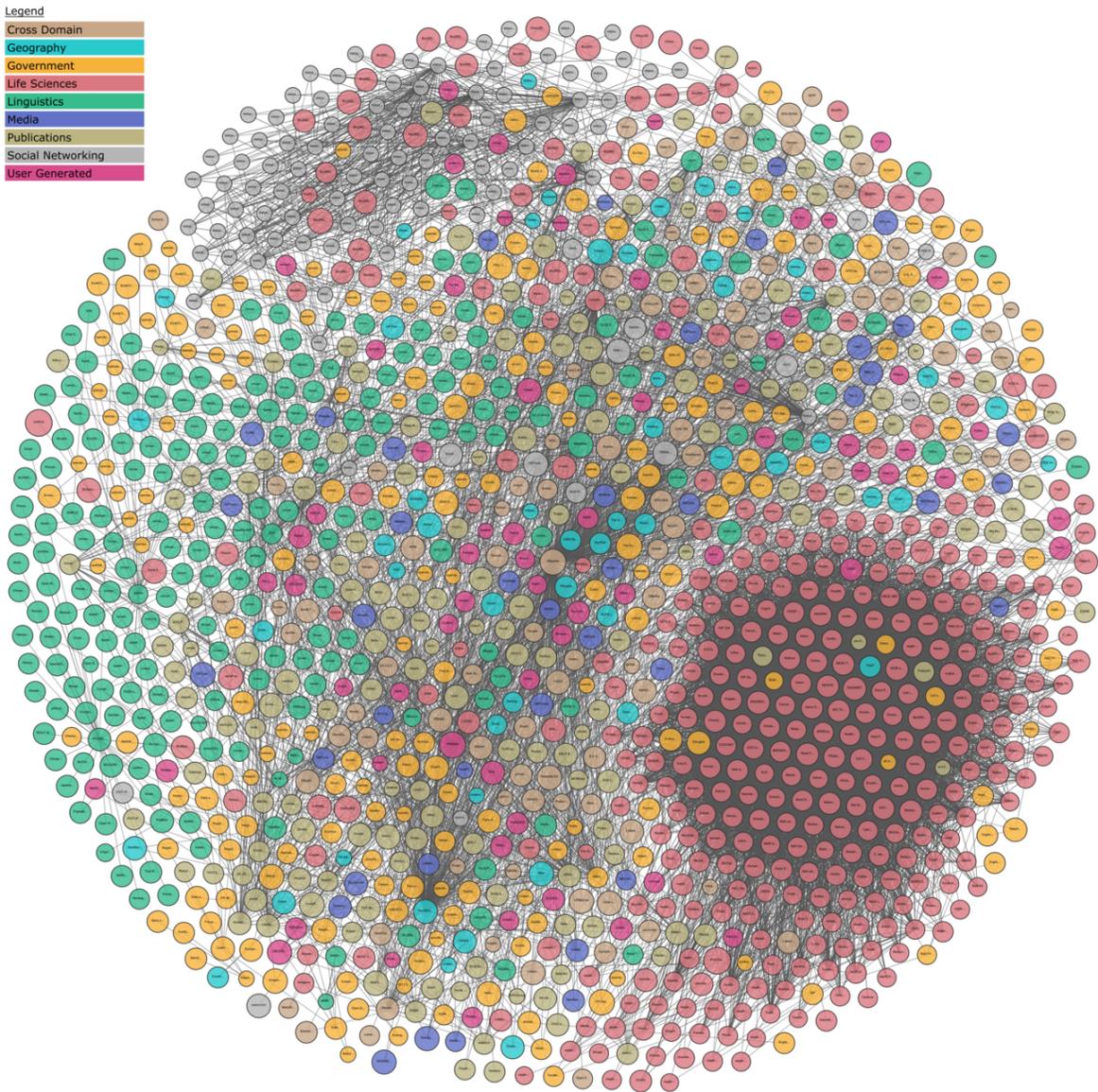
Des bases de connaissances telles que DBpedia [9], Wikidata [269] ou encore Yago [245], permettent ainsi aujourd'hui l'accès à de la connaissance formelle sur des millions d'entités générales (personnes, lieux...). Plusieurs initiatives ont aussi permis le développement d'ontologies et de bases de connaissances établies dans différents domaines spécialisés, e.g. médical, agronomie... La figure 2 présente une illustration populaire des principales bases de connaissances ouvertes et interconnectées exposées sur le Web Sémantique. A titre d'exemple de la quantité de connaissances aujourd'hui exposées dans ces bases, la base Uniprot repose, dans sa version d'avril 2023, sur près de 140 milliards d'éléments de connaissances (i.e. triplets RDF) sur les protéines au sens large [64].¹⁰

Ces développements, aujourd'hui probablement moins populaires que certains résultats amenés par l'Apprentissage Machine, sont tout autant impressionnants. La recherche effectuée en IA n'est naturellement pas menée en silo par ses différentes communautés - nous le verrons par exemple au travers de certains de nos travaux. Les travaux couplant l'apprentissage automatique et l'utilisation de représentations des connaissances sont fréquents depuis maintenant plusieurs années, e.g. pour la formulation d'approches neuro-symboliques.

L'IA, un domaine dynamisé par de fortes politiques nationales, européennes, et industrielles

Le sujet de l'IA, traité au regard d'enjeux économiques, écologiques, mais aussi géopolitiques et civilisationnels, comporte un aspect stratégique tout particulier pour les états. La France a maintenant depuis plusieurs années saisi l'importance du sujet. Dès 2018, les résultats du rapport Villani [225] soulignent publiquement la criticité du sujet et le positionnement favorable et particulier de la France du fait notamment de ses tissus académiques (recherche et formation), industriel et entrepreneurial. Dans la continuité de cette étude, des mesures fortes ont amené la création des Instituts Interdisciplinaires d'Intelligence Artificielle (3IA) et les propositions successives de nombreux appels à projet dans les domaines de l'IA. Des mesures semblables ont été prises au niveau européen. L'institution dans laquelle nous évoluons - qui ne s'inscrit pas dans un institut 3IA - a directement bénéficié de ce type de financements ; c'est aussi le cas de nos recherches.

10. <https://sparql.uniprot.org/>



The Linked Open Data Cloud from lod-cloud.net



FIGURE 2 *The Linked Open Data Cloud*. source <https://lod-cloud.net/>.

L'importance donnée à l'IA dans l'Informatique française et plus largement à l'échelle européenne a nécessairement contribué à préciser l'orientation de nos travaux. Il est à noter que beaucoup de laboratoires de Recherche et de R&D de grandes structures privées (e.g. GAFAM), voire d'industries, sont dirigés par des chercheurs de l'IA qui s'intéressent spécifiquement à l'apprentissage machine. Ce facteur contribue nécessairement à asseoir l'apprentissage machine comme technologie privilégiée de l'Informatique et de manière incidente à éveiller nos partenaires industriels sur l'utilisation de ces technologies et sur l'importance d'initier des collaborations académiques a minima de recherche appliquée sur ce sujet.

L'évolution graduelle de la place de l'apprentissage machine dans nos laboratoires et plus largement dans notre société est aussi illustrée graphiquement en Figure 1. Compte tenu de l'incidence de l'essor de l'apprentissage machine sur notre positionnement, nous précisons quelques éléments de contexte supplémentaires qui contribuent à la fois à expliquer le succès de l'approche, sa démocratisation, et sa large adoption. Cela nous permettra aussi de mentionner quelques notions que nous retrouverons tout au long du manuscrit :

- **Disponibilité de jeux de données.** La disponibilité et la réduction du coût des capteurs facilitent grandement la constitution de jeux de données. Les investissements importants dans la création et le partage de jeux de données font que nous disposons aujourd'hui de beaucoup de jeux de données indispensables pour les phases d'apprentissage de modèles d'apprentissage machine. Un nombre important d'entre eux sont partagés librement pour différentes problématiques (e.g. HuggingFace datasets [128], Kaggle [141]). Plusieurs répertoires de jeux de données sont ainsi accessibles publiquement ; nous incluons par ailleurs parmi ces ressources les environnements de simulation qui permettent d'entraîner certains modèles d'apprentissage par renforcement, e.g. composants de type moteurs physiques tels que Mujoco récemment rendu libre [256].
- **Développement des infrastructures de calcul ;** démocratisation des ressources de calcul spécifiques, e.g. à base de cartes graphiques (GPU), mise à disposition d'environnements d'exécution, e.g. Google colab, et de ressources de calcul au niveau national. Plusieurs travaux que nous exposerons ont par exemple été rendus possibles par plusieurs allocations d'heures de calcul sur le supercalculateur Jean Zay géré par le GENCI (Grand Equipement National de Calcul Intensif, allocation d'environ 50k heures AD011011309, R1, R2, R3).
- **Développement de bibliothèques dédiées et du partage de codes ;** de nombreuses initiatives ont permis la définition et la démocratisation de cadres de développement qui facilitent largement l'application et la recherche en apprentissage machine, e.g. Scikit-learn [212], Pytorch [210], Gym [40]. Un grand nombre de codes de calcul et implémentations qui facilitent grandement la recherche dans le domaine sont aussi gratuitement partagés sur le Web, cf. PaperWithCode [207] et GitHub.
- **Disponibilités de modèles pré-entraînés.** Plusieurs plateformes contribuent au partage de modèles pré-entraînés qui peuvent être réutilisés dans des contextes de recherche spécifiques, cf. HuggingFace, PyTorch/Tensorflow Hub. . .

Nombre de ces développements pourront être qualifiés de technologiques, ou de contributions d'ordre de l'Ingénierie. Ils expliquent pourtant à eux seuls en grande partie l'essor de l'apprentissage machine et de l'apprentissage profond. ¹¹ Ils contribuent ainsi à révolutionner la recherche en IA, qui bien qu'elle contienne une branche formelle, restera étroitement liée à la notion d'empirisme, en particulier en apprentissage machine.

11. Beaucoup de contributions (théoriques) importantes du domaine sont notamment antérieures aux développements des jeux de données et des ressources de calcul. L'expression de ces approches existantes sur d'importantes volumétries de données suffiront dans certains cas à l'obtention de certaines performances surprenantes.

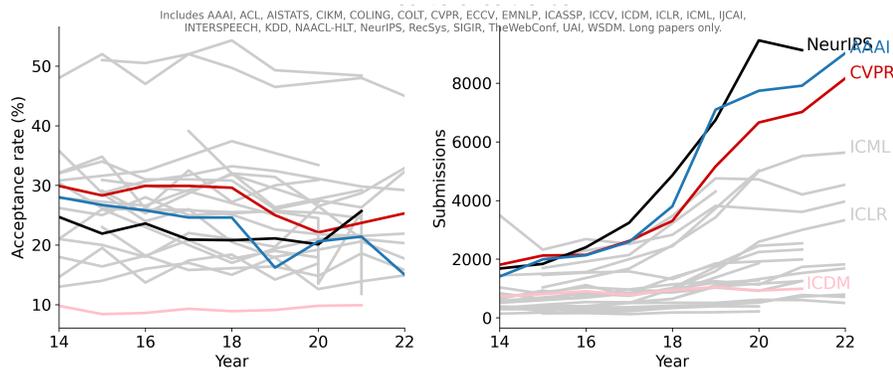


FIGURE 3 Évolution du nombre de publications dans les grandes conférences internationales du domaine de l'IA ces dernières années, et taux d'acceptation [62].

L'IA, un domaine de recherche particulièrement compétitif. La recherche dans le domaine de l'IA s'emballe littéralement. La figure 3 montre à titre d'exemple l'explosion du nombre de soumissions de papiers aux plus grandes conférences du domaine de l'IA (et la stagnation des taux d'acceptation). Pour illustrer la volumétrie de l'état de l'art dans le domaine, le papier qui présente l'architecture AlexNet [149], un des premiers résultats importants de l'apprentissage profond, a été cité plus de 120k fois depuis 2012; un autre papier populaire qui présente l'architecture de réseaux de neurones Transformer [265], populaire dans ce domaine, est lui cité plus de 80k fois depuis 2017 (source google scholar). Le risque de devenir spectateur d'un monde de la recherche en IA littéralement emballé est évoqué dans certaines discussions intimes entre chercheurs (et ces préoccupations n'appartiennent pas exclusivement aux petits laboratoires de province [257]).

On note de plus la diversité des approches qui méritent d'être appréciées pour suivre les développements de la littérature. Les systèmes d'IA mélangent aujourd'hui très souvent des cadres théoriques complexes et multiples. ChatGPT repose par exemple sur un modèle de type Transformer entraîné via des phases d'entraînement auto-supervisé, supervisé et par renforcement [203]. Suivre l'état de l'art est en soi aujourd'hui une véritable prouesse. Ajouté à cela, depuis quelques années, les difficultés rencontrées pour le recrutement des ressources humaines et l'accès aux infrastructures de calcul nécessaires pour réaliser certains travaux en apprentissage machine, empêchent aussi leur réalisation hors des cadres offerts par une poignée de structures pour la plupart privées. Cette particularité du domaine de l'apprentissage machine n'est pas négligeable dans un contexte recherche particulièrement compétitif.

Il était selon nous important de rappeler ces éléments de contexte en préambule de l'exposé de nos travaux. Nous ne souhaitons pas cependant conclure cette mise en contexte par une impression d'anxiété. La nature compétitive et débridée du domaine de l'IA s'explique par les perspectives favorables qu'il laisse envisager à nos sociétés; ces perspectives sont plus que jamais nombreuses et susceptibles de révolutionner des pans entiers de divers domaines. Les recherches que nous exposerons dans ce manuscrit s'inscrivent dans cette période semble-t-il singulière de l'étude de l'IA où, à l'image des débuts enthousiastes du domaine, la réalisation des programmes les plus ambitieux de l'IA n'a jamais été aussi plausible. Ce simple contexte contribue à lui seul à nourrir l'enthousiasme que nous exprimons dans la réalisation des travaux que nous exposerons par la suite.

1.2 Synthèse

Cette première partie introduit le contexte professionnel dans lequel ont été réalisés les travaux scientifiques que nous exposerons par la suite dans ce manuscrit. Après une brève parenthèse sur l'historique de notre orientation recherche, nous avons précisé le contexte institutionnel qui caractérise notre parcours. Celui-ci est intimement lié à l'Institut Mines Télécom (IMT) et plus particulièrement à l'école des mines d'Alès aujourd'hui appelée IMT Mines Alès. Cette institution à laquelle l'auteur est aujourd'hui rattaché en qualité de Maître-assistant en Informatique, nous permet de nous enrichir chaque jour d'activités de recherche, d'enseignement et de transfert auprès d'acteurs du tissu socio-économique régional ou national. Cette institution a ainsi hébergé au fil du temps les différentes structures - le laboratoire LGI2P et l'équipe KID, puis le centre CERIS et l'équipe I3A -, et les projets scientifiques en lien avec l'automatisation cognitive et l'Intelligence Artificielle auxquels nous avons contribué. Cette institution héberge aujourd'hui, pour partie, le projet Euromov Digital Health in Motion (Euromov DHM) mené en cotutelle avec l'Université de Montpellier pour lequel nous réalisons des travaux interdisciplinaires à l'interface du Numérique et en particulier de l'IA, des Sciences du mouvement humain, et de la Santé.

Nous avons aussi souhaité rappeler des éléments propres au contexte disciplinaire de l'Informatique et de l'IA que nous estimons importants - notamment pour comprendre les motivations de l'évolution du positionnement de la recherche que nous exposerons -, e.g. fort dynamisme de l'IA pour l'essentiel porté par l'apprentissage machine et l'apprentissage profond qui sont aujourd'hui sources de ruptures dans le domaine, fortes pressions multiformes (difficultés d'accès aux ressources humaines, de calcul. . .). Pour finir, nous avons exposé différentes responsabilités que nous occupons en lien avec notre activité de chercheur. Celles-ci illustrent notre implication à différents niveaux dans l'animation de la recherche, e.g. coresponsabilité du thème de recherche *Learning & Complexity* au sein de l'unité de recherche Euromov DHM. Elles illustrent aussi notre implication pour permettre la réalisation au quotidien d'une recherche dynamique, raisonnée et équilibrée, au sein de différents collectifs, e.g. représentant des enseignants-chercheurs, membre du réseau sécurité défense, représentant d'IMT Mines Alès auprès de la communauté Data-IA@IMT (Sciences des données & IA de l'IMT).

Nous espérons que les différents éléments d'information partagés dans cette section contribueront à une meilleure compréhension des travaux qui seront exposés par la suite.

Synthèse des contributions recherche

Sur les épaules des géants

Cette section vise à exposer le positionnement de nos travaux de recherche en Informatique, son évolution au fil des ans, ainsi qu'une synthèse des résultats obtenus jusqu'à présent. Nous présenterons différents axes de recherche que nous avons étudiés en distinguant trois volets :

1. L'étude des mesures sémantiques.
2. L'exploitation de connaissances dans des processus d'aide à la décision.
3. L'apprentissage de modèles prédictifs à partir de données.

Sommaire

2.1 Étude des mesures sémantiques	28
2.1.1 La notion de mesure sémantique	29
2.1.2 Synthèse des travaux de doctorat	34
2.1.3 Synthèse des travaux post-doctoraux sur les mesures sémantiques	35
2.1.4 Regard critique sur nos travaux et évolution du domaine	39
2.2 Exploitation de connaissances formelles dans des processus d'aide à la décision 44	
2.2.1 Résumé d'annotations terminologiques et conceptuelles	45
2.2.2 Estimation de la qualité de sources d'information et vérification d'assertions	59
2.2.3 Enrichissement de modèles prédictifs et d'inférence en TALN	67
2.2.4 Découverte de règles d'association à partir de données imparfaites	74
2.3 Apprentissage de modèles prédictifs à partir de données	85
2.3.1 Contributions au Traitement Automatique du Langage Naturel	85
2.3.2 Contributions à l'étude de problèmes d'optimisation combinatoire	102
2.3.3 Contributions disciplinaires en Apprentissage Machine	109

2.1 Étude des mesures sémantiques

Nous détaillons dans cette partie les travaux que nous avons réalisés sur le sujet des mesures sémantiques. Ils ont été initiés dans le cadre des études doctorales qui ont amené l’auteur à défendre une thèse en Informatique intitulée « *Knowledge-based Semantic Measures : From Theory to Applications* », « Mesures sémantiques à base de connaissance : de la théorie aux applicatifs » (2011-2014) [112]. Ces travaux de doctorat ont porté sur l’étude des mesures sémantiques dans des représentations des connaissances légères de types taxonomies.¹ Ils ont par la suite été approfondis dans de multiples travaux de recherche postdoctoraux menés en collaboration avec plusieurs collègues. Cela nous a notamment permis d’élargir notre champ d’étude aux mesures sémantiques capables de comparer des unités linguistiques telles que des mots, des phrases ou des textes. Nous notons cependant que le sujet des mesures sémantiques n’est plus depuis plusieurs années maintenant le point focal de nos recherches. Son côté transversal et l’importance des notions de similarités, proximités et autres distances (sémantiques) en IA, nous invitent cependant fréquemment à l’étudier directement ou indirectement.

Le caractère central des notions de distance ou de similarité fait qu’une très grande partie des travaux qui seront mentionnés dans ce manuscrit auraient très probablement pu être exposés au travers d’elles. Ce n’est pas le choix que nous faisons ici, afin en particulier d’être fidèles à la recherche que nous avons menée au fil des ans et à sa progression en termes de positionnement. Nous proposons cependant dans cette section - qui se distinguera des autres - une parenthèse sur cette notion de mesure sémantique qui fera écho à une multitude de recherches que nous avons menées, certaines étant par ailleurs détaillées dans ce manuscrit.

Nous débuterons cette section par un ensemble de précisions concernant la notion de mesure sémantique. Suivra une synthèse des travaux réalisés pendant nos études de doctorat. Nous détaillerons ensuite les contributions postdoctorales que nous avons proposées en lien direct avec cette problématique. Nous terminerons par une analyse critique de nos contributions accompagnée d’éléments d’informations sur l’évolution des travaux de recherche sur ce thème.

Collaborateurs :

Les travaux mentionnés dans cette section ont été réalisés avec de nombreux collaborateurs :

- Doctorants encadrés : Valentina Beretta (2015-2018), Massissilia Medjkoune (2015-2018), Gildas Tagny-NGompe (2015-2020), Jean-Christophe Mensonides (2015-2020), Pierre-Antoine Jean (2014-2017).
- Postdoctorants encadrés : Kaoutar Ghazi (2020), Naser Jelassi (2017).
- Stagiaires encadrés : Sébastien Marchal (2019), Jie Hi (2016).
- Collaborateurs d’IMT Mines Alès : Jacky Montmain, Sylvie Ranwez, Stéfan Janaqi, Abdelhak Imoussaten, Nicolas Fiorini, François Troussel, Jocelyn Poncelet.
- Collaborateurs externes : David Sánchez et Montserrat Batet (*Universitat Rovira i Virgili*, Espagne).

1. L’Annexe B précise ce que nous entendons par ontologies légères dans ce manuscrit.

2.1.1 La notion de mesure sémantique

Les notions de similarité et de distance sont centrales pour la caractérisation de nombreux phénomènes. Les sciences cognitives considèrent qu'elles sous-tendent nos facultés à classifier des objets, former des concepts, faire des généralisations. . . [262, 162, 110] Elles joueraient ainsi un rôle important dans l'expression d'un grand nombre de facultés cognitives en lien direct avec la prise de décision (e.g. mémorisation, apprentissage, raisonnement). Elles nous sont par ailleurs intuitives car nous comparons fréquemment des objets sur la base de différences exprimées à l'aide de mesures variées (longueur, volume, poids, prix. . .). Nous utiliserons par la suite le terme *mesure* pour faire référence aux notions de proximité, distance, similarité, dissimilarité. . . Cette économie de rigueur admise par la littérature facilitera largement la lecture (et la rédaction) des développements qui suivront. Notez aussi que par abus de langage, nous parlerons très souvent de mesures pour mentionner les outils de mesure que nous utiliserons pour mesurer une quantité d'intérêt.

Les *mesures* de similarité, proximité et de distance sont largement retrouvées dans de nombreuses modélisations formelles ; les notions de distance et de mesure étant par ailleurs clairement distinguées en Mathématiques. Une littérature colossale est effectivement dédiée à la comparaison d'objets d'intérêt dans des domaines variés. Il existe ainsi des mesures, définies de manière plus ou moins détachée d'un applicatif spécifique, pour comparer à peu près tout et n'importe quoi : séquences génomiques, textes, images, graphes, vecteurs, distributions de probabilité. . . [74]

Parmi cette diversité de *mesures*, celles dites *sémantiques* (*semantic measures*) regroupent un ensemble de techniques qui visent à comparer des entités de nature sémantique ou représentées sémantiquement - nous parlerons d'*entités sémantiques* par la suite. On retrouve parmi ces entités sémantiques des unités linguistiques telles que des mots, des phrases, des textes, ou encore toutes entités qui ont une représentation sémantique au sens où elles sont définies ou caractérisées par une représentation des connaissances, e.g. concepts d'une ontologie, des documents, gènes ou tous autres objets annotés par des concepts, ou plus généralement toutes ressources exprimées dans un graphe de connaissances.

Dans son appréciation générale, la notion de mesure sémantique n'est pas définie formellement. Comme nous le verrons, toute entreprise visant à la définir ne serait-ce qu'informellement s'avère même laborieuse et délicate pour être très honnête (la littérature semble s'accorder sur ce point). Elle renvoie pourtant à des choses très caractéristiques sur notre manière de manipuler des entités sémantiques. Autorisons-nous quelques petits tests :

1. Parmi les deux couples de termes suivants, (i) *thé* et *café*, et (ii) *café* et *mafé*², quel est celui qui est composé des termes les plus proches sémantiquement parlant ? Les gens s'accorderont généralement pour dire que les termes *thé* et *café* sont plus proches d'un point de vue sémantique car ils font tous deux référence à une boisson chaude et stimulante ; les termes *café* et *mafé*, qui sont pourtant plus proches d'un point de vue syntaxique car une seule lettre diffère entre eux, renvoient à des notions dans l'absolu *plus* distantes.
2. Quel concept vous évoque le couple de concepts **Banane** et **Animal** ? Allez-y, fermez les yeux, proposez un concept sans trop réfléchir. La plupart des gens penseront naturellement au concept précisé dans cette note³ ; aucune magie, l'association semble évidente. Vous auriez pourtant pu penser à une infinité d'autres concepts. Celui qui regroupe « les concepts mentionnés

2. plat à base de pâte d'arachide originaire du Mali.

3. Singe.

2.1. Étude des mesures sémantiques

par des noms communs de 6 lettres composés de deux lettres 'a' » aurait été tout aussi valide... Cependant, même si ce concept est discriminant et caractéristique de la représentation terminologique des concepts Banane et Animal, i.e. « b·a·n·a·n·e » et « a·n·i·m·a·l », son *évocation* n'était peut-être pas aussi évidente que celle de Singe direz-vous.

3. Complétez la phrase suivante : « Après l'avoir regardée manger son fromage, le chat a finalement attrapé la petite [?] ». Peu d'entre vous compléteront cette phrase par les mots plante, veste ou encore fièvre ; nous sommes d'accord.
4. Quelles sont les deux phrases qui vous semblent les plus proches sémantiquement parlant parmi les trois phrases suivantes : (i) « Luc écrit un mémoire », (ii) « Une femme rédige un manuscrit », et (iii) « Une femme travaille sa mémoire ». Cette question est naturellement sujette à interprétation. Les phrases (i) et (ii) seront pourtant très souvent jugées plus proches d'un point de vue sémantique car elles font toutes deux référence à une personne engagée dans un processus de création d'une production écrite.

Ces quelques tests mobilisent différentes capacités cognitives et mériteraient un développement conséquent pour être discutés en détail. ⁴ Nous nous accorderons sur le fait qu'ils mobilisent tous, à leur façon, une appréciation d'une forme de *relation* (ou d'*association*) entre des entités sémantiques. Le propre des mesures sémantiques est alors d'estimer de manière quantifiée *la force de la relation* qui unie deux entités comparées dans un contexte donné. Elles visent même dans la plupart des travaux à estimer *notre* appréciation (humaine) de cette force. Elles exploitent pour cela l'étude de *proxies sémantiques* qui permettent l'appréciation de la nature des entités comparées au sens de leur sémantique, à savoir des corpus de textes et/ou des bases de connaissances. La nature de la relation en question n'est pas clairement définie et peut traduire de choses très différentes en fonction des contextes d'étude (e.g. relation au sens taxonomique).

Notez que lorsque nous souhaitons mimer l'appréciation que fait l'Homme d'une telle relation, nous admettons par postulat le fait qu'il existe une appréciation consensuelle d'une mesure sémantique, i.e. le fait que l'on s'accorde sur un degré d'appréciation de la mesure sémantique évaluée. Cela est discutable en soi - et cela même si nous avons tous des réponses a minima semblables aux questions traitées ci-dessus. Une IA alignée sur des comportements humains dans des contextes d'étude correspondant aux exemples de tests mentionnés ci-dessus serait d'une grande utilité. Elle permettrait l'automatisation de traitements complexes relatifs à la manipulation d'entités sémantiques. Nous proposons dans l'encadré suivant la réponse proposée par ChatGPT 3.5 à une des questions que vous venez de traiter.

A - Réponse de ChatGPT 3.5 à la question (prompt) : « Quel concept vous évoque le couple de concepts banane et animal ? »

« Le couple de concepts "banane" et "animal" peut évoquer plusieurs idées ou associations d'idées, selon le contexte. Voici quelques interprétations possibles :

4. Notons que les affirmations que nous proposons quant aux réponses attendues aux questions proposées - pour lesquelles l'auteur est très confiant - ne reposent pas sur des tests supportés par l'implémentation d'une démarche scientifique rigoureuse.

1. *Alimentation animale* : Les bananes sont parfois données comme collation aux animaux domestiques tels que les chiens ou les singes. Certains animaux sauvages, comme les singes, sont également connus pour consommer des bananes dans leur habitat naturel.
2. *Symbolisme humoristique* : Le simple fait de penser à un animal en train de manger une banane peut être perçu comme comique, car les bananes sont souvent associées à des gestes humains plutôt qu'à des comportements animaux.
3. *Régime alimentaire végétarien ou herbivore* : Les bananes sont un aliment végétal, et les animaux herbivores, tels que les singes, ont un régime alimentaire principalement composé de végétaux. Ce couple de concepts pourrait donc évoquer des idées liées à l'alimentation herbivore.
4. *Écologie* : La culture et le commerce des bananes peuvent avoir des impacts environnementaux et sociaux, par exemple, dans les plantations de bananes. Le couple de concepts pourrait donc être lié à des préoccupations environnementales ou éthiques liées à la production et à la consommation de bananes.

Il est important de noter que la signification précise dépend du contexte dans lequel ces concepts sont utilisés, et les associations d'idées peuvent varier d'une personne à l'autre. Si vous avez un contexte spécifique en tête, n'hésitez pas à le partager pour une interprétation plus précise. »

Remarque : notez que même si le concept Singe est évoqué, le développement sur le symbolisme humoristique est pour le moins douteux...

B - Réponse de ChatGPT 3.5 à la question (prompt) : « En un mot, quel concept vous évoque le couple de concepts banane et animal ? » (en changeant de session bien entendu).

« *Monkey (singe).* »

Remarque : réponse attendue trouvée.

Un large spectre de la littérature en IA et plus particulièrement en TALN et en Représentation des Connaissances, mais aussi en Recherche d'Information et en Recommandation, est directement ou indirectement intéressé par cette notion de mesure sémantique. Cela s'explique par le rôle important que ces mesures semblent jouer d'un point de vue fondamental dans l'élaboration de nombreux processus cognitifs (ou par leur capacité à traduire certains aspects liés à des phénomènes cognitifs). Leur popularité en tant qu'objet d'étude s'explique aussi par leur rôle central comme brique élémentaire dans de multiples traitements fréquents en Informatique et à finalité pratique, e.g. comparaison pour la mise en correspondance, le classement ou le regroupement de documents, ressources d'une base de connaissances, ou toutes autres entités sémantiques.

Toute comparaison d'objets nécessite au préalable de les représenter. La *représentation cognitive* ou *représentation mentale* d'objets concrets ou abstraits chez l'Homme est largement étudiée depuis plusieurs décennies, en particulier en Sciences Cognitives et en Neurosciences. Ces travaux visent à comprendre comment sont encodées les caractéristiques visuelles, auditives, tactiles, olfactives, fonctionnelles, émotionnelles (...) de ces objets, ainsi que les approches engagées dans leurs comparaisons. Comme nous le verrons par la suite, une grande partie de la littérature sur les mesures sémantiques est elle aussi indissociable de la notion de représentation. A contrario des travaux menés en Sciences Cognitives, la question des propriétés conceptuelles ou perceptives des objets comparés n'est cependant généralement pas abordée dans cette littérature. Il ne s'agit pas non plus dans cette littérature de postuler des modèles de similarité qui satisferont certaines caractéristiques

de l'appréciation de similarité supposées ou observées chez l'Homme, comme cela est le cas en Sciences cognitives.⁵ Le sujet est fréquemment traité sous l'angle de l'approche à adopter pour représenter les objets d'intérêt de manière à les comparer d'un point de vue algorithmique. On admet alors très souvent une connaissance qui exprime la nature des objets comparés (e.g. représentation des connaissances), ou une stratégie capable de la distinguer, e.g. sémantique distributionnelle des termes (détaillée ci-après). Cette approche du problème permet, de manière silencieuse, de taire l'impossibilité soulignée dans la littérature de discuter dans l'absolu la notion de similarité indépendamment d'un contexte, entendu ici comme l'importance à donner aux dimensions des objets comparés. Nous faisons notamment référence aux propos populaires du philosophe Nelson Goodman, rapporté par exemple par Hahn [110], qui défend l'idée que la notion de similarité est « *an imposter, a quack* » car elle ne peut dans l'absolu se résumer à l'analyse indiscriminée des propriétés des objets comparés. Notez en effet que deux objets partagent, indépendamment de leur nature, une infinité de propriétés.

Un grand nombre de travaux en Sciences Cognitives postulent l'existence d'un *espace sémantique quasi-consensuel* dans lequel les objets d'intérêt sont représentés, manipulés et comparés de manière proche entre les individus. Notons en effet que même si un tel espace n'est pas à ce jour qualifiable d'un point de vue neurobiologique [30, 286], lorsque nous (vous et nous) parlons de Singe, nous faisons globalement référence à un même concept. Il est ainsi commun de considérer la notion abstraite d'espace sémantique en admettant une analogie confortable avec la notion mathématique d'espace.⁶ Là aussi, un manuscrit détaillé sur cette notion d'espace sémantique ne ferait très probablement qu'effleurer le sujet.

La très grande majorité des travaux sur les mesures sémantiques considèrent explicitement ou implicitement l'existence d'un tel espace sans en faire une discussion. Dans les approches à base de connaissances se sont très souvent les représentations des connaissances, dans lesquelles sont exprimés les objets comparés, qui définissent les espaces sémantiques qui seront étudiés. L'approche est différente pour la comparaison de termes, phrases, ou textes. On distingue les travaux de certains chercheurs qui construisent explicitement un espace sémantique pour des exercices de visualisation ou d'analyse de corpus par exemple [219, 267]. Il nous arrivera de procéder de cette façon pour comparer des textes à l'aide de caractéristiques sélectionnées de manière *ad hoc*. Nous inscrirons cependant la plupart de nos travaux sur des unités linguistiques analysées à l'aide de l'étude de corpus de textes, dans la lignée des contributions qui construisent des espaces sémantiques de manière indirecte en implémentant notamment la *sémantique distributionnelle* [120, 232] - cf. encadré. Nous ferons en particulier le lien avec la large littérature récente sur la construction de représentations sémantiques (plongement ou *embedding*) à l'aide d'approches neuronales.

On ne manquera pas tout de même de rappeler les propos de Tversky (1977) qui restent d'actualité : « *Both dimensional and metric assumptions are open to question* » [262]. Nous ne chercherons pas à faire la démonstration de l'existence d'un espace sémantique (latent) dans l'absolu, ou d'étudier ses propriétés : nous l'admettrons simplement comme postulat.

5. e.g. des modèles défendent l'idée que la similarité n'est pas symétrique : « *an ellipse is more similar to a circle than a circle is to an ellipse* » [262].

6. Analogie qui permet peut-être de rendre l'étude de ce sujet complexe plus familière et moins intimidante pour certains ; ce lien étroit entre mesure et espace fait par ailleurs référence au lien quasi-indissociable, pour ceux formés aux sciences formelles, entre les notions d'espace et de distance ou a minima de voisinage, e.g., espace métrique ou espace topologique dont la topologie repose sur une distance, à savoir plus largement le lien entre la notion d'espace topologique et de voisinage, ou encore le rôle important des distances dans les espaces vectoriels.

Sémantique et hypothèse distributionnelles

La sémantique distributionnelle est un cadre théorique linguistique étudié à la croisée de la Linguistique Informatique, de l'IA, et des Sciences Cognitives. Ce cadre est notamment issu des travaux de Harris [120] dans lequel l'hypothèse du même nom (distributionnelle) admet que le contexte linguistique suffit pour caractériser le sens d'un mot. L'hypothèse distributionnelle est d'ailleurs fréquemment résumée par la fameuse citation de Firth (1957) : « *You Shall Know a Word by the Company It Keeps* ». Autrement dit : les cooccurrences lexicales suffisent à définir la sémantique d'un mot. Du fait de sa formulation originelle proche de la notion de statistique - un mot est qualifié par ses observations et sa distribution dans un corpus de textes -, la sémantique distributionnelle se distingue d'autres cadres théoriques qui reposent sur des approches symboliques, e.g. approches fondées sur des données lexicales de type WordNet [194].

La sémantique distributionnelle est étudiée depuis plusieurs années. Il est par exemple possible d'inscrire dans une certaine mesure les approches traditionnelles en Recherche d'Information (de type TF-IDF, Okapi BM25), les modèles de type Vector Space Model (VSM) [77], Latent Semantic Analysis (LSA) [161], ou encore les modèles de langue neuronaux récents comme des implémentations de l'hypothèse distributionnelle [261, 169]. Certains auteurs associent la sémantique distributionnelle à une représentation vectorielle ; Lenci et Sahlgren définissent la sémantique distributionnelle comme suit « *The meaning of a linguistic expression is represented with a real-valued vector (embedding) that encodes its statistical distribution in contexts* » [169]. Nous recommandons aux lecteurs intéressés par ce sujet la lecture complémentaire de travaux en Linguistique Informatique qui discutent la notion de sémantique distributionnelle en tenant compte des aspects liés à son implémentation via des modèles traditionnels (à base de cooccurrences) ou des modèles de langue neuronaux, e.g. [192].

Les mesures sémantiques sont multiformes et mériteraient pour être discutées pleinement d'adopter une approche interdisciplinaire afin d'aborder leurs liens étroits avec de nombreuses problématiques, e.g. relation signifiant/signifié, biais culturels, mécanismes biologiques impliqués. Nous nous concentrerons cependant sur l'étude des contributions du domaine de l'Informatique. Les travaux que nous exposerons seront en ce sens très modestes. Comme nous l'avons mentionné, les mesures sémantiques sont souvent mal ou non définies d'un point de vue sémantique - il est souvent difficile de comprendre ce que veut dire une mesure spécifique, e.g. afficher que la proximité sémantique entre les termes *singe* et *banane* est de 0.7 n'a généralement de sens qu'en relatif. Nous n'élaborerons pas ici sur les aspects liés à la sémantique des mesures sémantiques ; nous ne discuterons pas en particulier les multiples et variables différences entre similarité, proximité ou distance sémantiques que peut admettre la littérature - cf. [117] pour nos travaux sur ce sujet. Nous éviterons dans ce contexte de réduire la discussion à une entreprise subjective de formalisation de la sémantique des mesures sémantiques. Nous ne discuterons pas non plus certaines réserves légitimes formulées quant à l'utilité de ces mesures.⁷ Nous admettons que les notions de similarité et de proximité sémantiques semblent encodées dans nos cerveaux tout en correspondant à des construits sociaux mobilisés dans de nombreux processus cognitifs. Nous adopterons généralement une attitude très pragmatique motivée par la nécessité par exemple de comparer ou de rapprocher des entités sémantiques, e.g. faire le lien entre un texte et une requête, regrouper ou annoter des textes. . .

7. « *the concept of semantic similarity is too broad to be useful* » [232] car, selon certains auteurs, il ne distingue pas le plus souvent les dimensions considérées pour la comparaison (synonymies, antonymies. . .).

2.1.2 Synthèse des travaux de doctorat

Nos premiers travaux sur ce sujet se sont concentrés sur des aspects liés à l'étude de modèles calculatoires des mesures sémantiques. Nous avons plus spécifiquement étudié différentes mesures sémantiques qui reposent sur l'analyse de connaissances exprimées dans des ontologies légères. Ces mesures sont retrouvées dans la définition d'algorithmes ou de traitements exploitant des représentations des connaissances. Elles sont aussi utilisées dans le cadre de la mise en place de techniques de raisonnement approché. Elles intéressent en particulier beaucoup de travaux académiques et d'industriels en lien avec la recherche d'information et la recommandation supportées par des ontologies, e.g. [282, 275]. Elles trouvent aussi plusieurs applications dans l'analyse de gènes, ou plus largement dans le domaine biomédical [215].

L'entreprise principale de nos travaux a porté, dans la continuité des recherches réalisées par Blanchard et collaborateurs [33, 32], à démontrer que malgré l'apparente hétérogénéité des mesures existantes, certaines propriétés communes des mesures permettent de distinguer une classification générale des approches, et un ensemble réduit de primitives abstraites. Nous soulignons dans la thèse défendue en 2014, qu'un grand nombre de mesures de la littérature (qui précèdent nos travaux), bien qu'elles soient définies indépendamment et sous-tendues par des cadres théoriques variés, ne sont que des expressions particulières de mesures paramétriques génériques fondées sur les primitives abstraites en question. Ces travaux transposent à plusieurs égards l'entreprise de Tversky sur les caractéristiques des similarités, avec une application aux mesures sémantiques, en particulier dans le cadre des représentations des connaissances de type ontologies légères. Nos résultats nous ont ainsi conduits à proposer un cadre théorique unificateur pour les mesures sémantiques à base de connaissances. Ce dernier permet : (i) d'étudier les propriétés théoriques des mesures, (ii) d'orienter l'utilisateur dans le choix d'une mesure adaptée à sa problématique, et si besoin (iii) d'exprimer de nouvelles mesures. Nous pensons que ce cadre théorique, élaboré dans la continuité des travaux précités, notamment rappelons-le de Blanchard et collaborateurs [33], a contribué à poser un regard nouveau sur ce domaine (environ 185 citations de la publication qui expose ce cadre [119] depuis 2014).

Un travail important d'outillage logiciel a aussi été réalisé pour les besoins de nos travaux de thèse. Les recherches que nous avons réalisées reposent en effet largement sur des évaluations empiriques. Nous avons donc proposé la Semantic Measures Library (SML), une librairie logicielle dédiée au calcul et à l'analyse des mesures sémantiques [112]. Cette librairie permet l'instanciation de centaines de mesures issues à la fois de la littérature et des fonctions paramétriques découlant du cadre unificateur introduit. Des utilitaires spécifiques ont de plus été développés pour permettre son utilisation sur des représentations des connaissances exprimées de manière non standard à l'heure de nos travaux, e.g. *Medical subject headings* (MeSH) [178]. La SML et les outils associés ont été supportés pendant des années à la suite des travaux de thèse (support actif jusqu'en 2018); ils ne sont plus aujourd'hui activement développés bien qu'ils soient toujours partagés, documentés, open source, et accessibles librement, e.g. au travers de la *marketplace* de l'IMT (<https://factory.euromov.eu/sml>).

Ci-dessous nos autres contributions algorithmiques et théoriques réalisées durant cette période :

- La définition d'une méthode permettant de comparer des ressources exprimées dans un graphe sémantique, e.g. de type RDF, avec une application pour la mise en place de systèmes de recommandation à base de contenu (i.e. qui se basent sur le descriptif des ressources, a contrario des approches qui se basent sur du filtrage collaboratif) [116].

- Une approche pour comparer des concepts retrouvés dans des taxonomies chevauchantes. Contribution réalisée en collaboration avec Montserrat Batet et David Sánchez de l'Université espagnole *Universitat Rovira i Virgili* (Tarragone) [14].
- Des optimisations algorithmiques pour le calcul de certaines mesures sémantiques (contributions visant à réduire la complexité calculatoire des mesures en question) [112].
- Une technique d'apprentissage supervisé afin de cibler les mesures sémantiques adaptées à un contexte applicatif en prenant en compte l'incertitude associée au jeu de test utilisé [131, 130].

2.1.3 Synthèse des travaux post-doctoraux sur les mesures sémantiques

Les travaux en lien avec les mesures sémantiques que nous avons réalisés suite à nos recherches de doctorat portent sur : (i) la réalisation d'un ouvrage dédié au domaine, (ii) la proposition de contributions techniques liées aux mesures, et pour l'essentiel (iii) des aspects de mise en oeuvre de ces mesures dans différents contextes, i.e. plusieurs tâches et domaines applicatifs. Nous nous proposons ici de mentionner quelques exemples de travaux réalisés ; certains d'entre eux seront détaillés par la suite dans ce manuscrit. Nous discuterons cependant dans cette section un exemple de contribution portant sur la définition d'un contenu informationnel, i.e. composante importante un grand nombre de mesures sémantiques.

Ouvrage de synthèse de l'état de l'art

Un large travail d'état de l'art a été réalisé dans le cadre de nos travaux de doctorat. Celui-ci portait exclusivement sur les mesures à base de connaissances et plus spécifiquement d'ontologies légères. Il a été étendu de manière à intégrer la prise en compte de mesures issues de l'analyse de corpus de textes, e.g. mesures proposées pour comparer des termes. Ce travail nous a amenés à proposer en 2015 un ouvrage de synthèse de l'état de l'art des mesures sémantiques intitulé *Semantic Similarity from Natural Language and Ontology Analysis* [117].⁸ L'ouvrage propose une couverture étendue du domaine - inédite à l'heure de sa parution - qui traite à la fois des mesures à base de connaissances et à base d'analyse de textes. Ce travail nous a permis d'étudier et d'exposer plus encore l'état de l'art lié aux mesures sémantiques en détaillant entre autres les divers applicatifs, les aspects logiciels, mais aussi en couvrant une partie des mesures utiles pour l'analyse d'unités linguistiques simples, les mots en particulier.

Un des aspects selon nous intéressant abordé dans cet ouvrage est l'importance de la notion de *représentation*. Une grande partie de l'histoire de la recherche sur la similarité dans les domaines des Sciences Cognitives admet l'exigence existentielle de représentations pour comparer des objets concrets ou abstraits. L'essentiel des travaux dans ce domaine portent alors sur la proposition de modèles de représentation mentale des objets (*spatial model*, *feature model*. . .) puis sur la définition de mesures qui seront utilisées pour comparer les objets en fonction de l'approche de représentation adoptée [117]. Cette notion de représentation était d'ores et déjà (implicitement) centrale dans l'état de l'art des mesures à base de textes mais était jusque-là peu considérée dans l'état de l'art des mesures à base de connaissances. Il est en effet intéressant de souligner que les travaux portant sur les mesures de ce type s'attachent essentiellement à expliciter les modalités de comparaison des objets traités, en considérant que ceux-ci sont représentés dans la connaissance qui les exprime, e.g. deux concepts d'une taxonomie seront comparés à partir du plus court chemin qui les relie dans la

8. Ouvrage de 238 pages rédigé en anglais et publié par Springer dans la collection *Synthesis Lectures on Human Language Technologies*; cité plus de 300 fois depuis 2015 (source google scholar). Une version gratuite *preprint* est proposée sur Arxiv [118].

taxonomie ; la représentation du concept comme un noeud du graphe orienté acyclique (DAG) associé à la taxonomie n'est pas discutée en soi. A contrario, beaucoup de travaux sur les mesures à base de textes se concentrent sur la construction de représentations qui seront ensuite comparées à l'aide de mesures usuelles pour la comparaison des représentations adoptées, e.g. similarité cosinus pour des représentations vectorielles, mesures ensemblistes pour des représentations ensemblistes. L'effort de modélisation était alors dans ce cas très souvent réalisé sur la construction des représentations - pour la très grande majorité des représentations vectorielles - et non pas sur la manière de comparer ces représentations.

Nous avons dans ce contexte étudié une classification des mesures au regard (i) des éléments qu'elles visent à comparer (e.g. concepts, termes); (ii) des sources d'information considérées (*proxies sémantiques*⁹, e.g. représentations des connaissances ou corpus de textes); (iii) des assumptions considérées (e.g. hypothèse distributionnelle); et (iv) de la forme canonique adoptée pour les représenter et les comparer. Une telle lecture de l'état de l'art qui donne une place importante à la notion de représentation est selon nous intéressante pour souligner le lien entre différentes contributions, et pour établir des liens interdisciplinaires entre différents travaux sur ce vaste sujet. Notons qu'en 2015, la notion de représentation avait à ce moment-là une place moins singulière que celle que lui ont graduellement accordée les travaux récents en apprentissage profond, cf. l'apprentissage de représentations [18].

Nous pensons que la critique de notre ouvrage proposée par Xiong [274] est juste (en libre accès¹⁰). En effet, même si l'ouvrage dispose d'un certain nombre de qualités, comme souligné dans la critique, il aurait mérité de proposer un traitement plus large et complet des techniques liées à l'évaluation de proximités sémantiques d'unités linguistiques à base de plongements notamment construits par des modèles neuronaux. La publication du livre a en effet été effectuée pendant la première phase des développements relatifs à l'utilisation de techniques d'apprentissage machine à base de réseaux de neurones pour la construction de représentations vectorielles utilisées pour l'estimation de proximité. Les auteurs n'étaient, par ailleurs, pas particulièrement familiers des modèles en question à l'heure de l'écriture de l'ouvrage. L'ouvrage que nous avons proposé mériterait aujourd'hui d'être largement actualisé, par exemple pour tenir compte des multiples développements du domaine en lien avec l'essor des techniques de représentation à base d'apprentissage machine. Nous évoquerons ce point spécifique par la suite dans cette section.

Utilisation de mesures spécifiques dans des domaines d'application variés

Différents travaux que nous avons menés portent sur l'utilisation et la proposition de mesures spécifiques dans des contextes d'indexation sémantique [95, 93, 94], de résumé sémantique [115, 114], de post-traitement de plongements de termes à l'aide de représentations des connaissances, i.e. *retrofitting* [103], ou encore entre autres pour la génération de jeux de données que nous avons utilisés pour évaluer des techniques de vérification d'assertions [25].

9. Nous définissons dans nos travaux un *proxy sémantique* comme toute source d'information à l'aide de laquelle des indications de la sémantique des éléments comparés, qui seront utilisées par une mesure sémantique, peuvent être extraites. Nous distinguons deux grands types de proxies sémantiques : (i) les textes non structurés ou semi-structurés (corpus de textes, vocabulaires contrôlés, dictionnaires. . .), et (ii) les représentations structurées (ontologies, thésaurus, vocabulaires structurés, taxonomies. . .).

10. via ce lien : <https://aclanthology.org/J16-4010.pdf>

Nous avons aussi travaillé sur la définition de mesures *ad hoc* pour comparer des termes au regard de spécificités liées aux corpus de textes étudiés, e.g. [190]. Un travail portant sur l'apprentissage supervisé d'une distance afin de comparer des documents a aussi été effectué - et aurait mérité d'être approfondi. La tâche d'apprentissage que nous avons proposée se base sur une formulation proche de l'apprentissage auto-supervisé dans le sens où l'on cherche à apprendre une quantification relative à des modifications effectuées de manière contrôlée sur un texte, cf. [250]¹¹.

Analyse empirique des mesures

Nous avons mené plusieurs études empiriques des mesures sémantiques dans différents contextes. A titre d'exemple, nous avons participé à une étude comparative de l'impact des mesures de similarité sémantique dans un algorithme de partitionnement (application allant du domaine biomédical à la détection de comportements de consommation) [218], et dans un contexte d'indexation par extraction et appariement textuels [190]. Nombre de ces évaluations empiriques ont été réalisées dans le cadre de collaborations industrielles, e.g. projet ARIA, en collaboration avec la société ESII pour la mise en place de bornes d'orientation [102], ou pour l'appariement de fiches descriptives de produits en collaboration avec la société Price Observatory.

Définition d'un contenu informationnel hybride

La notion de contenu informationnel - *Information Content* (IC) - est étroitement liée à la notion de mesure sémantique car de nombreuses mesures qui tirent parti de représentations des connaissances reposent sur son calcul. C'est en particulier le cas des mesures fondées sur la théorie de l'information [117]. A titre d'exemple, Lin [175] propose une mesure de ce type aujourd'hui populaire, qui modélise la similarité de deux concepts a et b d'une ontologie légère par : $sim_{Lin}(a, b) = \frac{2 \cdot IC(MICA(a, b))}{IC(a) + IC(b)}$, avec IC une fonction qui associe à tout concept une valeur réelle positive quantifiant son degré d'informativité - contenu informationnel - et $MICA(a, b)$ l'ancêtre commun de a et b le plus informatif au regard de cette notion d'IC.

Traditionnellement, l'IC d'un concept est défini comme fonction du nombre d'observations du concept et de ses spécialisations, e.g. dans un corpus de textes [117] [226]. Nous ne rentrerons pas dans les détails des nombreuses formulations existantes [113]. Cette modélisation peut se traduire de manière intrinsèque en formulant l'IC d'un concept comme inversement proportionnel au nombre de concepts qui le spécialisent dans la représentation des connaissances. Elle peut aussi être formulée de manière extrinsèque si l'on dispose de fréquences d'observations des concepts, en définissant l'IC comme inversement proportionnel au cumul des fréquences d'observations du concept et de ses spécialisations.

Un travail technique a été réalisé sur la notion d'IC, et plus spécifiquement sur les IC hybrides, c'est-à-dire ceux calculés via l'analyse de représentations des connaissances de type ontologies légères structurant les concepts, et à partir de données de fréquence d'observations de ses concepts (i.e. hybridation des approches intrinsèque et extrinsèque) [113]. L'objectif de cette proposition est de supposer un phénomène dit d'*inférence inductive* lors de l'évaluation de l'IC d'un concept. Nous

11. Cette contribution est détaillée dans la thèse de Tagny NGompe, section 5.3. Elle correspond, toute proportion gardée, à une expression spécifique de l'approche actuellement adoptée par certains gros modèles de langue populaires de type *denoising autoencoder*, e.g. BART [170]. Nous en reparlerons page 95.

permettons grâce à cette nouvelle formulation de ne pas simplement tenir compte des conséquences logiques exactes (déductives) exprimées dans la représentation des connaissances sous-jacente lors de l'évaluation de l'IC, mais aussi de considérer une hypothèse inductive qui nous semble rationnelle - et qui fait référence à une notion d'évocation conceptuelle.

En effet, les deux approches intrinsèque et extrinsèque existantes, rappelées ci-dessus, ne traduisent pas selon nous d'aspects importants pour l'évaluation de la quantité d'information portée par un concept. Ces modélisations ne tiennent pas compte en particulier des évocations amenées par un concept, autres que celles sous-tendues par une approche déductive simple du type, parler d'Algèbre implique de parler de Mathématiques. Les évocations implicites qui reposent sur des mécanismes d'inférence inductifs ne sont pas considérées par ces modèles, ce qui peut être limitant dans certains contextes applicatifs.

Prenons un exemple. Admettons que vous sachiez qu'une personne dispose de 1000 livres dans sa bibliothèque, chaque livre étant annoté par un unique concept. 900 livres sont annotés par le concept Mathématiques, i.e. qualifiés de manière explicite comme des livres de Mathématiques. Parmi les 100 restants nous avons 50 livres sur l'Antiquité, et 50 livres dits d'Algèbre (ce qui nous fait donc 950 livres de Mathématiques au total car nous considérons ici Algèbre \square Mathématiques). Les deux concepts Algèbre et Antiquité obtiendront la plupart du temps la même valeur d'IC si l'on se base sur l'approche de calcul usuelle. Nous estimons cependant que, dans certains contextes, l'IC du concept Antiquité mériterait d'être apprécié comme plus important que celui du concept Algèbre. Imaginons en effet que l'observation des concepts traduise des centres d'intérêt de la personne. L'Algèbre semblait être un centre d'intérêt très probable au-delà des observations explicites de ce dernier (i.e. 900 livres de Mathématiques); a contrario le centre d'intérêt sur l'Antiquité expose lui une nouvelle facette du profil et s'avère donc nettement plus informatif pour qualifier la personne, notamment dans des contextes de recherche de diversité - dans le sens où si nous devons ajouter un concept supplémentaire à celui de Mathématiques pour qualifier le profil, nous préférons dans beaucoup de contextes le concept Antiquité à celui d'Algèbre.

Afin de modéliser cette notion d'inférence inductive nous avons fait le choix de nous baser sur le cadre des fonctions de croyance [234].¹² Ce travail nous a par ailleurs permis de formuler dans un même cadre les IC extrinsèques traditionnels et ceux implémentant la notion d'inférence inductive que nous avons étudiée.¹³ Les résultats obtenus dans nos expérimentations traduisent empiriquement la pertinence de la proposition sur des jeux de données établis du domaine. Il convient bien entendu de préciser que la modélisation de cette hypothèse d'inférence inductive est sujette à discussion et que sa prise en compte - comme son exclusion - ne conviendrait pas à tous les contextes applicatifs.

12. Nous proposons d'explorer les premières formulations d'IC reposant (i) sur la notion de *plausibilité* - au sens des fonctions de croyance -, (ii) sur une formulation convexe mariant IC traditionnels (croyance) et IC se basant sur la plausibilité, ou encore (iii) sur la notion de probabilité pignistique.

13. Nos travaux exposent une erreur dans la formulation probabiliste originelle de la notion même d'IC ([226] cité plus de 5k fois), à notre connaissance non mentionnée auparavant dans l'état de l'art, cf. [113].

Comme nous le verrons, plusieurs de nos travaux feront référence à des propagations dans des taxonomies, de manière semblable à ce qui peut être fait dans le calcul d'IC. Cette modélisation à l'aide des fonctions de croyance sera aussi retrouvée dans différentes approches que nous avons proposées. Cela sera en particulier le cas, parmi les travaux discutés dans ce manuscrit, de ceux que nous exposerons sur le résumé sémantique [185], la vérification d'assertions [25], l'extraction de règles [155], ou l'inférence de connaissances [133].

2.1.4 Regard critique sur nos travaux et évolution du domaine

L'objectif n'est pas ici de faire une synthèse des critiques formulées sur nos travaux dans la continuité de celles exposées dans notre thèse. Nous présentons seulement quelques retours d'expérience forgés par l'épreuve du temps qui nous semblent intéressants. Nous proposons aussi une analyse de l'évolution du domaine de recherche. Nous soulignons à nouveau que post 2015, le choix a été fait d'élargir notre positionnement recherche et de ne pas centrer plus encore nos travaux pendant plusieurs années sur la *seule* notion de mesure sémantique - aussi vaste soit-elle. Aussi nous ne prétendons pas avoir une pleine connaissance et maîtrise de l'état de l'art des sept dernières années sur ce sujet, bien que notre curiosité et notre devoir professionnel nous invitent annuellement à avoir un regard sur les développements importants du domaine.

Les limites des cadres unificateurs.

Dans la continuité de travaux initiés par d'autres, et en particulier de ceux proposés par Blanchard et collaborateurs [33, 32], nous avons proposé un cadre unificateur qui souligne notamment le fait que beaucoup de « *nouvelles mesures* » - vocable utilisé pour les introduire dans la littérature - fondées sur des ontologies légères, ne comportent pas dans l'absolu de véritable originalité. Nous n'avions bien entendu pas espéré que le cadre proposé, largement perfectible, soit jugé définitif et susceptible d'étudier et d'exprimer toutes les mesures qu'il concerne. Nous espérions cependant qu'il contribue à stimuler et à encourager des travaux que nous estimions importants pour la communauté (e.g. sur les propriétés des mesures). Force est de constater que cet objectif n'a pas été atteint (malgré qu'il soit fréquemment cité).

Une piste d'explication du *faible* intérêt exprimé pour les cadres unificateurs réside probablement en partie dans la diversité des communautés et des profils de personnes intéressées par les mesures sémantiques. Certains de ces profils n'ont peut-être tout simplement pas la possibilité d'apprécier des initiatives unificatrices impliquant des cadres plus abstraits, et de composer avec l'ajout de complexité que leur utilisation induit (manque de temps, centres d'intérêt et objectifs disciplinaires différents. . .). Une autre cause de la faible adoption des cadres théoriques unificateurs existants réside aussi dans leur incapacité à englober la diversité des mesures, par exemple celles nécessitant la prise en compte de spécificités propres à la problématique traitée (e.g. prise en compte de contextes particuliers, e.g. [79, 45]), celles hybridant différentes informations (e.g. taxonomiques, issues de la ABox, définitions logiques. . ., comme par exemple dans [275]), où celles discutées ci-après découlant des travaux récents sur les modèles neuronaux. Des cadres unificateurs capables d'englober une diversité plus large de mesures, e.g. basés sur la notion aujourd'hui devenue centrale de représentation - au sens de plongement sémantique -, seraient très probablement plus prolifiques.

De la difficulté de fédérer autour d'une solution logicielle.

Maintenir une librairie logicielle est une entreprise complexe et très chronophage. Difficile ainsi, sans support d'Ingénierie, de jouer le rôle de *lead developer* et de *community manager* dans le temps. Il est aussi souvent plus simple, plus stimulant, et plus gratifiant de proposer une nouvelle alternative logicielle que de contribuer à un projet existant. Plusieurs raisons peuvent l'expliquer (maîtrise du code, du langage de programmation utilisé, ticket d'entrée associé à la prise en main d'une base de code importante). Nous pensons que la structuration en modules de la base de code (conséquente) supportant la librairie SML¹⁴ n'a pas contribué à faciliter l'intégration de nouveaux développeurs. Nous ne sommes pas non plus convaincus que l'utilisation du langage de programmation Java se soit avérée judicieuse dans le temps pour ce type de développements. On préférera très probablement maintenant des interfaces de programmation et du code Python associé éventuellement à du code C ou C++ pour les traitements les plus coûteux en temps de calcul.

Mesures sémantiques et apprentissage machine

La littérature sur les mesures sémantiques n'a bien entendu pas échappé aux développements récents en apprentissage machine et en apprentissage profond. De nombreux travaux ont ainsi été proposés afin d'intégrer de l'apprentissage machine dans les phases de construction des représentations des objets comparés, ou dans le calcul même de la mesure sémantique.

La taille réduite des jeux de données dédiés aux mesures sémantiques ne permet pas d'envisager l'apprentissage de représentations, ou directement de mesures, en se basant uniquement sur des données labélisées, e.g. les jeux de données les plus volumineux ne contiennent que quelques milliers, voire centaines de paires de mots associées à un score de similarité [117] [46]. Des exceptions sont bien entendu rencontrées dans certains contextes dans lesquels des données labélisées sont générées à l'aide de jeux de données existants de grande taille (dans le contexte de l'évaluation de la similarité de phrases, des jeux de données de paraphrases peuvent être utilisés pour générer des données labélisées [270, 142]). On retiendra cependant que, bien que l'apprentissage machine puisse être utilisé pour apprendre des mesures [243, 244], il intervient aujourd'hui la plupart du temps dans le calcul de représentations qui seront ensuite comparées à l'aide de mesures *ad hoc*.

Les développements récents de l'apprentissage profond sont étroitement liés à l'apprentissage de représentations. Ils ont donc naturellement eu une influence importante sur l'estimation de mesures sémantiques ; comme nous l'avons souligné, les sujets de la représentation des objets comparés, et de l'estimation de mesures sémantiques pour ces mêmes objets sont étroitement liés, s'ils peuvent être considérés comme distincts. Nous traiterons les développements en lien avec les mesures sémantiques à base de connaissances dans un second temps. Concentrons-nous tout d'abord sur les mesures portant sur des unités de langage telles que des mots, des phrases ou des textes.

14. <https://github.com/sharisper/slib/>

Les succès des développements récents que nous observons dans le TALN s'expliquent en grande partie par l'essor de l'implémentation de la sémantique distributionnelle et compositionnelle, en particulier à l'aide d'approches neuronales.¹⁵ Ces développements s'inscrivent dans la continuité d'une longue *tradition* en TALN et en Recherche d'Information notamment. En effet, les constructions de représentations vectorielles étaient déjà la règle pour la définition de mesures sémantiques à base de textes avant l'avènement de l'apprentissage profond. Les approches traditionnelles reposent sur des procédures arbitraires et *ad hoc*, e.g. à base de modèles vectoriels de type VSM, des modèles de type n-grams, ou de dérivés de ces types de modèles reposant sur une analyse fréquentiste des termes, e.g. Latent Semantic Analysis, Latent Semantic Indexing, Okapi BM25 (nous ne traiterons pas dans le détail cette vaste littérature [233]). Ces approches reposaient (déjà) sur une implémentation explicite ou implicite de l'hypothèse distributionnelle qui se base sur l'analyse de cooccurrences ; elles n'abordent cependant pas la tâche de construction des représentations comme une tâche d'apprentissage. Seuls quelques précurseurs, dont les travaux n'étaient pas très populaires il y a même encore dix ans en TALN, exploraient déjà il y a plus de vingt ans l'idée d'introduire un apprentissage explicite des représentations au niveau des mots [19] - l'apprentissage machine pour l'apprentissage de représentations de documents était déjà plus populaire, e.g. cf. Latent Dirichlet Allocation (LDA) [34].¹⁶ Cette approche gouvernée par de l'apprentissage de représentations s'est graduellement imposée.

Depuis le succès rencontré dans les années 2010 par les développements proposés par Collobert et al. [56] étendus par Mikolov et al. (cf. CBOW et Skip-gram, 2014 [193]), beaucoup de travaux ont porté sur l'utilisation de réseaux de neurones pour l'apprentissage de représentations vectorielles de mots à partir de larges corpus de textes [46]. Word2vec [193], GloVe [214], FastText [35], Transformer [265] et BERT [73] ont été des étapes marquantes de la première phase de cette littérature. A noter que chaque approche produit des plongements statiques ou contextuels - ces derniers étant plus riches que les statiques (e.g. gestion de la polysémie, représentation de n-grams) [258], même s'ils sont moins simples à réutiliser lorsque l'on souhaite par exemple comparer des mots sans contexte prédéfini. Ces approches se sont généralisées à l'apprentissage de représentations d'unités de langue plus complexes, e.g. phrases, textes. Elles reposent sur une gamme variée d'architectures de réseaux de neurones susceptibles d'encoder des informations contextuelles dans l'idée d'implémenter la notion de sémantique distributionnelle et compositionnelle, e.g. réseaux récurrents, Transformers, CNN (1D), auto-encodeurs bruités, le tout très souvent dans des contextes auto-supervisés, e.g. la tâche de prédiction de mots masqués étant souvent préférée aujourd'hui.

Cette approche de construction de représentations à l'aide de modèles neuronaux qui implémentent la sémantique distributionnelle et compositionnelle représente la rupture de ces dernières années dans le domaine d'étude des mesures sémantiques ; elle a plus généralement tout simplement révolutionné le TALN. Ces modèles neuronaux sont en effet très souvent nettement plus performants que les approches traditionnelles pour obtenir de *bonnes* représentations lorsque des jeux de données

15. Nous ne détaillons pas les aspects liés à l'approche compositionnelle pour nous focaliser sur l'aspect distributionnel ; ils sont cependant tout aussi importants, cf. travaux sur les approches neuronales de la sémantique compositionnelle [31, 195, 174, 75]. Le lecteur notera que des travaux récents questionnent le fait que les plongements obtenus à l'aide d'approches neuronales implémentent à proprement dit la sémantique distributionnelle [192]. Cette question étant ouverte, nous nous conformons ici à la vision consensuelle simplifiée, et peut-être galvaudée, qui consiste à admettre que toute forme d'analyse contextuelle de type analyse de cooccurrences s'inscrit dans le cadre d'une implémentation de l'hypothèse distributionnelle.

16. A noter que certaines approches qui n'utilisent pas explicitement un paradigme de l'apprentissage machine peuvent en arriver à extraire des informations semblables à celles extraites par de l'apprentissage. La distinction entre les approches à base d'apprentissage machine et les autres n'est, de plus, pas évidente dans certains cas.

importants sont constitués, e.g. cela est même aujourd’hui parfois le cas en recherche d’information, domaine probablement le plus prolifique à ce jour en termes de propositions de représentations d’unités linguistiques complexes [67]. Les approches traditionnelles, quoique souvent déclassées, restent bien entendu intéressantes lorsque peu de données labélisées sont disponibles et qu’il est impossible d’engager des modèles pré-entraînés (ce qui est encore le cas dans certains contextes appliqués). Des couplages intéressants avec des modèles plus traditionnels sont aussi étudiés, e.g. couplage avec des *topic models* [213] ou avec des approches de types TF-IDF [69].¹⁷ Les techniques d’utilisation de modèles pré-entraînés et d’apprentissage par transfert évoluant, la place qu’occupent les approches neuronales pour la construction des représentations ne cesse de croître. Cela représente une véritable rupture dans l’état de l’art.

Les performances obtenues par les approches neuronales à la fois sur du texte mais aussi sur d’autres modalités (e.g. images, voix) ont naturellement invité les chercheurs à étudier leurs applications pour la construction de représentations d’objets définis dans des modèles de connaissances (ontologies, graphes de connaissances. . .). Le caractère générique du cadre de l’apprentissage machine profond qui se base sur l’apprentissage (auto-supervisé) de représentations - conceptuellement et dans une certaine mesure de manière agnostique au type de données - le rend aussi très attractif. Qu’en est-il de l’usage de l’apprentissage machine dans la définition de mesures sémantiques à base de connaissances ? Il est lui aussi croissant et offre des perspectives séduisantes en termes d’amélioration des performances.

Il est notamment intéressant de noter que la plupart des mesures récentes à base de connaissances ne reposent plus sur des formulations arbitraires de similarités ou de distances qui reposent sur une représentation *ad hoc*, e.g. distance entre des noeuds d’un graphe, ou IC du MICA des concepts comparés. Nous observons là aussi une utilisation croissante des techniques d’apprentissage de représentations, i.e. plongement des éléments que l’on souhaite comparer, e.g. concepts, ou ressources d’un graphe RDF [151]. Ces plongements sont réalisés à l’aide de différentes techniques de plongements de graphes (*graph embedding*), parfois paramétrées de manière contextuelle [106, 244]. En effet, dans la continuité des travaux précurseurs de Hinton et al. sur l’encodage de représentations symboliques [123], de nombreux efforts ont été fournis pour encoder des informations symboliques structurées dans des graphes. Plusieurs techniques de plongement de graphes ont ainsi été proposées [106], e.g. RDF2Vec [228], TransE [38], ditMult [281]. Des techniques permettent aussi d’encoder des informations de représentations des connaissances plus riches, e.g. de type ontologies [51]. Ces approches permettent la construction de représentations vectorielles qui pourront servir l’estimation de scores de mesures sémantiques. Les résultats obtenus dans certaines études soulignent que les similarités (taxonomiques) traditionnelles, e.g. à base d’analyse d’IC, permettent dans certains cas l’obtention de meilleures performances que les mesures qui reposent sur la comparaison de représentations construites par des plongements de graphes [244]. La tendance de l’état de l’art est tout de même claire : l’essentiel des forces vives qui étudient ce sujet l’aborde sous l’angle de l’apprentissage machine.

17. Des travaux sur la nature des informations encodées par ces nouvelles approches par rapport aux approches traditionnelles de type TF-IDF, ou plus généralement issues de la Recherche d’Information, sont aussi menés, e.g. BERT [73] semble dans une certaine mesure encoder, entre autres, des informations semblables à l’IDF [54]. Notons aussi que des travaux caractérisent par ailleurs théoriquement les différences observées entre des représentations TF-IDF et celles obtenues via des techniques de plongement plus récentes, e.g. [66]. Ces résultats sont de plus à analyser au regard d’observations empiriques qui soulignent dans certains cas les différences importantes de comportements de différents modèles récents pré-entraînés et pourtant proches, e.g. sur du français [96].

Les performances obtenues par les approches neuronales sur du texte laissent difficilement imaginer que l'on puisse aujourd'hui encore aborder l'estimation de similarité ou de proximité sémantiques de manière complètement isolée des approches textuelles lorsqu'une hybridation est possible, e.g. pour comparer des unités de linguistiques. A titre d'exemple, les approches qui se basent uniquement sur WordNet sont aujourd'hui très généralement déclassées pour comparer des mots, et naturellement les hybridations sont généralement préférées, e.g [129]. Il reste cependant des cas pour lesquels des mesures traditionnelles sur des représentations des connaissances ou lexicales seront préférées. Celles-ci bénéficient entre autres de l'avantage d'explicitier des termes ou concepts de manière désambiguïsée, ce qui n'est pas le cas des représentations construites par les techniques d'apprentissage profond [46]. Les approches textuelles et à base de connaissances méritent d'être considérées comme complémentaires. Lorsque des hybridations sont possibles, des travaux ont souligné la complémentarité des informations contenues dans les différentes sources d'information/connaissance. Les représentations des connaissances peuvent de plus être utilisées pour encoder des contraintes ou relations axiomatiques qu'il serait difficile de voir encoder dans des représentations vectorielles issues de l'analyse de textes [240, 213, 129]. La prise en compte d'une expression de connaissance intelligible par l'Homme est aussi intéressante pour faciliter l'interprétabilité des résultats obtenus [239], e.g. le score obtenu peut plus facilement être analysé.

On note ainsi globalement une forte adoption des techniques à base d'apprentissage de représentations à l'aide de modèles neuronaux. Cette dynamique contribue à une forme de normalisation des travaux sur l'étude des mesures sémantiques et leur réorientation vers l'étude de la représentation des objets comparés. Il est intéressant de noter une fois de plus que l'effort de modélisation effectué sur le calcul de la mesure sémantique à proprement dit, s'est quasi-exclusivement déporté sur le calcul des représentations à partir desquelles sont calculés les scores des mesures sémantiques. Comme mentionné ci-dessus, les représentations sont généralement ensuite comparées à l'aide de distances standard, e.g. simple produit scalaire normalisé, similarité cosinus, norme L_p de la différence des vecteurs. Des exceptions existent bien entendu, e.g. définitions originales de distances entre textes à l'aide du problème du transport formulé sur les plongements des termes qui composent les textes [153, 72].

Rares sont aujourd'hui les contributions du domaine qui traitent de représentations autres que des représentations vectorielles, i.e. le modèle spatial prédomine très largement par rapport aux modèles à base de caractéristiques (*feature-based model*) ou d'autres modèles issus des Sciences Cognitives comme les modèles à base de transformation [117]. Cette normalisation a des avantages évidents car elle permet la transposition de contributions d'un sujet d'étude à un autre, indépendamment des spécificités des proxies sémantiques considérés pour le calcul des mesures (bases de connaissances, corpus de textes, couplage des deux). Elle s'accompagne aussi du désavantage de restreindre la recherche du domaine, en concentrant et focalisant par exemple la discussion de la représentation des objets comparés sur la construction de représentations vectorielles. En effet, ces développements sont généralement réalisés au détriment de l'étude d'approches alternatives de l'état de l'art, e.g. représentation à base de caractéristiques – *feature-based* - qui quoique marginales aujourd'hui, restent encore présentes et pertinentes pour des formulations *ad hoc* de mesures. Il est dans tous les cas clair que la transformation opérée ces dernières années par l'apprentissage machine - qui s'annonce chaque jour plus profonde et durable - a largement impacté l'étude des mesures sémantiques.

2.2 Exploitation de connaissances formelles dans des processus d'aide à la décision

Nous avons mené différentes recherches sur l'exploitation de connaissances formelles dans des processus d'aide à la décision, voire des processus décisionnels semi-automatisés. Elles ont notamment visé à étudier l'utilisation de représentations des connaissances légères de types taxonomies comme sources de connaissances *a priori* pour la réalisation de traitements variés. Nous centrerons notre exposé sur les travaux réalisés dans le cadre des finalités suivantes :

- La production de résumés sémantiques d'annotations terminologiques et conceptuelles.
- L'estimation de la qualité de sources d'information et la vérification d'assertions.
- La définition de modèles d'inférence en traitement automatique du langage naturel.
- La découverte de règles d'association à partir de données imparfaites.

Ces contributions s'inscrivent la plupart du temps dans des contextes de découverte de connaissances (inférence de faits et découverte de règles). Elles ont parfois été proposées pour des applicatifs spécifiques, e.g. analyse de corpus, analyse psychosensorielle, mise en place d'un outil de retour d'expériences (RetEx). Nous exposerons à la fois les contributions fondamentales et les applications que nous avons étudiées.

2.2.1 Résumé d'annotations terminologiques et conceptuelles

Cette section traite de la difficile question du résumé ou de la synthèse d'informations. Nous sommes souvent confrontés à la tâche complexe consistant à résumer l'information portée par un ensemble de données. Le rôle d'un expert ne consiste-t-il pas en effet très souvent à savoir faire bonne synthèse d'une grande quantité d'informations dans un contexte spécifique et au regard d'une connaissance métier ?

Des Statistiques à la Visualisation de données, un grand nombre d'outils théoriques et pratiques nous permettent d'ores et déjà de synthétiser un ensemble de données à l'aide d'indicateurs quantitatifs ou de représentations graphiques. La diversité et les spécificités des nombreux travaux menés sur ce sujet s'expliquent très souvent par la nature des données que l'on souhaite synthétiser, e.g. listes de valeurs numériques, collections de textes. L'état de l'art du domaine est du reste plus ou moins fourni et varié en fonction de la nature des données d'intérêt. Une majorité de travaux s'intéressent ainsi à l'étude de données numériques. Nous notons cependant que, compte tenu du rôle central que jouent le langage et les communications écrites et orales dans nos sociétés, il est étonnant que si peu d'outils ne soient disponibles pour synthétiser les données à caractère sémantique.¹⁸

S'intéresser à la problématique de production de résumés invite naturellement à définir ce que l'on appellera *bon* résumé. Le problème n'est pas simple pour les données numériques¹⁹ ; il est néanmoins nettement plus complexe dans le cadre du traitement de données sémantiques. Chacun verra ici s'immiscer le spectre de la subjectivité en ce qui concerne l'appréciation de la bonne nature contextuelle d'un résumé, et l'appréciation qualitative ou quantitative de sa qualité : le point de difficulté principal est ici identifié. Ce point contribue très probablement à expliquer la littérature plus modeste dédiée au vaste sujet du résumé d'informations à caractère sémantique.²⁰

Des technologies intéressantes de résumé sémantique de textes ont connu des développements importants ces dernières années, notamment en lien avec l'essor de l'apprentissage profond [143]. Elles visent à préserver l'important dans la sémantique portée par un texte en s'attachant naturellement à réduire son nombre de mots. Nous ne discuterons pas ici ces développements pour nous centrer sur l'exposé de nos contributions en lien avec une problématique spécifique de ce vaste domaine.

18. Nous entendons par *données à caractère sémantique* toute information susceptible d'être représentée par un texte composé d'un ou plusieurs mots (vu ici comme un élément d'une langue, et donc porteur de sens). Ce texte peut être le produit d'une transcription audio en texte et engager des structures plus ou moins complexes propres à une langue (conjugaisons, construction de phrases. . .).

19. e.g. en admettant que l'on aborde la problématique sous l'angle de l'étude d'indicateurs statistiques ; au-delà des limites de leur capacité de représentation, les estimateurs des simples indicateurs de position (e.g. moyenne) ou de dispersion (e.g. variance) peuvent comporter des biais.

20. Pour l'anecdote, à l'époque de ces travaux, nous avons eu une discussion intéressante avec un directeur de laboratoire en Informatique et en IA, qui exprimait un scepticisme marqué quant à la pertinence de mener des travaux sur un sujet si peu défini, et dont la formalisation inclurait nécessairement une part de subjectivité. Nous nous accordions alors sur le fait qu'il est nettement plus confortable de considérer un système formel admis, pour lequel le vrai - synonyme du bon - n'est pas sujet à discussion. L'entreprise que nous visions était en effet bien détachée de ce confort. La position de l'auteur a par ailleurs nettement évolué depuis.

Problématique. Nous souhaitons en particulier traiter le cas spécifique du résumé sémantique (i) d'un ensemble de concepts d'une taxonomie, et (ii) de termes issus d'un vocabulaire contrôlé. La tâche de résumé était traitée de manière décontextualisée : seules les annotations à résumer et une représentation des connaissances étaient alors disponibles. La tâche de résumé avait pour objectif de produire des synthèses sous la forme d'ensembles pondérés de concepts définis dans la représentation des connaissances considérée.

Nous le verrons par la suite, ces travaux ont plus précisément amené la définition d'approches algorithmiques capables de synthétiser des annotations terminologiques ou conceptuelles par l'analyse d'une représentation des connaissances simple de type ontologie légère structurant un ensemble de concepts sous la forme d'une taxonomie.

Collaborateurs :

- Doctorante encadrée : Massissilia Medjkoune (2015-2018).
- Stagiaires encadrés : Jie Hi (2016), Joyss Rakotonirina (2018).
- Ingénieur encadré : Joël Dichamp (2018).
- Collaborateurs d'IMT Mines Alès : Jacky Montmain, Pierre-Antoine Jean, Sylvie Ranwez, Nicolas Fiorini, Jean-Louis Fanlo (Chimiste), Stéphane Cariou (Chimiste).
- Collaborateur externe : Vincent Ranwez (Montpellier SupAgro).

Les travaux mentionnés dans cette section ont pour l'essentiel été menés dans le cadre de la thèse de Massissilia Medjkoune [185], et dans le cadre du projet CARNOT M.I.N.E.S CARATS I3AS (Intelligence Artificielle pour l'Automatisation d'Analyses Sensorielles) dont l'auteur était responsable scientifique.

Bien que nous admettions de traiter des ensembles d'annotations pondérés, nous traiterons ici le cas non pondéré - le lecteur intéressé par la généralisation se référera à nos travaux sur ce sujet, e.g. [115, 185].²¹ Voici un premier exemple évident de résumé : préciser que l'auteur de ce manuscrit répond à l'ensemble de concepts {Personne, Homme} n'est en soi pas plus informatif que de simplement dire que c'est un Homme ; en effet, si nous admettons $\text{Homme} \sqsubseteq \text{Personne}$, une synthèse correcte de {Personne, Homme} serait alors le singleton {Homme}. Même si cet exemple est trivial, le problème ne l'est pas dans l'absolu : quel est par exemple le résumé de {Vanille, Rose, Champignon, Bois} ? Progresser dans l'automatisation de la résolution de ce problème nécessitera, comme nous l'avons mentionné, d'affronter la très souvent fâcheuse question de la subjectivité. Toute solution ou angle d'attaque de la problématique, repose selon nous sur une définition *ad hoc* explicite ou implicite de ce que l'on entend par *bon* résumé. Deux approches non orthogonales seront intuitivement préférées. La première repose sur la définition d'un ensemble de contraintes requises pour exclure des résumés que l'on estime *a priori* non pertinents, e.g. *redondance* d'information. La seconde repose elle sur la considération d'une fonction objectif qui traduit directement ou indirectement la qualité d'un résumé, e.g. en fonction, entre autres, de sa taille. Nous proposons dans nos travaux une approche permettant, via la définition d'une fonction d'évaluation paramétrique, une quantification arbitraire de la qualité d'un résumé. Cette fonction objectif sera utilisée pour identifier des résumés pertinents par l'application d'une procédure d'exploration de l'espace de recherche des résumés.

21. Un article qui traite de ce sujet est joint au manuscrit, cf. Annexe A page 135 [115].

Contexte applicatif : nos contributions prennent en compte des structures de données, représentations des connaissances, et procédures génériques qui les rendent en soi agnostiques d'un applicatif spécifique. Elles ont cependant été proposées et évaluées dans un cadre spécifique : la mise en place d'une approche automatisée de l'analyse psychosensorielle lors de l'analyse d'odeurs. En effet, caractériser la qualité d'une odeur est une tâche complexe, chronophage et coûteuse pour un industriel. Cette tâche vise très généralement à identifier des descripteurs (concepts) en mesure de synthétiser au mieux la sensation olfactive ressentie par des évaluateurs experts au cours de séances d'analyse sensorielle. Les caractérisations des odeurs faites par ces experts sont exprimées via des descripteurs généralement définis dans un vocabulaire contrôlé, parfois structuré, qui est imposé par les industriels du domaine d'analyse concerné, e.g. une roue des odeurs. L'approche est semblable pour d'autres types d'analyses psychosensorielles. La Figure 1 propose un exemple de roue utilisée dans le cadre de dégustations de vins. Un expert en charge de la synthèse d'information récolte donc les annotations produites par différents experts en charge d'évaluer les produits d'intérêt, et les synthétise manuellement dans un résumé composé d'un ensemble réduit de descripteurs du référentiel utilisé. Un exemple de résumé est proposé en figure 2.

Cette approche de l'analyse psychosensorielle est dite orientée car elle repose sur (i) la subjectivité de l'opérateur en charge de produire le résumé, et (ii) l'apprentissage chronophage et technique d'un vocabulaire spécialisé par les évaluateurs experts de domaine. Ces deux aspects empêchent respectivement le traitement d'importants volumes d'évaluations et la mise en place d'analyses psychosensorielles incluant des cohortes de non-initiés. Afin de lever ces verrous, nous nous sommes intéressés au résumé automatisé d'ensembles de termes définis ou non dans un vocabulaire contrôlé et structuré. Il est intéressant de souligner qu'il était difficile pour les experts métier en charge de la production des synthèses ne serait-ce que de préciser les qualités d'un bon résumé ; l'approche adoptée par nos collaborateurs experts métier était nourrie d'analyses descriptives élémentaires intégrant l'analyse du référentiel utilisé, une forte expertise, et même parfois disons-le, une belle dose d'intuition.

Résumer un ensemble d'annotations conceptuelles

Nous souhaitons résumer un ensemble d'évaluations conceptuelles proposées par un ensemble d'évaluateurs. Nous notons C l'ensemble de concepts - nos descripteurs -, $\mathcal{O} = (\preceq, C)$ leur structuration sous la forme d'une taxonomie, et E l'ensemble d'évaluateurs ($n = |E|$). Chaque évaluateur $e_i \in E$ ($i \in \llbracket 1, n \rrbracket$), est associé à un ensemble de concepts $X_i \in \mathcal{P}(C)$. Nous noterons la séquence d'annotations à résumer par $\hat{X} = (X_1, X_2, \dots, X_n)$, et X l'ensemble de concepts de cette séquence : $X = \bigcup_{i=1}^n X_i$.

Nous nous sommes intéressés à une forme simple de résumé en étudiant les formulations du type $f : \hat{X} \rightarrow \mathcal{P}(C)$. Nous avons pour cela distingué un ensemble de propriétés qui nous semblaient consensuelles pour caractériser un résumé $Y \in \mathcal{P}(C)$ d'un ensemble de concepts X (en anglais) :

1. *Summarizing* : $|Y| \leq |X|$, avec le plus souvent $|Y| \ll |X|$. Un résumé doit avoir une représentation de taille inférieure ou égale à la représentation de l'information qu'il synthétise. La taille d'une représentation est ici définie en nombre de concepts.

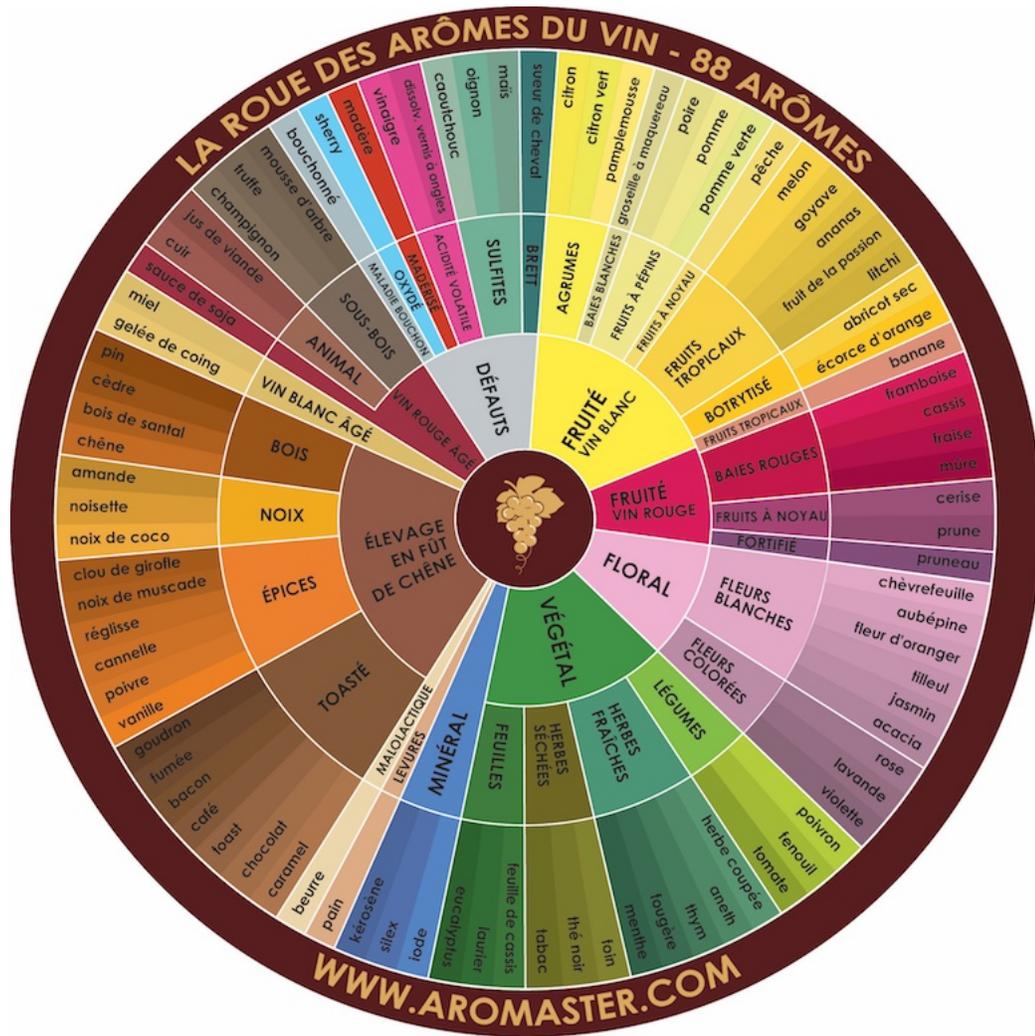


FIGURE 1 Exemple de roue utilisée en analyse psychosensorielle [84].

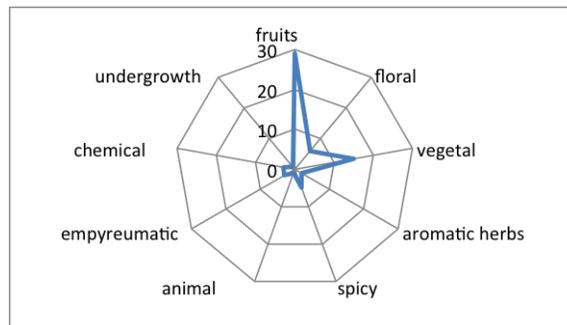


FIGURE 2 Exemple de radar utilisé en analyse psychosensorielle. Le poids donné à chaque dimension (concept de l'ontologie) correspond au nombre d'occurrences du concept ou d'une de ses spécialisations. Le résumé attendu par l'expert était ici {Fruit, Vegetal, Floral} [115].

2. *Faithful* : $\forall y \in Y, \exists x \in X$ tel que $x \preceq y$. Nous estimons que le résumé doit se faire *sans création d'information*, i.e. tout concept du résumé doit correspondre de manière explicite, au sens de \mathcal{O} , à une annotation fournie par un évaluateur.²²
3. *Non-total redundancy* : $\forall (x, y) \in Y^2, x \not\prec y \wedge y \not\prec x$. Deux concepts d'un résumé ne doivent pas être ordonnés dans \mathcal{O} car nous estimons que cela induit une redondance d'information.

Ces propriétés découlent de choix arbitraires motivés par l'objectif visé ; ils caractérisent notre modélisation de la problématique. Les propriétés de fidélité (*faithfulness*) et de non-redondance (*non-total redundancy*) peuvent en particulier être sujettes à discussion. Il pourrait par exemple être intéressant de s'autoriser à enfreindre la propriété de *faithfulness* afin de recouper des informations portées par des annotations, et alors intégrer dans un résumé des concepts qu'elles ne mentionnent pas explicitement. Cela sera notamment modélisé dans les techniques que nous étudierons pour le résumé d'ensembles de termes.

A noter que notre modélisation autorise la perte d'information par abstraction, mais aussi par omission. Il est en effet possible de ne pas intégrer la totalité des informations portées par toutes les annotations, i.e. même si nous admettons la propriété de fidélité, nous n'admettons pas la *couverture totale*, i.e. $\forall x \in X, \exists y \in Y$ tel que $x \preceq y$. Cela a un intérêt lorsque les annotations sont pondérées, pour s'autoriser à exclure des concepts qui correspondraient à des valeurs aberrantes. La notion de non redondance totale invite elle à exclure du résumé des concepts d'ores et déjà mentionnés explicitement ou potentiellement respectivement par spécialisation ou généralisation.²³ Certains contextes préféreront autoriser des formes de redondance d'information pour souligner certains aspects, e.g. préciser pour un produit : {Agrume, Citron}.

Nous notons $\mathcal{S} \subseteq \mathcal{P}(C)$ l'ensemble des sous-ensembles de C qui vérifient la propriété de *non-total redundancy*, et $\mathcal{S}^X \subseteq \mathcal{S}$ l'ensemble des résumés de la séquence d'annotations \hat{X} qui vérifient les propriétés de *summarizing* et *faithful*. Nous nous sommes alors attachés à l'étude d'une fonction de résumé S visant à résumer une séquence de n annotations $\hat{X} \in \mathcal{P}(C)^n$ par un résumé de \mathcal{S}^X :

$$S : \mathcal{P}(C)^n \rightarrow \mathcal{S}, \text{ avec } S(\hat{X}) \in \mathcal{S}^X$$

Le problème consistant à résumer une séquence d'annotations \hat{X} est alors abordé sous l'angle de la recherche de l'élément $Y \in \mathcal{S}^X$, qui sera qualifié de *meilleur* résumé de \hat{X} . Nous précisons quelques propriétés des espaces que nous manipulons avant de nous attarder sur l'appréciation qualitative de la qualité d'un résumé.

Un résumé $Y \in \mathcal{S}^X$ est dit couvrant d'une séquence d'annotations \hat{X} si tout concept de X est abstrait par au moins un élément du résumé : $\forall x \in X, \exists y \in Y$ tel que $x \preceq y$, i.e. respect de la propriété de couverture totale mentionnée ci-avant. Les résumés de \mathcal{S}^X couvrants de \hat{X} sont notés \mathcal{S}_{cov}^X . La figure 3 illustre les espaces de recherche mentionnés. Notons que nous pouvons seulement garantir $X \in \mathcal{P}(C)$ bien qu'il soit possible d'avoir $X \in \mathcal{S}_{cov}^X$.

22. Avec $x, y \in C$ et $x \preceq y$, nous considérons que x mentionne explicitement x et y .

23. On ne peut ajouter Agrume au résumé {Citron} car Citron mentionne explicitement Agrume, Citron \prec Agrume. Impossible aussi d'ajouter Citron au résumé {Agrume} car Agrume mentionne potentiellement, i.e. de manière plausible, Citron.

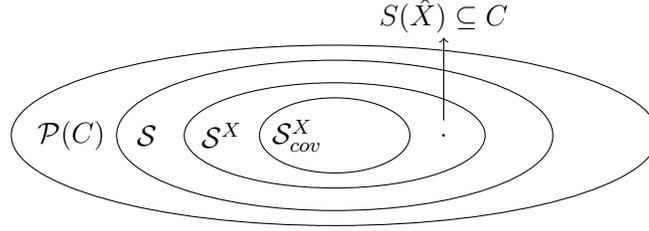


FIGURE 3 Représentation des espaces de recherche mentionnés. $\mathcal{P}(C)$ parties de C ; les éléments de $\mathcal{S} \subseteq \mathcal{P}(C)$ respectent la *non-total redundancy*; les éléments de $\mathcal{S}^X \subseteq \mathcal{S}$ respectent les propriétés de *summarizing* et *faithful* vis-à-vis de $X \subseteq C$; les éléments de \mathcal{S}^X_{cov} respectent de plus la *couverture totale* vis-à-vis de X .

De manière détachée du fait que les notions de pertinence et de qualité d'un résumé soient subjectives et potentiellement dépendantes d'un contexte, nous avons proposé quelques quantités qui nous paraissent d'intérêt pour les étudier. Notre approche de recherche de résumés repose finalement sur la définition d'une fonction objectif *ad hoc* qui mobilise l'évaluation de ces quantités :

$$S(\hat{X}) = \arg \max_{Y \in \mathcal{S}^X} \left(\Psi(Y, \hat{X}) - \mathcal{L}(Y, \hat{X}) \right) \quad (2.1)$$

$\Psi(Y, \hat{X})$ modélise la quantité d'information portée par \hat{X} représentée par Y et $\mathcal{L}(Y, \hat{X})$ explicite une pénalité imputable à l'abstraction induite, avec :

$$\mathcal{L}(Y, \hat{X}) = \Delta(Y, \hat{X}) + \lambda(Y) + \gamma(Y, \hat{X})$$

et :

- $\Delta(Y, \hat{X})$ la pénalité induite par le fait d'abstraire par Y l'information portée par \hat{X} – en intégrant les pénalités de type *perte*, *ajout* et *distorsion* d'information.²⁴
- $\lambda(Y)$ une fonction évaluant le caractère concis d'un résumé - au regard de la redondance d'information.
- $\gamma(Y, \hat{X})$ une fonction permettant d'ajouter des contraintes additionnelles lors de l'évaluation des résumés, e.g. pour exclure les résumés non couvrants.

Les formulations théoriques des différentes composantes de la fonction objectif (eq. 2.1) exploitent l'analyse de caractéristiques de la topologie du DAG sous-jacent à la représentation des connaissances qui sous-tend l'analyse. Nous adoptons dans certains cas des formulations fondées sur la théorie des fonctions de croyance (utilisation des masses, *belief* et *plausibility*) et sur la notion de théorie de l'information, au sens du contenu informationnel proposé par Resnik [226]. A titre d'exemple, nous modélisons :

– La quantité d'information de \hat{X} abstraite par un résumé Y par :

$$\Psi(Y, \hat{X}) = \sum_{x \in \mathcal{A}(X) \cap \mathcal{A}(Y)} bel(x) \cdot IC(x)$$

avec $bel(x)$ la croyance accordée à x au regard des observations représentées par les annotations, $IC(x)$ son contenu informationnel au sens d'une mesure topologique intrinsèque [117], et $\mathcal{A}(X \subseteq C) = \bigcup_{x \in X} \{y | x \preceq y\}$.

²⁴ Perte au sens de l'abstraction, ajout au sens de l'information plausible induite par l'abstraction, et distorsion au sens de l'omission de certaines annotations, i.e. résumés non couvrants.

– La perte d'informations (exactes ²⁵) évaluée dans le calcul de $\Delta(Y, \hat{X})$ par :

$$\Delta^{E-}(Y, \hat{X}) = \sum_{x \in \mathcal{A}(X) \setminus \mathcal{A}(Y)} bel(x) \cdot IC(x)$$

Nous ne détaillons pas l'ensemble de la modélisation; le lecteur intéressé consultera [115].²⁶ Nous précisons cependant que celle-ci autorise une configuration de la fonction objectif via une pondération de ses composantes; cela permet de préférer, voire d'exclure, d'entre elles lors de l'évaluation.

L'expression de cette fonction d'évaluation de la qualité des résumés nous permet de quantifier de manière *ad hoc* la pertinence d'un résumé au regard d'une séquence d'annotations, et d'un contexte donné via l'importance attribuée à chacune des composantes qui interviennent dans son évaluation. Reste à définir une procédure de recherche de l'espace des résumés en vue de maximiser cette expression. Le problème est complexe car, au-delà de la problématique induite par l'énumération des éléments de \mathcal{S}^X , une procédure de parcours capable de cibler ce sous-ensemble $\mathcal{S}^X \subset \mathcal{P}(C)$ doit au préalable être proposée.

Nous rappelons quelques propriétés de cet espace de recherche \mathcal{S}^X : le nombre de sous-ensembles de C est $2^{|C|}$, et dans le pire des cas, lorsque nous traitons des représentations des connaissances correspondant à des taxonomies avec une racine unique, nous avons la borne théorique maximale suivante $\forall X \subset C, |\mathcal{S}^X| \leq 2^{|C|-1} + 1$.²⁷ Appliquer une approche itérative naïve de manière exhaustive est naturellement impossible (et cela même si l'on se restreint aux résumés qui respectent les conditions étudiées, e.g. *faithfulness* et *non-total redundancy*).

Nous proposons dans nos travaux un algorithme pour distinguer \mathcal{S}^X , ainsi que des règles d'élagage qui servent à parcourir de manière non naïve cet espace de recherche [115]. La complexité de l'algorithme reste prohibitive dans le pire des cas. Cela restera vrai même lorsque des gains importants seront amenés par une restriction aux résumés couvrants (\mathcal{S}_{cov}^X). L'approche permet cependant en pratique, en fonction du temps alloué à la recherche, de trouver des solutions intéressantes au regard de la problématique de maximisation définie.

L'évaluation de résumés d'informations à caractère sémantique reste un problème ouvert. Nous ne disposons pas de jeux d'évaluations reconnus pour évaluer nos travaux. Nous avons alors réalisé une évaluation dans un contexte d'analyses psychosensorielles en admettant des résumés produits par un expert de domaine. Les résultats des évaluations qualitatives que nous avons menées soulignent que l'approche proposée permet de faciliter, voire d'automatiser, le processus de génération d'un résumé [185]. Les résumés proposés par l'expert de domaine ont en effet la plupart du temps été obtenus par l'approche automatisée. La procédure de recherche visant à maximiser la fonction objectif définie semble donc encourageante, même si des configurations spécifiques méritent d'être considérées dans certains cas. Par construction, l'approche ne distingue par un unique résumé, mais parcourt plutôt l'espace de recherche en tentant de manière itérative, avec élagage, de maximiser

25. Nous distinguons les informations mentionnées de manière exacte au sens de la taxonomie, i.e. avec $x \preceq y$, x mentionne de manière certaine y , des informations mentionnées de manière plausible : dire y peut vouloir dire x .

26. Article joint en Annexe page 135.

27. Lorsque $\mathcal{O} = (\preceq, C)$ correspond à un ensemble non ordonné raciné par un unique concept, e.g., avec $|C| = 100$ nous obtenons $|\mathcal{S}^X| = 6.3 \times 10^{29}$.

la fonction objectif. Cette approche permet alors d'obtenir *in fine* un classement de plusieurs résumés d'intérêt, ce qui peut par ailleurs être intéressant dans un contexte applicatif donné. Une implémentation Java open source (non supportée) de l'approche est proposée et partagée à la communauté : <https://github.com/sharispe/Conceptual-Summary>.

Une des principales limites de l'approche définie réside dans le temps de calcul nécessaire pour son application lors de l'utilisation de représentations des connaissances importantes, et de grandes séquences d'annotations. De multiples pistes d'amélioration ont été proposées dans nos travaux ; celles-ci mériteraient cependant d'être étudiées plus encore pour faire de réels progrès sur ce sujet. L'approche autorise par ailleurs la définition de configurations spécifiques pour distinguer l'importance à donner à chaque composante intervenant dans la fonction objectif. Cette flexibilité peut aussi induire une difficulté car, les composantes n'étant pas en soi indépendantes, l'incidence d'un paramètre sur le résultat final n'est pas toujours intuitive. Il est de plus complexe d'introduire les différents paramètres à un non expert.

Notre modélisation du problème admet de nombreux postulats sur la notion de résumé. Ceux-ci sont critiquables, ne serait-ce qu'en ce qui concerne les aspects liés aux contraintes que nous fixons pour un résumé (i.e. propriétés), ou les dimensions que nous admettons pour quantifier sa qualité. Au-delà des contributions techniques, e.g. sur l'optimisation des procédures de recherche de l'espace des résumés, il serait intéressant d'étudier les caractéristiques communes aux résumés générés par un opérateur humain en situation écologique. Nos études préliminaires en lien avec ce sujet soulignent de fortes variabilités inter-sujets. Nous avons aussi noté, en analyse psychosensorielle, le fréquent non-respect de certaines propriétés, et en particulier de la *faithfulness* (qui oblige tout concept du résumé à correspondre à une annotation conceptuelle). Cela s'explique par le fait que l'expert s'autorise une interprétation des résultats qui est motivée par son expertise métier. Nous notons cependant que, même si certaines propriétés prises en compte par notre modélisation empêcheront dans certains cas la production de résumés attendus, il est facile pour l'opérateur d'étudier les propositions candidates générées automatiquement et de les éditer conformément à ses attentes. Les potentiels biais induits par ce type d'approches (semi-)automatisées restent à être évalués [266].

Extension au cas d'annotations terminologiques

Nous avons étendu les travaux menés sur le résumé d'annotations conceptuelles de manière à permettre la synthèse d'annotations exprimées sous la forme de termes. L'objectif était dans ce cas aussi de produire un résumé conceptuel comme précédemment [114, 185]. Nous admettons alors des annotations exprimées sous la forme d'ensembles de termes, $\hat{X} = (X_1, X_2, \dots, X_n)$, avec ici chaque $X_i \subseteq T$, et T l'ensemble de termes.

Applicatif : Nous visions en particulier à permettre une généralisation du concept d'analyse psychosensorielle à des cohortes non expertes dont les membres n'auraient pas alors à prendre connaissance d'un ensemble de descripteurs spécifiques. Une approche de ce type permettrait en effet de *démocratiser* l'analyse psychosensorielle et d'offrir des perspectives d'analyses différentes, non dirigées par des experts de domaines, et qui impliqueraient des nombres d'évaluateurs plus importants. D'autres domaines d'application pourraient naturellement être intéressés, e.g. résumés conceptuels d'unités linguistiques (phrases, textes), indexations.

Les contributions proposées dans ces travaux traitent de l'utilisation de plongements sémantiques de mots pour la transcription de termes exprimés en langage naturel, en descripteurs conceptuels. Nous parlerons de *transcription conceptuelle* pour faire référence à la mise en correspondance d'un ensemble de termes avec un ensemble de concepts, ici définie par la fonction *annot* :

$$\text{annot} : \mathcal{P}(T) \rightarrow \mathcal{P}(C)$$

Pour résumer un ensemble de termes nous considérerons alors : (i) la transcription conceptuelle des termes en concepts, puis (ii) le résumé conceptuel des concepts tel qu'introduit précédemment.

Deux classes de modèles ont été étudiées pour la transcription conceptuelle :

1. La *transcription directe*, qui vise à mettre en correspondance chaque terme à transcrire conceptuellement avec un concept spécifique.
2. La *transcription indirecte* qui autorise à ce qu'un concept puisse être évoqué par la mention de plusieurs termes juxtaposés, e.g. {*drôle, film*} \rightsquigarrow Comédie, ou encore {*gras, cacahuete*} \rightsquigarrow BeurreDeCacahuetes.

L'approche de type transcription indirecte vise à répondre à la limitation induite par la prise en compte de la *faithfulness* lorsque l'on admet une transcription directe. Obliger toute annotation terminologique à avoir une correspondance directe avec un concept peut en effet s'avérer limitant. Dans certains cas plusieurs évocations sont utilisées pour exprimer un concept (relations $n : 1$ fréquentes lorsque les évocations sont formulées par des profanes). Nous étions aussi intéressés par l'étude de l'extension du cadre de manière à prendre en compte le fait qu'il faille parfois admettre l'impossibilité d'une mise en correspondance conceptuelle unique, et plutôt considérer une mise en correspondance avec plusieurs concepts (relations $1 : n$), e.g. lorsque la représentation des connaissances n'est pas suffisamment fournie.

Nous admettons dans nos travaux que chaque concept de la représentation des connaissances utilisée dispose d'une correspondance terminologique sous la forme d'un ou plusieurs termes permettant d'y faire référence (labels, e.g. via `rdfs:label`). Cette information est centrale pour permettre de faire le lien entre les espaces terminologique et conceptuel.

Les approches que nous avons étudiées reposent sur le calcul de similarités terminologiques (entre termes) et conceptuelles (entre concepts), mais aussi entre termes et concepts. Ce dernier type de calcul de similarité *terme-concept* nécessitait alors la construction de plongements sémantiques terminologiques et conceptuels dans un même espace. Différentes approches ont été proposées pour répondre à ce besoin. Elles tiennent compte de la connaissance exprimée dans les représentations des connaissances, mais aussi dans des structures de graphes proposées, induites par les notions de similarités sémantiques terminologiques. Nous avons en autres étudié une procédure de construction de graphes de termes et de concepts. Les relations entre les termes y sont définies en fonction des similarités terminologiques calculées à l'aide des plongements classiquement obtenus par l'analyse de textes (e.g. Word2vec était populaire pendant ces travaux). Ces graphes sont ensuite enrichis (i) de connaissances taxonomiques par l'ajout de la représentation DAG de \mathcal{O} , et (ii) de liens entre les termes et les concepts via l'utilisation des labels propres aux concepts. Une fois la construction du graphe incluant des termes et des concepts, une procédure de propagation *ad hoc* permet de construire des plongements enrichis pour chaque noeud (termes et concepts). C'est sur l'analyse de ces plongements que les procédures de transcriptions directe et indirecte se basent. Des techniques de construction de représentations hétérogènes, mariant termes et concepts, étaient étudiées dans la littérature au moment de nos travaux sur ce sujet : nos propositions s'inscrivent dans la lignée de ces recherches.

Les modes de calcul des scores d'appariement entre les termes et les concepts que nous avons étudiés sont assez directs. A titre d'exemple, la *similarité* entre un terme $t \in T$ et un concept $c \in C$, ici donc vue comme un degré d'appariement, est modélisée par une fonction $\sigma_{TC} : T \times C \rightarrow [0, 1]$:

$$\sigma_{TC}(t, c) = \max_{l \in T(c)} \sigma_{TT}(t, l)$$

avec $T(c)$ les termes associés au concept c dans l'ontologie et $\sigma_{TT} : T \times T \rightarrow [0, 1]$ une similarité terminologique qui étudie des plongements sémantiques (similarité de type cosinus ou reposant sur un simple produit scalaire).²⁸ Nous avons exploité ces définitions de degrés d'appariement pour la mise en correspondance d'un ensemble de termes vers un ensemble de concepts via des expressions simples de transcription directe, e.g., avec $X_i \subseteq T$:

$$\text{annot}(X_i) = \bigcup_{t \in X_i} \left\{ \arg \max_{c \in C} \sigma_{TC}(t, c) \right\} \quad (2.2)$$

On retrouve ici des formulations intuitives et usuelles semblables à celles aujourd'hui classiquement considérées en recherche d'information conceptuelle ou dans le domaine de l'alignement d'ontologies. La mise en correspondance conceptuelle amenée de manière indépendante et isolée pour chaque terme (eq. 2.2) est corrigée par l'approche par transcription indirecte. Les détails sont discutés dans [114].

Nous avons dans ces travaux utilisé des plongements de mots construits par des techniques de l'état de l'art à l'heure de l'étude en question (SkipGram et CBOW [193], GloVe [214]. . .). Nous avons par ailleurs réalisé des évaluations de l'incidence du choix des approches de plongement, de leurs paramétrages, et des corpus de textes utilisés lors de leur construction (par apprentissage auto-supervisé donc). Nos résultats ont par exemple montré que, dans le contexte étudié lié à l'analyse psychosensorielle, l'utilisation d'un corpus spécifique pour construire les plongements n'était pas différenciant, i.e. les performances n'étaient pas améliorées en considérant des plongements construits par l'analyse de restrictions de corpus constituées par la sélection de textes ou passages de textes contenant des termes du domaine de l'analyse psychosensorielle étudié (e.g. analyse d'odeurs). Cette observation mériterait d'être discutée car elle s'oppose à l'intuition et aux résultats récents qui soulignent l'incidence potentiellement marquée de la phase de sélection du corpus sur la performance des modèles de langue, e.g. [80, 109]. Nos résultats soulignaient de plus que les approches étudiées tendaient à faire des erreurs différentes. Cette dernière observation nous a notamment invités à travailler l'idée de coupler différents modèles (e.g. agrégation d'approches). Une implémentation Java open source (non supportée) de nos approches est proposée à la communauté : <https://github.com/sharisp/ICE>.

Ces travaux ont pour l'essentiel été menés entre 2014 et 2018. Des techniques de construction de plongement plus contemporaines mériteraient naturellement d'être utilisées aujourd'hui. Nous pensons en particulier à l'utilisation de plongements construits à l'aide de techniques de type Transformers, et de plongements à base de *Graph Neural Network* (GNN).

28. Nous étudions plutôt des formulations du type $\sigma_{TT}(t_1, t_2) = \frac{\rho(t_1) \cdot \rho(t_2)}{\|\rho(t_1)\| \|\rho(t_2)\|}$, avec $\rho : T \rightarrow \mathbb{R}^d$ la fonction de plongement sémantique ; une part importante du travail portait en effet sur la construction des représentations des termes et des concepts.

```

http://onto#Lemon ; fruit, yellow, acid
http://onto#Banana ; fruit, monkey, yellow
http://onto#London ; city, UK, capital
http://onto#Apple ; computer, fruit, juice
http://onto#Ferrari ; red, car, italy
http://onto#God ; Allah, HaShem, Brahma
http://onto#Blood ; red, life, murder
http://onto#Armstrong ; walk, space, moon
http://onto#Hulk ; green, hero, angry
http://onto#Salmon ; fish, canada, bear
http://onto#Sun ; star, burning, sky, yellow, hot, tanned
http://onto#Fire ; hot, burning, red, lighter, element
http://onto#Hair ; ponytail, head, curly, cut, brush
http://onto#Water ; liquid, blue, drink, refreshing, swim
http://onto#Pencil ; write, notebook, pen
http://onto#Dog ; pet, cat, bark

```

FIGURE 4 Exemple d'entrées du jeu de données partagé [114]; format, Concept : terme 1, terme 2. . . Pour chaque ligne l'ensemble des termes évoque le concept correspondant.

Exemple d'application *Pyramide* et comparaison avec ChatGPT : nous avons évalué nos approches de transcription conceptuelle indirecte dans le cadre d'une tâche visant à retrouver un concept évoqué par un ensemble de termes [114], e.g. { *fruit, yellow, acid* } \rightsquigarrow Lemon. Cette tâche correspond en quelque sorte à une version simplifiée du jeu Pyramide [137]. La figure 4 présente des exemples d'entrées du jeu de données constitué pour l'évaluation (et partagé sur le site de l'implémentation de l'approche). Les performances des approches étaient dans l'ensemble plutôt faibles, environ 30% de bonnes réponses pour le top-1, 50% pour le top-5.

Inutile de préciser que le contexte a nettement évolué depuis 2017. ChatGPT 3.5 (version 04/08/23) obtient par exemple de très bons résultats, sans même que l'on précise les concepts au choix (et sans représentation des connaissances donc). Il trouve la bonne réponse pour tous les exemples de la figure 4 excepté pour *Armstrong*, *Water* et *Pencil* pour lesquels il propose respectivement *moonwalk*, *blue sports drink*, et *writing*. La requête est formulée de la manière suivante : *Find the concept mentioned by : terme 1, terme 2. . .*

Exemple ; Find the concept mentioned by : hot, burning, red, lighter, element

Réponse de ChatGPT : *The concept mentioned by "hot, burning, red, lighter, element" is "fire".*

Fire is a chemical reaction that releases heat and light, resulting in a burning sensation. It often appears red or yellow when it burns intensely. A "lighter" is a device used to produce a flame for various purposes, typically by igniting a flammable gas or liquid. Fire is also considered one of the classical elements, along with earth, water, and air, in many ancient philosophical and mythological traditions. It has been essential for human development and is both a tool and a force of nature that can be controlled but also potentially destructive if not managed carefully..

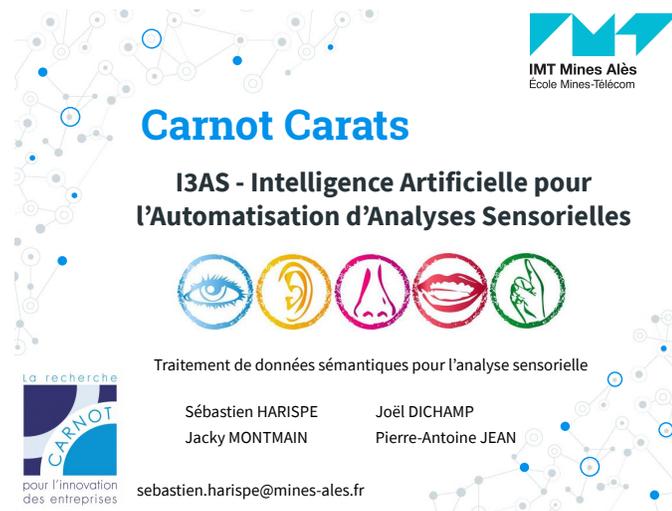


FIGURE 5 Plaquette du projet I3AS.

Élargissements

Un travail de consolidation des propositions étudiées dans ces travaux a été proposé ultérieurement dans le cadre du projet CARNOT M.I.N.E.S CARATS I3AS (Intelligence Artificielle pour l'Automatisation d'Analyses Sensorielles), cf. Figures 5 et 6. Ce financement nous a en particulier permis d'amener le développement d'un prototype de solution Web intégrant les différentes approches proposées dans le contexte de l'analyse psychosensorielle. Des techniques d'analyse supplémentaires ont aussi été étudiées pour permettre l'analyse de larges cohortes d'évaluateurs non experts en tenant compte de traits individuels, e.g. regroupement (*clustering*) d'individus par proximité des résumés, suivi de l'étude des propriétés partagées par les individus de chaque groupe. La partie basse de la Figure 7 illustre l'approche : (i) interface de récupération des annotations terminologiques non contrôlées, (ii) phase de traitement des termes (e.g. lemmatisation), (iii) mise en correspondance conceptuelle via le vocabulaire contrôlé et structuré, e.g. transcription directe ou indirecte, (iv) regroupement des individus au regard de la similarité sémantique de leurs annotations (i.e. mesures sémantiques entre groupes de concepts), (v) résumé sémantique des différents groupes, (vi) analyse des métadonnées associées aux évaluateurs pour une mise en correspondance de traits partagés par les évaluateurs d'un même groupe, (e.g. âge). Ces développements ont fait l'objet d'un dépôt à l'APP (L'Agence pour la Protection des Programmes).

I3AS : Intelligence Artificielle pour l'Automatisation d'Analyses Sensorielles

INSTITUT CARNOT CARATS

Les Industries de la Mode et du Luxe, dépendantes de productions d'excellence, accordent une importance toute particulière à la perception de leurs produits par des cibles clientèles toujours plus exigeantes. Afin d'assurer le respect des standards de qualité particulièrement élevés, ces industries tirent fréquemment parti d'analyses sensorielles. Ces analyses permettent une étude systématique des perceptions sensorielles (vue, ouïe, odorat, goût, toucher), et offrent la possibilité d'apprécier le ressenti de panels d'évaluateurs face à des maquettes produits susceptibles d'être commercialisées. Les approches d'analyse sensorielle standard requièrent un apprentissage chronophage et délicat qui assujettit de facto l'analyse sensorielle à l'utilisation exclusive et onéreuse de panels restreints d'experts entraînés et empêche ainsi la mise en place d'évaluations auprès de larges échantillons du public visé.

Basé sur les progrès récents de différents domaines relatifs à l'Intelligence Artificielle, le projet I3AS confié au Laboratoire d'Informatique et d'Ingénierie de la Production d'IMT Mines Alès s'est donné pour objectif d'implanter l'automatisation d'approches sensorielles non-orientées dans un service informatique à destination des industriels. Nous y proposons la mise en œuvre de techniques d'analyse de données permettant d'évaluer la proximité sémantique d'appréciations sensorielles retranscrites en langage naturel. Ces techniques sont couplées à l'utilisation de représentations des connaissances (ontologies), capables de modéliser l'expertise métier et de caractériser les composantes sémantiques qui donneront à la machine la possibilité d'apprécier les liens entre descripteurs qui méritent d'être intégrés. Le croisement de ces différentes technologies permet la définition d'une chaîne de traitements algorithmiques capables d'envisager une retranscription fidèle et pertinente de l'appréciation de larges panels non-spécialistes, voire néophytes.

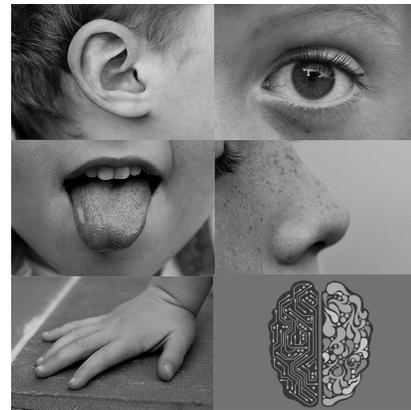


Image adaptée de Five Senses (Nicki Dugan Pogue – [Creative Commons](#))

FIGURE 6 Descriptif du projet I3AS.

I3AS - Intelligence Artificielle pour l'Automatisation d'Analyses Sensorielles

Traitement de données sémantiques pour l'analyse sensorielle par approches non contrôlées

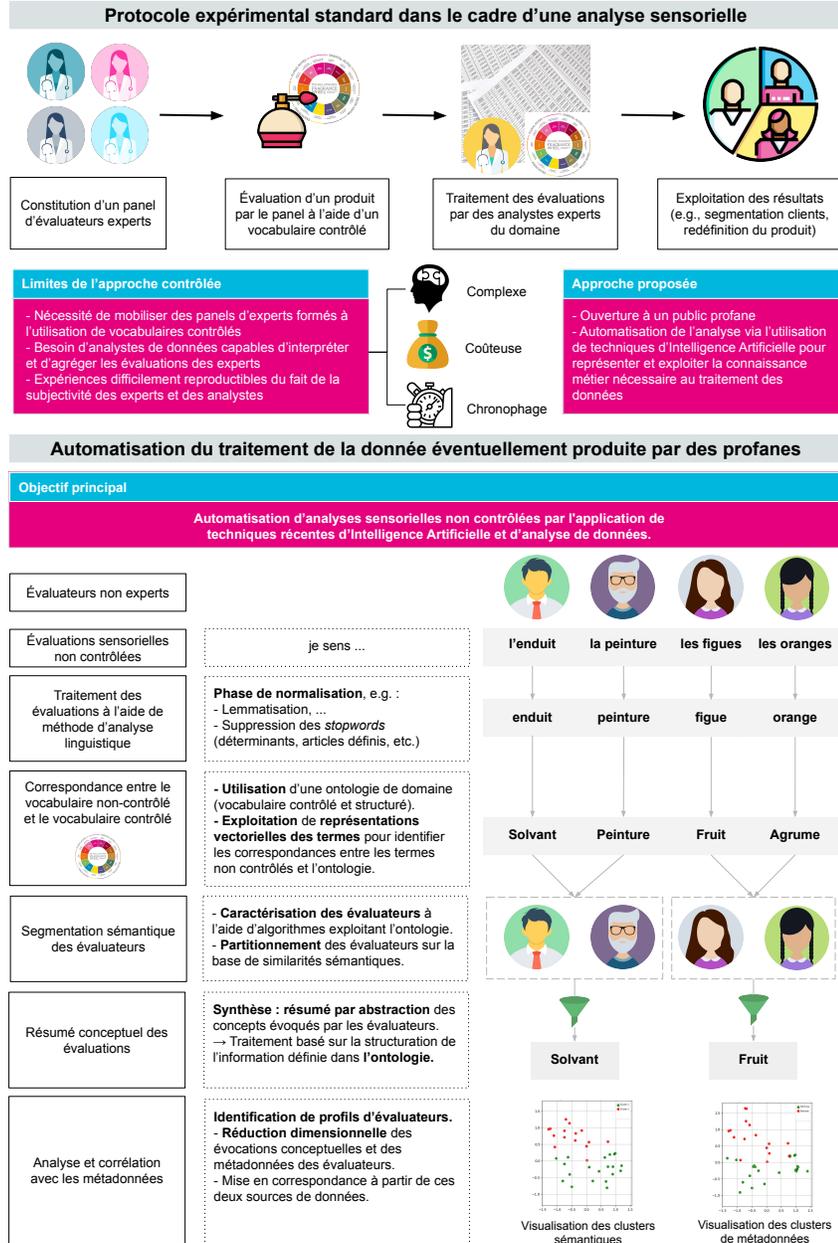
Auteurs

Sébastien Harispe
Pierre-Antoine Jean
Joël Dichamp
Jacky Montmain

Partenaires



IMT Mines Alès
est certifiée par



Contact : sebastien.harispe@mines-ales.fr

FIGURE 7 Illustration du projet CARATS I3AS.

2.2.2 Estimation de la qualité de sources d'information et vérification d'assertions

Nous présentons dans cette section nos travaux sur l'estimation de la qualité de sources d'information et la vérification d'assertions²⁹ ; ils s'inscrivent dans la continuité de la vaste littérature proposée par le domaine nommé *Truth Discovery* en anglais [29, 28]. L'originalité de nos contributions porte sur l'utilisation de représentations des connaissances comme sources de connaissances *a priori* en vue d'améliorer l'évaluation (i) de la véracité d'une source, et (ii) la vérification d'assertions. Nous avons en particulier étudié l'adaptation d'approches de l'état de l'art via l'utilisation de représentations des connaissances légères de type taxonomies, mais aussi de connaissances exprimées sous la forme de règles extraites de bases de connaissances.

Collaborateurs :

- Doctorante encadrée : Valentina Beretta (2015-2018).
- Stagiaire encadré : Moez El Gabsi (2017).
- Collaborateurs d'IMT Mines Alès : Sylvie Ranwez, Andon Tchechmedjiev.
- Collaborateurs externes : Isabelle Mougnot (Université Paul Valéry, Montpellier), Konstantin Todorov (LIRMM, Montpellier), Katarina Boland et Luke Lo Seen (GESIS, Cologne, Allemagne).

Les travaux mentionnés dans cette section ont pour l'essentiel été menés dans le cadre de la thèse de Valentina Beretta [21].

Cadre formel des travaux

Tout comme les approches contemporaines du domaine de la vérification d'assertions, nos travaux se basent sur le postulat *dual* : (i) plus une source d'information affirme des vérités, plus on lui accordera du crédit ; (ii) plus une information est partagée par des sources fiables, plus nous considérerons qu'il y a des chances qu'elle soit vraie. On note ici, par construction du postulat, l'interdépendance communément retrouvée entre la notion de confiance des sources d'information, et la notion de vérité des informations qu'elles véhiculent.

Le cadre formel que nous avons étudié dans l'essentiel de nos travaux admet des assertions exprimées à l'aide d'une logique bivalente, i.e. une assertion est soit vraie, soit fausse. Les assertions ne portent de plus que sur des prédicats dits fonctionnels, c'est à dire qu'ils n'acceptent qu'une unique valeur précise de vérité (e.g. lieu de naissance, âge).

Nous reprendrons la terminologie du domaine et nommerons par *entité* le sujet d'intérêt sur lequel porte une assertion dont on souhaitera évaluer la valeur de vérité. Les assertions seront de type triplet (entité, prédicat, valeur) semblables aux triplets RDF (sujet, prédicat, objet) classiquement retrouvés en Web Sémantique et en Représentation des Connaissances, e.g. (*picasso*, *bornIn*, *malaga*). Nous réserverons le terme *assertion* à tout triplet pour lequel la valeur de vérité est discutée. On parlera de *fait* pour faire référence à un triplet vrai. Nous nommerons de plus *description* un couple (entité, prédicat) pour lequel on recherchera la valeur de vérité, e.g. (*picasso*, *bornIn*).

29. En Représentation des Connaissances une assertion vaut vérité ; ici nous aurons des assertions potentiellement contradictoires formulées par plusieurs sources, il nous faudra distinguer la vraie.

Plus formellement nous noterons :

- E, D, V, S les ensembles disjoints d'entités, descriptions, valeurs et sources d'informations.
- $V_d \subseteq V$ l'ensemble des valeurs exprimées par les sources d'informations pour une description $d \in D$; nous noterons aussi v_d un élément de V_d .
- $S_{v_d} \subseteq S$ l'ensemble des sources d'information qui affirment la valeur v_d , pour la description d donc.
- $V_s \subseteq V$ l'ensemble des valeurs proposées par une source $s \in S$.

L'objectif sera alors d'exprimer une fonction :

$$tf : D \times V \rightarrow \{true, false\}$$

L'état de l'art distingue souvent pour cela deux fonctions : la première pour discuter la fiabilité d'une source (*source trustworthiness*), et la seconde pour évaluer la confiance à donner à une assertion (*value confidence*), avec respectivement $t : S \rightarrow [0, 1]$ et $c : D \times V \rightarrow [0, 1]$. Cela nous permet d'admettre la définition suivante : *Soit D, V, S les ensembles de descriptions, valeurs et sources. L'objectif de la vérification d'assertions est de trouver pour chaque description $d \in D$, la valeur de vérité $v_d^* \in V_d \subseteq V$. Cette recherche se base sur une estimation de la fiabilité des sources d'information (fonction t) et sur une estimation de la confiance attribuée aux assertions (fonction c) [21].* Nous retrouvons dans cette modélisation communément admise dans l'état de l'art les difficultés évoquées par le postulat mentionné ci-avant en termes de dépendance circulaire.

Le modèle *Sums* est un standard de la littérature [209]. Il implémente une approche itérative, qui alterne l'estimation des fonctions propres à l'évaluation de la fiabilité d'une source et de la confiance d'une assertion.

$$t^i(s) = \alpha^i \sum_{v_d \in V_s} c^{i-1}(d, v_d) \quad (2.3)$$

$$c^i(d, v) = \beta^i \sum_{s \in S_{v_d}} t^i(s) \quad (2.4)$$

avec i le numéro de l'itération lors de la phase d'estimation et α^i et β^i deux facteurs de normalisation.³⁰ L'initialisation des valeurs s'effectue généralement de manière uniforme. La sélection de la valeur dite vraie pour une description reposera alors sur une procédure de maximisation après convergence des valeurs ou suite à la réalisation d'un nombre arbitraire d'itérations.³¹

Différentes approches ont été proposées selon que l'on considère des degrés variables de fiabilité entre sources, l'indépendance des assertions ou des sources, ou encore la possibilité d'interdépendances temporelles, et même le cas de vérités non statiques (i.e. variables en fonction du temps). . . [172, 29, 28] Nous nous sommes dans nos travaux concentrés sur l'adaptation de certaines de ces approches pour la prise en compte de connaissances *a priori*. Comme nous le verrons, ces adaptations nécessiteront des modifications importantes du cadre formel de base, mais aussi la considération de postulats supplémentaires.

30. $\alpha^i = \left(\max_{s' \in S} \sum_{v_d \in V_{s'}} c^{i-1}(d, v_d) \right)^{-1}$ et $\beta^i = \left(\max_{v_d \in V} \sum_{v'_d \in S_{v'_d}} t^i(s) \right)^{-1}$.

31. Nous disposons de preuves théoriques qui garantissent la convergence de certains modèles.

Prise en compte d'un ordre partiel sur les valeurs éligibles

La littérature qui s'intéresse à la vérification d'assertions dans le cadre de prédicats fonctionnels considère que les valeurs V_d de toute description $d \in D$ sont indépendantes au sens de l'implication, i.e. $\forall (v_1, v_2) \in V_d \times V_d, v_1 \not\leq v_2$. Ce cadre est restrictif, en particulier lorsque nous cherchons à traiter la valeur de vérité d'assertions extraites de sources d'informations exposées sur le Web, e.g. articles, blogs. Il est en effet fréquent que des valeurs de niveaux de précisions différents soient retrouvées. Considérez par exemple que nous nous intéressons au lieu de naissance de Pablo Picasso, i.e. à la valeur de vérité de la description (`picasso, bornIn`). Une extraction de valeurs candidates extraites de textes issus du Web pourrait nous amener à distinguer les valeurs suivantes : `europa`, `spain`, `malaga` et `granada`. Picasso étant originaire de Málaga, les valeurs `Europe`, `spain`, `malaga` sont correctes car nous admettons implicitement la relation suivante `malaga` $\prec_{\text{Part-of}}$ `spain` $\prec_{\text{Part-of}}$ `europa`.³² Les cadres classiques de la vérification d'assertions ne permettent pas de traiter sérieusement ces cas car ils admettent que toutes les valeurs sont indépendantes ; seules `malaga` et `granada`, que nous savons être deux villes distinctes, mériteraient alors d'être considérées comme l'expression de deux points de vue spécifiques non conciliables. Le fait d'admettre un ordre sur les valeurs induit de fortes modifications par rapport au cadre initial. Il serait nécessaire de formaliser le sujet d'étude en termes de modèles au sens de la Représentation des Connaissances [111]. Nous ne proposerons cependant pas de traitement plus formel dans cette section afin de limiter la taille et la technicité de ce manuscrit. Le lecteur admettra :

- Une vision ensembliste des valeurs proposées pour une description, i.e. chaque élément de V_d peut être vu comme un ensemble de valeurs d'un type d'intérêt, e.g., en admettant une définition en extension nous aurons `City` \equiv {`granada`, `malaga` . . . }, et une instance de la classe `City` correspondra au singleton de l'instance associée, e.g. {`malaga`} pour `malaga`.
- Des représentations des connaissances de type POSET susceptibles d'être définies par le croisement de multiples prédicats, e.g. `rdfs:type`, `rdfs:subClassOf`, `rdfs:partOf`. Les triplets suivants (`malaga, rdfs:partOf, spain`), (`spain, rdfs:type, Country`), (`Country, rdfs:subClassOf, Location`) permettent de manière simplifiée d'admettre la relation d'ordre : `malaga` \prec `Country` \prec `Location`.
- Des raisonnements en monde fermé au regard de la connaissance exprimée par le POSET, ce qui permet une définition des valeurs en extension, comme précisé ci-dessus pour la vision ensembliste des valeurs.

Il est important de souligner que l'existence d'un ordre partiel sur les valeurs candidates n'induit pas une modification de la nature fonctionnelle du prédicat considéré. Nous considérons une flexibilité dans les notations si l'on admet une restriction sur le codomaine d'un prédicat, e.g., (`born in, rdfs:range, City`), avec (`picasso, bornIn, spain`), équivaudra à exprimer : `picasso bornIn some (partOf spain)`.

Nos travaux s'intéressent au cas où une représentation des connaissances structure les valeurs formulées par les sources d'information pour chaque description sous la forme d'un ordre partiel. Cette représentation des connaissances aura autorité pour l'analyse réalisée. Nous admettrons en outre que si pour une description $d \in D$ donnée, nous avons deux valeurs $v_1, v_2 \in V_d$ avec $v_1 \leq v_2$, alors l'évocation de v_1 appuie en quelque sorte v_2 , et cela bien que l'expression de la valeur v_1 par une source ait été faite de manière détachée de toute représentation des connaissances.

32. Le codomaine (*range*) des prédicats n'est généralement pas restreint en vérification d'assertions.

Les modélisations que nous avons étudiées reposent sur une interprétation des fonctions de croyance [234] dans le cadre de notre étude. Les adaptations sont illustrées sur le modèle *Sums* introduit ci-dessus [209]. Nous avons en particulier proposé différents schémas de modification qui adaptent, en tenant compte de la structuration des valeurs, les ensembles de sources d'informations et de valeurs qui seront pris en compte lors de la phase itérative visant à estimer les fonctions t et c , cf. équations 2.4 et 2.3 [24]³³. Ces modifications reposent sur la modélisation de l'analyse des POSET à l'aide des fonctions de croyance que nous avons proposée dans nos précédents travaux sur le résumé sémantique, e.g. [113]. Nous illustrons les modifications proposées avec un exemple simple. Celui-ci vise à modifier l'ensemble des sources lors de l'estimation de la confiance d'une valeur pour une description, i.e. dans l'équation 2.4 l'ensemble S_{v_d} est substitué par $S_{v_{d+}} \triangleq \cup_{v'_d \preceq v_d} S_{v'_d}$. Cela consiste, assez naturellement, à considérer qu'une source supporte le fait que Picasso soit né en Espagne si elle mentionne que sa ville de naissance se situe en Espagne. Dans le cadre d'une interprétation plausibiliste nous admettons l'expression alternative $S_{v_{d++}} \triangleq \cup_{v'_d \in V^P(v_d)} S_{v'_d}$ avec $V^P(v_d) \triangleq \{v'_d | (v'_d, v''_d) \in V_d^2, v'_d \preceq v_d \wedge v''_d \preceq v'_d\}$. Cela consiste à intégrer la prise en compte du crédit attribué à toute source qui affirme une valeur qui n'est pas en contradiction avec la valeur proposée (en admettant une vision de type monde fermé).³⁴

Sélection de la valeur de vérité d'une description Le cadre classique de la vérification d'assertions appliquée à l'étude de prédicats fonctionnels admet une unique valeur de vérité. Nous l'avons vu, la prise en compte d'un ordre partiel nous invite à sortir de ce cadre. Cela oblige alors à définir une procédure pour sélectionner une valeur que nous souhaiterons *in fine* proposer comme la valeur vraie la plus informative pour une description. Une simple procédure de type maximisation de la confiance d'une valeur ne serait pas suffisante car, pour toute description, nous aurions $v' \preceq v \implies c(v', d) \leq c(v, d)$, i.e. la valeur qui maximise la confiance serait la racine du POSET, e.g. **Thing** ou **Location**, qui ne serait pas alors informative. Nous avons proposé différentes procédures paramétrables de sélection pour traiter ce point. Celles-ci se basent sur l'exploitation de la connaissance exprimée dans le POSET et la confiance estimée des assertions. Elles s'intéressent par exemple à identifier le ou les sous-POSET contenant les meilleures valeurs candidates. Ces valeurs seront ensuite analysées et triées au regard de différents critères proposés (e.g., contenu informationnel, fiabilité moyenne des sources qui les supportent), en fonction des spécificités du contexte applicatif concerné. Elles permettent notamment, en fonction du paramétrage configuré, de retourner des réponses imprécises et de regrouper, puis trier des valeurs qui correspondent à des réponses alternatives.

L'évaluation des approches proposées a été réalisée sur plusieurs jeux de données synthétiques produits à partir de DBpedia [9] et de la Gene Ontology [63]. Les évaluations ont été menées en admettant différents schémas de génération de sources d'informations afin de simuler différents comportements possibles des sources (en fonction du niveau d'expertise qui induira en autres des réponses plus ou moins justes et précises). Nos tests soulignent que la prise en compte de l'ordre partiel sur les valeurs à l'aide des procédures proposées pour (i) l'évaluation de la fiabilité des sources et la confiance des assertions, ainsi que (ii) la sélection des valeurs, permet d'obtenir des résultats intéressants. Ces résultats sont meilleurs que ceux qui peuvent être obtenus sans la prise en compte de la connaissance sur la structuration des valeurs - ce qui n'a rien d'étonnant car les approches

33. Cet article est joint au manuscrit, cf. Annexe A page 162.

34. Prenons l'exemple d'une source qui proposerait la réponse `MedievalCity`; si nous admettons `granada` \preceq `MedievalCity` et `granada` \preceq `spain`, alors la fiabilité de cette source d'information serait considérée pour estimer la confiance de la valeur `spain` - les valeurs `MedievalCity` et `spain` ne sont en effet pas contradictoires car elles peuvent toutes les deux faire référence à `granada`, i.e. en vision ensembliste : `granada` \in `spain` \cap `MedievalCity`.

classiques considèrent que les valeurs sont indépendantes. Les approches étudiées permettent de plus de proposer des résultats plus riches en fonction de la nature des solutions candidates que l'on souhaite étudier (valeurs disjointes, ordonnées). Elles permettent aussi d'identifier comme pertinentes des valeurs présentes dans la représentation des connaissances bien que celles-ci ne soient pas ou que peu mentionnées par des sources d'informations. Les implémentations et les jeux de données proposés dans le cadre de ces travaux sont disponibles sur <https://github.com/lgi2p/TDSelection>. Ces résultats illustrent l'intérêt d'étudier la proposition de modélisations permettant d'intégrer de la connaissance *a priori* pour la vérification d'assertions et l'évaluation de la confiance à donner à des sources d'information.

Prise en compte de motifs récurrents dans les bases de connaissances

La section précédente introduit une approche permettant d'évaluer les valeurs d'une description en tirant parti de connaissances sur les valeurs à évaluer exprimées sous la forme d'un POSET. Cette connaissance est pour l'essentiel tirée de la connaissance classiquement retrouvée dans la TBOX d'une ontologie.³⁵ Ces travaux ont été étendus afin d'intégrer une autre forme de connaissance issue de bases de connaissances, à savoir des motifs récurrents identifiés par l'analyse de la ABox [23, 26, 27].³⁶ L'intuition sur laquelle repose l'étude que nous avons menée est simple : il existe des motifs qui supportent l'observation de certains faits. A titre d'exemple, savoir qu'une personne née en Espagne parle l'espagnol ne vous étonnera pas. Cela s'explique plus généralement par le fait que nous admettons classiquement que la règle suivante sera très souvent vérifiée : $\text{isBornIn}(x, y) \wedge \text{hasOfficialLanguage}(y, z) \implies \text{speaks}(x, z)$. Ainsi, si l'on revient à nouveau sur l'exemple discuté dans la section précédente sur la description (`picasso, bornIn`), savoir que Pablo Picasso parle l'espagnol peut nous conforter dans l'idée qu'il soit en effet né en Espagne.³⁷ Nous avons ainsi étudié comment intégrer dans nos modèles l'information portée par des règles d'association, afin d'augmenter l'appréciation des assertions qu'elles supportent, et cela en fonction de leur qualité.

Il est important de noter que la connaissance utilisée dans la section précédente était exacte, et que son exploitation adoptait pour l'essentiel une approche strictement déductive, i.e. fondée sur de simples conséquences logiques du type, puisque `granada` \preceq `spain` et $(x, \text{bornIn}, \text{granada})$, alors $(x, \text{bornIn}, \text{spain})$.³⁸ A contrario, l'approche qui se base sur l'exploitation de règles d'association adopte un raisonnement approché, et plus particulièrement de type raisonnement inductif : les observations de motifs en lien avec une assertion ne pourront que la supporter, sans qu'il n'y ait de conséquence logique en tant que telle. On note ici le lien avec les approches d'Apprentissage Machine pour lesquelles nous exposerons nos travaux dans les prochaines sections.

Nous avons utilisé différentes approches existantes d'extraction de règles pour détecter des motifs récurrents dans des bases de connaissances. L'approche AMIE+ [99] a été utilisée ; elle permet d'extraire des règles sous la forme de clauses de Horn. Nous avons par ailleurs intégré la prise en compte de métriques permettant de critiquer le crédit à donner à une règle dans l'objectif du

35. Même si comme nous l'avons vu, le POSET en question peut être construit en tenant compte de prédicats autres que `rdfs:subClassOf`, e.g. `rdfs:type`, `partOf`.

36. Un article sur ce sujet est joint au manuscrit, cf. Annexe A page 145 [23].

37. Si l'on considère une formulation probabiliste, cela revient à admettre l'expression suivante :

$$P((x, \text{bornIn}, \text{spain}) | (x, \text{speaks}, \text{Spanish})) \geq P((x, \text{bornIn}, \text{spain})).$$

38. Si l'on admet que le POSET est consensuel parmi les sources (i.e. admis par les sources), et que les valeurs exprimées par les sources correspondent bien à celles du POSET. Cela ne vaut pas par ailleurs pour les modélisations alternatives à base de plausibilité.

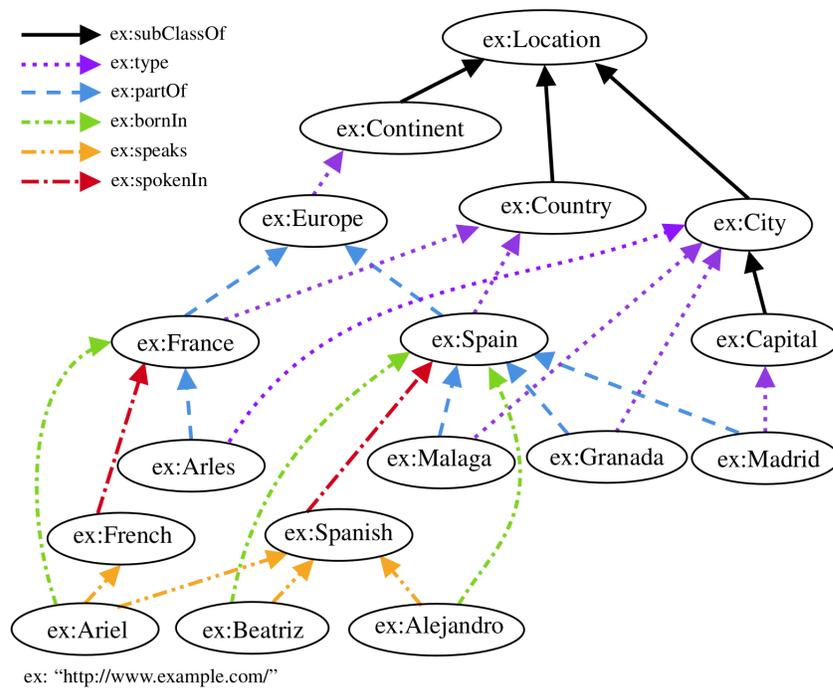


FIGURE 8 Exemple de graphe de faits étudié dans nos travaux. La structuration formelle des valeurs et les règles extraites du graphe nous permettent de critiquer la véracité d'une assertion. Source [26].

raisonnement inductif visé. Ces métriques, de type confiance, support, confiance PCA (*Partial Completeness Assumption*), permettent de quantifier différents aspects utiles pour critiquer l'importance à donner à une règle en fonction du nombre d'observations qui la supportent et la contredisent par exemple, éventuellement en tenant compte d'une assomption de type monde ouvert. Nous avons en particulier proposé une agrégation des scores de confiance PCA et de support qui repose sur la méthode Empirical Bayes [229] [26]. Cette méthode permet, en admettant des *priors*, d'ajuster les évaluations fondées sur peu d'observations en tenant compte du risque d'observations par chance. Ainsi, toutes les règles qui supportent une assertion sont utilisées pour calculer un facteur d'amplification (*boosting factor*) tenant compte du score agrégé précité. Ce facteur intègre la prise en compte des règles qui supportent l'assertion au regard de la totalité des règles éligibles, i.e. les règles qui auraient pu être vérifiées pour le prédicat et l'entité en question - les notions de règles éligibles et qui supportent une assertion étant définies précisément dans nos travaux [26, 23]. Une adaptation a aussi été proposée pour intégrer la prise en compte de la relation d'ordre entre les valeurs lors de l'évaluation des règles.³⁹ Le modèle intégrant l'ordre partiel sur les valeurs s'exprime comme suit :

$$boost_{PO}(d, v_d) \triangleq \frac{\alpha_b + \sum_{r \in R_d^{v+}} score(r)}{\alpha_b + \beta_b \sum_{r \in R_d} score(r)}$$

avec $score : R \rightarrow [0, 1]$ une fonction d'agrégation qui quantifie l'importance à donner à une règle $r \in R$, R_d l'ensemble des règles éligibles, et R_d^{v+} l'ensemble des règles vérifiées dont la tête porte sur (x, p, v') , $v' \preceq v_d$. Sur la base de ce facteur d'amplification nous avons redéfini l'évaluation de la confiance d'une assertion, e.g. utilisée dans l'approche SUMS, par :

$$c_{RULES\&PO}^i(v_d) \triangleq \beta \left((1 - \gamma) c_{PO}^i(v_d) + \gamma boost_{PO}(d, v_d) \right)$$

Les expériences menées sur des jeux de données synthétiques montrent l'apport d'une approche intégrant des modalités de raisonnement inductif. Dans nos évaluations, les modèles les plus performants correspondaient la plupart du temps aux modèles intégrant de la connaissance sous la forme de règles d'association et d'ordre partiel sur les valeurs. Les jeux de données utilisés et les implémentations proposées sont disponibles à l'adresse suivante : <https://github.com/lgi2p/TDwithRULES>.

Nos approches ont été éprouvées dans des contextes réels de *Slot-Filling* en utilisant DBpedia et des extractions issues du Web via DBpedia Spotlight [186] - contexte de recherche de valeurs pour le prédicat `dbo:birthPlace`. La figure 9 présente deux pages Web qui proposent des informations contradictoires sur le lieu de naissance de Picasso. Deux expérimentations ont été effectuées pour comparer la méthode Sums intégrant de la connaissance *a priori* via l'approche proposée, et les autres méthodes de l'état de l'art, cela en fonction de deux techniques utilisées pour l'extraction des valeurs des textes. Nous avons observé des augmentations de performances de l'ordre de 18% et de 14% par rapport à l'approche Sums que nous avons modifiée. Notons tout de même que les approches les plus performantes de l'état de l'art restent compétitives avec l'approche Sums augmentée; elles mériteraient elles aussi d'être étudiées afin d'intégrer dans leur traitement la prise en compte de connaissances connues. Nos travaux démontrent empiriquement que l'ajout de connaissances issues de bases de connaissances permet d'envisager une amélioration notable des performances des systèmes de vérification d'assertions.

39. Cette information n'étant pas prise en compte par AMIE+ lors de l'extraction des règles.

2.2. Exploitation de connaissances formelles dans des processus d'aide à la décision

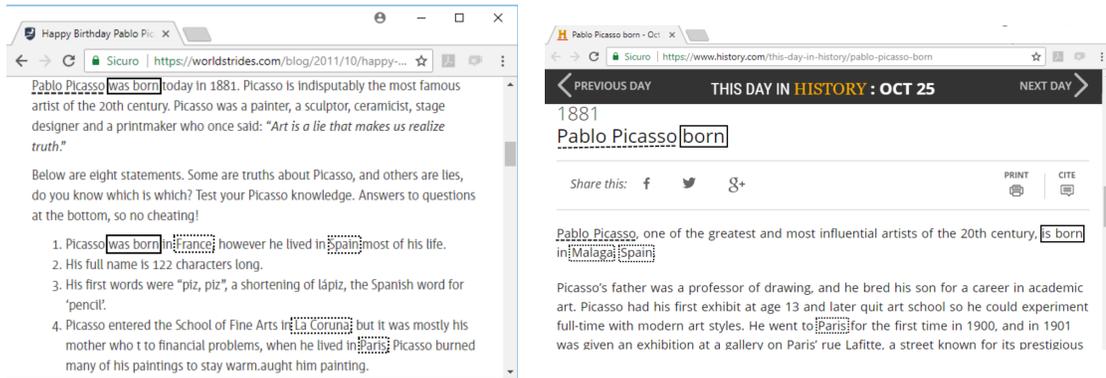


FIGURE 9 Deux pages Web proposant des informations contradictoires sur le lieu de naissance de Pablo Picasso. Source [21].

Les perspectives qui pourraient être données à ces travaux sont nombreuses. Au-delà de l'adaptation d'approches autres que l'approche *Sums* étendue dans nos travaux, et dans la continuité de la prise en compte de l'information portée par des règles d'association, il pourrait être intéressant d'inclure des informations issues de techniques d'inférence de faits fondées sur l'analyse de bases de connaissances [152, 260]. L'idée intuitive serait alors d'augmenter le crédit accordé à un fait (et ainsi la confiance des sources qui le supportent) si celui-ci semble probable au regard d'une technique d'inférence de faits.⁴⁰

40. Notez que Picasso est bien né à Málaga en Espagne et parlait espagnole.

2.2.3 Enrichissement de modèles prédictifs et d'inférence en TALN

Nous présentons dans cette section les travaux que nous avons menés sur la prise en compte de connaissances *a priori* dans des techniques de Traitement Automatique du Langage Naturel.

Étude des techniques de *retrofitting*

Nous avons réalisé plusieurs travaux sur l'utilisation de techniques de prise en compte de connaissances pour l'enrichissement de plongements sémantiques de termes obtenus de manière usuelle via des techniques d'apprentissage auto-supervisé [103]. Ces travaux, pour partie réalisés dans le cadre de collaborations industrielles, se sont plus spécifiquement concentrés sur la prise en compte de sources de connaissances dans une approche dite de *retrofitting*.

Collaborateurs :

- Posdoctorante encadrée : Kaoutar Ghazi (2020).
- Stagiaire encadré : Sébastien Marchal (2019).
- Collaborateurs d'IMT Mines Alès : Andon Tchechmedjiev, Nicolas Sutton-Charani, Pierre-Antoine Jean.
- Collaborateurs externes : Gildas Tagny-Ngompe et Philippe Lergenmuller (société ESII, Montpellier).

Les travaux mentionnés dans cette section ont pour l'essentiel été menés dans le cadre du projet ARIA (détaillé ci-après) réalisé avec la société ESII et dont l'auteur était responsable scientifique académique.

Le *retrofitting* correspond à un ensemble de post-traitements qui visent à améliorer des plongements de termes en tenant compte de la connaissance exprimée dans des sources de connaissances de type lexiques, ontologies légères ou même dans des bases de textes (e.g. exemples de paraphrases). Il vise à contraindre les plongements de manière à ce qu'ils répondent à des attendus définis dans la source de connaissance considérée (e.g. relations de synonymie, antonymie ou hyperonymie retrouvées dans FrameNet [12] ou WordNet [194], ou encore à l'aide de bases de données de paraphrases telles que PPDB [100]). On parle aussi parfois de phase de raffinement de l'espace sémantique latent.

Parmi la diversité d'approches existantes pour adapter des plongements existants, les approches de *retrofitting* se basent généralement sur la prise en compte de fonctions objectif qui permettront, dans un contexte d'apprentissage supervisé, d'apprendre de manière conjointe les plongements des termes et les relations conceptuelles ou lexicales retrouvées dans les sources de connaissances utilisées. De manière intéressante, le *retrofitting* permet alors de s'affranchir de la *seule* considération de l'hypothèse distributionnelle implémentée via une approche statistique que se base sur l'analyse de cooccurrences contextuelles. Ce traitement permet de pallier certaines limitations des approches classiques qui, pour de nombreux chercheurs, n'encodent que des *melting-pot* de relations semantico-lexicales entre termes. A titre d'exemple de limitation des plongements standard, même si les termes *angry* et *offence* évoquent tous deux l'émotion *anger*, leurs plongements auront dans certains cas une similarité cosinus plus faible que celle observée pour les termes *angry* and *enjoy* [235]. Pour pallier ces limitations, les approches de *retrofitting* visent à intégrer de la connaissance peu exprimée voire inexistante dans des textes. Cela rend aussi bien entendu ces approches intéressantes pour des applications relatives à des domaines pour lesquels peu de données textuelles existent.

Nos travaux d'évaluation des approches de retrofitting ont été effectués dans un contexte de classification de textes courts, et plus particulièrement pour l'étude de problématiques rencontrées dans le développement de bornes vocales d'orientation avec la société ESII (cf. projet ARIA). Les textes courts correspondaient dans notre étude aux transcriptions écrites des expressions vocales de personnes souhaitant se faire orienter par les bornes en question. Seule la partie académique de cette étude sera discutée ci-après. Sur cette partie, nos travaux ont porté sur l'évaluation de trois techniques compatibles avec les exigences du projet : deux techniques de retrofitting proposées par Faruqui et al. [88] et Mrkšić et al. [196], et l'approche Paragram [105] qui vise à agréger des plongements construits de manière dissociée à partir de textes et de représentations des connaissances. Nos évaluations ont été réalisées dans plusieurs contextes, e.g. plongements non contextuels (Glove [214]. . .) et contextuels (Flair [3], RoBERTa [180]), entraînés à l'aide de différents corpus, et cela pour deux jeux de données propres à la classification de textes courts en anglais (*HuffPost headlines* et *Product Listing On Amazon India*, cf. [103]). Le retrofitting a été effectué en considérant PPDB, FrameNet, et WordNet (avec ou sans restriction aux synonymes). Deux méthodes de classification de textes ont aussi été évaluées : (i) un classifieur standard (de type XGBoost, forêt aléatoire, SVM. . .) sur la base de plongements des textes effectués par agrégation des représentations des termes, (ii) un classifieur de type réseau de neurones intégrant une partie bi-LSTM pour la construction d'un plongement du texte fourni à une tête de classification de type MLP.

Les résultats obtenus via une validation croisée permettent dans certains cas spécifiques d'observer que le retrofitting tend à améliorer les systèmes de classification de textes courts (e.g., jusqu'à +9 points d'*accuracy* pour des plongements MUSE [167] avec classifieur LSTM/RNN). De fortes dégradations des performances sont cependant observées dans d'autres cas. Ces résultats laissent penser que le retrofitting peut parfois induire des phénomènes de perte ou d'oubli de caractéristiques présentes dans les plongements initiaux, bien que celles-ci soient importantes pour la tâche traitée (e.g. observation de dégradations pouvant aller jusqu'à -17 points d'*accuracy*). Ainsi nos résultats ne nous permettent pas d'établir que l'utilisation de sources de connaissances spécifiques, semble-t-il pertinentes *a priori*, via une approche de retrofitting, amène un gain de performance systématique dans les contextes étudiés. Ils soulignent a contrario qu'une considération dissociée des phases d'entraînement exploitant l'information textuelle, puis les sources de connaissances sélectionnées, induit souvent une dégradation des plongements obtenus par la simple analyse de corpus et l'implémentation de l'hypothèse distributionnelle. Ainsi, bien que le retrofitting permette dans certains cas d'améliorer notablement les plongements de mots au regard des performances des systèmes qui les utilisent, e.g. système de classification, son utilisation mérite d'être finement contrôlée et évaluée de manière contextuelle.

La prise en compte efficiente de connaissances pour l'enrichissement des plongements sémantiques obtenus par l'analyse de textes offre de nombreuses perspectives intéressantes. Ce sujet reçoit encore aujourd'hui une attention particulière, e.g. dans le contexte du Meta-Embedding⁴¹ [36]. Les plongements statiques, i.e. non contextuels, généralement moins performants que les plongements contextuels, e.g. issus de Transformers [265], restent très populaires dans le cadre de ressources limités (textes, ressources de calcul). La proposition de techniques capables d'améliorer leur performance au regard d'une tâche précise représente alors un véritable enjeu. De manière plus générale, toute approche capable d'adapter de manière orientée et sans dégradation des plongements de mots peut contribuer à permettre de lever certaines limitations des techniques de plongement. Cela permettrait notamment de lutter contre des biais introduits par la prise en compte de certaines

41. Technique de mariage de plusieurs plongements issus de différentes sources d'informations/de connaissances, sans accès à ces sources.

données, e.g. biais induits par l'usage statistique des genres dans de larges corpus de textes, cf. « *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* » [37]. Des résultats intéressants ont entre autres été publiés récemment sur le domaine des similarités sémantiques [288], dans l'utilisation de ce type de techniques pour encoder l'aspect émotionnel des mots [235], ou dans des domaines spécifiques tels que le domaine biomédical [92, 43]. Nous ne pouvons qu'encourager les efforts d'évaluation de ces approches orientées par de la connaissance. Elles laissent entrevoir un meilleur contrôle des plongements au regard d'une tâche visée. Comme nous l'avons souligné, ces approches s'avèrent de plus généralement moins coûteuses en termes de ressources de calcul par rapport aux approches de l'état de l'art à base de Transformers par exemple.

Sur un sujet proche, nous avons aussi mené des travaux sur l'utilisation couplée de bases de connaissances et de ressources textuelles dans des contextes de classification. En lien avec nos travaux sur le *fact-checking* et la vérification d'assertions, nous avons étudié les incidences de la prise en compte d'un graphe de connaissances sur les performances d'un modèle visant à prédire le domaine (*topic*) d'une assertion [22]. Nous avons pour cela utilisé le graphe de connaissances ClaimsKG [253] qui propose et structure des informations diverses sur des assertions, e.g. source, date, affirmation au format textuel. . . En comparant la performance de modèles fondés sur des plongements construits via l'analyse de ClaimsKG (plongement des noeuds des *claims*), de leur représentation textuelle, ou des deux types d'information, nous montrons notamment que, dans le contexte étudié, les techniques de plongement de graphes de connaissances peuvent s'avérer complémentaires des techniques de plongement de textes, même si les augmentations de performances - non systématiques - restaient généralement modestes.

Extraction de connaissances

Nous avons mené différents travaux d'extraction de connaissances à partir de textes. Certains d'entre eux seront discutés dans la seconde partie de ce manuscrit en lien avec l'Apprentissage Machine. Nous nous concentrons ici sur ceux qui mobilisent des connaissances exprimées dans des bases de connaissances. Nous présentons en particulier une approche que nous avons proposée pour inférer des connaissances via l'analyse de relations extraites de textes. Cette étude exploratoire a été menée dans un contexte de découverte de connaissances à l'aide d'approches semi-automatisées et guidées par l'analyse de larges corpus de textes. Elle s'inscrit dans le domaine d'étude dit de l'inférence textuelle (*textual inference*) [183], qui n'adopte pas une approche formelle de l'inférence telle que l'on peut l'entendre en inférence de connaissances à partir de bases de connaissances. Nous étions en ce qui nous concerne intéressés par des approches de type modèles d'inférence hybrides mariant extractions de textes et bases de connaissances formelles, e.g. *Knowledge Vault* [76]. A noter, comme nous le verrons ensuite, que ces approches se distinguent d'approches qui visent l'enrichissement (ou le peuplement de bases de connaissances), e.g. via des techniques de liage d'entités existantes dans une base, et d'extraction de relations entre ces entités, e.g. [181].

Collaborateurs :

- Doctorant encadré : Pierre-Antoine Jean (2014-2017).
- Collaborateurs d'IMT Mines Alès : Jacky Montmain, Sylvie Ranwez.
- Collaborateur externe : Patrice Bellot (LSIS, Marseille).

Les travaux mentionnés dans cette section ont pour l'essentiel été menés dans le cadre de la thèse de Pierre-Antoine Jean [132].

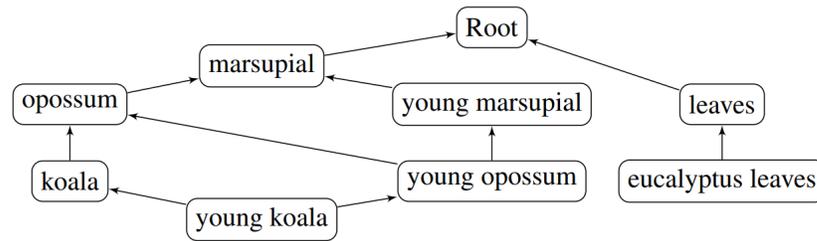


FIGURE 10 Exemple d'ordre partiel construit à partir des syntagmes 'young koala' et 'eucalyptus leaves' issus de la déclaration (*young koala, eat, eucalyptus leaves*) et d'une taxonomie existante dans laquelle nous avons $Koala \prec Opossum \prec Marsupial$ [133].

Différents systèmes d'extraction de relations permettent d'analyser des textes afin de distinguer des relations susceptibles d'être intéressantes dans un processus de découverte de connaissances. Nous proposons ci-dessous un exemple de relation extraite par le système REVERB [85] : à partir du texte « *The mineralization of the psammoma bodies is induced principally by the collagen fibers synthesized by the meningocytes [. . .]* » la relation suivante sera extraite : (*the mineralization of the psammoma, induced by, the collagen fibers*) [132]. Ces systèmes engagent des techniques plus ou moins complexes, certaines étant parfois de simples heuristiques, e.g. pour l'extraction des syntagmes nominaux ou l'identification d'entités provenant de Wikipedia.

Suite aux phases d'extraction, les relations obtenues par ces systèmes - du type (*sujet, prédicat, objet*) - font cependant très souvent référence à des sujets, objets, voire prédicats différents. Cela empêche le croisement naïf des relations extraites sur la base d'une mise en correspondance directe des éléments constitutifs des relations. Il est donc difficile dans ce contexte de croiser l'information portée par les différentes relations extraites. Les deux observations (*young koala, eat, eucalyptus leaves*) et (*koala, eat, leaves*) seront jugées sans lien, excepté en ce qui concerne leur prédicat. Pour lever cette limitation, nous avons cherché à croiser ces informations en intégrant la prise en compte de connaissances *a priori*. Ces connaissances, couplées à des règles d'analyse *ad hoc* que nous proposons, ont notamment été exploitées afin de structurer les objets d'étude des relations (e.g. syntagmes nominaux) dans un ordre partiel. La Figure 10 présente un exemple simple qui illustre la création d'un *graphe de syntagmes*. Celui-ci correspond à un ordre partiel construit à partir des syntagmes 'young koala' et 'eucalyptus leaves' issus de la relation (*young koala, eat, eucalyptus leaves*), et d'une taxonomie existante dans laquelle $Koala \prec Opossum \prec Marsupial$ [133]. La structuration obtenue intègre des implications syntaxiques du type 'young koala' $\prec_T koala$ et des implications conceptuelles exprimées dans une taxonomie.⁴² Les règles de construction proposées permettent de structurer à la fois les sujets, les objets, et aussi les prédicats.

La structuration des constituants des relations nous permet d'envisager la construction d'une structuration des relations. Différentes règles de structuration sont là aussi considérées pour obtenir un ordre partiel sur les relations (cf. Figure 11). La figure 12 propose un exemple d'ordre partiel structurant les objets et valeurs de la relation (*koala, eat, eucalyptus leaves*) en intégrant de la

42. La notation \prec_T souligne que l'ordre induit par les implications syntaxiques que nous admettons, et celui défini dans la base de connaissances méritent d'être distingués. L'implication syntaxique découle simplement ici de la propriété d'inclusion entre les mots des entités [108]. Les éventuels modificateurs syntaxiques sont perçus comme un moyen de spécialiser l'entité.

	$b \equiv d$	$b \prec d$	$d \prec b$	$\neg(b \preceq d) \wedge \neg(d \preceq b)$
$a \prec c$	$s_1 \prec s_2$	$s_1 \prec s_2$	$s_1!s_2$	$s_1!s_2$
$c \prec a$	$s_2 \prec s_1$	$s_1!s_2$	$s_2 \prec s_1$	$s_1!s_2$
$a \equiv c$	$s_1 \equiv s_2$	$s_1 \prec s_2$	$s_2 \prec s_1$	$s_1!s_2$
$\neg(a \preceq c) \wedge \neg(c \preceq a)$	$s_1!s_2$	$s_1!s_2$	$s_1!s_2$	$s_1!s_2$

FIGURE 11 Exemples de règles de construction pour deux déclarations $s_1 = (a, p, b)$ et $s_2 = (c, p, d)$ avec a, b, c et d des syntagmes et p un prédicat. Nous notons $s_2 \prec s_1$, la déclaration s_1 qui subsume la déclaration s_2 (au sens de l'implication), $s_1 \equiv s_2$ lorsque les deux déclarations sont équivalentes et $s_1!s_2$ lorsque s_1 ne peut être ordonnée avec s_2 [133].

connaissance existante. Cette simple structuration permet d'initier une phase de découverte de connaissances en nous invitant à critiquer la pertinence des relations induites, e.g. (*opossum, eat, leaves*) est potentiellement induite par le fait que nous observons (*koala, eat, eucalyptus leaves*) et qu'un koala est un opossum. ⁴³

Il reste alors à évaluer la pertinence des différentes relations induites. Nous proposons que cette évaluation soit elle aussi guidée par la structuration de la connaissance que nous avons construite. Nous définissons pour cela un modèle de propagation qui permettra *in fine* de critiquer la confiance à accorder à une relation. Ce modèle repose sur une stratégie de propagation de l'information proche de celle adoptée dans le cadre théorique des fonctions de croyance (la *belief* sert par exemple ici le principe de généralisation au sens de l'induction). Cette dernière intègre le degré d'incertitude exprimé dans le texte, un aspect sur lequel nous reviendrons en Section 2.3.1. Elle permet d'estimer la confiance à accorder à une relation du fait des observations disponibles. Différents schémas d'exploration de connaissances visant à gérer la dualité généralité/informativité, en incluant ici la dimension incertitude, sont notamment discutés. Des modèles de sélection tenant compte des profils des utilisateurs sont aussi proposés. Ceux-ci permettent d'intégrer leur niveau de connaissance et leur manière de rechercher de l'information en préférant, en fonction des contextes, des restitutions abstraites et des supports plus ou moins marqués pour les relations mises en avant. La figure 13 présente l'approche générale que nous avons proposée et étudiée.

Exemple d'extraction [132] :

Phrase : *La kératose solaire favorise la survenue de mélanomes.*

Relation extraite : (*kératose solaire, favorise, mélanome*)

Relations générées (i.e. à évaluer, interprétation existentielle) :

- (*kératose solaire, favorise, cancer de la peau*)
- (*kératose solaire, favorise, cancer*)
- (*lésion cutanée, favorise, mélanome*)
- (*lésion cutanée, favorise, cancer de la peau*)
- (*lésion cutanée, favorise, cancer*)

43. Dans le cas de l'inférence inductive, ces relations auront une interprétation existentielle et non pas universelle (i.e. $\exists x \in \text{opossum}, \exists y \in \text{leaves} (x, \text{eat}, y)$). Nous ne précisons pas ici ce point important discuté dans [132].

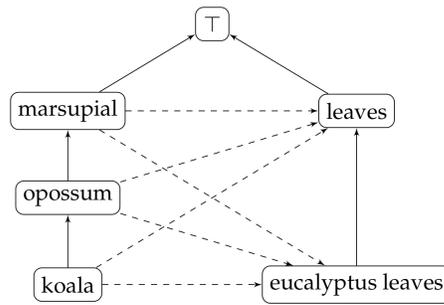


FIGURE 12 Exemple de relations induites par la relation initiale (*koala, eat, eucalyptus leaves*).

Ces travaux ont été exploités dans le cadre de la mise en place de systèmes de réponse à un questionnaire exploitant des données textuelles en langue anglaise issues du Web. Les résultats obtenus permettent de faire la démonstration de l'apport de la prise en compte de la connaissance *a priori* pour notre modélisation.

Les structurations manipulées, e.g. pour structurer les sujets et les objets, mariaient concepts, modificateurs associés aux entités, e.g. adjectifs, et même parfois des termes non désambiguïsés. L'approche étudiée se base sur une décomposition syntaxique et taxonomique des sujets et objets qui est détaillée dans [132]. La procédure de construction du graphe de syntagmes nominaux a été proposée en tenant compte de particularités syntaxiques de la langue anglaise (adaptations possibles pour d'autres langues). Nous rappelons que cette approche exploratoire, qui peut surprendre par la naïveté des constructions qu'elle permet, a été adoptée dans un contexte de définition d'approches de découverte de connaissances, semi-automatiques, et guidées par l'analyse de larges corpus de textes. Elle illustre les multiples difficultés à lever pour coupler l'usage de représentations des connaissances et de données non structurées porteuses de sémantiques de natures différentes.

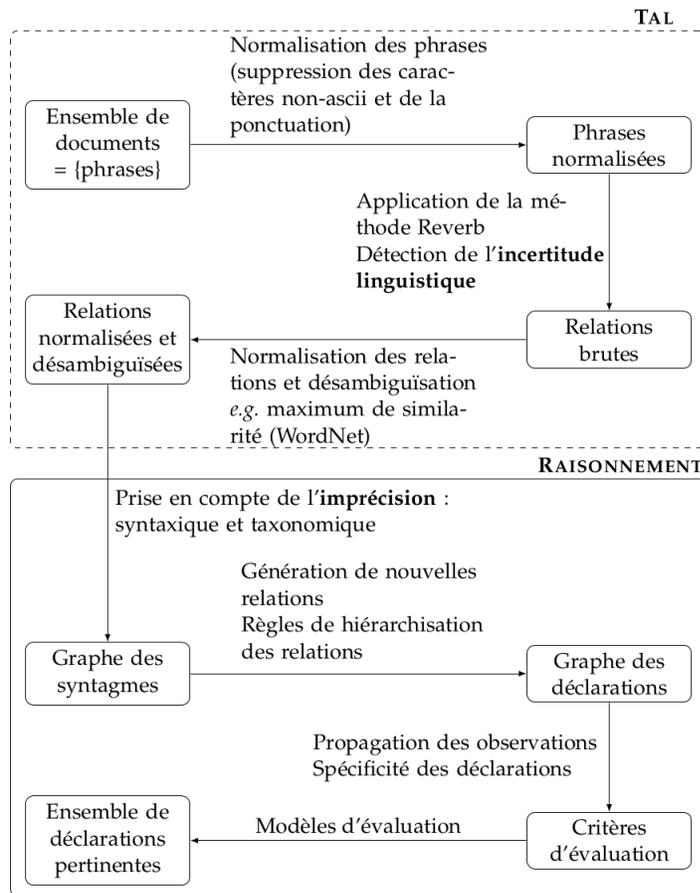


FIGURE 13 Approche générale.[132]

2.2.4 Découverte de règles d'association à partir de données imparfaites

Nous sommes capables de tirer des enseignements de nos expériences vécues, et ces enseignements peuvent s'exprimer dans certains cas sous la forme de règles qui pourront être utilisées dans des processus décisionnels, e.g. en vue d'anticiper des situations potentiellement préjudiciables. Identifier des règles pertinentes à partir d'observations est cependant un problème complexe [287, 242]. Comme nous allons le voir dans les développements qui suivent, le problème est suffisamment complexe pour rester ouvert même lorsque l'on se restreint à l'extraction de règles dans des systèmes formels. La discussion de la finalité même, i.e. de ce que l'on souhaite identifier, n'est pas triviale. Dans la continuité des travaux sur l'inférence textuelle, et plus particulièrement sur l'extraction de relations à partir de textes que nous avons exposés dans la section précédente, nous présentons ci-après nos travaux sur l'extraction de règles à partir de bases de connaissances.

Ces travaux ont pour l'essentiel été réalisés dans le cadre de l'étude d'une approche de Retour d'Expérience (RetEx) proposée en vue d'inférer des enseignements de façon semi-automatisée par l'analyse d'expériences passées. Nous introduisons dans un premier temps l'approche générale adoptée pour la mise en place du système de RetEx avant de centrer plus spécifiquement nos propos sur quelques aspects techniques de notre méthode d'extraction de règles.

Collaborateurs :

- Doctorante encadrée : Cécile L'Héritier (2017-2020).
- Stagiaire encadrée : Mengling Zhan (2018).
- Collaborateurs d'IMT Mines Alès : Abdelhak Imoussaten, Gilles Dusserre (Pharmacien).
- Collaborateur externe : Benoît Roig (Chimiste, Université de Nîmes).

Les travaux mentionnés dans cette section ont pour l'essentiel été menés dans le cadre de la thèse de Cécile L'Héritier [158].

Contexte de l'étude

Le Retour d'Expérience (RetEx) vise à enrichir une *organisation apprenante*, à l'échelle de l'individu ou du groupe, par la valorisation de connaissances issues d'expériences passées. Il suscite naturellement un fort intérêt de la part de décideurs conscients du caractère stratégique des connaissances liées à leur organisation (e.g. optimisation des processus, réduction des accidents). Le RetEx propose pour cela un cadre conceptuel et différentes démarches plus ou moins formelles, pour aborder les volets d'étude allant du recueil à la représentation et à l'exploitation des connaissances. Nos travaux se sont concentrés sur l'automatisation du RetEx en vue d'inférer, à partir d'expériences passées formalisées et potentiellement entachées d'incertitude, des connaissances générales utiles pour de futures prises de décisions stratégiques. Ils ont plus spécifiquement porté sur la définition d'une approche semi-automatisée de RetEx qui se base sur le couplage de techniques de représentation des connaissances, d'analyse multicritère, et d'analyse de données. Comme nous allons le voir, l'approche que nous proposons permet (i) la valorisation de données non structurées et hétérogènes, (ii) de nature imprécise, (iii) dans un contexte d'observations limitées (peu d'expériences), et (iv) intégrant de manière tant que possible efficiente l'expertise subjective des acteurs du RetEx. Le RetEx permis par notre système s'exprime sous la forme de règles d'association formelles qui

serviront potentiellement les futurs processus de prise de décision au sein de l'organisation ciblée. L'évaluation du système a été réalisée dans le cadre d'une illustration de mise en oeuvre dans le domaine humanitaire, aussi nous parlerons ci-après de missions pour faire référence à la notion d'expérience.⁴⁴

La figure 14 présente le schéma général de l'approche étudiée et ses trois phases principales :

1. **Représentation et formalisation des connaissances** : constitution d'une base regroupant les connaissances sur les missions. Les connaissances sont tirées d'entretiens et de phases d'extraction sur des données semi-structurées, e.g., formulaires, notes textuelles, fichiers tableur [154]. L'objectif de cette étape est de structurer et de formaliser tant que possible la connaissance propre aux missions (localisations, ressources), mais aussi celle relative au contexte d'intervention (e.g. contextes géopolitiques, équipements). Nous nous sommes en particulier intéressés à la formalisation du triptyque *Contexte, Stratégie* et *Évaluation* pour caractériser les missions et leurs évaluations. La représentation des connaissances choisie s'appuie sur des vocabulaires contrôlés propres au domaine étudié⁴⁵ et sur les différents langages standardisés par le W3C que sont RDF, RDFS et OWL. La figure 15 présente une illustration de graphe RDF structurant quelques éléments de connaissance relatifs à une mission humanitaire. Nous admettons la présence de connaissances imprécises sur le sujet ou l'objet d'un triplet. Si l'on considère la figure 18, on admettra par exemple l'expression de connaissances sur une mission *m* du type (*m*, *aPourTheatre*, *x*) avec (*x*, *rdf:type*, *paysAfricain*) - on notera par ailleurs (*m*, *aPourTheatre*, *paysAfricain*). Nous ne détaillerons pas plus précisément les processus susceptibles d'être engagés pour la constitution d'une telle base de connaissances. Ceux-ci appartiennent à l'état de l'art, ne sont que partiellement automatisés, et restent nécessairement dépendants du domaine étudié et bien entendu de la nature des données, de leur état de numérisation, de leur structuration, de leur qualité. . .
2. **Évaluation et identification des critères d'intérêt** : identification des critères importants pour orienter l'extraction de connaissances [156]. L'objectif est d'aider les décideurs à préciser les connaissances susceptibles d'être intéressantes afin de restreindre tant que possible l'espace de recherche des facteurs de succès ou d'échec d'expériences passées. Nous centrons nos efforts de découverte des connaissances sur les *Évaluations* obtenues au regard des *Stratégies* adoptées dans des *Contextes* rencontrés. La discussion liée aux critères d'évaluation est alors centrale. Nous avons proposé une approche permettant d'identifier, dans un contexte spécifique, le critère influençant le plus le processus de décision au regard de l'évaluation globale d'un ensemble de missions. L'approche se base sur Electre tri, une méthode de surclassement issue de l'analyse multicritère qui repose sur la comparaison par paires des alternatives (alternatives à classer et de référence) [91]. Ce processus de comparaison d'alternatives est souvent plus intuitif pour les décideurs car il leur est généralement plus naturel de comparer les forces et faiblesses de deux alternatives plutôt que de leur assigner un score arbitraire. Cette approche d'évaluation leur permet ainsi de rester dans leur référentiel métier sans les contraindre à avoir recours à des artefacts pour se prononcer en faveur d'une alternative. Nous soulignons en aparté que, à notre connaissance, l'identification de la contribution des critères à la performance globale n'avait jusque-là pas été traitée pour les méthodes de surclassement. Nous ne détaillerons pas les aspects techniques de ce travail dirigé par A. Imoussaten et exposé dans [156]. Ce traitement nous permet d'identifier les critères qui semblent revêtir un intérêt

44. Ce projet a été réalisé en collaboration non contractuelle avec l'organisation Médecins Sans Frontières (MSF).

45. e.g. MOAC, *Management of a Crisis Vocabulary* pour le domaine humanitaire [179].

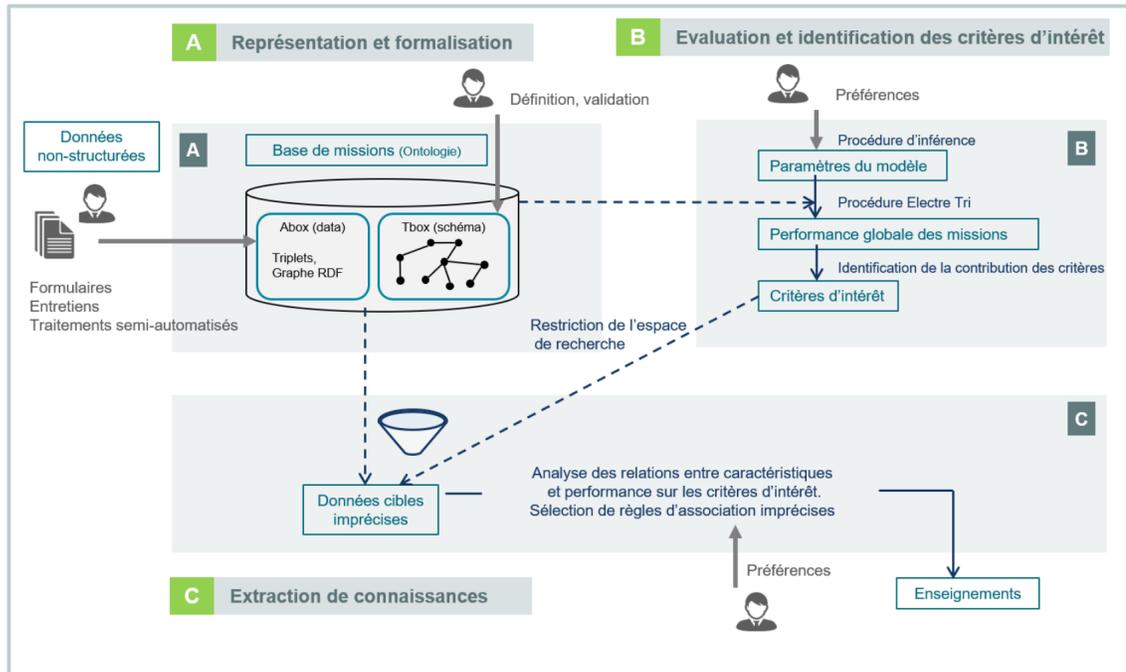


FIGURE 14 Schéma général de l'approche de Retex proposée. Source : [158].

particulier pour le décideur pour l'extraction de connaissances ; nous nous concentrerons ci-après sur l'identification de règles qui portent sur ces critères, e.g. si le critère « intégrité physique et mentale des opérateurs sur le terrain » est implicitement jugé plus important que le critère « intégrité physique et mentale des autochtones », nous considérerons un biais visant à préférer l'identification de règles relatives au premier critère.

- 3. Extraction de connaissances sous la forme de règles d'association :** analyse des relations entre les caractéristiques des missions et les performances sur les critères d'intérêt pour la sélection de règles d'association [155].⁴⁶ Nous proposons une procédure permettant de distinguer les règles les plus pertinentes au regard d'informations caractérisant leur intérêt vis-à-vis d'un contexte d'étude et des critères pré-identifiés en phase 2. Cette procédure tire parti de la connaissance formalisée dans la base de connaissances (phase 1) et permet une interaction étroite avec les décideurs lors de la phase subjective et contextuelle de sélection des règles pertinentes. Cette partie est détaillée ci-après.

Méthode d'extraction de règles d'association

Une règle d'association r , notée $r : X \rightarrow Y$, est une quasi-implication entre un antécédent X et un conséquent Y , de sémantique informelle : si X , alors nous avons souvent Y .⁴⁷ A une règle pourront être associées des mesures qui préciseront le crédit à lui accorder et son degré de validité au regard du contexte d'extraction.

46. Cet article discuté ci-après est joint au manuscrit, cf. Annexe A page 173.

47. Une règle d'association n'est pas une règle logique, i.e. qui n'accepte aucun contre-exemple. On parlera dans certains cas de règle statistique.

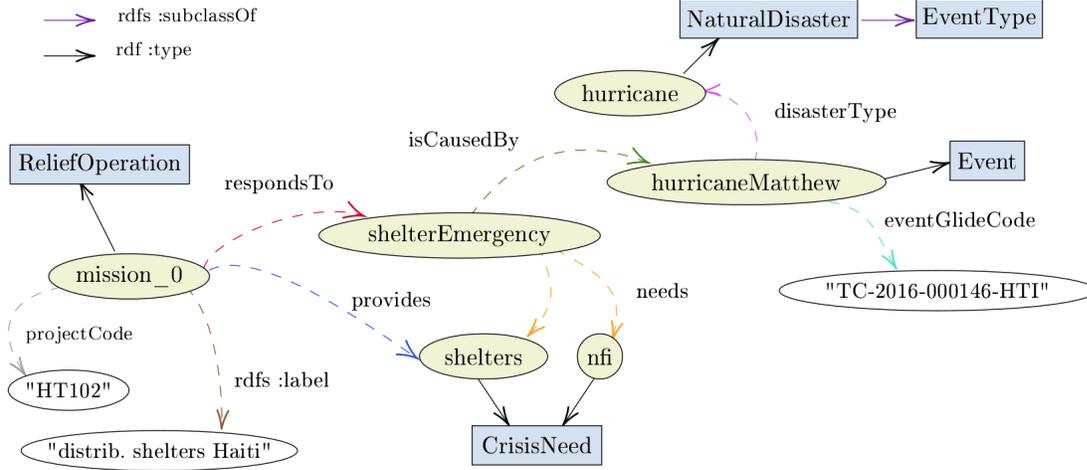


FIGURE 15 Illustration de graphe RDF structurant des éléments de connaissance relatifs à une mission humanitaire. Source : [158].

La notation $r_{X,Y}$ sera utilisée pour faire référence à une règle $r : X \rightarrow Y$. Nous admettrons de plus $N = \{1, \dots, n\}$ un ensemble d'attributs de domaines de valeurs discrets respectifs $\Theta_i, i \in N$. Nous aurons alors $X \in \prod_{i \in I} \Theta_i, I \neq \emptyset, I \subset N$ et $Y \in \prod_{j \in J} \Theta_j, J \neq \emptyset, J \subset N \setminus I$ (I et J formant une partition de N). Nous traitons dans notre cas des conjonctions de valeurs d'attributs, i.e. r équivaut à :

$$r_{XY} : \bigwedge_{i \in I} X_i \rightarrow \bigwedge_{j \in J} X_j$$

Nous nous autoriserons à noter $r_{XY} : \prod_{i \in I} X_i \rightarrow \prod_{j \in J} X_j$ pour faire le lien avec les cadres théoriques que nous mobiliserons.

Représentation d'une règle

Notez que, par souci de simplification, contrairement aux règles extraites dans certaines bases de connaissances, nous n'explicitons pas la plupart du temps les prédicats : une règle peut alors être vue comme un vecteur de valeurs d'attributs. Considérons l'exemple de projets humanitaires définis par 3 attributs [156] : Pays, Saison et ConditionSanitaire. Leurs domaines de valeurs respectifs sont : $\Theta_1 = \{\text{mali, haiti, indonesie}\}, \Theta_2 = \{\text{pluies, sécheresse}\}$ et $\Theta_3 = \{\text{malnutrition, paludisme, choléra}\}$. Une règle de la forme $r : X \rightarrow Y$, i.e. $r_{X,Y}$, pourrait alors être $\text{mali} \wedge \text{pluies} \rightarrow \text{paludisme}$. Elle pourrait suggérer que toute activité au Mali lors de la saison des pluies tend à induire des cas de paludisme. L'antécédent est $X = (\text{mali, pluies}) \in \Theta_1 \times \Theta_2$; le conséquent est $Y = \text{paludisme} \in \Theta_3$. Nous admettrons par ailleurs que cette règle peut s'écrire dans une forme vectorielle $[\text{mali, pluies, paludisme}]$, comme instance du patron de règle prédéfini $\langle \text{Pays, Saison, ConditionSanitaire} \rangle$.⁴⁸

Plus généralement, nous admettrons que la sémantique d'une règle est définie par une sélection (i) d'attributs et (ii) de leurs relations (éventuellement indirectes). Ces deux éléments ne seront pas explicités dans l'expression vectorielle d'une règle. On peut cependant considérer de manière implicite qu'une requête SPARQL définit des domaines d'exploration de règles sur lesquels nous

48. En FOL on pourrait avoir $\text{estLocaliséeAu}(\text{?mission, mali}) \wedge \text{aPourMétéo}(\text{mali, pluies}) \rightarrow \text{perturbéePar}(\text{?mission, paludisme})$.

```

SELECT ?Pays ?Saison ?MoyenTransport ?EvaluationReactivité WHERE {
  ?m rdf:type ex:Mission .
  ?m aPourTheatre ?Pays .
  ?m ex:SaisonIntervention ?Saison .
  ?m ex:aPourMoyenTransport ?MoyenTransport .
  ?m ex:aPourEvaluation ?e .
  ?e ex:evalReactivité ?EvalReactivité
}

```

FIGURE 16 Requête SPARQL utilisée pour extraire les attributs étudiés pour la découverte de règles.

nous concentrerons, e.g. un patron de graphe RDF qui permet pour chaque observation (e.g. mission) d'obtenir le vecteur des valeurs $\langle \text{Pays, Saison, MoyenTransport, EvalReactivité} \rangle$. Un exemple de requête est proposé Figure 16. Cette approche correspond à l'utilisation de *templates* ou schémas, classiquement retrouvés dans le domaine de la découverte de règles d'association [146, 101]. Notez que, du fait de la notation adoptée, les variables n'apparaissent pas ; les règles ainsi formulées seront en pratique implicitement connectées et fermées (cf. [160]).

Base d'observations

Notre processus d'extraction de règles considère une base d'observations pour un vecteur d'attributs d'intérêt prédéfini tel que $\langle \text{Pays, Saison, MoyenTransport, EvalReactivité} \rangle$. Nous admettrons l'existence d'une base de connaissances composée d'observations associées à des vecteurs de ce type, e.g. :

```

{
  [Mali, Eté, Camion, Bon],
  [Haïti, Eté, Camion, Mauvais]
}

```

Dans l'extraction *classique* de règles l'analyse porte sur des informations précises telles que la valeur de l'attribut i est $X_i \in \Theta_i, i \in N$. Nous avons souhaité travailler sur un cadre plus général qui permet l'expression d'imprécisions sur les valeurs d'attributs, i.e. $X_i \subseteq \Theta_i$. Ainsi lorsque $X_i \subset \Theta_i$, avec $|X_i| > 1$ nous modélisons l'imprécision sur la valeur de l'attribut i , et lorsque $X_i = \Theta_i$ nous exprimons l'ignorance.

Exemple [156] : Cela nous permet de modéliser l'expérience suivante : « *La mission à Haïti fait suite à une catastrophe de type géophysique, les délais d'approvisionnement ont été mauvais voire très mauvais* ». Si l'on considère qu'une catastrophe de type géophysique fait référence à $\{\text{tsunami, seisme}\}$, nous pouvons exprimer l'expérience par $\{\text{haiti}\} \times \{\text{tsunami, seisme}\} \times \{\text{tresMauvais, mauvais}\}$, i.e. $\{\{\text{haiti}\}, \{\text{tsunami, seisme}\}, \{\text{tresMauvais, mauvais}\}\}$ si l'on considère le vecteur d'attributs $\langle \text{Pays, TypeCatastrophe, EvalApprovisionnement} \rangle$. Cela correspond à exprimer que l'expérience vécue est un élément de l'ensemble suivant $\{ [\text{haiti, tsunami, tresMauvais}], [\text{haiti, tsunami, mauvais}], [\text{haiti, seisme, tresMauvais}], [\text{haiti, seisme, mauvais}] \}$ sans pouvoir être plus précis sur l'élément de connaissance en question. La Figure 17 présente une base d'observations pour le patron $\langle \text{DisasterType, Season, Environment, HumanCost} \rangle$.

	Disaster Type	Season	Environment	Human Cost
d_1	{ <i>earthquake</i> }	{ <i>autumn</i> }	{ <i>rural</i> }	{ <i>medium</i> }
d_2	{ <i>tsunami</i> }	{ <i>autumn</i> }	{ <i>urban</i> }	{ <i>medium</i> }
d_3	{ <i>epidemic</i> }	-	{ <i>urban</i> }	{ <i>veryHigh</i> }
d_4	{ <i>earthquake, epidemic, tsunami</i> }	{ <i>spring</i> }	-	{ <i>high, veryHigh</i> }
d_5	{ <i>epidemic</i> }	{ <i>spring</i> }	{ <i>urban</i> }	{ <i>high</i> }
d_6	{ <i>epidemic</i> }	{ <i>spring, summer</i> }	-	{ <i>high, veryHigh</i> }
d_7	{ <i>epidemic</i> }	{ <i>spring, summer</i> }	{ <i>urban</i> }	{ <i>high, veryHigh</i> }
d_8	{ <i>epidemic</i> }	{ <i>spring, summer</i> }	{ <i>urban</i> }	{ <i>veryHigh</i> }
d_9	{ <i>earthquake, epidemic, tsunami</i> }	{ <i>summer</i> }	{ <i>rural</i> }	{ <i>high</i> }
d_{10}	{ <i>epidemic</i> }	{ <i>summer</i> }	{ <i>urban</i> }	{ <i>high</i> }
d_{11}	{ <i>epidemic</i> }	{ <i>summer</i> }	{ <i>urban</i> }	{ <i>veryHigh</i> }
d_{12}	{ <i>earthquake</i> }	{ <i>winter</i> }	{ <i>rural</i> }	{ <i>high, medium, veryHigh</i> }
d_{13}	{ <i>earthquake</i> }	{ <i>winter</i> }	{ <i>rural</i> }	{ <i>low</i> }
d_{14}	{ <i>earthquake, epidemic, tsunami</i> }	{ <i>winter</i> }	{ <i>rural</i> }	{ <i>high</i> }

FIGURE 17 Exemple de base d'observations pour 4 attributs. Source : [158].

Vers la notion de règle imprécise

La considération de connaissances imprécises nous invite à généraliser la notion de règle :

$$r_{X,Y} \text{ avec } X = \prod_{i \in I} X_i, X_i \subseteq \Theta_i \text{ et } Y = \prod_{j \in J} Y_j, Y_j \subseteq \Theta_j \quad (2.5)$$

Pour rappel, les domaines de définition des valeurs des attributs, i.e. Θ_k ($k \in N$), sont discrets et composés d'éléments distingués.

Nous notons \mathcal{R} l'ensemble des règles définies par la formule 2.5, i.e.

$$\mathcal{R} = \{r_{X,Y} | X \in \prod_{i \in I} 2^{\Theta_i}, Y \in \prod_{j \in J} 2^{\Theta_j}\}$$

L'objectif sera alors de réduire \mathcal{R} à un sous-ensemble de règles valides et pertinentes pour le domaine concerné. Notons cependant que l'espace de recherche des règles est très important, nettement plus grand que celui rencontré dans le traitement de règles précises ; en effet, sans contrainte, nous considérerons *a priori* $\mathcal{R} = \prod_{k \in N} 2^{\Theta_k}$.

Parenthèse technique : pour être précis nous admettons en pratique des observations dans $2^{\prod_{k \in N} \Theta_k}$ et non pas dans $\prod_{k \in N} 2^{\Theta_k}$. Cela nous permet de modéliser des connaissances imprécises du type « *La mission à Haïti utilisait des camions Renault au diesel, ou des camions Peugeot à l'essence* ». Pour le vecteur d'attributs <Pays, Transport, Carburant> cela correspond à exprimer que l'expérience vécue est un élément de l'ensemble suivant { [haïti, camion_Renault, diesel], [haïti, camion_Peugeot, essence] }. Cette connaissance ne peut être exprimée si l'on considère des connaissances exprimées dans $\prod_{k \in N} 2^{\Theta_k}$; on serait obligé dans ce cas de considérer par exemple l'éventualité de [haïti, camion_Renault, essence] qui est pourtant incompatible avec la connaissance véhiculée initialement. Cependant, même si nous admettons des observations dans $2^{\prod_{k \in N} \Theta_k}$, nous nous concentrerons sur la découverte de règles définies dans $\prod_{k \in N} 2^{\Theta_k}$ de manière à ne pas avoir à composer avec une explosion de l'espace de recherche.⁴⁹ Cela induira cependant une

49. Avec 6 attributs de 4 valeurs, nous obtenons un espace d'observations de taille $\prod_{i=1}^6 2^4 = 2^{24}$ dans le premier cas, et de taille $2^{4^6} = 2^{4096}$ dans le second !

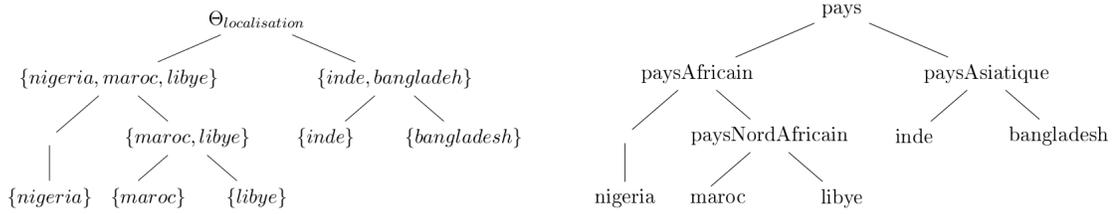


FIGURE 18 Représentation des connaissances de type taxonomie. Source : [158].

forte augmentation (i) de la complexité algorithmique de certains traitements, et (ii) de la technicité de l'implémentation pratique de l'approche. Nous éluderons ci-après les technicités induites par cet aspect et le lecteur pourra conceptuellement considérer des observations dans $\prod_{k \in N} 2^{\Theta_k}$ s'il le souhaite.

Nous nous sommes dans notre cas intéressés aux règles pour lesquelles l'antécédent et le conséquent sont des ensembles identifiés dans le sens où l'antécédent correspond aux caractéristiques des expériences, et le conséquent correspond aux performances observées sur les critères d'intérêt identifiés via l'approche mentionnée ci-avant (cf. phase 2). Nous considérons de plus que le conséquent est un unique attribut (de performance dans notre étude). On retrouve ici notre vecteur générique d'attributs $\langle \text{Attribut}_1, \text{Attribut}_2, \dots, \text{Attribut}_{\text{Eval}} \rangle$. Bien que nous souscrivions à cette restriction par la suite, l'approche proposée s'applique au cas général, e.g. conséquents composés de plusieurs attributs.

Connaissance a priori

Nous utilisons une représentation des connaissances qui ordonne partiellement les valeurs précises et imprécises qui peuvent être prises par les attributs, i.e. $\mathcal{O} \subseteq \mathcal{O}^* = (\preceq, \bigcup_{i \in N} 2^{\Theta_i})$. Cette représentation peut être la simple taxonomie de classes précisées dans la TBOX de notre base de connaissances. Elle correspondra cependant très généralement à une structure issue d'extractions d'instances, de classes et de leurs relations. La figure 18 illustre une représentation possible des valeurs de l'attribut Localisation : $\Theta_{\text{localisation}} = \{\text{nigeria}, \text{maroc}, \text{libye}, \text{inde}, \text{bangladesh}\}$. En adoptant une approche de raisonnement en monde fermé, cette représentation exprime que $\text{paysAfricain} = \{\text{nigeria}, \text{maroc}, \text{libye}\}$ et donc que mentionner paysAfricain correspond nécessairement à mentionner une de ces trois instances. A noter que si l'on veut s'autoriser à raisonner en monde ouvert, nous pouvons admettre une instance fictive supplémentaire pour tout noeud (*concept*) associé à un sous-ensemble non singleton, e.g. en définissant $\text{paysAfricain} = \{\text{nigeria}, \text{maroc}, \text{libye}, \text{autre_paysAfricain}\}$, la mention d'un paysAfricain ne ferait ainsi pas nécessairement référence à un pays présent dans notre base.

Sémantique d'une règle

Il est simple de comprendre une règle de la forme $X \rightarrow Y$ lorsque X et Y sont précis, i.e. X et Y correspondent à des éléments de l'ensemble de discernement. L'extension de la notion de règle à des cas pour lesquels les éléments de l'antécédent ou du conséquent sont imprécis mérite cependant d'être discutée. De manière assez directe, nous pouvons admettre une interprétation universelle en considérant que pour une règle $r_{X,Y} : \bigwedge_{i \in I} X_i \rightarrow \bigwedge_{j \in J} Y_j$, $X_i \subseteq \Theta_i$, $Y_j \subseteq \Theta_j$, cela correspond à admettre comme vraies toutes les implications induites par $\prod_{i \in I} X_i \times \prod_{j \in J} Y_j$. Cela pose cependant souci dans le cas de prédicats fonctionnels ; à titre d'exemple $\{\text{haiti}\} \times \{\text{tsunami}\} \times \{\text{mauvais}$,

`tresMauvais`} consisterait à admettre deux conséquences différentes sur un même prédicat fonctionnel, i.e. $haiti \wedge tsunami \rightarrow mauvais$ et $haiti \wedge tsunami \rightarrow tresMauvais$. Nous étudierons en particulier une interprétation existentielle d'une règle : toute règle admet dans l'absolu au moins une vérité dans l'ensemble identifié par ladite règle, i.e. au moins une quasi-implication vraie parmi $\prod_{i \in I} X_i \times \prod_{j \in J} Y_j$. L'objectif sera cependant pour nous d'identifier via ce moyen détourné une règle générale $r_{X,Y} : \bigwedge_{i \in I} X_i \rightarrow \bigwedge_{j \in J} Y_j, X_i \subseteq \Theta_i, Y_j \subseteq \Theta_j$ qui maximise le nombre de règles précises valides généralisées par $r_{X,Y}$, i.e. qui peuvent s'écrire sous la forme $\bigwedge_{i \in I} x_i \rightarrow \bigwedge_{j \in J} y_j, x_i \in \Theta_i, y_j \in \Theta_j$.

A noter de plus que la notion de règle existentielle que nous admettons se distingue de celle parfois retrouvée en représentation des connaissances - pour laquelle les travaux n'étudient pas la découverte de règles à partir de données, et où les règles visent l'expression de connaissances non définies explicitement dans une base de connaissances [254].⁵⁰ La notion de règle imprécise que nous étudions renvoi plutôt aux travaux relatifs à l'extraction de règles associatives en tenant compte d'une hiérarchie d'items (e.g. [287, 90]⁵¹), ou à la notion d'extraction contextuelle d'*itemset* dans une base de connaissances [222], même si dans ce cas les informations traitées sont seulement précises.

Synthèse du contexte d'étude :

- l'antécédent et le conséquent sont définis (sous-ensembles d'attributs, e.g. `<Pays, Saison, MoyenTransport, EvaluationReactivité>`); et le conséquent ne concerne qu'un attribut.
- les attributs prennent des valeurs discrètes dans un ensemble fini et connu.
- nous admettons des valeurs imprécises d'attribut.
- les informations manquantes sur les valeurs d'attribut sont traitées de la même manière que l'ignorance totale.
- nous admettons une connaissance *a priori* identifiant une restriction de la structuration sous forme de POSET des valeurs que peuvent prendre les attributs, i.e. \mathcal{O} , cf. Figure 18.

Évaluation des règles

Nous nous sommes dans un premier temps intéressés à la définition de mesures d'intérêt objectives pour évaluer les règles. Nous avons dans cette étape exclu toute mesure subjective qui nécessiterait la mise en place d'interactions plus approfondies et contextuelles avec l'utilisateur du système. Nous avons pour cela proposé de centrer l'évaluation sur des critères classiquement utilisés dans la littérature, à savoir : (i) le support, et (ii) la confiance. Ces derniers ont été adaptés au cas des règles imprécises. Nous avons aussi souhaité intégrer la prise en compte (iii) d'évaluations indirectes, et (iv) du degré d'imprécision des règles.

La notion de support a été modélisée à l'aide de la *Belief* (nombre d'observations implicites ou explicites de la règle), $supp(r_{X,Y}) = Bel(X \times Y)$. La confiance, traditionnellement évaluée comme une probabilité conditionnelle (nombre d'observations de l'association antécédent/conséquent, par rapport au nombre de fois où l'antécédent est observé), a été modélisée par une croyance

⁵⁰. e.g. avec *projectField* un prédicat qui relie un projet à son domaine d'étude, on peut exprimer la connaissance « tout chercheur d'un domaine d'expertise réalise un projet dans ce domaine » par : $\forall X \forall Y (Researcher(X) \wedge hasExpertise(X, Y) \rightarrow \exists Z (projectField(Z, Y) \wedge isMemberOf(X, Z)))$ - adapté de [254].

⁵¹. A noter que la sémantique de ces règles - parfois dites incertaines dans la littérature - n'est à notre connaissance pas discutée dans ces travaux.

$A \rightarrow \bar{B}$	(-)	si <i>vérifiée</i> tend à invalider $A \rightarrow B$
$\bar{A} \rightarrow B$	(-)	si <i>vérifiée</i> tend à diminuer la pertinence de $A \rightarrow B$ (i.e. B vrai dans n'importe quel contexte)
$\bar{A} \rightarrow \bar{B}$	(+)	si <i>vérifiée</i> tend à conforter la croyance de $A \rightarrow B$

FIGURE 19 Illustration de l'interprétation à donner aux règles impliquant le complémentaire de l'antécédent ou du conséquent d'une règle d'intérêt. Source : [158].

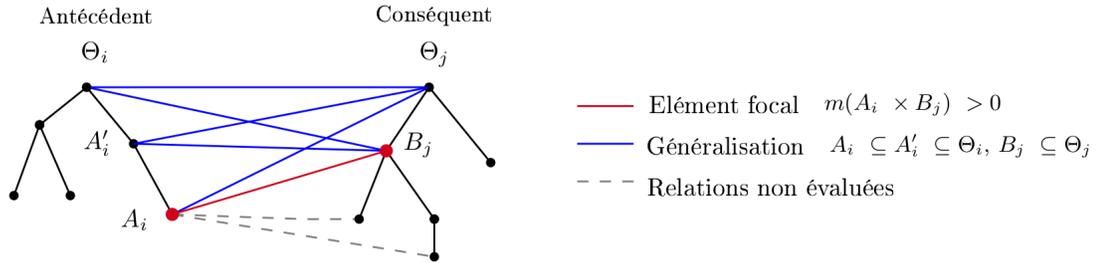


FIGURE 20 Structuration des règles sous la forme d'un ordre partiel. Source : [158].

conditionnelle en adoptant la formalisation proposée par Fagin et Halper [86] :

$$\begin{aligned} \text{conf}(r_{X,Y}) &= \text{Bel}(Y|X) \\ &= \frac{\text{Bel}(X \times Y)}{\text{Bel}(X \times Y) + \text{Pl}(X \times \bar{Y})} \end{aligned}$$

avec Pl la fonction de plausibilité et \bar{Y} le complémentaire de Y .

Différentes mesures indirectes de la pertinence éventuelle d'une règle ont aussi été évaluées en tenant compte de l'analyse des règles suivantes $r : \bar{X} \rightarrow \bar{Y}$, mais aussi $r : X \rightarrow \bar{Y}$, $r : \bar{X} \rightarrow Y$ (en intégrant leur support).⁵² Un processus d'évaluation global a été distingué sur la base des indicateurs de support et de confiance attribués à ces règles (illustration Figure 19).

Nous avons aussi intégré la prise en compte de la spécificité d'une règle en considérant une formulation dérivée de la notion de contenu informationnel intrinsèque proposée dans le cadre des mesures sémantiques (IC de Seco [117]. Celui-ci repose dans notre cas sur une formulation qui tire parti de l'ordre partiel défini sur les règles, cf. Figure 20).

Une approche itérative d'élagage de l'espace de recherche \mathcal{R} est considérée. Nous considérons dans un premier temps des restrictions intuitives de \mathcal{R} au regard de la connaissance disponible, à savoir :

1. Restriction aux règles supportées par des observations, i.e : $\mathcal{R}_r = \{r_{X,Y} \in \mathcal{R} | \exists Z \subseteq X \times Y, m(Z) > 0\}$, avec $m : 2^\Theta \rightarrow [0, 1]$ une fonction de masse standard, cf. Figure 20.

⁵². Contrairement au cadre probabiliste, dans le contexte d'une modélisation crédibiliste les règles $r : X \rightarrow \bar{Y}$, $r : \bar{X} \rightarrow Y$ fournissent des informations supplémentaires, cf. [158] pour une discussion détaillée.

2. Restriction aux concepts exprimés dans la représentation des connaissances. Pour un domaine $\Theta_i, i \in N$, il existe des sous-ensembles de Θ_i qui ne correspondent pas à des *concepts a priori* utilisés par le domaine concerné. Ceux-ci ne seront pas retrouvés dans la base de connaissances et dans sa TBox que nous notons \mathcal{O} . A titre d'exemple, dans la Figure 18, aucun concept n'est exprimé pour mentionner $\{\text{inde, libye}\}$. Nous appliquons donc une restriction pour nous intéresser aux règles pour lesquelles les éléments de l'antécédent et du conséquent sont retrouvés dans l'ontologie, i.e. : $\mathcal{R}_{r,t} = \{r_{X,Y} \in \mathcal{R}_r \mid \forall Z \in \{X_i\}_{i \in I} \cup \{Y_j\}_{j \in J}, Z \in \mathcal{O}\}$.⁵³

Nous nous sommes ensuite intéressés à la définition d'un processus de sélection de règles. Ce processus tient compte de l'ensemble des métriques et de leur potentielles relations conflictuelles (e.g. dualité support, spécificité). Deux étapes sont distinguées :

1. **Élagage des règles fondé sur la relation de dominance** ; nous évaluons des relations de dominance par l'analyse des mesures d'intérêt étudiées (e.g. support, confiance) : une règle r_1 domine une règle r_2 , si et seulement si il n'existe pas une mesure pour laquelle r_2 est préférable à r_1 , et il en existe au moins une pour laquelle r_1 l'est par rapport à r_2 .
2. **Élagage des règles en appliquant Electre I**. Nous utilisons une approche d'analyse multicritères de surclassement pour identifier indirectement la préférence du décideur. L'objectif est ici aussi, si besoin, de distinguer les règles surclassées au regard des préférences du décideur afin d'élaguer plus encore la restriction de l'espace de recherche obtenue jusque-là.

Une illustration détaillée de l'approche été proposée dans le cadre d'une application au domaine de l'humanitaire [158]. Les évaluations des différentes composantes de l'approche sont détaillées dans les contributions afférentes [159, 156, 154, 157, 155, 158]. Une implémentation C++ de l'approche est par ailleurs partagée à la communauté <https://github.com/sharispe/KRE>.⁵⁴

Discussion

Les pistes d'amélioration de ces travaux sont nombreuses :

- Étude des complexités algorithmiques induites par les généralisations et l'optimisation des traitements. La complexité algorithmique des approches étudiées dépend largement de l'expressivité des règles et des bases d'observations. Des avancées importantes en termes de réduction de complexité seront requises pour une mise en oeuvre sur d'importantes bases d'observations.
- Généralisation au cas des attributs continus. Smets propose une généralisation des fonctions de croyance aux valeurs réelles [241]. Elle ne permet pas cependant à elle seule d'explicitier de manière directe la modalité de propagation de l'information qui mériterait d'être considérée lorsque les observations portent sur des attributs hétérogènes. Un travail spécifique mériterait d'être réalisé afin de permettre l'intégration de ce type de données.
- Intégration de métriques permettant d'inclure des appréciations subjectives. L'approche de surclassement que nous avons étudiée se base sur l'analyse de métriques objectives. La littérature propose aussi différentes mesures subjectives qui permettent de critiquer la pertinence de règles dans des contextes spécifiques, en incluant de manière importante les utilisateurs finaux dans le processus de sélection [146].

⁵³. Il est possible de lever cette restriction qui peut s'avérer limitante, en particulier dans un processus de découverte de connaissances ou lorsque la TBox est pauvre en concepts.

⁵⁴. Le code ne comporte que la partie liée au calcul des métriques, i.e. la partie la plus coûteuse en temps de calcul. La partie impliquant Electre I se base sur un outil existant.

Une discussion qui nous semble particulièrement intéressante porte sur la sémantique des règles que l'on cherche à extraire. Il n'est pas en soi satisfaisant, pour une règle : r_{XY} , d'étudier une interprétation existentielle du type : il existe au moins une quasi-implication dans $X \times Y$. Dire par exemple `PaysSahel` \implies `risqueElevé` correspond, au regard du sens commun, à exprimer une règle *universelle* de type *pour tout* : $\forall x \in \text{PaysSahel}, \text{evalRisque}(x, \text{risqueElevé})$. Permettre l'extraction de ce type de règle nécessiterait de reconsidérer largement notre approche. Il nous faudrait entre autres revisiter l'expression du support que nous proposons (i.e. via la croyance, *belief*). Le cadre théorique des fonctions de croyance, en tout cas dans l'utilisation que nous en faisons, ne serait très probablement pas adapté. Qui plus est, comme nous l'avons mentionné dans cette section, dépasser une interprétation disjonctive des règles imprécises (i.e. l'interprétation existentielle et non pas universelle) nécessitera de prendre en compte le caractère inconsistant de certaines *règles*, ou de préciser une interprétation des règles consistante avec certaines propriétés des prédicats, e.g. leur caractère fonctionnel `PaysSahel` \implies `{risqueElevé,risqueTrèsElevé}` n'impliquera pas l'existence de plusieurs niveaux de risque pour un même pays, mais plutôt $\forall x \in \text{PaysSahel}, \text{evalRisque}(x, \text{risqueTrèsElevé}) \vee \text{evalRisque}(x, \text{risqueElevé})$. Nous ne développons pas plus encore les perspectives qui mériteraient d'être données à ces travaux. L'extraction de ce type de règles universelles via l'analyse de bases de connaissances n'est cependant, à notre connaissance, pas traitée dans la littérature.⁵⁵ L'extraction de ce type de règles semble cependant importante dans le contexte d'une augmentation à la fois des volumes de données disponibles et de leur niveau de structuration.

55. Les approches existantes ne tiennent pas compte de la structuration des valeurs, excepté en autorisant la substitution par des valeurs relatives à des concepts, e.g. [2]. Elles se restreignent très souvent à l'étude de la ABox, et à des faits pour lesquels les sujets et objets correspondent à des instances.

2.3 Apprentissage de modèles prédictifs à partir de données

Nous présentons dans cette partie nos contributions relatives à l'Apprentissage Automatique. Celles-ci s'intéressent principalement à deux paradigmes de l'apprentissage machine que sont l'apprentissage supervisé et l'apprentissage par renforcement. Elles s'inscrivent dans différents domaines d'étude tels que le Traitement Automatique du Langage Naturel, la Recherche Opérationnelle, et les Sciences du Mouvement Humain. Comme nous le verrons tout au long de cet exposé, nos travaux ont un fort ancrage applicatif. Cependant, même si la plupart de nos contributions sont proposées dans des contextes applicatifs particuliers, et tiennent compte de spécificités propres aux domaines traités (e.g. nature des données), certaines d'entre elles n'en sont pas moins génériques. Ces dernières seront alors introduites dans une section distinguée.

Nous adopterons la structuration suivante pour présenter ces travaux :

1. Contributions au Traitement Automatique du Langage Naturel : cette section aborde nos travaux sur l'étude de techniques d'apprentissage automatique pour l'analyse de textes. Nous traiterons de différentes tâches - e.g. classification de textes, extraction d'entités, analyse d'arguments, indexation - étudiées avec différents modèles d'apprentissage automatique, e.g. réseaux de neurones, champs aléatoires conditionnels (CRF). Ces travaux ont pour la plupart été proposés pour l'analyse de textes juridiques, en particulier liés à la jurisprudence française.
2. Contributions à l'étude de problèmes d'optimisation combinatoire : cette section présente nos travaux exploratoires sur la résolution de problèmes de tournées de véhicules à l'aide d'apprentissage profond et par renforcement. Nous nous intéresserons entre autres au cadre de l'optimisation combinatoire neuronale (*Neural Combinatorial Optimization*), et à évaluer l'utilisation de modèles d'apprentissage automatique en substitution des approches exactes et heuristiques classiquement utilisées pour aborder cette classe de problèmes d'optimisation.
3. Contributions disciplinaires en Apprentissage Automatique : cette section présente nos contributions sur l'interprétabilité de modèles de types réseaux de neurones et l'optimisation du mécanisme d'attention utilisé dans certains d'entre eux.

2.3.1 Contributions au Traitement Automatique du Langage Naturel

Analyse Automatique de corpus judiciaires

Nos travaux dans ce domaine portent sur l'étude de différentes méthodes d'analyse sémantique de corpus de décisions jurisprudentielles. Ils traitent pour l'essentiel de problématiques de classification de textes et d'extraction d'informations.

Collaborateurs :

- Doctorants encadrés : Gildas Tagny NGompe (2015-2020), Charles Condevaux (2018-2022), Pierre-Antoine Jean (2014-2017).
- Stagiaire encadrée : Anastasiia Prysiazniuk (2017).
- Collaborateurs d'IMT Mines Alès : Jacky Montmain.
- Collaborateurs externes : Stéphane Mussard (économiste, Université de Nîmes), Guillaume Zambrano (juriste, Université de Nîmes).

Les travaux mentionnés dans cette section ont pour l'essentiel été menés dans le cadre des thèses de Gildas Tagny NGompe [250] et de Charles Condevaux [57].

La jurisprudence française fait référence à l'ensemble du corpus de décisions judiciaires prononcées en France. Elle représente l'interprétation des lois faite par les cours de justice françaises pour résoudre les contentieux jugés. Elle est donc importante pour comprendre la prise de décision des juges, ce qui la rend par conséquent indispensable pour les juristes. Son analyse permet d'évaluer le risque judiciaire et d'anticiper des décisions de justice, e.g. au regard d'un contexte d'étude particulier et d'une stratégie de défense. En effet, en pratique, pour anticiper l'issue d'un procès, un juriste recherche souvent des affaires similaires à celle qu'il doit traiter afin d'évaluer les chances que son client obtienne un verdict qui lui sera favorable. La jurisprudence est aussi par essence centrale afin d'étudier l'application du droit en France, e.g. étudier la variabilité des décisions de justice dans les territoires, ou à partir de caractéristiques des parties engagées.

La jurisprudence correspond quantitativement à des centaines de milliers, voire millions de documents textuels non ou semi-structurés, qui forment le corpus de décisions de justice.⁵⁶ Une analyse manuelle exhaustive est par conséquent impossible. Dans ce contexte, les juristes travaillent généralement sur des extractions identifiées à l'aide de moteurs de recherche payants qui ne permettent alors qu'une appréciation partielle de la jurisprudence. Cette situation empêche aussi tout particulier d'évaluer simplement le risque judiciaire pour un cas d'intérêt. Au-delà de sa difficulté d'accès, la jurisprudence est de plus prononcée et rédigée dans un langage judiciaire difficilement accessible pour le non spécialiste. Dans ce contexte, nous avons étudié des méthodes pour automatiser l'analyse des décisions de justice en vue de structurer la jurisprudence et de faciliter son utilisation à des fins d'analyses descriptives et prédictives. Nous traitons pour l'essentiel des problématiques d'indexation et de classification de demandes en particulier pour la mise en place de systèmes de justice prédictive. Nous avons pour cela concentré nos travaux sur l'automatisation de tâches importantes pour l'analyse métier de la jurisprudence, à savoir :

1. La segmentation des décisions et la détection d'entités d'intérêt. Les décisions sont organisées en plusieurs sections qui détaillent différentes informations sur le cas traité, les demandes formulées, et les décisions prononcées. La détection des sections qui structurent les décisions permet de faciliter de multiples traitements d'extraction d'informations utiles pour l'analyse de la jurisprudence, e.g. extraction des entités juridiques d'intérêt (localisation de la cours, parties, mentions aux citations de lois). L'extraction de ces informations permet de proposer une première structuration de la jurisprudence à l'aide de laquelle pourront être réalisées des analyses descriptives de la jurisprudence. Elle permet aussi d'envisager la mise en place de moteurs de recherche avancés, e.g. recherche par cours, par région. . . L'identification d'entités juridiques d'intérêt peut aussi servir la tâche d'anonymisation.
2. L'identification des demandes des parties. Il existe différentes classes de demandes usuelles, e.g. Licenciement, Divorce. Les analyses de la jurisprudence sont souvent effectuées à l'aide de restrictions de corpus portant sur le type de demandes. Il convient donc d'identifier a minima la classe des demandes formulées dans une décision. L'automatisation de la caractérisation des demandes vise à faciliter la compréhension des réclamations des parties.
3. La détermination du sens du résultat. L'issue d'une décision pour une demande spécifique, e.g. d'une classe particulière, est une information importante pour l'analyse de la jurisprudence. Au-delà d'une analyse manuelle des cas de rejet ou d'acceptation par un juriste, la disponibilité de cette information permet aussi d'envisager l'utilisation de techniques automatisées d'analyse.

56. Un exemple de décision est proposé en annexe B.

Nous centrerons notre exposé sur les travaux liés à l'identification des demandes. Des éléments relatifs aux autres sujets traités ne seront que rapidement évoqués. Nous illustrerons la finalité de nos travaux par la présentation d'une application portant sur l'analyse descriptive d'un grand corpus de décisions judiciaires françaises. Les études menées sur ce domaine se basent pour la plupart sur des résultats expérimentaux obtenus sur des données réelles annotées manuellement. Ces données correspondent à des corpus de décisions françaises constitués durant nos études (par nos collaborateurs juristes), ou des corpus existants rendus disponibles par la communauté, e.g. décisions de la Cour de Justice de l'Union Européenne et de la Cour Européenne des Droits de l'Homme.

Identification des demandes des parties

La notion de demande est fondamentale dans le domaine juridique et en particulier pour l'étude de la jurisprudence. L'argumentation proposée pour une affaire a en effet généralement pour buts : (i) de faire accepter les demandes d'une partie, et (ii) de faire rejeter celles de la partie adverse. Ainsi, l'identification des demandes des parties est centrale dans les analyses réalisées par les juristes. Elle permet notamment, conjointement avec la décision prononcée en termes d'acceptation ou de rejet d'une demande, de juger de la pertinence d'une prétention dans un contexte donné.

Un premier travail réalisé avec les experts juristes a consisté à modéliser la notion de demande au regard des données susceptibles d'être extraites des décisions. Cinq attributs ont été distingués : l'objet de la demande (e.g., dédommagement), la norme qui la fonde (e.g., article de loi), les quantum demandé et obtenu (e.g. sommes d'argent), ainsi que le sens du résultat (acceptation ou rejet).

En accord avec les juristes impliqués dans le projet, nous avons considéré la notion de classe de demandes afin de regrouper les demandes fondées sur une même norme et qui ont un même objet, e.g. troubles du voisinage, licenciement, dommages-intérêts, déclaration de créance.⁵⁷ Les demandes d'une même classe portent en quelque sorte sur des prétentions semblables. Un ensemble de classes a ainsi été distingué par nos collaborateurs juristes.

Différentes tâches ont été proposées pour traiter l'identification des demandes des parties. Ces dernières correspondaient généralement à des tâches d'apprentissage automatique supervisé :

1. La classe d'une demande et le sens du résultat sont identifiés par la résolution de tâches de classification. La première tâche vise à savoir si une demande d'une classe est présente dans une décision, et la seconde si une demande d'une classe précise a reçu une acceptation ou un rejet.⁵⁸
2. Les quanta sont identifiés via l'analyse *ad hoc* de sections distinguées par un modèle d'étiquetage de séquences.⁵⁹
3. Les différentes informations extraites sont ensuite appareillées pour constituer une demande distincte.

57. Exemple de classe définie par un objet et une norme identifiés : dommages-intérêts pour concurrence déloyale, avec pour objet dommages et intérêts, et norme, l'article 1382 du code civil. La notion de classe de demande se base sur la nomenclature des affaires civiles NAC+ qui distingue plus de 500 classes.

58. Il est rare que deux demandes d'une même classe soient formulées dans une même décision, cf. [250] page 74 pour des statistiques descriptives précises.

59. un modèle CRF semblable à celui présenté dans [202] est utilisé pour le sectionnement et l'extraction d'information des décisions, e.g. date, cf. section *autres travaux* page 94.

Nous présentons ci-après les traitements de classification de demandes, et d'identification du sens du résultat. Les autres traitements proposés pour la caractérisation des décisions ne sont pas présentés. En amont de la définition des approches d'apprentissage machine, nous avons proposé et implémenté un protocole pour l'obtention d'une base de décisions annotées à partir de laquelle ont été constitués nos jeux d'entraînement et d'évaluation. Des outils d'annotation ont été développés à cet effet et des juristes ont pris en charge le travail d'annotation qui a permis de distinguer entre 70 et 400 décisions par classe traitée [250, 60].

Classification des demandes

Puisqu'une décision de justice peut faire référence à plusieurs demandes de natures différentes, nous étions confrontés à un problème de classification multi-classes et multi-labels selon la terminologie de l'apprentissage machine. Une modélisation de type classification binaire a cependant été considérée afin d'éviter de devoir modifier le système de traitement de toutes les classes suite à l'ajout d'une nouvelle classe. Différentes approches ont alors été étudiées pour obtenir un classifieur par classe [251, 250, 61, 57].

Une des approches que nous souhaitons étudier visait l'obtention de bonnes performances tout en permettant d'identifier des passages de la décision susceptibles d'appuyer la classification fournie. Nous avons pour cela adopté une approche pragmatique en testant en première intention des modélisations simples qui se basent sur l'état de l'art exhaustif en classification de textes. A noter que le faible nombre d'observations labélisées par classe n'encourageait pas à engager des techniques d'apprentissage profond (sans pré-entraînement). La définition d'une approche *ad hoc* d'identification de caractéristiques a de plus été encouragée par le fait que, en pratique, un juriste arrive généralement très facilement et rapidement à déterminer une classe de prétentions à l'aide de certaines expressions ou de certaines lois - la technicité du domaine empêche cependant souvent la réalisation de la tâche par des non experts. L'approche complète que nous avons proposée est détaillée dans [61, 57]; nous nous contenterons d'en présenter les grandes lignes.

La tâche de classification des demandes a été décomposée en trois étapes classiquement retrouvées en classification de textes longs lorsque l'on dispose de peu d'observations labélisées : (i) identification d'un vocabulaire discriminant pour chaque classe, (ii) identification à l'aide de ce vocabulaire de passages des décisions potentiellement discriminants pour une classe à traiter (contexte de classification binaire), (iii) construction de représentations des décisions sur la base de l'analyse de l'occurrence des termes dans les passages sélectionnés, (iv) classification des représentations vectorielles par un classifieur standard. Des précisions sur les traitements réalisés pour chacune de ces étapes sont fournies ci-après.

Des modélisations fondées sur des métriques de l'état de l'art ont été utilisées pour identifier les vocabulaires discriminants, e.g. différentiel de la fréquence de termes au sein et à l'extérieur d'une classe, divergence de Kullback-Leibler, gain d'information. . . Le choix de la métrique (parmi celles testées) n'a par ailleurs eu que peu d'incidence sur le vocabulaire discriminant identifié pour une classe. Nous distinguons les termes « *voisin, travaux, troubles, habitation, propriété, huissier, propriétaire, astreinte, réparation, inconvénients, maison, constat, terrain, anormaux, construction* » pour la classe Troubles du voisinage, et les termes « *sérieuse, réelle, salarié, préavis, congés, salaire, rupture, payés, emploi, afférents, indéterminée, compensatrice, ancienneté, poste, durée, indemnité, rémunération* » pour la classe Licenciement. Les vocabulaires propres aux classes semblent intéressants mais restent imparfaits (e.g. *sérieuse* et *réelle* ne semblent pas en tant que tels discriminants de la classe Licenciement pour le non initié).

Litige	Par assignation en date du 20 février 2012, monsieur et madame M. ont fait assigner leurs voisins devant le tribunal de grande instance de Périgueux afin que le tribunal les condamne principalement à faire cesser le trouble anormal de voisinage subi par la présence de batraciens introduits dans une mare créée au pied de leur immeuble, ordonne d'une part la construction d'un mur pour avoir coupé une haie mitoyenne et d'autre part l'arrachage de bambous surplombant leurs panneaux photovoltaïques.
Motif	Il est donc établi en tout état de cause que les époux P. ont créé une mare sur leur propriété à moins de 10 m de la maison des époux M. alors qu'ils sont propriétaires d'un terrain d'une superficie de plus de 15 hectares.
Dispositif	Ordonne aux époux P. de combler leur mare située à moins de 10 mètres de l'habitation M. sous un délai de 4 mois après le prononcé du présent arrêt, ce sous astreinte provisoire de 150 euros par jour de retard et pendant un délai de 2 mois.

FIGURE 21 Phrases identifiées pour la classe Troubles du voisinage dans trois sections, source [57].

Une fois le vocabulaire identifié, celui-ci est utilisé pour sélectionner les passages caractéristiques de la classe étudiée dans une décision. Il est alors exploité pour déterminer si une phrase d'une décision appartient à une demande de la classe traitée. Nous adoptons pour cela une pondération des phrases tenant compte des métriques précitées : une phrase se voit attribuer un score par simple agrégation des valeurs de chacun de ses termes (max, moyenne. . .). Une sélection des phrases les plus discriminantes au regard de cette étude est ensuite réalisée, e.g. à l'aide d'une règle de sélection *ad hoc* de type top 10% des phrases ayant les scores les plus importants, comparaison par rapport aux scores moyens des phrases de la décision. . . La figure 21 présente des exemples de phrases sélectionnées - le sectionnement des différentes parties de la décision est réalisé par le modèle graphique mentionné précédemment (CRF) [201, 202].

L'approche de classification des décisions se base exclusivement sur les phrases identifiées comme potentiellement associées à la classe traitée. A noter que cela exclut de fait l'exploitation de passages de la décision susceptibles de faciliter l'identification de la négation de la classe. Les représentations des textes sont construites par une technique naïve qui évalue la présence et l'absence de chaque élément du vocabulaire identifié (caractéristique binaire). Différents classifieurs ont ensuite été testés (régression logistique, forêt aléatoire, classifieur bayésien naïf, SVM. . .).

Une des évaluations que nous avons faite pour ce type d'approches se base sur l'étude de 9 classes de demandes [57]. Le pré-traitement réalisé se limite à la suppression de la ponctuation et de la casse, et à la substitution des sommes d'argent par un symbole spécial. La tokenisation est réalisée à l'aide du séparateur espace (simples unigrams). Le cadre d'une validation croisée dans un contexte d'évaluation binaire (absence ou présence de la classe) a été utilisé. Les résultats obtenus en termes de précision, rappel et f-mesure, sont presque parfaits. Ils montrent même que de simples classifieurs linéaires permettent de classifier de manière satisfaisante les représentations des décisions construites. La tâche s'avère donc triviale et peut être traitée parfaitement dans la majorité des cas testés - les résultats sont semblables dans le cas d'une modélisation multi-classes. Ces

résultats ont été validés sur d'autres jeux de données dédiés à l'étude d'autres classes de demandes. Les résultats obtenus avec une variante de cette approche soulignent que la détection de la classe de demande ne comporte pas de réelle difficulté après identification des vocabulaires propres aux classes de demandes (f-mesure de plus de 0.98) [250].

Les résultats obtenus pour cette partie soulignent que l'utilisation d'approches traditionnelles mérite d'être préférée à l'utilisation de gros modèles dans le traitement de tâches facilement automatisables.

Détermination du sens du résultat

Nous présentons maintenant les travaux menés pour identifier le sens d'une décision, i.e. l'acceptation ou le rejet d'une demande d'une classe spécifique, en admettant qu'une seule demande du type en question ne soit mentionnée dans la décision. Cette tâche est nettement plus difficile que la précédente - une approche *ad hoc* naïve fondée sur (i) la sélection de passages de la décision par l'utilisation de termes associés à la classe de demande traitée, et (ii) l'identification du sens du résultat à partir de l'identification de termes fréquemment utilisés pour mentionner le sens d'un résultat permet d'obtenir des performances de f-mesure entre 0.33 et 0.72.

De nombreux algorithmes standard de classification ont été comparés sur des dizaines de représentations vectorielles usuelles de l'état de l'art (e.g. 55 méthodes de représentations vectorielles de textes sont comparées dans une de nos études [250]). Les résultats obtenus ont montré, sans réelle surprise, que la performance d'un classifieur est dans ce cas d'étude principalement influencée par (i) le très faible nombre de données labélisées qui défavorise largement certains modèles (ceux sensibles aux valeurs aberrantes, ou qui requièrent beaucoup d'observations pour converger), (ii) le fort déséquilibre entre les classes acceptation et rejet, (iii) l'hétérogénéité des demandes dans un même document, avec des demandes de types différents pour lesquelles des décisions opposées à celle de la classe étudiée sont souvent formulées.

Une approche de filtrage semblable à celle présentée ci-avant pour classifier les décisions a été utilisée pour pallier le problème induit par la présence de plusieurs demandes différentes dans un même document [57]. L'objectif est alors d'identifier les passages qui mentionnent l'énoncé de la décision pour la classe étudiée. La sélection des passages peut permettre des augmentations de performances pouvant aller jusqu'à 0.2 point de f-mesure.⁶⁰ Les performances et les degrés d'amélioration enregistrés restent fonction de la classe de demandes traitée.

Deux adaptations du modèle de régression Gini-PLS ont aussi été étudiées en vue de traiter le problème du faible nombre de données annotées [251, 250]. Le modèle Gini-PLS, proposé par S. Mussard (collaborateur) & Souissi-Benrejeb [199], correspond à une extension de l'analyse des moindres carrés partiels (PLS - *Partial Least Square*). Il vise à réduire la sensibilité aux valeurs aberrantes, ce qui peut être intéressant lors du traitement de petits jeux de données labélisées. Notre travail a porté sur l'adaptation du modèle, initialement proposé dans un contexte de régression, à la classification (plus spécifiquement une généralisation du Gini-PLS et une combinaison du modèle à la régression logistique ont été proposées). Ces deux adaptations permettent d'obtenir des performances proches de celles obtenues à l'aide d'arbres de décision. Ces travaux sont développés dans [251, 250].

60. Dans ce cas l'expérimentation avec et sans sélection des passages a été réalisée en considérant des représentations des textes construites par l'agrégation des plongements des termes des passages. Les plongements, de type FastText [139], étaient entraînés sur un corpus d'un million de décisions.

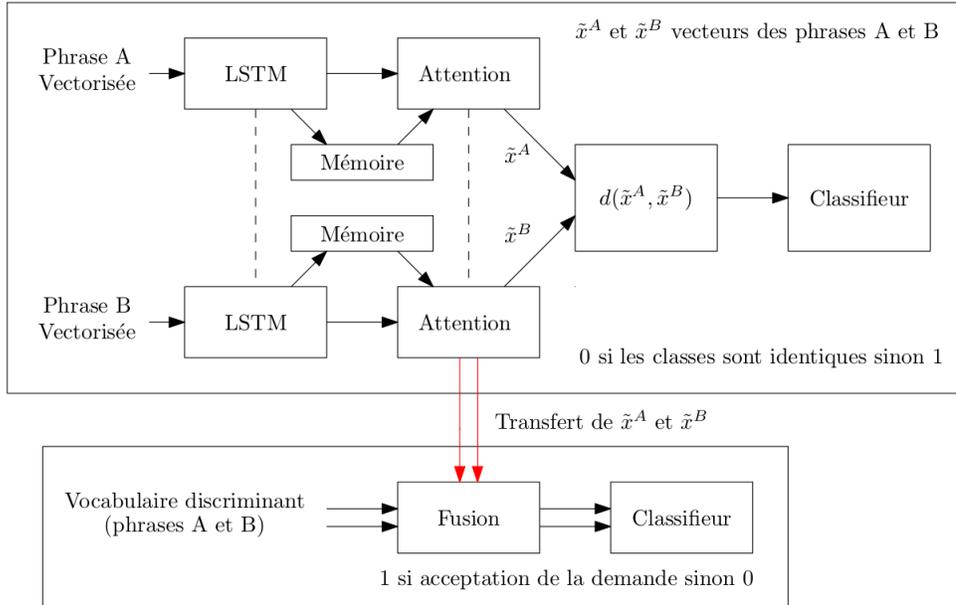


FIGURE 22 Réseau siamois étudié dans [60].

Le problème induit par le manque de données labélisées est fréquent. Le domaine juridique n'est pas épargné car, même si de larges corpus existent, les jeux de données labélisées sont le plus souvent rares et de très petite taille car coûteux à construire. Ce contexte empêche l'utilisation directe de techniques à base d'apprentissage profond supervisé. Celles-ci s'avèrent pourtant très performantes pour les tâches de classification pour lesquelles l'on dispose de jeux de données labélisées de taille importante. Nous avons donc axé nos travaux sur les techniques permettant de tirer parti des modèles d'apprentissage profond avec peu de données labélisées. Ces travaux ont plus spécifiquement porté sur l'étude de réseaux de neurones récurrents avec attention, dans un contexte de classification *one-shot* faiblement supervisée [61, 60].

L'approche *one-shot* que nous avons adoptée vise à apprendre une forme de distance entre les éléments à classer (ici des décisions), de manière à ce que les éléments d'une même classe soient jugés moins distants que des éléments de classes différentes. Ainsi, contrairement à ce qui est fait en apprentissage supervisé standard, on ne cherche pas à apprendre directement une correspondance entre une donnée fournie en entrée et une classe i.e., dans notre cas le lien entre une décision (textuelle) et une décision de justice pour une classe donnée. On cherche plutôt à apprendre si deux décisions (textuelles) contiennent toutes les deux l'acceptation d'une demande d'une même classe. L'objectif visé permet donc de générer beaucoup de données d'apprentissage en se basant sur un nombre limité de données labélisées propres au problème initial.⁶¹ L'implémentation de l'approche repose dans notre cas sur un réseau siamois [41] qui permettra l'estimation de la fonction de distance entre deux décisions ou passages de décisions. L'approche générale est introduite dans ses grandes lignes ci-après ; les détails sont exposés dans [61, 60, 57].

61. Avec n données labélisées nous pouvons générer $(n(n-1))/2$ comparaisons.

La figure 22 présente l’architecture considérée. Les données labélisées utilisées pour l’entraînement du réseau appartiennent à $\mathcal{D} \times \mathcal{D} \times L$, avec \mathcal{D} l’ensemble des décisions, et $L = \{0, 1\}$ les labels possibles, e.g. une donnée labélisée (A, B, y) avec $A, B \in \mathcal{D}$ deux décisions et $y = 0$ si A et B appartiennent à la même classe. L’architecture se compose de deux sous-réseaux symétriques qui, bien qu’ils partagent les mêmes paramètres (poids), traitent des entrées différentes, resp. A et B . La partie haute du schéma illustre la sous-architecture dédiée à l’apprentissage des représentations \tilde{x}^A et \tilde{x}^B . L’apprentissage est réalisé en entraînant un classifieur qui, à partir de ces représentations, doit distinguer si les éléments de la paire (A, B) traitée appartiennent ou non à la même classe. Cette partie repose sur un réseau récurrent (cellule LSTM) couplé à un mécanisme d’attention pour la construction de \tilde{x}^A et \tilde{x}^B . Ces représentations sont ensuite comparées pour obtenir une représentation $\tilde{x}^{d(A,B)}$ qui sera finalement l’entrée du classifieur de cette partie, chargé de discuter si A et B appartiennent à une même classe.⁶²

Cette composante de l’architecture pourrait directement être utilisée pour la prédiction de la classe d’appartenance d’une décision d’intérêt. Durant la phase d’inférence, une observation peut en effet tout à fait être comparée à d’autres observations labélisées des classes *acceptation* et *rejet*. Une technique de vote pourrait alors être utilisée pour l’attribution d’une classe à l’observation (e.g. vote majoritaire). Nous avons cependant préféré l’utilisation d’un second classifieur qui exploitera la représentation construite (e.g. \tilde{x}^A), car nous avons observé que les modélisations de ce type sont généralement plus stables et plus performantes. Cette seconde partie de l’architecture, illustrée par la partie basse de la figure, vise à classifier indépendamment A et B en tenant compte de leur représentations respectives \tilde{x}^A et \tilde{x}^B . Cette partie intègre par ailleurs une connaissance relative au vocabulaire discriminant de chaque classe (pour la décision A cela sera une simple concaténation de \tilde{x}^A et d’une représentation vectorielle traduisant l’observation des éléments du vocabulaire dans A). La représentation ainsi construite sera finalement dans notre cas traitée par un modèle linéaire. Le vocabulaire est construit automatiquement (50 n-grams qui maximisent un indicateur spécifique); il pourrait aussi correspondre à une connaissance *a priori*. L’entraînement de la totalité du réseau est finalement réalisé par minimisation de deux entropies croisées, respectivement dédiées à la tâche one-shot (appartenance des décisions à une même classe), et à la classification *acceptation/rejet*. Ce type de modélisation consiste dans l’absolu à considérer une tâche annexe de manière jointe à la tâche de classification traitée.

Nos évaluations du réseau siamois ont été réalisées pour 4 classes de demandes pour lesquelles nous avons constitué des jeux de données allant de 74 à 100 décisions labélisées. Parmi les tests effectués, nous avons aussi dans ce cas comparé l’incidence du choix de différents plongements lexicaux (FastText [139], ELMo [216], Flair [3], CamemBERT [184]). Sur ce point, nos résultats montrent que des plongements obtenus à l’aide de techniques moins complexes (e.g. FastText par rapport à CamemBERT), que l’on peut spécialiser à moindre frais sur de données spécifiques à la tâche de manière auto-supervisée, permettent d’obtenir des résultats compétitifs.⁶³

62. Techniquement dans notre cas, $\tilde{x}^{d(A,B)}$ est obtenu en considérant chaque dimension de l’espace latent indépendamment, e.g. par calcul de la valeur absolue de la différence de chaque composante de \tilde{x}^A et \tilde{x}^B .

63. Il est fréquent aujourd’hui de ne pas être en mesure de faire du *fine-tuning* auto-supervisé sur des corpus spécifiques à une tâche, du fait de la taille des modèles de langue - le plus souvent par manque de ressources de calcul ou à cause du coût induit par les calculs. Des modèles plus modestes restent souvent intéressants.

Plusieurs approches d'augmentation de données de l'état de l'art ont aussi été étudiées pour cette architecture. Une approche visant à générer des observations diversifiées par traduction s'est avérée intéressante et pourrait l'être dans d'autres contextes. L'idée est simple : il s'agit de traduire chaque entrée à traiter dans d'autres langues, puis de traduire à nouveau ces traductions dans la langue initiale. Cela permet dans notre cas de diversifier les décisions labélisées. A noter que cette approche permet d'améliorer nos résultats bien que dans notre cas la terminologie du domaine juridique soit parfois spécifique, et les tournures de phrases très souvent pour le moins particulières (voire clairement alambiquées).

Les résultats obtenus montrent que le fait de combiner un apprentissage one-shot avec le type d'architecture siamoise étudiée permet d'obtenir des modèles performants (scores de 0.81 à 1 de f-mesure). Ils soulignent aussi globalement l'apport de l'approche proposée par rapport à une approche classique de type TF-IDF (couplée à un classifieur standard). Cela est d'autant plus vrai lorsque des techniques d'augmentation de données sont utilisées. Il est en effet important de préciser qu'une approche de type TF-IDF permet de rivaliser avec l'approche one-shot pour certaines classes. Les résultats obtenus nous interpellent aussi sur le risque de surapprentissage auquel nous expose l'utilisation de ces techniques. Même si des résultats intéressants ont été observés, nous notons en effet la sensibilité des performances aux conditions de configuration (e.g. hyperparamètres du réseau ou de l'entraînement).

Autres aspects étudiés pour cette tâche : un aspect intéressant de l'approche repose aussi sur la possibilité d'étudier les scores d'attention calculés. La figure 23 présente un exemple de résultats pour différentes classes. Cela permet de mettre en évidence des termes qui semblent importants pour la classification. ⁶⁴

⁶⁴. Il y aurait beaucoup à dire sur ce point, e.g. le terme *légitime* est utilisé de manière contextuelle pour justifier des classifications différentes.

Au vu de ces éléments, la Cour estime que l'appelant justifie d'un intérêt légitime à porter désormais le prénom 'Michael'.	Acceptation
Attendu que la créance de l'expert comptable étant établie en son existence et son montant, elle sera admise pour la somme de 7.232,18 euros TTC, (tva au taux de 18,6 %) au passif de la procédure collective de la SNC.	Acceptation
Dès lors, même à supposer que le kilométrage réel de la voiture ait été supérieur de plus de 78000 km, cet élément ne suffit pas à caractériser un manquement à l'obligation de délivrance conforme.	Rejet
Pas plus que devant le premier juge, Colette B. ne produit à hauteur de Cour des preuves pertinentes attestant d'un intérêt légitime au changement de son prénom.	Rejet

FIGURE 23 Pondérations liées à l'attention [57].

Autres travaux

Étiquetage de séquences pour la segmentation des sections qui structurent les décisions de justice, et la détection d'entités juridiques, et de citations de lois.

Comme nous l'avons précisé, les décisions de justice sont des documents textuels semi-structurés voire non structurés (pour rappel un exemple est proposé en annexe B). En effet, bien que la structure d'une décision soit codifiée, les juges et les greffiers tendent à adopter différents modèles de rédaction ce qui induit en pratique une forte variabilité des transcriptions écrites des décisions prononcées par différentes cours (juridictions et localisations). Une décision en jugement civil est traditionnellement composée de quatre parties, cf. [250, 57] : (i) l'*en-tête* qui précise le numéro de référence, la date, le nom des parties, la localisation de la cour, (ii) le *litige* qui expose les faits, prétentions (demandes) et moyens des parties, (iii) les *motifs* qui visent à présenter les éléments de faits et de droit censés motiver la décision, (iv) le *dispositif* qui présente les conclusions pour chaque prétention.

La nature faiblement structurée des décisions rend difficile l'analyse de larges corpus de décisions et limite alors l'étude de la jurisprudence. Pour pallier ce problème, nous avons étudié différentes techniques de segmentation de textes et d'extraction d'information afin de détecter les sections qui structurent les décisions, d'extraire des entités juridiques d'intérêt (e.g. localisation de la cour), ou encore d'identifier les mentions aux citations de lois. Ces travaux ont amené la proposition de plusieurs modèles graphiques probabilistes de type modèles de markov cachés (HMM) et champs

aléatoires conditionnels (CRF). Au-delà de la proposition de modèles spécifiques à la tâche, nos travaux discutent la définition et la sélection de caractéristiques, la représentation des éléments traités (e.g. segments), et l'impact de la volumétrie de données labélisées sur la performance des modèles [201, 202].

Prise en compte de textes longs : les textes juridiques sont généralement de taille importante (plusieurs milliers de mots). Une approche d'approximation du mécanisme d'attention utilisé par des approches neuronales contemporaines, e.g. Transformer, a été étudiée dans le cadre de l'analyse des textes juridiques [57, 58]. Elle sera discutée en Section 2.3.3.

Étude des circonstances factuelles de classes de décisions. Un travail important du juriste consiste, durant l'analyse de la jurisprudence, à identifier les principaux contextes dans lesquels les demandes d'une classe donnée d'intérêt sont généralement formulées. Ce travail se base sur l'analyse de décisions proches d'un cas à traiter. Une de nos études sur l'analyse de la jurisprudence propose une métrique de distance entre décisions. Celle-ci se base sur l'apprentissage d'un modèle capable d'évaluer le degré de transformation nécessaire pour transformer une décision en une autre ; l'apprentissage du modèle est défini de manière supervisé à l'aide de transformations *ad hoc* appliquées de manière contrôlée sur des décisions. La distance entre deux décisions est ensuite évaluée comme fonction du degré de transformation inféré entre les deux décisions comparées [250]. Nous retrouvons ici une expression spécifique de l'approche actuellement adoptée par certains gros modèles de langue populaires de type *denoising autoencoder*, e.g. BART [170].

Recherche d'articles de loi pertinents pour une étude de cas (COLIEE 2017). L'objectif de ce travail était de formuler une approche permettant de faciliter l'analyse d'un cas juridique pratique décrit en langage naturel. Nous cherchions pour cela à retrouver un ou plusieurs articles du code civil susceptibles de permettre à un expert de prendre une décision sur la légalité du cas décrit (ici la requête). Cette étude a été réalisée dans le cadre de la compétition COLIEE 2017 (*Competition on Legal Information Extraction/Entailment*). Le code civil en question était le code civil japonais traduit en anglais. L'évaluation correspondait à celle d'un système de recherche d'informations textuelles, i.e. requête textuelle qui décrit le cas d'étude, et les documents résultats correspondaient aux articles du code civil.

La méthode que nous avons étudiée se base sur une réorganisation du code civil et la génération automatisée de descriptifs d'articles de loi permettant de tenir compte des références directes et indirectes entre les articles. Une forme d'augmentation des données était en quelque sorte réalisée avant la phase d'indexation. Une méthode standard de pondération de type BM25 était ensuite utilisée pour la mise en correspondance requête-articles à l'aide des scores d'appariement obtenus entre la requête et les descriptifs construits pour chaque article. La méthode est décrite dans l'annexe A de [132]. Elle a été classée 4ème sur 17 participants (0.4 points de différence de f-mesure par rapport au premier). Une implémentation de l'approche est proposée (<https://github.com/PAJEAN/Coliee2017>).

Application à l'analyse descriptive d'un grand corpus de décisions judiciaires françaises. Les travaux proposés dans le cadre de l'étude de techniques de TALN appliquées au domaine juridique ont amené le développement de différents outils informatiques utiles pour les juristes (outils d'annotation et d'analyse de la jurisprudence). Un exemple détaillé d'analyse rendue possible par les outils de classification et d'extraction d'information mentionnés est proposé dans la thèse de Tagny NGompe [250]. A titre d'illustration, la figure 24 présente une visualisation d'une analyse permise par les outils que nous avons proposés. Une restriction du corpus global de décisions est

2.3. Apprentissage de modèles prédictifs à partir de données

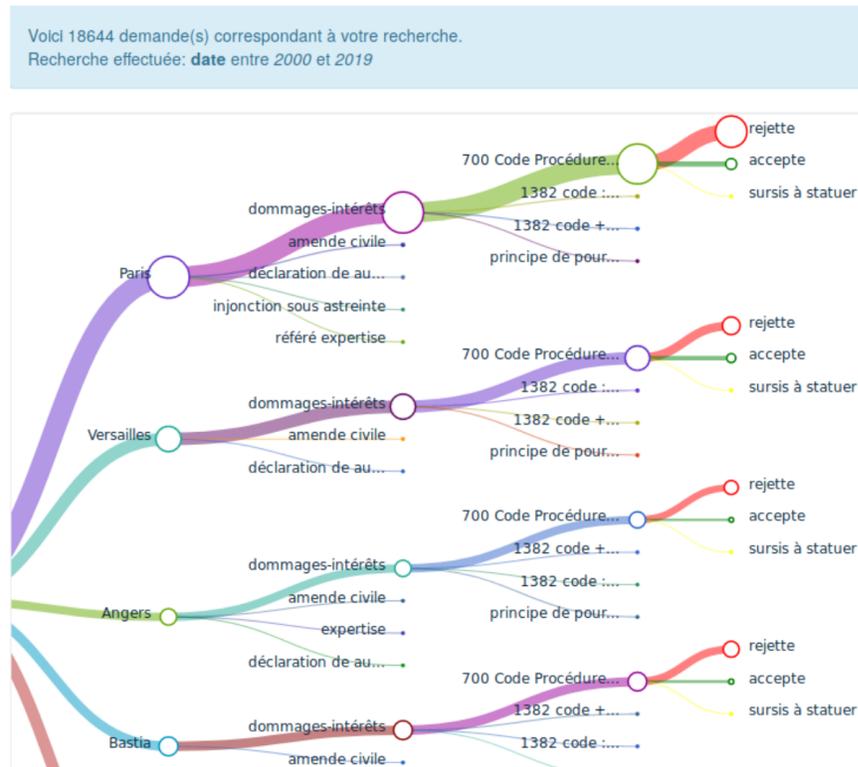


FIGURE 24 Comparaison des décisions prises dans différentes cours de justice. source : [250].

appliquée sur la base (i) d'une période précise (2000-2019), (ii) de cours de justice d'intérêt (Paris, Versailles, Angers, Bastia. . .), et d'une classe de demandes précise « *frais irrépétibles* »⁶⁵ - ces données sont extraites par les outils d'extraction d'information et de classification que nous avons proposés. Cette représentation permet d'observer graphiquement que les traitements des demandes de ce type sont semblables en relatif dans les différentes cours (les proportions rejet/acceptation sont du même ordre de grandeur). L'étude et la mise en oeuvre de ce type de visualisations ont été effectuées dans le cadre du stage de M1 d'Anastasiia Prysiazniuk.

Une partie des travaux menés sur ce sujet ont été consolidés dans le cadre du programme *Traitement Automatisé de la Jurisprudence* (TAJ) financé par la BPI, et réalisé par certains de nos collaborateurs. Les résultats obtenus et les réflexions menées ont aussi contribué à la définition d'un projet aujourd'hui implémenté dans l'ANR LAWBOT : Apprentissage Profond pour la modélisation prédictive de la jurisprudence.⁶⁶

65. objet : dommages-intérêt, fondement : Art. 700 du code de procédure civile.

66. Porteur Stéphane Mussard (Unîmes), collaboration Unîmes, LIG, LAMPS, IRIT, début 12/2020 - 48 Mois, 500k€. <https://anr.fr/Projet-ANR-20-CE38-0013>. J. Montmain et l'auteur, en qualité de membres d'IMT Mines Alès, ont participé à la rédaction de la première version du projet déposé auprès de l'ANR. Nous n'avons pas participé aux dépôts suivants suite au choix d'une réorientation stratégique des positionnements des laboratoires d'IMT Mines Alès.

Autres exemples de contributions en TALN

Nous proposons dans cette section un aperçu d'autres travaux réalisés en TALN.

Détection d'expressions d'incertitude

Les produits de systèmes d'extraction d'informations ou de connaissances servent souvent des traitements automatisés qui nécessitent une pleine maîtrise des données qu'ils exploitent, e.g. dans une finalité d'aide à la décision. A titre d'exemple, des études soulignent que plus de 10% des phrases issues de résumés scientifiques (articles de MEDLINE) véhiculent une forme d'incertitude [132]. Cela pose bien entendu souci si cette incertitude n'est pas prise en compte lors des phases automatisées d'exploitation des données extraites de ce type de textes. Nous avons dans ce contexte réalisé des travaux sur la détection d'expressions d'incertitudes linguistiques fréquemment rencontrées dans des textes [136, 135, 134]. Ces travaux ont été réalisés dans le cadre plus large de l'étude présentée en Section 2.2.3 page 69 sur la découverte de connaissances à partir de textes (inférence textuelle).

Collaborateurs :

- Doctorant encadré : Pierre-Antoine Jean (2014-2017).
- Collaborateurs d'IMT Mines Alès : Jacky Montmain, Sylvie Ranwez.
- Collaborateur externe : Patrice Bellot (LSIS, Marseille).

Les travaux mentionnés dans cette section ont pour l'essentiel été menés dans le cadre de la thèse de Pierre-Antoine Jean [132].

Deux grandes catégories d'incertitudes linguistiques sont généralement distinguées [249] : l'incertitude au niveau du discours et l'incertitude sémantique. L'incertitude au niveau du discours traduit d'un manque d'information, potentiellement intentionnel, qui rend la proposition mentionnée vague ou ambiguë, e.g. « *Ce n'était semble-t-il pas vraiment tous des manifestants conventionnels* », la construction de la phrase ne rend pas le propos clair. L'incertitude sémantique fait quant à elle référence aux propositions qui explicitent une incertitude sur la proposition énoncée : « *Il devait y avoir environ 100 manifestants* ». Nous ne détaillons pas plus les différents types d'incertitudes qui sont distingués dans ces deux grandes catégories [132].

Le problème de détection de l'incertitude linguistique est dans notre cas abordé comme un problème d'apprentissage supervisé et plus particulièrement de classification binaire. Les différents types d'incertitudes linguistiques ont été étudiés à l'aide de trois jeux de données existants (WikiWeasel [87], Bioscope [248] et SFU corpus [147]). Nous ne disposons pas cependant de bases importantes de données annotées. La méthode de détection de l'incertitude linguistique proposée se base alors sur une analyse statistique de différentes caractéristiques lexicales et syntaxiques identifiées. Elle repose sur une construction de représentations vectorielles des phrases à classer à partir de caractéristiques évaluées comme d'intérêt pour la tâche, e.g. taille de la phrase. Elle exploite par ailleurs le fait que certains jeux de données labélisés annotent, en plus du label traduisant la présence ou l'absence d'incertitude, les termes qui marquent l'expression d'incertitude. La méthode repose alors sur une estimation, à partir de la base annotée, de la probabilité qu'un terme traduise une expression

d’incertitude. Ces scores intègrent par ailleurs une appréciation du degré de confiance à apporter aux estimations au regard du nombre d’observations disponibles pour leurs calculs. Ils seront utilisés pour calculer des valeurs de caractéristiques exploitées comme composantes de la représentation vectorielle qui sera traitée par un classifieur standard de type SVM.

La méthode étudiée obtient des résultats compétitifs sur des jeux de données reconnus. Ils améliorent de 2 points de f-mesure les résultats obtenus par rapport à la référence du domaine sur WikiWeasel (en 2016); un comparatif avec FastText était aussi largement en faveur de l’approche proposée, cf. [132] page 67. Nos résultats surpassent aussi les approches testées à CoNLL 2010 (évaluation réalisée en 2016, postérieure au challenge en question; les approches proposées à CoNLL appartenaient cependant à la même catégorie d’approches, pas d’apprentissage profond). Une implémentation de notre approche réalisée par Pierre-Antoine Jean est disponible à l’adresse suivante : <https://github.com/pajeau/uncertaintyDetection>.

Analyse d’arguments

L’extraction et l’analyse automatique d’arguments à partir de textes (*argument mining*) laissent envisager l’automatisation d’un grand nombre de traitements intéressants pour la caractérisation de profils ou l’étude de relations publiques (communications) [165]. De telles techniques permettraient notamment d’étudier le positionnement de personnes qui échangent par écrit sur certaines questions (ou pour lesquelles nous disposons d’une transcription écrite d’un échange verbal). Elles permettraient aussi d’étudier les motivations des positionnements exprimés.

Nous avons proposé une étude sur l’analyse automatique d’arguments via des techniques d’apprentissage supervisé exploitant le paradigme de l’apprentissage multi-tâches [189, 188, 187]. Nous nous sommes intéressés à l’analyse automatisée de la micro-structure argumentative qui fait référence à la manière dont différents composants argumentatifs interagissent. Cela consiste à identifier plusieurs composants tels que les conclusions majeures, mineures et les prémisses, et à étudier leurs liens (relations de support et d’attaque) [78]. L’identification de cette micro-structure argumentative correspond à la construction d’un graphe dont la génération automatisée peut être décomposée en quatre étapes : (1) Délimitation des frontières des composants argumentatifs, (2) Détermination du type de chaque composant argumentatif, (3) Détermination de l’existence d’un arc entre chaque paire de composants argumentatifs, et (4) Étiquetage des arcs existants comme relation de support ou d’attaque. Nous nous sommes intéressés à ces quatre étapes.

Collaborateurs :

- Doctorant encadré : Jean-Christophe Mensonides (2015-2020).
- Collaborateur d’IMT Mines Alès : Jacky Montmain.
- Collaboratrice externe : Véronique Thireau (économiste, Université de Nîmes).

Les travaux mentionnés dans cette section ont pour l’essentiel été menés dans le cadre de la thèse de Jean-Christophe Mensonides.

Nos travaux ont porté sur l'étude d'une approche à base d'apprentissage profond dans un contexte d'apprentissage multi-tâches pour pallier les difficultés induites par le manque de données labélisées. Ainsi, au-delà de l'entraînement sur les tâches propres à l'identification de la micro-structure argumentative, nous avons entraîné notre modèle sur trois tâches annexes : l'étiquetage morpho-syntaxique (EMS), du chunking⁶⁷, et l'analyse de dépendances grammaticales (ADG). Cette approche s'inscrit dans la continuité des travaux proposés par Hashimoto et al. sur l'apprentissage multi-tâches à l'aide de modèles neuronaux [121].

Nous ne détaillerons pas l'architecture étudiée ; celle-ci repose globalement sur l'application de plusieurs couches bi-LSTM, propres aux tâches traitées, sur des plongements de mots pré-entraînés de manière auto-supervisée. Les représentations construites sont ensuite fournies à une tête de classification standard propre elle aussi à la tâche traitée. L'aspect multi-tâches est implémenté en considérant que lors du traitement de certaines tâches, les entrées intègrent les représentations construites pour la réalisation d'autres tâches. La figure 25 illustre l'interdépendance des tâches.

Les résultats obtenus rivalisaient avec ceux obtenus par les approches de l'état de l'art. Si l'on se focalise sur les tâches propres à l'analyse d'arguments, ils étaient légèrement meilleurs que l'état de l'art sur certaines d'entre elles, et légèrement moins bons sur d'autres - à noter que certaines approches comparées se basaient sur des caractéristiques choisies par des experts et qu'aucune ne traitait à notre connaissance toutes les tâches à l'aide du même modèle. Les résultats obtenus soulignent globalement l'apport modeste amené par le fait de traiter les tâches d'analyse d'arguments via une approche multi-tâches (DeICAn DetCA, DetArcs. . . dans la figure 25). Des tests réalisés sur la même architecture dans un contexte d'entraînement standard, i.e. chaque tâche d'analyse d'arguments traitée indépendamment, montrent des résultats de performances semblables (+0.2 de macros f1-scores par rapport à une approche mono-tâche du même type). L'apport des tâches annexes (EMS, chunking, ADG) s'est cependant avéré important : nous notons jusqu'à +37% de performance sur une tâche d'analyse d'arguments. L'utilisation d'un nombre variable d'époques pour chaque tâche est aussi nettement recommandée. Les résultats obtenus contribuent à faire la démonstration de la possibilité de développer un système complet d'analyse d'arguments reposant sur un modèle unique.

Challenges Nous mentionnons quelques challenges en lien avec le TALN auxquels nous avons participé. Les collègues qui ont collaboré avec nous dans le cadre de ces challenges sont mentionnés dans les publications associées.

- **DEFT 2019**, Défi Fouille de Textes (atelier TALN-RECITAL) : indexation et appariement textuel [190]⁶⁸ - porteur Jean-Christophe Mensonides. Le corpus était relatif aux domaines biomédical et clinique, en langue française. La première tâche traitée visait à annoter des descriptions de cas cliniques avec des mots-clefs d'un vocabulaire contrôlé. Un ensemble limité d'exemples d'annotations était partagé. Les méthodes que nous avons étudiées reposent sur des techniques à base de TF-IDF - celles-ci sont éprouvées dans le domaine de la Recherche d'Information et du TALN, et s'avèrent généralement robustes sur de petits jeux de données. Nous avons dans notre cas tenu compte d'adaptations à la nature spécifique du corpus et à la spécificité de la tâche. Des traitements de normalisation des textes à base de similarités sémantiques ont été étudiés afin de limiter l'incidence de l'utilisation de synonymes dans les

67. Tâche d'analyse syntaxique superficielle.

68. https://deft.lisn.upsaclay.fr/actes/actes_DEFT_CH_PFIA2019.pdf

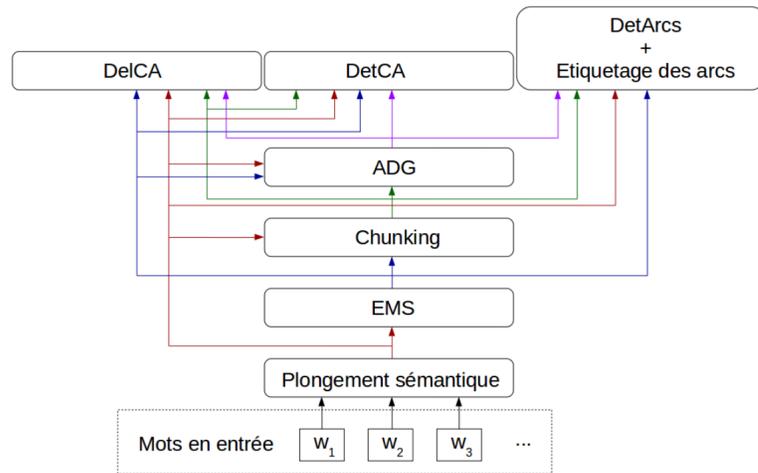


FIGURE 25 Architecture multi-tâches proposée. EMS, chunking et ADG correspondent aux tâches prétextes ; DelCA (Délimitation des composants argumentatifs), DetCA (Détermination des composants argumentatifs) et DetArcs (Détermination des arcs) correspondent aux tâches propres à l'identification de la micro-structure argumentative. Source : [187].

annotations et les textes. Une adaptation de l'approche TF-IDF a en particulier été formulée pour tenir compte de ce pré-traitement. La seconde tâche traitée visait à rapprocher des cas cliniques et des discussions ; cette dernière a été abordée comme une problématique d'estimation de similarité de documents. Nous avons dans ce cas proposé une estimation de similarité via des constructions de représentations (adaptées de LSA et TF-IDF) pour *in fine* calculer une matrice de distances normalisées entre les cas cliniques et les discussions. Une étude comparative de l'impact du choix de la distance et de la manière de représenter les textes sur la performance de l'approche a par ailleurs été menée. Les approches proposées pour les deux tâches ont obtenu de très bons scores à l'évaluation.

- **HASOC 2019** : *Automatic Hate Speech Detection* [191] - porteur Jean-Christophe Mensonides. Les tâches traitées consistaient à détecter l'expression de propos haineux dans des publications partagées sur des réseaux sociaux (classifications binaire et multi-classes en trois langues : anglais, allemand et hindi). Une analyse des hashtags et des emojis sur les données annotées a mis en évidence le fait que les emojis sont généralement de très bons indicateurs d'appartenance à une classe ; étonnamment par contre, la plupart des hashtags que l'on pourrait considérer offensants ne se sont pas avérés discriminants (ils sont souvent réutilisés dans des réponses). Les meilleurs résultats que nous avons obtenus reposaient sur une approche fondée sur un modèle FastText [139] appliqué à des données prétraitées pour gérer les spécificités de ce type de textes (e.g. emoji, URL). Le système n'utilisait pas de plongements de mots pré-entraînés. Il s'est avéré inadapté voire très mauvais pour l'anglais et l'allemand. Il a cependant atteint la seconde place dans le cas de l'hindi (différence de 0.003 de f-mesure par rapport au premier). Ce résultat souligne l'utilité des approches traditionnelles de TALN lorsque le nombre de ressources disponibles est limité, e.g. pour envisager des tâches auto-supervisées.

- **COLIEE 2017** - *Competition on Legal Information Extraction/Entailment*⁶⁹ : Recherche d'articles de lois pertinents pour une étude de cas (texte anglais) - porteur Pierre-Antoine Jean. Ce challenge en lien avec le traitement de textes juridiques est discuté page 95, cf. [132] Annexe A.
- **BioASQ 2015** : indexation de publications du domaine biomédical à l'aide de concepts du MeSH [178] - porteur Nicolas Fiorini. Cette tâche a été traitée comme une tâche d'annotation conceptuelle et fait référence aux travaux sur les mesures sémantiques mentionnés en première partie de ce manuscrit [95].

69. <http://webdocs.cs.ualberta.ca/~lijmiyoung2/COLIEE2017>

2.3.2 Contributions à l'étude de problèmes d'optimisation combinatoire

Nous présentons dans cette section nos travaux sur l'étude de techniques d'apprentissage automatique pour le traitement de problèmes d'optimisation combinatoire de type problèmes de tournées de véhicules.

Collaborateurs :

- Doctorant encadré : Ali Yaddaden (2019-2023).
- Collaborateur d'IMT Mines Alès : Michel Vasquez (Recherche Opérationnelle).

Les travaux présentés ont pour l'essentiel été menés dans le cadre de la thèse d'Ali Yaddaden (soutenue début novembre 2023) [280, 278, 277, 5].

Cadre général des travaux

Le domaine de la Recherche Opérationnelle étudie depuis des dizaines d'années une multitude de problèmes d'optimisation combinatoire de forte complexité et liés à des enjeux sociétaux, économiques, et environnementaux importants, e.g. problèmes logistiques tels que la gestion de stocks ou la gestion de flottes de véhicules. Un problème standard, que nous aborderons ci-après, est le problème du voyageur de commerce (*Travelling Salesman Problem*, TSP [65]) : comment desservir un ensemble de villes en minimisant la distance totale parcourue. Les problèmes d'optimisation combinatoire se caractérisent par le fait qu'il est impossible d'évaluer par simple énumération toutes les alternatives pour distinguer les solutions candidates et critiquer leurs pertinences en vue de trouver une solution optimale⁷⁰ ; l'objectif est alors de proposer une méthode optimisée afin de permettre l'identification rapide d'une solution optimale ou proche de l'optimum.

De nombreuses approches spécifiques ont ainsi été proposées pour aborder des problèmes variés d'optimisation combinatoire ; celles-ci reposent sur la formulation de stratégies explicites de résolution : un algorithme est défini de manière à pouvoir appliquer un ensemble d'instructions machine qui permettront de générer une solution candidate - que nous espérons optimale ou de qualité - pour une instance donnée d'un problème spécifique. Il est important de souligner que, même si des classes générales d'approches de résolution existent (heuristiques, métaheuristiques. . .), les algorithmes les plus performants sont très généralement, si ce n'est tout le temps, spécifiques du problème traité, voire de caractéristiques des instances du problème en question. De plus, ces algorithmes reposent sur des techniques heuristiques optimisées et le plus souvent raffinées, qui ont été définies de manière chronophage par des experts de la Recherche Opérationnelle. Cette approche générale, qui prévaut en Recherche Opérationnelle aujourd'hui, requiert alors une formulation manuelle et complexe de techniques de résolution pour tout problème ou variante d'un problème traité. Cela induit une difficulté pour l'utilisation pratique de ce type d'approches. De plus, même si ce paradigme de recherche d'approches de résolution a permis des progrès importants au fil des décennies pour de multiples problèmes, il peine aujourd'hui à améliorer les résultats des approches existantes – en particulier pour les problèmes les plus largement étudiés. Nous soulignons bien entendu que malgré ces limites, des gens brillants ont fait et continuent à faire de très belles contributions au domaine de la Recherche Opérationnelle à l'aide de ce paradigme.

70. nous avons $n!$ solutions candidates pour un TSP de n villes.

Une approche alternative radicalement différente existe cependant. Elle vise, tant que possible, à considérer la recherche des heuristiques performantes pour la résolution d'un problème comme un problème d'apprentissage machine. Il s'agit alors, à l'aide de données relatives au problème traité, e.g. collection d'instances du problème, de chercher à apprendre l'algorithme de résolution, i.e. une alternative à celui manuellement explicité dans le cadre du paradigme traditionnel de résolution de la Recherche Opérationnelle. Nous chercherons *simplement* dans notre cas à proposer l'approche qui permettra d'apprendre l'algorithme susceptible de résoudre un problème d'intérêt. Cette approche trouvée, il s'agira finalement de l'appliquer afin d'apprendre, à partir d'instances du problème en question, un algorithme qui pourra alors être utilisé pour le résoudre [280]. Ce paradigme de résolution par apprentissage machine peut bien entendu être couplé au paradigme traditionnel : il s'agira dans ce cas de ne déléguer à la phase d'apprentissage que la résolution d'une partie identifiée d'une décomposition du problème [20].

Un parallèle peut par ailleurs être fait entre cette progression vers l'apprentissage d'heuristiques que nous évoquons, et la progression vécue par l'apprentissage automatique dans l'apprentissage de représentations qui sous-tend l'essor de l'apprentissage profond. Il était en effet classique il y a quelques années encore - nous l'avons vu précédemment dans nos travaux en TALN notamment - de définir manuellement des représentations des instances des problèmes à traiter par la sélection de caractéristiques d'intérêt, e.g. caractéristiques pour représenter des textes via des plongements *ad hoc*. La recherche *ad hoc*, technique, manuelle et chronophage de caractéristiques d'intérêt, voire d'algorithmes de résolution, appartiendra-t-elle au passé de la Recherche Opérationnelle dans quelques années ? Difficile de répondre à cette question aujourd'hui.

L'étude du paradigme visant à *apprendre des heuristiques* est aujourd'hui encouragée par l'obtention, via l'apprentissage automatique, de ruptures dans le traitement de plusieurs problèmes à forte combinatoire. Nous pensons particulièrement aux problèmes de type jeux de stratégie pour lesquels des progrès importants ont été obtenus ces dix dernières années, e.g. performances suprahumaines pour le jeu de Go dès 2016 [237], puis de manière connexe aux développements liés à AlphaStar [268], ChatGPT [205]. . . Ces progrès invitent à étudier les approches à base de réseaux de neurones entraînés par renforcement pour éviter toute dépendance à des données labélisées, e.g. des instances pour lesquelles nous connaissons l'optimal. C'est le positionnement que nous avons adopté. Il s'inscrit dans la continuité de travaux récents et exploratoires dans ce domaine, en particulier sur la définition de solveurs neuronaux dans le cadre de ce que l'on appelle l'Optimisation Combinatoire Neuronale – *Neural Combinatorial Optimization* (NCO) [16].

Alors que des progrès importants ont été réalisés ces dernières années, les approches à base d'apprentissage machine ne permettent pas seules, à ce jour, de rivaliser avec les approches standard de la Recherche Opérationnelle. A titre d'exemple, dans le domaine des problèmes de tournées de véhicules (*Vehicle Routing Problem*, VRP [217]), les meilleurs systèmes NCO, bien qu'il soit très compétitifs sur de petites instances, peinent à obtenir des performances raisonnables par rapport aux meilleures métaheuristiques pour des instances de grande taille – les comparatifs des approches de l'état de l'art que nous avons proposés sur des jeux d'instances établis du domaine le soulignent, e.g. sur CVRPLib [5] [263]. ⁷¹ Cependant, même si les systèmes de type NCO n'obtiennent pas à ce jour des performances de rupture en Recherche Opérationnelle, nous sommes convaincus que

71. Les modélisations qui prévalent dans le domaine sont de plus, par construction, aujourd'hui souvent fortement dépendantes de la taille des instances (e.g. de complexité quadratique par rapport à cette taille). Dans le cas de succès, le problème du passage à l'échelle induira par ailleurs son lot de challenges.

cette ligne de recherche mérite d’être sérieusement étudiée et que les travaux exploratoires sur ce sujet doivent être encouragés. L’étude de ce nouveau paradigme peut par ailleurs contribuer à ce que le domaine de la Recherche Opérationnelle bénéficie aussi plus largement de la dynamique qui caractérise aujourd’hui l’apprentissage machine.

Nos travaux se sont pour l’heure dédiés à l’étude de l’apprentissage profond et par renforcement pour la résolution de deux problèmes de tournées de véhicules : le problème de tournées de véhicules avec contraintes de capacités (*Capacitated VRP*, CVRP), et le problème de covoiturage (*Ride-Hailing Problem*, RHP [150]). Le premier, largement étudié dans l’état de l’art, vise à trouver des tournées pour satisfaire des clients qui expriment des demandes quantifiées, et cela en tenant compte des limites de capacité de transport des véhicules de la flotte gérée. Le second est une variante relative à la gestion de flotte de taxis de personnes nécessitant la prise en compte d’incertitudes, e.g. variabilité des durées des trajets, localisations des clients inconnues à l’avance.

Notre ligne principale de recherche concerne l’étude de solveurs basés sur des réseaux de neurones profonds entraînés à l’aide de l’apprentissage par renforcement sur de grands jeux de données d’instances non labélisées, i.e. d’instances non résolues, ou à l’aide de simulateurs dans le cadre du RHP. Elle se focalise pour l’essentiel sur l’étude du cadre NCO pour la formulation d’approches simplement fondées sur des réseaux de neurones (approche de type *end-to-end*) ; quelques travaux sont aussi réalisés sur l’hybridation entre des réseaux de neurones et des approches heuristiques ou exactes. Nous avons essentiellement étudié des modélisations reposant sur des approches par renforcement de type *policy gradient* et *deep Q-learning* (espaces d’action discrets) [247, 173]. Celles-ci reposent sur des architectures de réseaux de neurones de type Encodeur-Décodeur exploitant des réseaux récurrents (RNN), adaptés au plongement de graphes (GNN), et qui reposent généralement sur le mécanisme d’attention, e.g. Transformers [265]. Nous détaillerons quelques exemples par la suite. Pour rendre le propos plus concret, dans le cas d’un TSP le lecteur peut à ce stade considérer une approche de résolution itérative qui ajoute à chaque itération un client à un ordre de traitement qui sera finalement notre tournée. Le réseau de neurones peut servir dans ce cas à modéliser une probabilité conditionnelle utile pour cette construction itérative de la solution candidate, e.g. la probabilité qu’un client spécifique, sachant les clients déjà visités et des informations sur le problème (e.g. localisations des clients), soit le prochain à visiter pour tendre vers la solution optimale.⁷²

Nos travaux portent pour l’heure principalement sur l’obtention de modèles performants pour un domaine spécifique, e.g. TSP, CVRP, RHP. Nous ne travaillons pas en tant que tel sur la définition de modèles de manière détachée d’une classe identifiée de problèmes d’optimisation. Nous tentons cependant d’étudier comment améliorer les performances des modèles et réduire les coûts de calcul des phases d’entraînement. Pour cela, nous cherchons à exploiter des similitudes entre des problèmes d’une même classe (e.g. transfert du TSP au CVRP, étude de tâches prétextes⁷³). Nous privilégions aussi l’étude empirique des capacités de généralisation de solveurs neuronaux agnostiques à la taille des instances traitées. Nous privilégions dans ce contexte des modélisations à base d’attention qui

72. Cette approche autoregressive adopte une stratégie similaire à celle des heuristiques dites constructives.

73. Pré-entraînement à la résolution du TSP sur des instances de CVRP dans le cadre d’entraînement au CVRP.

permettent d'éviter que le nombre de paramètres des modèles étudiés soit fonction des tailles des instances traitées – l'objectif visé étant d'arriver à entraîner un modèle performant sur des instances de petites tailles, et de conserver ses performances lors du traitement d'instances de plus grandes tailles.⁷⁴

Nous illustrerons nos travaux au travers de (i) l'étude d'une stratégie d'apprentissage par transfert du TSP au VRP, et de (ii) la formulation d'une approche hybride couplant un réseau de neurones à une méthode exacte. Le lecteur non familier des approches de type NCO et de leurs applications au VRP peut faire référence à la synthèse sur ce sujet que nous avons présentée à CNIA 2020 [280] et à la thèse d'Ali Yaddaden qui propose un état de l'art détaillé [5].

Apprentissage par Transfert du TSP au CVRP

Nous cherchons à étudier le potentiel de l'apprentissage par transfert dans le contexte de la résolution de problèmes d'optimisation combinatoire basée sur le cadre NCO. Nous étudions pour cela une approche par transfert du problème du TSP au problème de CVRP - le problème du CVRP étant une généralisation du TSP. Cela consiste dans notre cas à évaluer empiriquement s'il est bénéfique de considérer une phase d'entraînement dédiée à la résolution du TSP (tâche source) avant d'initier la phase d'entraînement propre à la résolution du CVRP (tâche cible). Plusieurs améliorations potentielles liées aux modèles et à l'entraînement sont notamment visées : (i) l'amélioration de la performance des modèles sur la tâche cible, (ii) l'obtention de propriétés intéressantes en termes de capacité de généralisation à des instances différentes (e.g., nombre ou répartitions des villes différentes dans le cadre du VRP), (iii) la réduction des coûts d'entraînement. Les résultats obtenus pour l'heure sont exposés dans [276, 277, 5].

Nous considérons le cas de TSP et CVRP 2D euclidiens, avec des graphes sous-jacents complets (toutes les villes/clients sont connectés), et des capacités homogènes des véhicules pour le CVRP (cela équivaut à considérer une flotte composée d'un unique véhicule). Nous étudions par ailleurs une approche qui vise à utiliser un même modèle et une même stratégie de résolution itérative pour traiter les deux tâches. A chaque étape, l'approche globale de résolution sélectionnera l'action à réaliser (i.e. un noeud du graphe qui correspond à la ville à visiter pour le TSP, et à un client ou au dépôt pour le CVRP). Nous estimons pour cela la probabilité conditionnelle mentionnée dans la section précédente via un réseau de neurones. Plus précisément, la probabilité qu'une politique stochastique (notre approche) génère la solution candidate Y pour une instance X est définie par : $P_\theta(Y|X) = \prod_{t=1}^T p_\theta(y_t|y_1, \dots, y_{t-1}, X)$, avec y_i le noeud sélectionné à l'itération i (T équivaut au nombre de villes dans le TSP). La probabilité p_θ sera modélisée par notre réseau de neurones de paramètres θ . Conceptuellement, nous aborderons l'entraînement du réseau comme l'identification des paramètres θ optimaux au regard de la minimisation de l'espérance $\mathbb{E}_{X \sim \mathcal{X}, Y \sim P_\theta(\cdot|X)}[L(Y, X)]$ avec $L(Y, X)$ le coût de la solution Y pour l'instance X tirée d'une distribution \mathcal{X} . Une stratégie de génération (max, aléatoire, *beam search*) utilisera notre estimation probabiliste (p_θ) pour générer des solutions candidates. Nous modélisons la probabilité conditionnelle p_θ par une adaptation de l'*Attention Model* [148]. Cette adaptation a été proposée afin de permettre le traitement des tâches source et cible sans ajout de paramètre. La figure 26 illustre l'architecture générale de type encodeur-decodeur paramétrique et différentiable ; seules les grandes lignes sont mentionnées ci-après, les détails sont proposés dans [277].

⁷⁴. Ce que font naturellement les heuristiques proposées manuellement par des experts de domaine, mais que les heuristiques apprises peinent à réaliser.

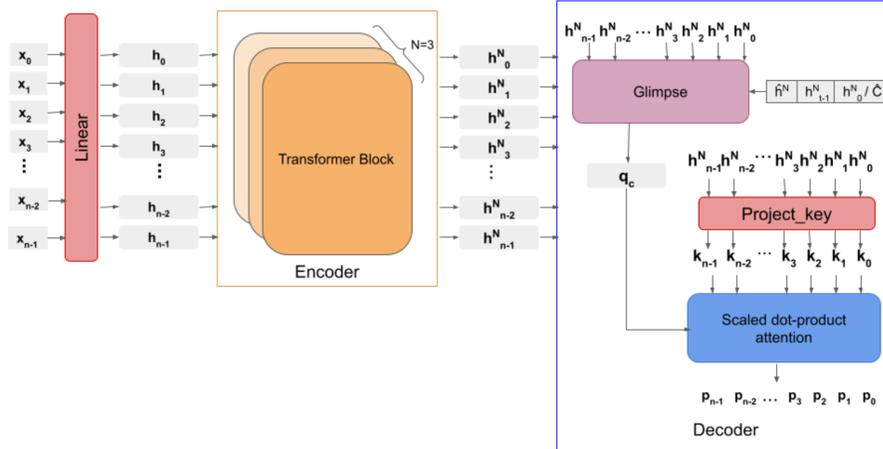


FIGURE 26 Illustration de l'architecture, adaptée de l'Attention Model [148], étudiée dans nos travaux pour estimer p_θ [277].

L'encodeur vise à construire les représentations des nœuds (villes, clients ou dépôt) à partir de leurs simples coordonnées cartésiennes. Il repose essentiellement sur plusieurs blocs Transformer standard munis d'une auto-attention multi-têtes [265]. L'encodeur que nous avons testé dans cette étude, composé de 3 blocs, n'est pas très profond par rapport à ce que l'on observe dans certaines littératures (e.g. TALN). Le décodeur est ensuite utilisé pour distinguer le prochain nœud à visiter. Nous confrontons pour cela une représentation du contexte de prise de décision à une itération donnée, avec les représentations des nœuds. La représentation du contexte de décision tient compte d'informations associées à l'instance (i.e. graphe sous-jacent), au nœud de départ et au dernier nœud sélectionné – ces informations correspondent essentiellement à des concaténations des représentations des nœuds construites au préalable par l'encodeur. Un mécanisme de *glimpse* permet la création d'une requête (q_c) en tenant compte de cette représentation du contexte. Il effectue pour cela une combinaison des représentations des nœuds non traités du graphe, pondérée par les scores d'attention que lesdits nœuds obtiennent avec la représentation de ce contexte de décision. On obtient à ce stade une représentation de notre requête, q_c , que l'on souhaitera confronter aux représentations des nœuds éligibles restants. Pour cela, un mécanisme d'attention est de nouveau utilisé afin de mettre en correspondance la représentation de la requête et celles des nœuds afin de générer la distribution de probabilité catégorielle p_θ que l'on cherche à estimer. Un masquage contextuel est utilisé pour exclure les nœuds traités pendant les différentes étapes de traitement.

La fonction de perte exploitée durant l'entraînement repose sur l'évaluation de la longueur des tours générés par une stratégie stochastique de génération des solutions qui se base sur l'estimation de p_θ . L'entraînement est effectué par descente de gradient avec une estimation du gradient effectuée par une approche acteur-critique fondée sur REINFORCE [271]. En fonction des contextes d'évaluation, entre 800k et 3.2 millions d'instances de 20 à 50 villes/clients sont utilisées pour les phases d'entraînement – ce sont des instances de très petites tailles par rapport à celles traitées classiquement par l'état de l'art du domaine hors apprentissage machine. Nous avons comparé l'incidence du transfert dans différents contextes. Les résultats obtenus varient en fonction des conditions testées ; ils montrent que le transfert du TSP au CVRP permet généralement (i) de réduire les temps d'entraînement des modèles – les plongements appris sur le TSP sont informatifs pour le

CVRP⁷⁵ -, et (ii) d'obtenir dans certains cas des modèles plus performants. De manière intéressante, nos résultats montrent aussi que, dans le pire des cas étudiés, le transfert ne semble pas induire de dégradation des performances. Aucune propriété intéressante en termes de généralisation ne semble être apportée de manière consistante par le transfert du TSP au CVRP. L'impact très certainement important de l'hyperparamétrage reste à être discuté. Plus généralement, des expérimentations supplémentaires méritent bien entendu d'être réalisées de manière complémentaire sur d'autres contextes d'évaluation et à l'aide d'autres modélisations du paradigme NCO (e.g. architectures différentes pour estimer p_θ).

Hybridation entre Apprentissage Automatique et méthodes exactes

Nous l'avons souligné en avant-propos, les approches NCO peinent aujourd'hui à rivaliser avec les meilleures approches de la Recherche Opérationnelle traditionnelle. L'état de l'art et les résultats obtenus dans nos multiples expérimentations attestent de cela pour les problèmes de VRP. En complément de l'étude des approches NCO qui correspondent aux solveurs neuronaux *purs* (de bout en bout), nous étudions aussi la définition d'approches hybrides. Une des idées que nous testons est intuitive et appartient à l'état de l'art : (i) décomposer un problème de manière *ad hoc* en plusieurs sous-problèmes, (ii) distinguer les sous-problèmes que l'on sait résoudre de manière satisfaisante (sans apprentissage donc), pour finalement (iii) ne traiter que les sous-problèmes restants à l'aide d'approches NCO. Nous présentons un exemple de nos travaux sur ce type d'approches dans le cadre de la résolution du problème du CVRP.⁷⁶ Les détails seront retrouvés dans [278, 279, 5].⁷⁷

L'approche que nous proposons est dénommée NOFSS pour *Neural Order-First Split-Second*. Elle peut s'appliquer à une variété de VRP mais nous nous concentrerons ici sur le cas du CVRP. La modélisation repose sur une décomposition du problème en deux phases : (i) ordonnancement, et (ii) découpage – cette approche *order-first split-second* est d'ores et déjà largement étudiée dans la littérature traditionnelle du VRP [221]. La formulation que nous proposons dans le cadre d'une hybridation mobilisant de l'apprentissage machine repose sur un réseau de neurones et un algorithme exact de programmation dynamique. Dans la première étape, le réseau de neurones est utilisé pour générer une permutation des clients à desservir (*order-first*). Il s'agit ensuite de distinguer les retours au dépôt pour faire de cette permutation une solution candidate, en respectant les contraintes liées à la capacité du véhicule vis-à-vis des demandes des clients. Ce découpage sera réalisé de manière optimale dans la seconde étape (*split-second*). La permutation sera en effet traitée par un algorithme exact (*Split* [220]) qui définira les retours dépôt et produira ainsi une solution candidate.

L'approche NOFSS appliquée au CVRP est motivée par le postulat discutable qu'il sera plus facile pour le réseau d'apprendre un espace latent lui permettant de représenter le problème sans prise en compte des retours dépôt. La modélisation adoptée pour la génération des permutations est proche conceptuellement de celle introduite dans la section précédente, excepté qu'une instance de CVRP est dans ce cas considérée comme un TSP, et que l'on démarre du dépôt (cf. encodeur-décodeur). Différentes évaluations ont été réalisées à l'aide de plusieurs architectures, en utilisant notamment des GNN pour la partie encodage et des RNN pour la partie décodage. La figure 27 illustre l'architecture complète.

⁷⁵. Pas de transfert *zero-shot* cependant, i.e. sans surprise, un modèle entraîné sur le TSP n'est pas directement opérationnel pour traiter le CVRP.

⁷⁶. Comme précédemment : CVRP 2D euclidien avec graphe complet et capacités homogènes.

⁷⁷. Un article joint au manuscrit traite de ce sujet, cf. Annexe A page 203 [278].

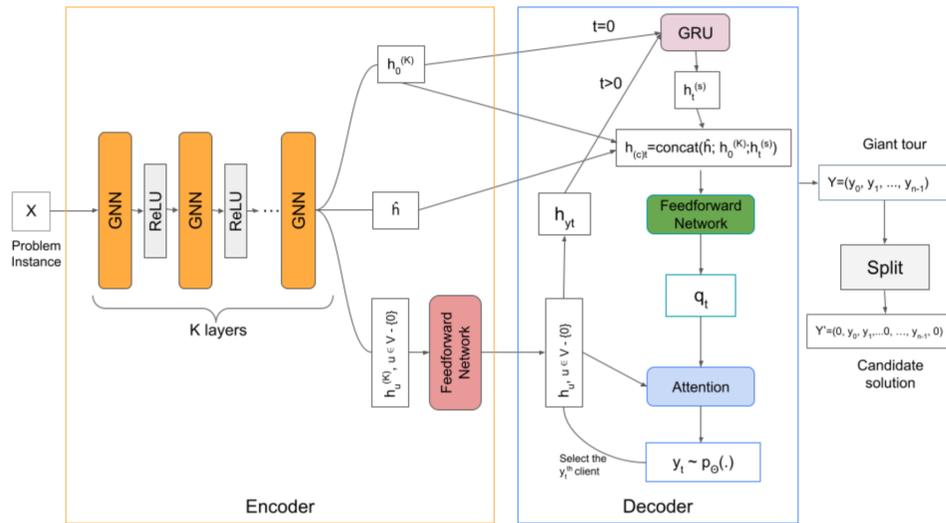


FIGURE 27 Illustration d’une implémentation de l’approche *Neural Order-First Split-Second* (NOFSS) que nous avons proposée dans le cas du CVRP [277]. L’encoder repose sur un GNN pour construire le plongement du graphe qui représente l’instance du problème (\hat{h}), et ceux de ses noeuds ($h_i^{(K)}, i \in V$). Ces plongements seront par la suite utilisés par le décodeur qui générera itérativement un tour géant (Y). On utilise pour cela (i) une cellule récurrente de type GRU pour représenter à chaque itération la requête (q_t) en tenant compte des choix précédents, et (ii) un mécanisme d’attention pour l’appariement entre la requête et les noeuds restants. Le tour géant sera ensuite traité par l’algorithme *Split* pour définir de manière optimale les retours au dépôt en tenant compte des contraintes de capacité.

L'évaluation de l'approche a été réalisée dans des contextes standard d'évaluation pour le domaine, sur des instances de 20 à 100 clients. Les résultats obtenus montrent que l'approche NOFSS permet d'apprendre un modèle capable de traiter le CVRP en apprenant une représentation indirecte de la solution (i.e. partielle, et dans l'espace des permutations des clients). Ce sont à notre connaissance les premiers résultats de ce type dans la littérature. NOFSS obtient de meilleurs résultats que des heuristiques manuelles simples de type *Sweep* ou *Nearest Neighbor*. L'approche NOFSS rivalise par ailleurs avec les approches NCO pures tout en permettant des gains notables en termes de temps d'exécution à l'entraînement et en inférence. De moins bonnes performances sont cependant enregistrées par rapport à l'approche RFCS (*Route-First Cluster-Second algorithm* [15]) qui est aussi une approche *order-first split-second* qui elle n'exploite pas d'apprentissage et qui vise la résolution explicite du TSP. L'approche NOFSS montre aussi des propriétés intéressantes en termes de généralisation, en particulier en ce qui concerne la généralisation à des tailles d'instances plus grandes.

Les codes de calcul développés et utilisés dans le cadre de nos expérimentations sur l'optimisation combinatoire neuronale sont disponibles sur GitHub (contributeur principal Ali Yaddaden, cf. <https://github.com/AYaddaden>).

2.3.3 Contributions disciplinaires en Apprentissage Machine

Nous présentons dans cette section une sélection de deux récentes contributions disciplinaires en Apprentissage Machine. Celles-ci permettront au lecteur d'apprécier quelques-unes de nos formulations techniques susceptibles de trouver un intérêt pour la communauté au-delà du contexte applicatif spécifique pour lequel elles ont été formulées. Le premier travail que nous présenterons porte sur l'approximation du mécanisme d'auto-attention - un mécanisme important dans un grand nombre d'architectures de réseaux de neurones aujourd'hui très populaires, e.g. Transformers [265, 176]. Le second travail que nous mentionnerons traite du sujet de l'interprétabilité des modèles prédictifs de manière conditionnée à une prédiction pour une entrée donnée.

Collaborateurs :

- Doctorant encadré : Charles Condevaux (2018-2022).
- Collaborateur externe (étude sur l'explicabilité) : Stéphane Mussard (Université de Nîmes).

Les travaux mentionnés dans cette section ont pour l'essentiel été menés dans le cadre de la thèse de Charles Condevaux [57].

Approximation du mécanisme d'auto-attention

Nous présentons dans cette section le travail que nous avons réalisé sur l'adaptation du mécanisme d'auto-attention afin de permettre le traitement d'entrées de grandes tailles, e.g. longues séquences.

La très grande majorité des approches d'apprentissage profond qui se distinguent aujourd'hui dans la littérature dédiée au traitement de données représentées sous la forme de séquences de tokens (e.g. textes) reposent sur des architectures de réseaux de neurones de type Transformer [265, 176].⁷⁸ Cela est notamment le cas dans le domaine du traitement du langage naturel et en particulier pour la plupart des gros modèles de langue (LLM - *Large Language Model*) et leurs dérivés (GPT-* et ChatGPT [42, 203], Sparrow [104], Palm 2 [8], Chinchilla [125], Bart [170], Llama [259], Claude[11], Falcon [6]. . .).

La pierre angulaire de ces architectures Transformers est résolument le mécanisme d'attention [10]. Ce mécanisme est utilisé afin de construire graduellement des représentations contextuelles des éléments de la séquence ; ces dernières seront utiles pour la résolution de la tâche traitée, e.g. construire un plongement de la séquence afin d'effectuer sa classification. Plus généralement, ce mécanisme d'attention est aujourd'hui fréquemment utilisé pour construire des plongements pour des entrées qui peuvent être représentées comme des ensembles d'éléments, e.g. graphes représentés comme des ensembles de nœuds interconnectés dans les approches de plongement de graphes [168], instances d'un problème d'optimisation combinatoire spécifique dans les approches de type *Neural Combinatorial Optimisation* [280], ou comme nous l'avons déjà précisé, un ensemble ordonné de tokens correspondant à une séquence représentant un texte [265], ou du code [171].

Le mécanisme d'attention est une opération simple qui repose sur une phase d'appariement entre les éléments de deux ensembles A et B .⁷⁹ Cette phase d'appariement vise le calcul de scores d'appariement, appelés scores d'attention, entre chaque élément de A et de B (chaque paire de $A \times B$ se voit attribuer une valeur réelle pour score). Le mode de calcul du score d'attention entre deux éléments peut varier mais il repose très généralement sur une forme de produit scalaire normalisé entre des représentations vectorielles des éléments. Ces scores seront exploités lors de la construction des représentations des éléments de A . De manière plus précise, le plongement d'un élément $a \in A$ produit par le mécanisme d'attention - $\rho(a) \in \mathbb{R}^d$ - sera très généralement modélisé par une somme pondérée de plongements existants associés aux représentations des éléments de l'ensemble B ; le score d'attention sert alors de facteur de pondération lors de l'agrégation, i.e. $\rho(a) = \sum_{b \in B} score(a, b) \rho_V(b)$, avec $\rho_V(b) \in \mathbb{R}^d$, un plongement existant pour tout $b \in B$.

De nombreuses variantes du mécanisme d'attention existent dans la littérature [48]. Une des expressions retrouvées dans de nombreux encodeurs, décodeurs, et dans les Transformers est l'auto-attention - plus spécifiquement l'auto-attention multi-têtes (MHA, Multi-Head Attention [265]). Dans ce cas les deux ensembles impliqués dans la phase d'appariement sont les mêmes ($A = B$). On cherche ainsi à représenter un élément d'un ensemble en tenant compte des autres éléments présents dans l'ensemble. Ainsi, dans une formulation simplifiée de l'auto-attention, les

78. La publication qui introduit l'architecture Transformer en 2017 a été citée prêt de 89k fois en environ 7 ans.

79. Comme nous le verrons ci-après, des informations supplémentaires sur les éléments des ensembles peuvent aussi être appréciées, e.g. ordre sur les éléments lorsque les ensembles font référence à des séquences. Cela permet de traiter des entrées pour lesquelles un ordre sur les éléments mérite d'être considéré, et cela même si la formulation initiale de l'attention est invariante à la permutation des éléments traités.

plongements de l'ensemble des éléments de A seront calculés comme suit :

$$\text{Attention}(A) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

avec $\text{Attention}(A) \in \mathbb{R}^{n \times d}$ le produit du mécanisme d'attention, qui correspond à une matrice dont chaque ligne correspond à un plongement dans \mathbb{R}^d d'un élément spécifique de A . Les matrices $Q, K, V \in \mathbb{R}^{n \times d}$ sont des représentations intermédiaires des éléments de l'ensemble A qui ont été construites à partir de simples transformations linéaires de représentations connues des éléments de A , e.g. de simples *one-hot encoding* ou des représentations plus complexes. La fonction *softmax*, appliquée par ligne, joue le rôle de *max* différentiable. La notion de têtes multiples dans la MHA traduit le fait que l'auto-attention est parfois réalisée plusieurs fois - autant de fois que de têtes définies. Le plongement final généré par le mécanisme est la concaténation du résultat produit par chaque tête.

Même si la formulation originale de l'auto-attention est invariante à l'ordre des éléments, i.e. elle est définie pour des ensembles, une connaissance supplémentaire est souvent considérée pour encoder la position des éléments lorsque l'ensemble fait référence à une séquence. Même si des variantes existent, cela est généralement fait en considérant des plongements de position en plus des plongements propres aux éléments lors du calcul de Q, K et V . Cette approche d'auto-attention adaptée au traitement de séquences est intéressante pour implémenter l'hypothèse distributionnelle en traitement automatique du langage naturel, et ainsi étudier la sémantique d'un terme comme fonction de son voisinage.

Nous nous sommes dans nos travaux focalisés sur l'optimisation du mécanisme d'auto-attention lorsque les entrées correspondent à des séquences. Dans ce contexte, le mécanisme d'attention comporte plusieurs avantages par rapport aux alternatives existantes. En effet, à titre d'exemple, les réseaux de neurones récurrents très populaires pour traiter des séquences avant la proposition des architectures Transformers sont sujets à des problèmes de calcul du gradient lorsque les séquences sont longues (suppression ou explosion des gradients). Ils induisent aussi par ailleurs des temps de calcul prohibitifs.⁸⁰ Ces problèmes peuvent être levés dans le cas de l'attention, e.g. le calcul n'est pas dans ce cas directionnel et peut être parallélisé ce qui amène généralement une réduction du temps de calcul non négligeable lors d'exécutions sur GPU.

Dans un contexte d'auto-attention appliqué à une séquence de n éléments, le calcul classique des scores d'attention est cependant de complexité quadratique par rapport à la taille de la séquence, i.e. de l'ordre de n^2 d'un point de vue calculatoire et de stockage mémoire. Pour une fonction de calcul de score d'attention symétrique de type produit scalaire, $(n \times (n - 1))/2$ scores d'attention doivent être calculés. Cela représente une limitation évidente lors du traitement de longues séquences. Nous avons été confrontés à cette limitation dans différents contextes applicatifs abordés dans nos travaux, en particulier lors du traitement de décisions de justice composées de milliers de tokens (éléments des séquences), et dans le traitement de grosses instances de VRP. Ce problème est clairement identifié dans la littérature et induit des limitations importantes dans la mise en oeuvre applicative des architectures à base d'attention.

⁸⁰. Le calcul de la représentation d'un élément i , $1 < i < n$, d'une séquence de n éléments dépend du calcul de la représentation des éléments $j < i$ lors d'un encodage en avant (*forward*), ou $j > i$ lors d'un encodage à l'envers (*backward*), les deux étant généralement réalisés en pratique (approches bidirectionnelles, contextes non autoregressifs).

Le succès rencontré par les architectures qui tirent parti du mécanisme d'attention a fait que de nombreuses équipes de chercheurs se sont concentrées sur la réduction de la complexité du mécanisme d'attention. Pléthore de contributions ont ainsi été proposées pendant la période de notre étude active de ce sujet. On distingue notamment les travaux qui visent à réduire la complexité du calcul exact de l'attention, de ceux qui portent sur son approximation. L'état de l'art sur ce sujet est mentionné dans une contribution liée à nos travaux sur ce thème [58] et dans la thèse de Charles Condevaux [57]; la thèse d'Ali Yaddaden expose aussi un état de l'art intéressant sur ce sujet [5].

L'approche que nous avons étudiée vise à approximer le calcul de l'attention, et non pas à optimiser son calcul exact. Nous nous sommes intéressés au cas où la proximité des éléments d'une séquence fournie en entrée importe dans la phase de construction des plongements, et en particulier lorsque les éléments proches dans la séquence ont généralement plus d'importance pour résoudre la tâche traitée. On observe en effet très souvent en traitement du langage naturel un effet de majoration du score d'attention entre les éléments proches dans la séquence [55]; cet effet est fréquent dans un grand nombre de contextes applicatifs, e.g. traitement de séries temporelles.

De manière distinguée d'une grande partie de l'état de l'art de la période d'étude, nous nous sommes aussi intéressés à la proposition d'une technique d'approximation de l'attention qui permet la réutilisation de modèles pré-entraînés. En effet, par définition, de nombreuses formulations optimisées de l'auto-attention induisent une forte dépendance du modèle à la taille des séquences traitées. Cela se traduit alors par l'impossibilité d'utiliser un modèle entraîné sur une taille de séquence spécifique, sur des tailles de séquences plus grandes. Compte-tenu des impacts écologiques et économiques liés à l'entraînement des modèles, nous nous sommes proposés de travailler sur l'étude d'approximations du mécanisme d'auto-attention qui permettent d'obtenir de bonnes propriétés d'adaptation et de généralisation lorsque l'on augmente la taille des séquences. Nous souhaitons ainsi qu'un modèle entraîné avec l'approximation en question puisse être réutilisé sur des séquences de plus grandes tailles, sans nécessiter un réentraînement complet. Nous nous sommes aussi centrés sur la formulation d'approximations capables d'extrapoler à des séquences de plus grandes tailles, des modèles existants reposant sur l'auto-attention standard et pré-entraînés sur des séquences de plus petites tailles. La formulation de ce type d'approximations permettrait de réutiliser des modèles entraînés avec l'expression originale de l'auto-attention, dans des cas qui requièrent des spécialisations (*fine tuning*) avec prise en compte de séquences plus longues.

Nous avons pour cela proposé l'approche LSG attention (*Local Sparse Global attention*) qui permet une approximation de l'auto-attention linéaire par rapport à la taille des séquences traitées. Nous ne détaillerons pas ici l'approche dont les éléments techniques seront retrouvés dans [58].⁸¹ Pour les familiers du domaine, l'attention LSG s'inscrit plus particulièrement dans la continuité des approches de type modèles éparses, e.g. Longformer [17], Big Bird [284], et modèles à base de regroupement tels que Reformer [145] (pas de prise en compte de récurrence ou de factorisation). En synthèse, l'attention LSG repose sur le postulat (i) qu'une bonne approximation de l'attention nécessite la prise en compte d'informations locales (hypothèse d'importance de la proximité dans la séquence) et (ii) que les informations contextuelles globales peuvent être intégrées au calcul par la prise en compte de plongements de blocs de la séquence (information éparse - *sparse*). Nous intégrons aussi à notre mécanisme la possibilité d'accorder de l'attention à des éléments globaux supplémentaires que nous ajoutons à la séquence et qui porteront cette information globale (*tokens* fictifs). Cela permet de faciliter le flux d'information pour le calcul de l'attention et du gradient lors

81. Cet article est joint au manuscrit, cf. Annexe A page 215.

de l'entraînement. Nous avons étudié différentes formulations de l'attention LSG en considérant des approches d'approximation de l'attention locale à base de convolutions, de techniques de sélections de clés (qui se basent sur l'évaluation de la norme des représentations), ou de techniques de regroupement (*clustering*).

Les évaluations que nous avons menées dans des contextes de TALN montrent que l'attention LSG permet l'obtention d'approximations rapides en pratique, et qui obtiennent de bonnes performances de généralisation dans des contextes de tâches de classification et de résumé sur de longs documents. De manière intéressante, nous avons aussi montré que l'attention LSG obtient de bonnes propriétés d'extrapolation et permet alors de réutiliser, dans des contextes qui nécessitent la prise en compte de longues séquences, des modèles pré-entraînés à l'aide de l'attention classique. Un travail important de mise à disposition d'outils permettant d'utiliser l'approche proposée en pratique a été réalisé⁸²; des modèles existants (BERT, RoBERTa, BART . . .) extrapolés à de plus longues séquences à l'aide de l'attention LSG ont aussi été partagés sur la plateforme Huggingface.⁸³ Ce travail important de développement et de mise à disposition de nos travaux à la communauté a été permis par Charles Condevaux.

Comme nous l'avons souligné, du fait de son importance dans les architectures contemporaines de réseaux de neurones, le mécanisme d'attention et plus particulièrement d'auto-attention font l'objet de beaucoup de travaux. Bien que pléthore d'approches aient été proposées dans la littérature, il est difficile aujourd'hui d'avoir une appréciation claire des avantages et des inconvénients de chaque proposition d'un point de vue pratique. Des initiatives visant à permettre une comparaison des approches sont à encourager, cf. *Long Range Arena* ou *Comprehensive Attention Benchmarking* [252, 285]. Un travail consistant à positionner l'attention LSG au regard de ces contextes d'évaluation mériterait d'être réalisé.

Il est probable que nous observions une stabilisation des architectures des gros modèles de langue (*Large Language Model*, LLM) et que les approches de type Transformers restent encore pour de nombreuses années populaires pour aborder des problématiques nécessitant le traitement de séquences (cela même si des alternatives intéressantes sont proposées, e.g. RetNet [246]). Dans ce contexte, il nous semble tout à fait important d'étudier les techniques d'approximation de l'attention ou plus largement tout mécanisme capable d'extrapoler des modèles à des entrées de plus longues tailles, e.g. LSG Attention. Ces techniques, en permettant de considérer des contextes réduits pendant les phases d'entraînement, et de réutiliser des modèles existants, permettront de limiter les coûts écologiques et économiques afférents à l'entraînement. Elles contribueront aussi à démocratiser l'accès à des LLM performants issus de modèles open source, libres, et entraînés de manière transparente (cf. encadré).

Les LLM payants les plus récents permettent aujourd'hui la prise en compte de contextes de tailles très importantes, e.g. contexte de 32k pour ChatGPT entreprise (GPT-4 32K)(prompt + réponse), et de 100k pour Claude. Ces modèles reposent souvent sur des architectures non détaillées qui exploitent très certainement des formes de linéarisation de l'attention, ou des approximations très efficaces comme l'attention *sparse* proposée par FlashAttention 1 et 2 [68]. Ils utilisent aussi par ailleurs très probablement des astuces lors des phases d'inférence pour

82. Utilitaire de conversion de *checkpoints* de modèles Transformers.

https://github.com/ccdv-ai/convert_checkpoint_to_lsg

83. <https://huggingface.co/ccdv>. Une adaptation du modèle BART a par exemple été téléchargée plus de 1000 fois, cf. <https://huggingface.co/ccdv/lsg-bart-base-4096>.

permettre des formes d'extrapolation. Les modèles ouverts ne permettent généralement pas la prise en compte de contextes longs - même si quelques initiatives visant à partager des versions extrapolées existent [58] [52]. A titre d'exemple, Falcon-180B, un LLM libre et populaire à l'heure de cette rédaction, utilise un contexte de seulement 2048 tokens en se basant sur une formulation exacte et optimisée de l'attention (FlashAttention, et de l'attention multi-requêtes pour éviter de démultiplier les représentations dans chaque tête)[6]. Llama 2 est lui entraîné en tenant compte de tailles de contextes de 2024 et 4096 tokens [259]. Ces modèles mériteraient d'être étendus à de plus longues séquences pour pouvoir être utilisés dans un plus grand nombre d'applications. Les techniques d'interpolation de position, adaptées pour les méthodes reposant sur des techniques de plongement de position *rotationnel* ont par exemple récemment permis d'étendre LLaMa à des contextes de 32k tokens [52]. Nous devons libérer du temps de travail afin d'éprouver les techniques d'extrapolation permises par l'attention LSG sur ces nouvelles architectures.

Interprétabilité en apprentissage profond

Nous présentons dans cette section le travail que nous avons réalisé sur l'interprétabilité des modèles d'apprentissage machine.

Rappelons tout d'abord quelques éléments de contexte de ces travaux. Nous disposons de modèles d'apprentissage machine d'ores et déjà capables d'adresser des tâches loin d'être triviales. Le dynamisme du domaine de l'apprentissage machine et l'essor du paradigme de l'apprentissage profond laissent par ailleurs imaginer de nouvelles ruptures à court terme. Le contexte est tel que nous évoquons aujourd'hui la substitution de l'opérateur humain dans de nombreuses tâches ce qui laisse imaginer la suppression ou a minima la transformation profonde de professions entières. Les risques civilisationnels potentiels induits par ces technologies sont aussi évoqués sérieusement. Au-delà des désaccords potentiels sur le caractère catastrophiste de certaines postures ou analyses [197], il nous faut admettre que nombre de discussions que nous partageons aujourd'hui sur ces sujets appartenaient, il y a encore quelques années, au domaine de la science-fiction. Le bouleversement semble acté, sa dynamique interroge. Beaucoup de questions importantes méritent alors de trouver réponses. Comment assurer notamment une bonne intégration des technologies à base d'apprentissage machine dans notre société, en particulier afin qu'elles restent justes et au service de l'Humain dans le respect de son écosystème ? L'importance aujourd'hui donnée à cette vaste question est pour le moins symptomatique du fait que notre société vit une véritable transition.

Les transformations amenées par l'évolution des technologies issues de l'apprentissage machine sont en effet profondes à bien des égards. En quoi l'essor de l'apprentissage machine change-t-il la donne par rapport aux approches traditionnelles ? L'automatisation galvanisée par l'Informatique est traditionnellement, et encore largement aujourd'hui, pour l'essentiel fondée sur l'exploitation de procédures algorithmiques définies de manière explicite. Cela rend possible la compréhension de la résolution du problème, ou de la prise de décision en des termes métiers. Les données d'entrées - ou caractéristiques (*features*) du problème - considérées pour aborder le problème traité, sont très souvent intuitives pour le décideur. La manière de les exploiter l'est souvent tout autant. A titre d'exemple, l'évaluation de l'accès à un prêt bancaire reposera sur l'appréciation de caractéristiques telles que le taux d'endettement, l'âge et le salaire d'une personne. Le décideur comprendra une prise de décision automatisée reposant sur une procédure explicite, e.g. M. X se voit refuser un prêt car il a un taux d'endettement supérieur à 35%. Le processus de prise de décision est seulement codé

pour être exécuté automatiquement ; sa formulation n'est en revanche pas déléguée : le métier est en pleine maîtrise de la logique décisionnelle, il l'explique et seule son exécution est automatisée. La procédure de prise de décision peut ainsi, avant toute utilisation, et avec plus ou moins de difficultés, être comprise, discutée, et évaluée au regard d'attentes légales. Chose importante, cette procédure peut aussi être confrontée à un corpus idéologique et à un système de valeurs d'intérêt.

Les décisions sont cependant aujourd'hui de plus en plus déléguées à des systèmes que l'on ne sait expliquer ; ces systèmes sont même dans certains contextes qualifiés de boîtes noires. Cela est en particulier vrai dans le cadre de l'apprentissage profond où la décision repose sur des modèles qui dépendent de milliards de paramètres en partie utilisés pour représenter de manière difficilement interprétables les objets sujets de la prise de décision. Richard Feynman disait "*What I cannot create I do not understand*.", ce qui n'empêche de créer collectivement ce que l'on ne comprend pas. Vous conviendrez en effet qu'il est nettement plus complexe d'expliquer la décision finale lorsque M. X est plongé dans un espace latent dont seule le nombre de dimensions est défini, et que la décision de l'octroi ou non d'un prêt, est ensuite prise via l'analyse de cette représentation par une tête de classification - e.g. simple modèle linéaire. Même si l'objet sur lequel porte la décision est représenté par des caractéristiques explicites pour le métier, la procédure implémentée par le prédicteur qui exploite les valeurs de caractéristiques dans un contexte de décision donné sera obscure, inaccessible en tant que telle par l'Homme. Il est bien entendu possible d'explicitier le processus de prise de décision (prérequis afin de pouvoir l'exécuter de manière automatisée). Un réseau de neurones n'est en soi qu'une fonction paramétrique. Il est donc tout à fait possible d'exprimer la forme fonctionnelle développée d'un prédicteur qui se base sur ce type de réseaux. Il s'agit pour cela d'explicitier la fonction et les valeurs de ses paramètres. Cela ne constitue cependant pas une explication intelligible pour notre espèce : résumer une décision à un simple calcul induit une perte de la sémantique de la logique décisionnelle sous-jacente.

Les modèles d'apprentissage machine que nous étudions sont finalement généralement mobilisés dans des chaînes de traitements ou des systèmes d'aide à la décision en vue d'aider, voire d'automatiser le traitement de problématiques complexes. Cette quête de l'automatisation, partielle ou totale, nous invite naturellement à aborder la question de la compréhension des décisions prises ou aidées par de l'apprentissage machine. Cela est d'autant plus vrai lorsque l'utilisation de l'apprentissage machine peut induire des risques médicaux, juridiques, ou plus largement éthiques et moraux (e.g. risques de biais). La question de la compréhension des décisions déléguées à la machine est suffisamment importante pour que des initiatives gouvernementales nous obligent aujourd'hui dans certains cas, en fonction des législations, à garantir ce droit à l'explication. Les outils théoriques et pratiques capables de nous amener une meilleure compréhension des modèles d'apprentissage machine sont ainsi naturellement les bienvenus. Ils contribueront aussi à l'amélioration des modèles (e.g. limitation des risques d'attaques adverses, *adversarial attacks*) et participeront plus généralement à une meilleure compréhension du domaine d'un point de vue recherche.

Nous ne traiterons pas largement le sujet vaste, complexe, et par nature interdisciplinaire de l'interprétabilité et de l'explicabilité des modèles d'apprentissage machine au sens large – cf. *eXplainable AI* (XAI), sous-domaine de l'IA. Nous n'aborderons pas notamment la complexe question de la définition d'une explication.⁸⁴ Nous adopterons dans notre cas une démarche modeste visant à étudier le sujet sous un angle restreint et résolument technique. Notre travail porte en

84. La littérature parle souvent d'*intellectual oversight* mais il n'y a pas de consensus sur le sens à lui accorder.

effet pour l'essentiel à ce jour sur l'étude de techniques d'interprétation *post hoc* locale à l'aide de méthodes d'attribution, agnostiques du modèle et des données utilisées pour son entraînement, et cela pour des méthodes d'apprentissage automatique supervisé. Notre positionnement ainsi qu'un exemple de contribution dans ce domaine seront précisés par la suite.

La notion d'interprétabilité *post hoc* signifie que le prédicteur que nous étudierons sera entraîné - nous n'étudierons pas le comportement d'un modèle dans l'absolu. La nature agnostique de l'approche au modèle et aux données, souligne le fait que (i) nous n'exploiterons pas de caractéristiques propres au modèle sur lequel repose le prédicteur étudié, et (ii) nous n'aurons pas un accès aux données utilisées lors de la phase d'entraînement du modèle qui a distingué le prédicteur.

La notion d'interprétation locale vise à expliquer la prédiction produite par un prédicteur et cela dans le contexte spécifique d'une entrée donnée. De manière plus générale, l'on souhaite savoir pourquoi le prédicteur a produit une décision spécifique pour l'entrée en question. L'interprétation locale se concentre ainsi sur l'explication d'une prédiction spécifique conditionnée à une entrée.

L'interprétation globale vise un objectif plus ambitieux que l'interprétation locale ; elle cherche à caractériser un prédicteur indépendamment d'une entrée spécifique, e.g. en distinguant des propriétés qui qualifieront globalement le modèle, e.g. règles de classification. Si l'on considère à nouveau l'exemple de la classification d'images, il pourrait s'agir de distinguer des motifs généraux discriminants en tenant compte d'invariants. Pour l'exemple de l'attribution d'un prêt, il s'agirait alors de distinguer la règle « si taux d'endettement supérieur à 35% alors refus du prêt ». L'interprétation globale de modèles complexes est difficile aujourd'hui. Nos travaux se sont pour l'instant intéressés à l'étude de la problématique de l'interprétabilité locale.

Revenons donc à la notion d'interprétation locale. Plusieurs approches peuvent être distinguées pour tenter d'expliquer une décision de manière locale. La première dite d'attribution vise à identifier l'importance donnée à chaque caractéristique par le prédicteur pour la prise de décision (i.e., *feature importance*) ; c'est l'approche que nous avons étudiée. Dans un contexte de classification d'une image, il s'agirait par exemple d'identifier les pixels qui s'avèrent discriminants pour sa classification ; si l'on considère différentes caractéristiques qualifiant M. X, il s'agirait de distinguer celles qui expliquent l'acceptation ou le refus du prêt, e.g. taux d'endettement pour un refus. D'autres méthodes d'interprétation locale adoptent une approche différente. On distingue entre autres celles qui tentent de générer des explications contrefactuelles, en évaluant la perturbation minimale de l'instance étudiée qui amènera une modification de la prédiction associée. D'autres méthodes cherchent aussi à substituer le modèle localement par un modèle plus simple à comprendre (*surrogates approaches*). Notez que dans ces deux derniers cas l'on vise à produire des informations sur la frontière de décision locale du prédicteur. Toutes les approches mentionnées comportent bien entendu des avantages et des limites (expressivité, fidélité. . .). Nous nous restreindrons par la suite aux méthodes d'attribution. Le lecteur intéressé par les approches à base d'explications contrefactuelles ou de modèles de substitution - et leurs limites - consultera par exemple [164, 163].

Les méthodes d'attribution visent à étudier la contribution de chaque caractéristique de l'entrée traitée pour la prédiction obtenue. On cherche dans ce cas à attribuer aux différentes caractéristiques fournies en entrée une partie du score attribué à une classe. Nous restreindrons notre exposé au cas de la classification mais ce cadre général s'étend aux problématiques de régression. Il s'agit alors d'assigner à chaque caractéristique un score d'attribution : une valeur réelle qui traduira son degré de contribution dans une prise de décision spécifique.

La modélisation de la problématique d'attribution que nous avons étudiée repose sur une formulation, standard dans l'état de l'art, qui se base sur la théorie des jeux coopératifs. Cette formulation définit plus précisément que le score d'attribution à associer à chaque caractéristique peut être calculé par la résolution d'un problème d'attribution dans un contexte de jeu coopératif. Les caractéristiques sont vues comme les joueurs du jeu coopératif pour lesquels il s'agit de partager un gain : dans notre cas le score d'une classe (logit ou probabilité).⁸⁵

L'état de l'art aborde quasi-exclusivement ce problème de rétribution à l'aide de la valeur de Shapley. Cette valeur comporte en effet l'avantage de disposer d'un certain nombre de propriétés jugées intéressantes au regard de l'objectif d'attribution. Le respect de ces propriétés caractérise par ailleurs la valeur, e.g. efficacité (la somme des attributions équivaut à la quantité à attribuer), ou l'axiome du joueur nul (si l'incidence d'un joueur est nulle il ne recevra aucune attribution). Une vaste littérature existe en théorie des jeux sur ce type d'indice, et plus spécifiquement sur la valeur de Shapley. Nous ne la détaillerons pas ; le lecteur intéressé fera référence à [98, 204, 272, 224].

Bien qu'intéressante à plusieurs égards, la valeur de Shapley induit des difficultés d'un point de vue calculatoire. Son calcul repose en effet sur l'évaluation de l'ensemble des sous-ensembles de joueurs, appelés coalitions, au nombre de 2^N pour N joueurs. Cette complexité exponentielle empêche le calcul exact de l'indice de Shapley et son utilisation pour la problématique d'attribution en apprentissage machine (il est fréquent d'avoir des centaines, milliers, voire des millions de caractéristiques). Une vaste littérature s'est alors intéressée à approximer la valeur de Shapley dans des contextes spécifiques d'attribution en apprentissage machine. Différentes alternatives d'approximation ont ainsi été proposées, e.g. [182, 7, 4, 50, 231]. Leur comparaison repose la plupart du temps sur des gains en termes de réduction de complexité algorithmique. Il est en effet important de souligner qu'il n'existe que peu de travaux sur l'évaluation des résultats produits par les méthodes d'attribution. Cela s'explique notamment par le fait qu'une approche axiomatique gouverne de nombreux pans de la littérature sur ce sujet. La valeur de Shapley, appliquée à cet exercice, est souvent considérée comme admise dans le sens où la pertinence des attributions qu'elle produit est selon nous peu discutée dans la littérature. Il s'agit alors très souvent dans les travaux sur le sujet, d'optimiser son calcul ou son approximation, et non pas de discuter la pertinence des attributions réalisées. A noter que l'utilisation de l'indice de Shapley est implicitement justifiée par les fondements axiomatiques de la valeur, bien que des approximations soient utilisées en pratique et que le respect de ces fondements ne soit plus vérifié dans ce cas. . . [231]

Nous avons souhaité adopter une approche différente en s'autorisant à reconsidérer certaines propriétés respectées par la valeur de Shapley du fait du contexte spécifique d'utilisation dans le cadre de la problématique d'attribution. Deux axes d'étude principaux ont alors été considérés dans nos travaux [59]⁸⁶ :

1. La proposition de techniques d'attribution alternatives à la valeur de Shapley.
2. La proposition d'un cadre d'évaluation des techniques d'attribution dans le contexte de la classification d'images.

L'étude d'alternatives à la valeur de Shapley est motivée par les difficultés induites par son calcul exact, mais aussi par certaines critiques possibles de son utilisation dans le contexte de l'attribution. Le calcul (éventuellement approché) de la valeur de Shapley d'une caractéristique repose sur une moyenne (non pondérée) des contributions marginales de la caractéristique en question pour chaque

85. La fonction de caractéristique au sens de la théorie des jeux correspond donc à la prédiction pour la classe étudiée (ou le logit correspondant).

86. Cet article est joint au manuscrit, cf. Annexe A page 187.

coalition, i.e. moyenne des différences sur le score final amené par l'ajout de la caractéristique à chaque sous-ensemble de caractéristiques l'excluant. Le calcul de la valeur de Shapley considère donc que toutes les coalitions contribuent de manière semblable à l'évaluation de l'attribution finale. Cela peut être critiquable dans certains contextes. Prenez par exemple le cas du traitement d'images ; la même importance sera donnée aux évaluations de la contribution marginale d'un pixel i dans le cas de l'étude des deux coalitions suivantes : (i) coalition composée d'un sous-ensemble aléatoire de pixels, et (ii) coalition composée d'un sous-ensemble de pixels sélectionnés dans le voisinage direct du pixel i (e.g., la coalition correspond à une extraction de l'image autour du pixel i). Il semble cependant évident qu'une dépendance spatiale entre les pixels mérite d'être prise en compte pour étudier la sémantique d'un pixel (et donc sa contribution). La prise en compte de propriétés spatiales dans l'analyse de données visuelles est par ailleurs établie au niveau biologique chez l'animal [177]. Il pourrait donc être souhaitable que tous les ensembles de pixels ne soient pas traités de manière équivalente lors de l'évaluation.⁸⁷ Cette variabilité dans la cohérence de certaines coalitions au regard d'un applicatif spécifique, ou d'un type de données, encourage à étudier à elle seule des formulations alternatives.

Nous nous sommes focalisés sur l'étude d'une famille de valeurs appelées ESL values pour *Efficient Linear and Symmetric values* – aussi appelées valeurs LES. Cette famille, qui peut s'exprimer sous la forme d'une expression paramétrique, généralise la valeur de Shapley [224].⁸⁸ Certaines valeurs ESL induisent des complexités calculatoires nettement plus faibles que celle de la valeur de Shapley. Elles peuvent même dans certains cas être linéaires par rapport au nombre de caractéristiques. Nous nous sommes aussi intéressés à étudier des formulations nouvelles de valeurs de manière à garantir le respect de propriétés jugées d'intérêt, tout en conservant une faible complexité algorithmique. Nous avons pour cela proposé une méthode d'attribution nommée FESP (*Fair-Efficient-Symmetric-Perturbation*). Celle-ci est de complexité linéaire par rapport au nombre de caractéristiques⁸⁹, et respecte différentes propriétés jugées d'intérêt. Tout comme la valeur de Shapley, la valeur FESP respecte en particulier la symétrie et la propriété de *Fair treatment* qui précise que si l'apport d'une caractéristique i est systématiquement supérieur à celui d'une autre caractéristique j , alors l'attribution de i sera supérieure à celle de j . Nous ne développons pas ici plus précisément les aspects théoriques de ces contributions. A noter que, l'auteur n'étant pas expert en théorie des jeux, ces aspects de nos travaux ont été largement portés par Stéphane Mussard et Charles Condevaux. Nous ne discuterons pas non plus les détails sur lesquels repose l'implémentation de l'approche – le masquage d'une caractéristique est notamment traité par une technique d'occlusion au niveau de super-caractéristiques.⁹⁰ L'approche FESP, et d'autres alternatives à Shapley, ont été comparées à différentes approches de l'état de l'art en considérant plusieurs contextes d'étude et types de données, e.g. images, textes. FESP et d'autres valeurs ESL alternatives à Shapley que nous proposons d'étudier se sont avérées performantes au regard de la modalité d'évaluation que nous avons proposée et que nous détaillons ci-après.

87. C'est l'hypothèse adoptée dans les CNN qui reposent sur les notions de convolution et de *pooling* exprimées via des noyaux qui évaluent l'information localement.

88. La valeur de Shapley est la seule valeur ESL qui satisfait l'axiome du joueur nul.

89. Elle se base sur la seule étude des coalitions extrêmes.

90. Dans le cas du traitement d'images, nous masquons des superpixels pour ensuite définir l'attribution d'un pixel comme la moyenne des attributions de ses superpixels. Différentes techniques sont étudiées en fonction du type de données.

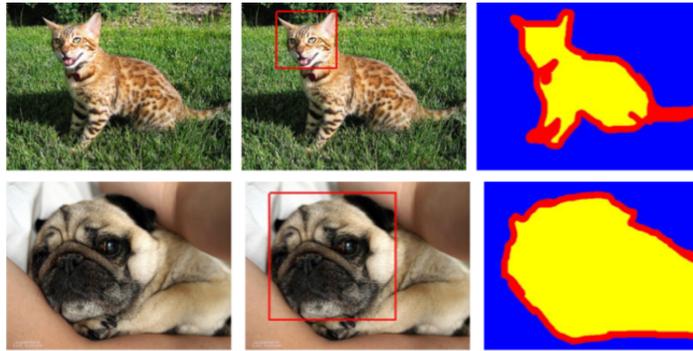


FIGURE 28 Exemple de photo avec masque de segmentation utilisé pour identifier les pixels attendus, source [255].

L'étude d'un cadre d'évaluation visait à critiquer la qualité du résultat d'attribution d'une méthode au regard de la finalité visée. Il est en effet en soi facile de modéliser le problème d'attribution à l'aide d'une approche quelconque, e.g. via l'indice de Shapley ou une autre valeur ESL, voire de manière alternative. Le résultat obtenu est-il cependant conforme à l'objectif fixé initialement, à savoir expliquer le modèle ? La question n'est pas triviale car la nature des modèles prédictifs étudiés fait que nous ne connaissons pas à l'avance l'attendu en termes d'attribution. Cela peut s'expliquer par plusieurs raisons et en particulier par le manque de compréhension des modèles prédictifs. Il nous faut aussi admettre qu'une restitution sous la forme d'attributions peut même dans certains cas ne pas être en mesure de traduire la complexité de la prise de décision ; les simples scores d'attributions peuvent dans l'absolu ne pas être suffisamment expressifs.

Afin d'étudier cette question de l'évaluation, nous avons proposé une approche intuitive fondée sur l'étude de tâches simples de classification d'images pour lesquelles nous estimons avoir une bonne maîtrise globale de l'attendu en termes d'attribution. Nous partons pour cela du principe (i) que la tâche sélectionnée est simple à résoudre pour les modèles étudiés (ceux-ci permettant d'obtenir des valeurs d'*accuracy* proches de 100% sur un jeu de test), et (ii) que nous connaissons les caractéristiques discriminantes de la classe attendue pour chaque observation. Plutôt que de considérer un contexte totalement artificiel avec des données générées, nous nous sommes tant que possible attachés à utiliser des données réelles. A titre d'exemple, nous proposons d'admettre une tâche de classification binaire d'images pour laquelle nous disposons des masques de segmentation permettant d'identifier les pixels directement associés aux classes. Un test a été réalisé afin de distinguer des images contenant des chiens ou des chats pour lesquelles nous disposons des masques des chats et chiens (Oxford-IIIT Pet Dataset[208, 255]⁹¹, cf. Figure 28). Les pixels des masques sont alors considérés comme discriminants pour les classes correspondant aux labels. Autrement dit, à la question : quelles sont les caractéristiques (pixels) de l'image jugées importantes pour expliquer son label ? Nous estimons que l'attendu correspond aux pixels du masque. Des variantes ont aussi été définies en générant des images composites en maîtrisant les parties des images créées dans lesquelles la classe du label était représentée, e.g. génération d'une image composée de quatre images juxtaposées sélectionnées aléatoirement, et pour lesquelles les classes sont connues et différentes ; les attendus sont dans ce cas les pixels de l'image de la classe traitée. Cette approche générale pourra dans une certaine mesure être généralisée à d'autres types de données, e.g. à l'aide de textes annotés traités dans le cadre d'une tâche de classification de textes.

91. VGG 16 obtient des performances de l'ordre de 100% d'*accuracy* sur ce type de tâches simples.

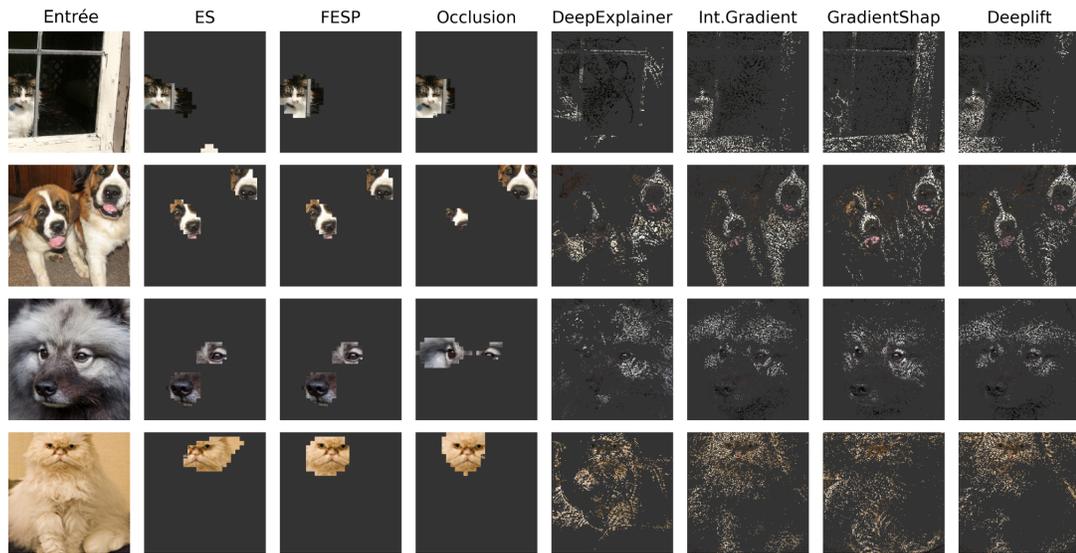


FIGURE 29 Exemple de résultats obtenus par différentes approches dont celles étudiées, Equal Surplus (ES) et FESP - source [59].

Dans le cadre de l'étude d'un modèle de classification performant, nous considérons alors le postulat suivant pour l'évaluation : la performance d'une méthode d'attribution peut être évaluée en fonction de sa capacité à distinguer les pixels attendus (e.g. des masques). Une approche d'attribution peut donc être critiquée sur la base des scores d'attribution assignés aux pixels appartenant aux masques. L'évaluation des méthodes d'attribution peut alors se baser sur les métriques classiquement utilisées en classification, e.g. si une image contient k pixels discriminants attendus, nous regardons la proportion de ces pixels retrouvés parmi ceux ayant les plus forts scores d'attribution assignés par la méthode d'attribution évaluée. Nous avons aussi proposé d'évaluer indirectement la méthode d'attribution en analysant la performance du modèle prédictif en ne considérant que les top- k pixels qu'elle distingue. En effet, puisque la méthode d'attribution juge que ce sont les pixels les plus à même d'expliquer la classification, de bons scores de classification sont alors censés être obtenus par le modèle prédictif en utilisant seulement ces pixels. Une illustration graphique des résultats obtenus sur quelques exemples dans un contexte de traitement d'images sont proposés en Figure 29 - nous proposons Equal Surplus (ES), et FESP dans nos travaux. Les résultats obtenus par nos approches soulignent l'intérêt de les étudier. Ils permettent aussi d'apprécier la, semble-t-il, pertinence de coupler certaines approches qui ont tendance à extraire des informations de natures différentes et complémentaires (partie la plus discriminante et forme globale). Un exemple de couplage d'approches est proposé en Figure 30.

L'approche d'évaluation proposée comporte bien entendu certaines limites. L'annotation peut être imparfaite car des objets hors du masque peuvent aussi, dans l'absolu, s'avérer discriminants, e.g. jouets dans notre cas. Nous considérons cependant que les tâches traitées sont sélectionnées afin que cela puisse être considéré comme marginal. De manière plus générale, nous devons reconnaître que le postulat sur lequel se base notre protocole d'évaluation admet une hypothèse implicite d'alignement à l'Humain - voire une forme d'anthropomorphisme des modèles d'IA à base d'apprentissage. Notre approche se base en effet sur le postulat que les caractéristiques identifiées comme attendues sont censées être discriminantes. Un modèle peut cependant tout à fait être performant en adoptant une

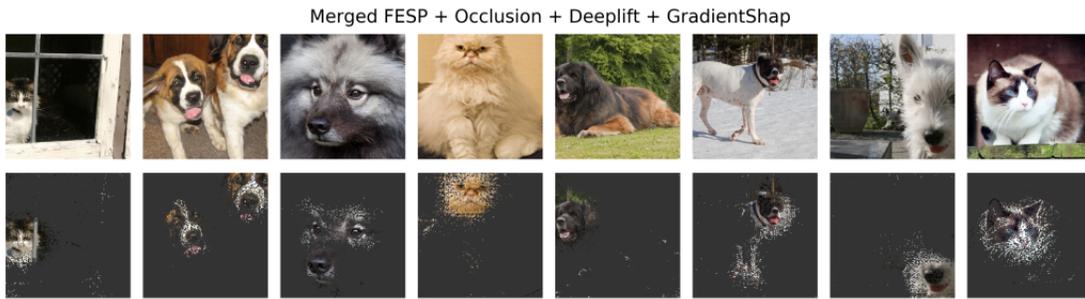


FIGURE 30 Exemple de résultats obtenus en couplant les approches FESP, Occlusion, Deeplift [236] et GradientShap [182] - source [59].

approche différente incompatible avec notre approche d'évaluation.⁹² Dans ce cas l'évaluation de la méthode d'attribution serait bien entendu tout simplement erronée. Nous estimons cependant effectuer une proposition argumentée, pour laquelle nous explicitons les limites, et qui a le mérite de permettre une critique des méthodes d'attribution.

Une implémentation open source de l'approche FESP et du protocole d'évaluation est partagée (https://github.com/ccdv-ai/fesp_es). Elle permet d'utiliser l'approche FESP avec différents types de données et plusieurs modèles d'apprentissage machine (e.g. d'apprentissage profond via la librairie Transformers d'Huggingface [273], mais aussi des modèles traditionnels implémentés dans Scikit-learn [212]). L'implémentation de la procédure d'évaluation permet de reproduire les résultats obtenus, et de l'appliquer sur de nouvelles données pour comparer d'autres approches. Ce travail d'implémentation a été réalisé par Charles Condevaux.

⁹². Imaginons dans notre contexte de classification binaire, un réseau qui apprenne à classer une image comme étant de la classe chat (resp. chien), par négation. La classe chat (resp. chien) s'explique du fait qu'il n'observe pas de chien (resp. chat) dans l'image. L'attribution théorique qui correspondrait à un tel réseau ne serait-elle pas une attribution uniforme sur l'ensemble des pixels ? Notons que la question de la validité des explications locales se pose aussi pour les approches non fondées sur une méthode d'attribution, e.g. pour des explications contrefactuelles [164].

Projet de recherche

Cette partie discute quelques éléments de notre projet recherche, ainsi que son adéquation avec le positionnement de l'unité de recherche Euromov Digital Health in Motion dans laquelle celui-ci sera réalisé.

Sommaire

3.1 Propos généraux	123
3.2 Apprentissage Machine appliqué	125
3.3 Transfert de connaissances de l'Homme vers la Machine	128
3.4 Transfert de connaissances de la Machine vers l'Homme	130

3.1 Propos généraux

Les travaux exposés dans ce manuscrit couvrent un spectre large de l'Intelligence Artificielle. Ils adoptent la plupart du temps un positionnement résolument appliqué, tourné vers l'Industrie, et cela dans des domaines variés (santé, logistique, droit, sport, analyse psychosensorielle. . .). Cette particularité constitue une force car elle permet l'émergence d'idées enrichies par les différentes expertises rencontrées à la croisée de ces multiples chemins. Elle assure par ailleurs des travaux (et un quotidien) stimulants du fait d'applications variées et de la diversité des échanges avec des individualités et expertises parfois sensiblement différentes. Elle constitue aussi une faiblesse non négligeable à l'heure d'une surspécialisation naturelle des travaux à des sujet précis dans chaque domaine disciplinaire.

Bien que conscients des forces et faiblesses d'un tel positionnement et sans chercher à opposer différentes visions de la recherche, nous ne souhaitons pas proposer un programme de recherche centré sur un objet de recherche totem, unique et cloisonné. Nous tenons à souligner l'importance fondamentale de positionnements recherche très spécifiques d'un point de vue disciplinaire. Cela ne correspond cependant pas à la recherche que nous souhaitons pratiquer. Nous préférons donc défendre l'idée d'une recherche tournée vers la résolution de problèmes, basée sur le mariage des cadres théoriques existants de l'Intelligence Artificielle, et sur la formulation d'approches novatrices lorsque nécessaire. Un grand nombre d'équipes de recherche hétérogènes d'un point de vue disciplinaire se démarquent en adoptant des approches de ce type. Accordons-nous de plus à penser que le domaine de l'Intelligence Artificielle est suffisamment vaste pour permettre l'expression de formes variées d'interdisciplinarité et de pluridisciplinarité, notions chères aujourd'hui à de nombreux organes d'évaluation de la recherche.

Nous sommes aussi défenseurs du droit au changement, d'ordre professionnel en particulier. Il nous paraîtrait très présomptueux de pouvoir affirmer ce que sera notre recherche dans 10 ans. Une tentative d'esquisse à une échéance de 5 ans nous semble d'ores et déjà un pronostic suffisamment risqué compte-tenu de la rapidité de l'évolution de nos domaines, et de la dépendance de notre recherche aux opportunités de financement (sans évoquer les contraintes d'ordre stratégique induites parfois par nos tutelles). Peut-être que certains diront que cette vision traduit une forme d'immaturité ou d'instabilité scientifique ; d'autres penseront comme nous que la recherche appliquée évolue, que les cycles de développement de la recherche se font de plus en plus courts, et que les modes de fonctionnement en équipe-projet sur une durée relativement courte sont probablement plus adaptés (e.g., à l'image des équipes-projets d'INRIA). Dans la continuité des précisions générales apportées sur notre positionnement recherche, nous partageons ci-après quelques éléments du programme de recherche qui, nous l'espérons, guidera notre recherche.

Nous avons exposé tout au long de ce manuscrit une progression de notre recherche de l'utilisation de représentations des connaissances pour des traitements multiples (résumé conceptuel, extraction de règles. . .), vers l'étude d'aspects appliqués et fondamentaux de l'apprentissage machine. Dans la continuité de cette progression, le programme de recherche que nous implémenterons s'axera de manière préférentielle sur l'étude des problématiques liées à l'apprentissage machine à base de réseaux de neurones, dans des contextes d'apprentissage supervisé et par renforcement. Comme nous le verrons, cela ne nous empêchera pas de positionner certains travaux futurs à l'interface de l'apprentissage machine et de la représentation des connaissances, en considérant les représentations des connaissances comme vecteurs de transfert de connaissances ou plus généralement comme lingua franca à l'interface de l'Homme et de la Machine. Ce choix de réduction du spectre recherche étudié, déjà largement opéré, s'explique par la compétitivité du domaine de la recherche en IA aujourd'hui, en particulier en apprentissage machine, et l'impossibilité en ce qui concerne l'auteur, de maintenir un niveau d'excellence satisfaisant sur des pans de recherche aussi vastes.

Nous l'avons souligné à plusieurs reprises dans ce manuscrit, dans un monde en pleine transition numérique, le développement de l'Intelligence Artificielle (et de l'Apprentissage Automatique) nourrit à la fois les plus grands fantasmes et les plus grandes craintes. L'Homme se retrouve en effet aujourd'hui dans une situation où la Machine offre de grands espoirs en termes d'amélioration des services et d'automatisation de tâches complexes. Pour de nombreux applicatifs sensibles, un contrôle de la Machine est cependant critique, et plus largement, une collaboration étroite entre la Machine et les opérateurs humains doit être envisagée – impossible en effet pour certains applicatifs d'imaginer (aujourd'hui) substituer l'Homme par la Machine ; impossible aussi d'imaginer l'utilisation seule

de techniques d'apprentissage automatique lorsque les données d'entraînement sont extrêmement limitées. Un axe important du programme de recherche que nous souhaitons porter propose alors, dans la continuité des travaux menés à ce jour, de nous intéresser à l'étude d'un panel de techniques permettant d'augmenter la collaboration Homme-Machine par notamment le couplage de méthodes d'Apprentissage Automatique, de Fusion d'Information et d'analyses d'incertitudes, le tout dans des contextes nécessitant le traitement de données multi-modales (e.g., physiologiques, de mouvement, psychosensorielles. . .). Nous pensons qu'un tel positionnement général est à la fois pertinent vis-à-vis des problématiques offertes à notre domaine de recherche, et cohérent au regard de la variété des travaux que nous avons menés et des expertises que nous avons développées jusqu'à présent. Nous insisterons sur les aspects de notre programme directement en lien avec l'apprentissage machine. De manière générale, trois composantes non orthogonales peuvent ainsi être distinguées dans le programme de recherche que nous discuterons :

1. L'application de l'apprentissage machine à des domaines multiples dans une démarche pluridisciplinaire, avec des applications privilégiées aux domaines de la Santé et du Sport.
2. Le transfert de connaissance de l'Homme vers la Machine dans la définition de modèles d'apprentissage machine, lors des phases d'entraînement, mais aussi ensuite, lors de leurs utilisations post-entraînement.
3. Le transfert de connaissances de la Machine vers l'Homme par l'analyse de modèles entraînés et performants.

Certains aspects des volets liés aux transferts de connaissances de l'Homme vers la Machine, et de la Machine vers l'Homme ont par ailleurs fait l'objet d'un projet de groupe de recherche discuté en interne au sein d'IMT mines Alès.¹

3.2 Apprentissage Machine appliqué

La recherche que nous avons réalisée jusqu'à présent avec différents collaborateurs a presque toujours été nourrie par des exigences afférentes à des problèmes concrets d'un domaine applicatif particulier, e.g. augmentation des performances des systèmes existants, réduction des temps de calcul, cela pour des domaines variés tels que le domaine juridique ou le domaine médical.

Du recueil des besoins auprès des experts de domaine, à l'implémentation sur des données réelles, en passant par la formalisation de la problématique et l'adaptation des techniques à ses spécificités, la recherche appliquée mobilise un large spectre de compétences. Elle permet de plus d'échanger avec une pluralité de profils et d'experts enrichissants. Cette recherche, gouvernée par une finalité concrète et visible à court terme, est particulièrement stimulante. Elle est de plus en totale adéquation avec les missions institutionnelles qui nous sont confiées par l'Institut Mines Télécom (IMT), et plus spécifiquement avec celles en lien avec la formation des ingénieurs de demain, et le transfert des techniques de recherche de l'état de l'art vers le monde de l'Industrie. Ce type de recherches appliquées restent complémentaires d'une recherche fondamentale, et est aussi, d'une manière ou d'une autre, en synergie étroite avec la recherche aujourd'hui réalisée dans de nombreuses équipes académiques de qualité.

1. Projet HMCKE – *Human-Machine Collaboration & Knowledge Exchange* – qui mobilisait plusieurs collègues, et pour lequel l'auteur était positionné comme responsable. Ce groupe de recherche n'a pas été formé car la décision fut prise d'effectuer une collaboration avec l'Université de Montpellier dans le cadre de la création de l'unité de recherche Euromov DHM.

La recherche appliquée a par ailleurs une place particulière dans la jeune science Informatique. L'histoire récente de l'apprentissage machine nous enseigne l'importance de l'étude conjointe de problématiques théoriques et pratiques. A titre d'exemple, les développements récents de l'apprentissage profond sont très largement permis par des prouesses d'ingénierie, e.g. calcul GPU, traitements de grosses volumétries de données, développement de cadres de programmation (e.g. PyTorch [210]), des initiatives souvent pluridisciplinaires de constitution de jeux de données, e.g. Imagenet [71]. Des développements tels que ChatGPT seront qualifiés par certains de simples démonstrations d'ingénierie² ; ils bousculent et surprennent cependant aujourd'hui la recherche [197]. Cela souligne par ailleurs l'intrication particulière entre les problématiques d'ordres théorique et appliqué en Informatique. Il nous semble important dans ce contexte, comme nous l'avons souhaité jusqu'à présent, que les travaux de recherche appliquée que nous mènerons permettent le partage, aussi ouvert que possible, de jeux de données et d'outils logiciels qui permettront de reproduire notre recherche et de dynamiser celles connexes. Une importance toute particulière sera accordée à ce point dans notre programme de recherche - dans le respect des principes FAIR - *Findable, Accessible, Interoperable, Reusable* -, récemment aussi explicité dans le contexte de développement de modèles en IA [127].

L'usage massif du numérique, et celui exponentiel de l'Intelligence Artificielle, nous impose de plus d'aborder avec le plus grand sérieux les problématiques liées aux impacts écologiques des technologies que nous contribuons collectivement à développer et à démocratiser au travers de nos travaux. Nous l'avons aussi souligné dans ce manuscrit, les implications sociétales de l'adoption large de ces technologies seront potentiellement profondes. Traiter des impacts écologiques et sociétaux de ces technologies nécessite une pleine maîtrise de leurs réalités applicatives. Cela nécessite aussi de mener une recherche ouverte sur d'autres champs d'étude, en lien étroit notamment avec les experts des Sciences Humaines et Sociales - tels que nos collaborateurs experts en sciences et techniques des activités physiques et sportives (STAPS) d'Euromov DHM. L'expertise que nous avons développée dans la mise en application des techniques à base d'apprentissage machine, nous positionne favorablement pour ce type de travaux. Différentes actions ont par ailleurs été initiées sur la thématique « IA et Humain », dans le cadre du projet Européen ENFIELD, et au sein de l'axe thématique dédié de la communauté Data Analytics & AI de l'IMT. Ces implications portent pour nombre d'entre elles sur la mise en place d'actions communautaires censées dynamiser les échanges et les travaux sur ce vaste sujet. Dans ce contexte, travailler au lien avec les GDR de notre communauté nous semble aussi important, e.g. GDR RADIA.³ Comme nous le verrons plus loin, les sujets que nous nous proposons d'étudier d'un point de vue recherche sur les aspects « IA et Humain » portent plus particulièrement sur le Transfert de connaissances de l'Homme vers la Machine, ainsi que celui de la Machine vers l'Homme.

Dans le contexte d'Euromov DHM, les domaines applicatifs que nous pensons privilégier sont le sport et le médical au sens large. Les travaux que nous menons actuellement avec le MHSC (Montpellier Hérault Sport Club) sur l'évaluation du niveau de forme de joueurs de foot élite (Ligue 1) à l'aide de modèles prédictifs représentent un bon exemple de mariage interdisciplinaire à l'interface de la pratique et de la théorie (thèse ED 463 SMH, I. Diouron). C'est aussi le cas des travaux que nous menons sur l'utilisation de techniques d'apprentissage par renforcement pour l'étude de mouvements humains sur lesquels nous reviendrons ci-après (thèse ED 166 I2S, A. Ouhssain). Les dépôts de projets nationaux et européens récents en lien avec l'étude de la lombalgie

2. Yann Lecun dira de ChatGPT sur France Inter « [c'est] de la bonne ingénierie, c'est très bien ficelé, mais au niveau de la science et la technologie sous-jacente, ce n'est pas du tout révolutionnaire », ce qui n'est en soi pas faux.

3. Raisonnement, Apprentissage, et Décision en Intelligence Artificielle. <https://gdr-radia.cnrs.fr>.

chronique, ou encore l'étude de la charge cognitive en milieu écologique (chez le radiologue) illustrent la dynamique initiée. ⁴ Cette orientation stratégique nous permettra de traiter des aspects multimodaux incluant le traitement de données physiologiques et de capture de mouvements. Elle amènera aussi très probablement une implication moins marquée de nos travaux sur le traitement de données textuelles. Nos premiers résultats sur l'étude de données médicales illustrent en toute modestie l'apport que peut amener l'apprentissage machine pour la résolution de certaines tâches complexes liées au mouvement humain qui sont encore peu étudiées aujourd'hui, e.g. développement d'un système à base d'apprentissage profond pour la détection de comportements d'échappement à la douleur à partir de données Mocap et d'électromyogramme (EMG, activité électrique des nerfs et muscles) - première place au challenge AffectMove 2021 [223].

Comme mentionné, dans la continuité des travaux passés, ce programme de recherche s'ancre naturellement dans une recherche appliquée. Traiter des problèmes de recherche appliquée impose cependant très souvent d'isoler l'essence de la difficulté rencontrée afin de pouvoir l'étudier au travers d'un cadre formel adapté et épuré de tout applicatif. A l'image des travaux récents que nous avons menés sur l'interprétabilité, sur l'approximation du mécanisme d'attention, ou encore sur la formulation d'approches à base d'apprentissage profond et par renforcement pour la résolution de problèmes d'optimisation combinatoire, une place importante sera alors consacrée à l'étude des mécanismes qui sous-tendent les approches, neuronales par exemple, que nous mobiliserons dans nos recherches appliquées. Ces travaux de recherche nous semblent tout à fait pertinents pour (i) améliorer les systèmes, et (ii) permettre leur pleine maîtrise en considération des enjeux écologiques, sociétaux et économiques, en lien avec les technologies de l'IA, e.g. coût au sens général induit par l'entraînement et l'utilisation de modèles d'apprentissage machine. L'auteur souhaite dans ce contexte réserver un temps recherche non négligeable (40%) à l'étude d'aspects fondamentaux susceptibles d'avoir un impact concret sur l'application de l'IA dans nos sociétés.

A court terme au moins, nos travaux sur l'étude des aspects fondamentaux du mécanisme d'attention seront poursuivis. Les recherches que nous avons initiées sur l'optimisation combinatoire neuronale seront aussi étendues. Des travaux sur la compréhension de l'encodage des connaissances dans les réseaux de neurones seront aussi initiés. Comme nous le verrons après, ce dernier aspect est en lien étroit avec les axes de recherche liés au transfert de connaissances entre l'Homme et la Machine que nous détaillerons ci-après. La notion d'encodage de l'information concerne aussi par ailleurs des problématiques très appliquées, e.g. en lien avec l'adaptation de modèles entraînés pour le traitement de tâches spécifiques, cf. par exemple la démultiplication des applications de la notion de LoRA (*Low Rank Adaptation*) pour réduire le coût de la spécialisation de modèles pré-entraînés [126].

Le programme de recherche que l'auteur souhaite porter lui permettra aussi, dans une moindre mesure, de consacrer une partie de son temps recherche (10%) à l'étude d'approches exploratoires qui semblent prometteuses bien qu'elles ne trouvent pas aujourd'hui d'adoption large dans le domaine pratique. Nos travaux sur l'optimisation combinatoire neuronale s'inscrivent dans ce registre. A titre d'exemple supplémentaire, nous débutons des travaux sur l'étude de la modélisation de données temporelles, rythmiques et de synchronisations sociales par réseaux de neurones impulsifs (participation à l'ANR JCJC ModPuls portée par Patrice Guyot [238], thèse ED 166 I2S, T. Velletaz).

4. Nous faisons ici référence aux dépôts d'ANR avec le CHU de Nîmes et au dépôt du projet HumAIne (*Human-AI experience for futureproof healthcare*) sur l'appel HORIZON-CL4-2022-HUMAN-01-02 avec différents collaborateurs européens et pour coordinateur principal Philips. L'auteur était responsable du *workpackage* 4/6 dédié à la mesure de la charge cognitive à partir d'approches d'apprentissage machine et de données multimodales.

Nous détaillons ci-après les deux axes structurants de notre programme de recherche sur les aspects Transfert de connaissances Homme-Machine.

3.3 Transfert de connaissances de l'Homme vers la Machine

Plusieurs contextes d'étude nous invitent à travailler sur le transfert de connaissances de l'Homme vers la Machine ; deux d'entre eux nous semblent particulièrement importants : (i) faciliter l'apprentissage de modèles, et (ii) tenter de contrôler les modèles obtenus. Comment améliorer le transfert de connaissances portées par des agents humains ou des sources d'informations pertinentes vers des agents machines, e.g. au travers de l'utilisation de représentations formelles de connaissances ? Voilà la question à laquelle nous souhaitons contribuer de répondre en nous centrant sur le domaine de l'apprentissage machine et plus particulièrement sur l'étude de réseaux de neurones. Nos travaux passés sur l'utilisation de représentations des connaissances pour la mise en place de systèmes d'aide à la décision, pour certains à base d'apprentissage machine, nous positionnent favorablement pour ce type d'études.

Le transfert de connaissances de l'Homme vers la Machine vise dans un premier temps à faciliter l'apprentissage de modèles en fournissant des connaissances connues par les experts de domaines, ou qui ne peuvent être retrouvées dans les données d'entraînement. Cela laisse en particulier envisager une réduction des temps de calculs et des quantités de données nécessaires lors des phases d'entraînement, ainsi qu'une réduction de la complexité des modèles nécessaires. Un tel transfert de connaissances permettrait aussi de contrôler les modèles obtenus en les contraignant à apprendre des connaissances – potentiellement non explicitées dans les données, voire contradictoires aux données - que nous jugeons importantes afin qu'ils soient finalement alignés à un système de valeurs consensuel pour le contexte applicatif visé.

L'étude du transfert de connaissances de l'Homme vers la Machine peut aussi être étudié indépendamment de la phase d'apprentissage, e.g. dans une optique d'adaptation post-entraînement à l'image de l'approche LoRA que nous avons mentionnée ci-avant. Les travaux en lien avec l'utilisation d'une mémoire externe, e.g. sous la forme de graphes de connaissances, sont aussi intéressants afin de permettre l'utilisation de modèles entraînés dans des contextes susceptibles d'évoluer dans le temps (contexte non stationnaire, *environment and distribution shift*).

Le sujet d'étude est par définition très vaste. La nature des connaissances partagées et le mode de partage méritent d'être discutés.⁵ Les travaux que nous mènerons porteront de manière préférentielle sur le transfert de connaissances exprimées sous la forme de représentations des connaissances formelles de type ontologies. Le rôle prépondérant aujourd'hui joué par le langage naturel dans l'interface Homme-Machine nous invitera très probablement à élargir notre vision et à concilier les différentes approches. Un des objectifs que nous visons est la proposition de techniques qui permettent de préciser de manière formelle des invariants dans les données afin d'éviter d'avoir à les apprendre, ou risquer de les rater. De telles techniques pourraient se substituer à des techniques d'augmentation de données ou de définition d'architectures qui encodent des propriétés souhaitées. A titre d'exemple, il n'est pas aujourd'hui possible de préciser une invariance à une caractéristique du problème traité de manière explicite dans une instruction spécifiée avant ou pendant l'entraînement, ou même après, pendant l'inférence. On ne peut pas par exemple préciser que l'orientation d'une

5. En effet, la sélection des corpus de textes utilisés dans un contexte d'entraînement de LLM peut notamment être vue comme une forme de contrôle de la connaissance qui pourrait être apprise par un modèle.

image n'a pas d'incidence sur la détection d'un objet d'intérêt, ou que seules les positions relatives des villes importent pour la recherche d'une tournée optimale. Ces connaissances sont aujourd'hui généralement implicitement encodées dans les architectures mobilisées ou dans des phases de prétraitements des données (augmentation de données, normalisation).⁶

De manière concrète, nous avons sur ce sujet initié un projet qui vise à encoder de la connaissance *a priori* dans les modèles génératifs de mouvements humains à base d'apprentissage profond. L'idée est d'exploiter des propriétés connues du mouvement humain issues des Sciences du Mouvement Humain pour faciliter l'obtention de modèles génératifs réalistes. Le projet est réalisé dans un contexte interdisciplinaire au sein d'Euromov DHM, et en collaboration avec des collègues informaticiens et un chercheur en Science du Mouvement Humain (thèse ED 166 I2S, A. Oussain, début 01/24, dans la continuité de plusieurs stages encadrés sur des modèles génératifs de mouvements humains). Les paradigmes de l'apprentissage supervisé et par renforcement sont considérés dans ces travaux. Les connaissances sont à ce jour exprimées sous la forme de contraintes intégrées dans la définition de la notion de récompense et de la fonction de perte utilisées lors de l'apprentissage. Nous étudions aussi la prise en compte de ces connaissances dans la formulation d'une fonction objectif indépendamment de toute donnée d'entraînement, e.g. les saccades peuvent être détectées et pénalisées, la notion de complexité du mouvement est aussi caractéristique en science du mouvement et mériterait d'être intégrée aux modélisations. La prise en compte de connaissances exprimées sous la forme d'ontologies sera étudiée par la suite dans ce contexte précis.

La question de la pertinence de l'étude de procédures de transfert de connaissances de l'Homme vers la Machine se pose dans certains cas, notamment lorsque la finalité est une amélioration de la performance des systèmes. Certains lecteurs auront probablement en tête la fameuse citation de Fred Jelinek : « *Every time I fire a linguist, the performance of the speech recognizer goes up* » [97]. Il convient en effet de souligner que certains traitements bas niveau semblent particulièrement accessibles à la Machine sans plus de prise en compte de connaissances que de simples labels, i.e. le paradigme supervisé (incluant pour nous l'auto-supervisé) classique est suffisant lorsque nous sommes en mesure de disposer de suffisamment de données. Notons tout de même que dans certains cas des millions d'images d'un concept peuvent être requises pour qu'un modèle de classification d'images (non entraîné) apprenne un concept, ce qui rend l'approche intellectuelle largement critiquable bien qu'elle puisse s'avérer satisfaisante à certains égards du point de vue de l'objectif de résolution de problèmes.

Au-delà de cet aspect, nous sommes cependant convaincus que l'impossibilité pour la Machine d'avoir aujourd'hui accès à un grand nombre d'informations et de connaissances - e.g. connaissances générales de domaine, informations contextuelles et sensorielles - oblige pour de nombreux contextes l'encodage de connaissances *a priori* sous une forme ou une autre afin d'enrichir les modèles. Les gains potentiels sur les coûts d'entraînement des systèmes (réduction des temps de calcul, diminution de la taille des modèles, réduction de la volumétrie des données requises pour l'entraînement) justifient aussi ces travaux. De manière plus générale, tout laisse à penser que la relation Homme-Machine deviendra de plus en plus étroite à l'avenir. La qualité de cette relation, comme toute relation, sera conditionnée par la qualité de la communication entre les différentes parties. Des travaux de ce type contribueront en quelque sorte à étudier les modalités d'échange permettant d'assurer la qualité d'une telle communication.

6. Pour rappel, nous entraînons aujourd'hui des modèles à l'aide de millions de données prémodifiées pour expliciter indirectement ce type d'invariance, e.g. rotation d'images en classification d'images.

3.4 Transfert de connaissances de la Machine vers l'Homme

Le second axe de recherche que nous privilégierons traitera du transfert de connaissances portées par des agents Machine vers des agents humains afin d'être en mesure d'implanter pleinement la collaboration Homme-Machine.

Nous sommes aujourd'hui en mesure d'entraîner des modèles d'apprentissage machine capables d'exprimer des compétences extrêmement riches. Des modèles sont d'ores et déjà suprahumains dans la réalisation de certaines tâches nécessitant un engagement cognitif important, e.g. l'exemple des jeux de plateau de type Échec ou Go est probablement le plus connu [237], plus récemment des résultats semblables ont aussi été obtenus pour des jeux stratégiques nettement plus complexes, e.g. StartCraft 2 [268], Diplomacy [13]. L'amélioration graduelle des modèles est attestée, e.g. augmentation des capacités de généralisation, des capacités de transferts entre domaines, réduction du nombre de données nécessaires pour les phases d'entraînement [200]. L'Intelligence Artificielle Générale (AGI) est même promise par certains dans peu de temps, même si le sujet est source de désaccord. Il n'est cependant pas illusoire que dans un avenir proche (moins de 50 ans) des modèles à base d'IA supplantent l'Homme dans beaucoup de domaines et pour de nombreuses tâches aujourd'hui inaccessibles à la Machine, y compris dans des domaines qui demandent une action physique dans notre environnement. Il est ainsi estimé que la révolution amenée par l'avènement de ce type de modèles aura un impact pour le moins comparable à celui des révolutions agricole et industrielle [39].

Substituer l'Homme pour la réalisation de tâches pénibles, dangereuses, non critiques et peu stimulantes ne posera de problème à personne. Qu'en sera-t-il des tâches critiques ? Nos sociétés n'accepteront très probablement pas que la prise de décision soit complètement déléguée dans tous les contextes, que nos meilleurs médecins, dirigeants, PDG, chercheurs, artistes, ou encore généraux soient des Intelligences Artificielles (cf. [197]). Aurons-nous cependant d'autre choix que de céder à l'injonction technologique lorsque la démonstration de la supériorité de la Machine sera implacable, lorsque son évidence économique sera patente et que l'IA sera la norme ? La question est abyssale.

Le développement de systèmes suprahumains ne se réduit pas à une entreprise stimulante sans réelle finalité applicative. Il semble évident que des systèmes intelligents ayant accès à un périmètre décisionnel aujourd'hui sacralisé seront déployés dans le futur, et que l'Homme aura à interagir et à collaborer avec eux. Il devra pour cela les aligner à des systèmes de valeurs établis et admis, s'assurer de leur confiance, et trouver sa place sans être réduit à un simple exécutant. Au-delà d'une maîtrise totale du contrôle de ces technologies, et du maintien de notre place d'Homme, nous espérons que peu à peu une interaction enrichissante se développera avec ce type d'agents.⁷

Beaucoup de problèmes méritent d'être traités avant qu'un mariage étroit des intelligences humaines et artificielles s'opère dans de nombreux pans de nos sociétés. Au-delà de la discussion importante sur le caractère souhaitable d'un tel mariage, les notions de confiance et d'interaction avec l'opérateur machine seront centrales. Dans ce contexte, les travaux sur l'interprétabilité et l'explicabilité des systèmes à base d'IA nous semblent essentiels. Plus largement, comprendre une IA au sens de la relation apprise entre les données qu'elle traite et les résultats qu'elle produit, ou encore au sens du

7. Après avoir rapetissé le monde, et accéléré le temps ([140]), il n'est pas à exclure que la technologie rende finalement l'Homme plus bête, i.e. moins intelligent par rapport à un référentiel qui ne sera plus fixé par le génie humain, mais plutôt par quelque chose capable d'exprimer une intelligence suprahumaine, quelque chose d'Intelligent certes, mais aussi d'artificiel, de désincarné.

modèle du problème qu'elle considère de manière sous-jacente, nous semble tout à fait critique. Sur ce point précis, des travaux ont par ailleurs montré que l'utilisation de technologies à base d'IA peuvent amener des comportements humains de type imitation [266], i.e. le modèle décisionnel appris par la Machine est en quelque sorte transféré vers l'Homme, ce qui nous encourage à étudier ce qui est réellement encodé dans ces modèles d'apprentissage en termes de règles décisionnelles. Dans ce contexte, les travaux que nous avons initiés sur l'interprétabilité locale dans des contextes de méthodes d'attribution méritent d'être étendus, en particulier en ce qui concerne la sémantique d'une interprétation / explication, et l'évaluation des résultats produits. Des travaux sur l'interprétabilité globale des réseaux de neurones méritent aussi d'être encouragés. De manière plus générale, les méthodes d'interprétabilité prenant en compte des connaissances *a priori* - e.g. sur les modèles, les données, ou les attentes utilisateurs (e.g. [138]) - lors des traitements de génération d'explications nous semblent importants, et possibles à inscrire dans la continuité des travaux que nous avons d'ores et déjà réalisés.

Les travaux sur le transfert de connaissances de la Machine vers l'Homme serviront naturellement les recherches sur le problème de l'alignement, à savoir l'étude de mécanismes permettant de garantir que les objectifs de systèmes à base d'IA répondent aux attentes et valeurs de l'Homme [200]. Nous aimons cependant à considérer cette problématique d'explicabilité et de transfert de connaissances de manière dissociée d'une potentielle perte de contrôle d'agents suprahumains susceptibles, aux yeux de certains, de nous avilir. Le risque n'est bien entendu pas négligeable et mérite d'être traité avec le plus grand sérieux.⁸ L'idée qui consisterait à s'enrichir de ces agents intelligents nous semble aussi cependant très intéressante. Dans ce contexte, au-delà de travaux portant sur l'interprétabilité des décisions produites par les systèmes, nous souhaiterions étudier des techniques visant à distinguer des stratégies apprises par des IA en vue de permettre un transfert de connaissances de la Machine vers l'Homme.

Des travaux ont d'ores et déjà fait la démonstration que des approches profondes sont capables d'étendre notre corpus de connaissances, en distinguant par exemple des algorithmes optimisés et inconnus auparavant, spécialisés pour la résolution de tâches précises, e.g. AlphaTensor a permis d'identifier des algorithmes de multiplication de matrices inconnus qui supplantent les approches existantes, fruits de dizaines d'années de recherche [89]. Plusieurs initiatives semblables sont aujourd'hui étudiées dans de nombreux domaines, e.g. Physique, Mathématiques, Pharmaceutique. Tout laisse à penser qu'à une approche de découverte de connaissances gouvernée par les données, i.e. approche *data-driven* dans la mouvance du *Big Data*, s'ajoute peu à peu une approche de découverte de connaissances basée sur l'étude de modèles entraînés à partir de ces données (Dira-t-on approche *AI-driven* dans le futur?).

L'objectif à court terme de notre programme de travail sur ce volet porte sur l'étude d'agents entraînés dans des environnements virtuels et dans des contextes de prise de décision séquentielle. Ces modèles seront entraînés à l'aide de techniques d'apprentissage par renforcement.⁹ Nos travaux porteront sur la définition de procédures d'analyse des agents entraînés dans des contextes spécifiques. Pour cela, nous souhaiterions caractériser, en des termes intelligibles par l'opérateur

8. L'auteur fait partie des sceptiques quant à notre capacité à contenir collectivement une technologie telle que l'IA.

9. avec éventuellement des phases d'apprentissage supervisé pour initier l'apprentissage, e.g. *imitation learning*, et des approches de RLHF - *Reinforcement Learning from Human Feedback* - impliquant des opérateurs humains pour traiter l'alignement comme fréquemment considéré dans les modèles contemporains.

humain, des comportements Machine par l'analyse des trajectoires produites par le système dans des contextes d'étude choisis. Les travaux que nous avons réalisés sur l'extraction de règles et sur la production de résumés conceptuels pourraient par ailleurs servir aux traitements nécessaires de synthèse d'information lors de l'analyse des trajectoires.

Les perspectives potentielles offertes par ce type de travaux sont légion, en particulier dans le domaine sportif. A titre d'exemple, des travaux de l'état de l'art ont porté sur l'apprentissage de stratégies de saut en hauteur dans des environnements virtuels via de l'apprentissage par renforcement [283]. Les perspectives envisageables pour l'étude de sports collectifs sont aussi nombreuses, notamment pour la recherche de stratégies de jeu, e.g. étude de conformation de jeux ou de positionnements stratégiques à adopter [53, 264]. L'environnement dont nous bénéficions à Euromov DHM nous semble tout à fait propice au développement de ce type de projets.

Nous souhaitons terminer ce manuscrit en soulignant que les travaux et le projet de recherche exposés dans ce manuscrit sont le fruit de riches collaborations locales, nationales et internationales, à la fois académiques et industrielles. Leurs limites découlent de celles de l'auteur ; les éventuelles qualités que pourrait leur accorder le lecteur seraient à attribuer à nos brillants collaborateurs, experts scientifiques de domaines disciplinaires ou externes de l'Informatique, ou étudiants-chercheurs que l'auteur à eu la chance d'encadrer.

Annexes



Sélection d'articles

- Sébastien Harispe, Massissilia Medjkoune et Jacky Montmain. Summarizing conceptual descriptions using knowledge representations. 2016 IEEE Symposium Series on Computational Intelligence (SSCI). Déc. 2016. [↗](#) (Article conférence)
- Valentina Beretta, Sébastien Harispe, Sylvie Ranwez et Isabelle Mougenot. Truth selection for truth discovery models exploiting ordering relationship among values. Knowledge-Based Systems 159 (nov. 2018). [↗](#) (Article journal)
- Valentina Beretta, Sébastien Harispe, Sylvie Ranwez et Isabelle Mougenot. Combining Truth Discovery and RDF Knowledge Bases to Their Mutual Advantage. The Semantic Web – ISWC 2018. The Semantic Web – ISWC 2018, Information Systems and Applications, incl. Internet/Web, and HCI series. Monterey, Californie, United States, oct. 2018. [↗](#) (Article conférence)
- Cécile L'Héritier, Sébastien Harispe, Abdelhak Imoussaten, Gilles Dusserre et Benoit Roig. Selecting Relevant Association Rules From Imperfect Data. 13th international conference on Scalable Uncertainty Management (SUM 2019). Compiègne, France, déc. 2019. [↗](#) (Article conférence)
- Charles Condevaux, Sébastien Harispe et Stéphane Mussard. Fair and Efficient Alternatives to Shapley-based Attribution Methods. ECMLPKDD 2022 - The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Grenoble, France, sept. 2022. [↗](#) (Article conférence)
- Ali Yaddaden, Sébastien Harispe et Michel Vasquez. Neural Order-First Split-Second Algorithm for the Capacitated Vehicle Routing Problem. OLA 2022 - International Conference on Optimization and Learning. Communications in Computer and Information Science. Syracuse, Italy : Springer International Publishing, juill. 2022. [↗](#) (Article conférence)
- Charles Condevaux et Sébastien Harispe. LSG Attention : Extrapolation of pretrained Transformers to long sequences. PAKDD 2023, The 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Osaka, Japan, mai 2023. [↗](#) (Article conférence)

Summarizing Conceptual Descriptions using Knowledge Representations

Sébastien Harispe*, Jacky Montmain, Massissilia Medjkoune
 École des mines d'Alès – LGI2P/EMA Research Centre,
 Parc scientifique G. Besse, 30035 Nîmes, France
 Email: {firstname.name}@mines.ales.fr

Abstract—Summarizing a body of information is a complex task which mainly depends on the ability to distinguish important information and to condense notions through abstraction. Considering a knowledge representation partially ordering concepts into a directed acyclic graph, this study focuses on the problem of summarizing several human descriptions expressed through sets of concepts. We formally define the problem of summarization in this context and we propose a model mimicking a Human-like Intelligence for scoring alternatives with regard to a specific objective. Several interesting theoretical results related to this problem (e.g. for optimization) are also given. Finally, the evaluation of the proposed approach performed in the domain of odor analysis highlights the benefits of our proposal and shows how it could be used to automatize time-consuming expert summarizing processes. Source code implementing the proposed approach as well as datasets are made available to the community.¹

I. INTRODUCTION

The ability of abstracting specific observations by means of distinguishing general patterns or concepts is central for the emergence of complex conceptual processes, e.g. learning. It is indeed one of the essential abilities defining intelligent agents that are able to analyze potentially complex and faintly related situations to acquire knowledge from these analyses [1], [2]. Abstraction is tightly linked to the process of analyzing observations through the lens of structured categories defined w.r.t. specific properties of the observations. This ability human have explains for instance that any child burning himself with a wood-fired oven will most often extract knowledge from this experiment and learn that the wood is not the cause of the unfortunate experiment, but rather the heat caused by its combustion. Therefore, thanks to this capacity of abstraction, children will identify similar situations (e.g. a lighter flame) as potentially harmful and take extra care while being confronted to them. Similarly, if a person tells you that a specific candy smells `Lemon`, and another person that the same candy smells `Orange`, you'll naturally and easily be able to abstract these two descriptions by summarizing the provided information to `Citrus Fruit`. Studying processes related to abstraction is therefore of major importance for Artificial Intelligence and has been central to numerous areas of research related to this domain - machine learning among others [2]. Interestingly, a growing number of knowledge representations (e.g. ontologies

or even taxonomies) today formally express specific domain-knowledge, and provide concept organizations that can be automatically analyzed. These knowledge representations, based on the notion of abstraction, therefore offer the opportunity to design Human-like intelligent procedure taking advantage of prior knowledge regarding concept organizations [3]. Mimicking Human-like intelligence able to process such knowledge representations to abstract knowledge in a meaningful way is however still a complex, important and open challenge.

In this study we focus on the problem of summarizing a set of conceptual descriptions expressed in the form of sets of concepts. The notion of concept here refers to the traditional notion of *concept* or *class* used in Knowledge Representation; concepts are assumed to be partially ordered with regard to specific properties they share – knowledge representations are domain-specific. Figure 1 shows an example of concept ordering for odor evaluation. It specifies for instance that the concept `Rose` refers to the concept `Floral`. Thanks to the ordering of concepts, intuitively, considering the two following sets of annotations corresponding to odor sample descriptions (`{Rose, Mint}`, `{Violet, Orange}`) several summaries could be proposed, e.g. `{Floral, Fresh}`, `{F-Class}`.

Even if summarizing is an intuitive process in which an evaluator wants to abstract as much information as possible while keeping the resulting abstraction meaningful and informative, the relevance of a summary is context-dependent and no formal consensual definition of *best* summary can be given - even if general constraints on the definition of a summary could be defined, as we will see in Section II. Indeed, considering aforementioned example, summarizing the conceptual descriptions by one of both summaries could make sense depending on how much we want the summary to be concise and on how much information loss we accept.

The paper is organized as follows: Section II introduces theoretical notions on which is based our approach as well as notations; it next formally defines the summarizing problem; Section III introduces the proposed model for evaluating summaries; Section III introduces algorithms enabling to use the model for searching for relevant summaries and discusses interesting properties of the search space; Section V presents results related to evaluation; State-of-the-art is proposed in Section VI. Finally, section VII summarizes our results and distinguishes perspectives.

¹Source code and datasets are publicly available at <https://github.com/sharispe/Conceptual-Summary>.

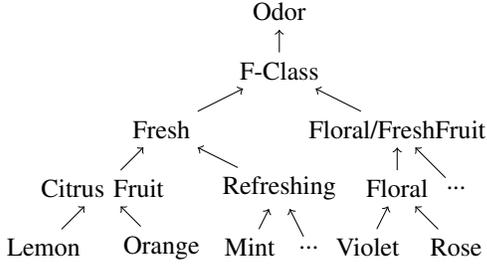


Fig. 1. Example of partial ordering of concepts related to odor evaluation

II. PROBLEM SETTING: FORMALIZATION

Prior to discuss the technical aspects of our contribution and to formally define the notion of summary, we introduce some notations and notions on which is based our approach.

A. Notations and preliminary notions

In this study we consider an *a priori* domain knowledge on the form of a set of concepts C partially ordered into a poset $O = (\preceq, C)$. $\forall(x, y) \in C^2, x \prec y$ means that x implies y , e.g. in the context of odor evaluation it means that smelling Rose implies smelling a Floral compound – cf. to Figure 1.

To further ease the introduction of formal definitions, we define the following functions $\mathcal{D} : \mathcal{P}(C) \rightarrow \mathcal{P}(C)$, $\mathcal{A} : \mathcal{P}(C) \rightarrow \mathcal{P}(C)$, with $C' \subseteq C$:

$$\mathcal{D}(C') = \bigcup_{c \in C'} \bigcup_{c' \preceq c} \{c'\} \quad \mathcal{A}(C') = \bigcup_{c \in C'} \bigcup_{c \preceq c'} \{c'\}$$

The commonly used notion of *Information Content* (IC) of concepts refers to concepts' degree of specificity; several models have been proposed. Some of them are based on the analysis of topological properties of the partial order (intrinsic models), sometimes mixed with additional prior knowledge w.r.t. concept usage (extrinsic models) [4]. For any IC function we have: $\forall(x, y) \in C^2, x \prec y \implies IC(x) > IC(y)$. An example of simple intrinsic IC formulation is [5]:

$$IC(c) = 1 - \frac{\log(|\{x|x \preceq c\}|)}{|C|} \quad (1)$$

The IC function which is used hereafter refers to this expression - any intrinsic IC with $IC : C \rightarrow [0, 1]$ could be used.

When occurrences over concepts are available, we consider the *mass function* $m : C \rightarrow [0, 1]$, satisfying $\sum_{c \in C} m(c) = 1$; $m(c)$ corresponds to the number of observations of c among the total number of observations. The belief and plausibility functions $bel : C \rightarrow [0, 1]$ and $pl : C \rightarrow [0, 1]$ proposed in the Dempster-Shafer theory are next defined such as:²

$$bel(c) = \sum_{x \preceq c} m(x) \quad (2)$$

$$pl(c) = \sum_{x \in C, \mathcal{D}(x) \cap \mathcal{D}(c) \neq \emptyset} m(x) \quad (3)$$

²We adopt the classical formalism used for defining the *mass*, *belief* and *plausibility* functions w.r.t knowledge representations. Note however that these definitions are not rigorous and are only used to lighten the formalism; they should be understood as the rigorous definitions proposed in [6].

B. Problem Formalization

The aim of the study is to summarize the information given by a set of evaluators (E) providing conceptual evaluations, such as each evaluator $e_n \in E$ is associated to a set of concepts $X_n \in \mathcal{P}(C)$. For convenience, we will always denote the sequence of annotations to summarize $\hat{X} = (X_1, X_2, \dots, X_n)$, and X the set of concepts mentioned in the sequence of annotations \hat{X} , such as:

$$X = \bigcup_{i=1}^n X_i$$

The simplest type of summary function f can be defined such as $f : \hat{X} \rightarrow \mathcal{P}(C)$. Nevertheless, in accordance to the process of summarization in most contexts of use, we want any summary $Y \in \mathcal{P}(C)$ of a set of concepts X to respect specific properties:

- 1) *summarizing* $|Y| \leq |X|$, with most often $|Y| \ll |X|$
- 2) *faithful* $\forall y \in Y, \exists x \in X$ such as $x \preceq y$
- 3) *non-total redundancy* $\forall(x, y) \in Y^2, x \not\prec y \wedge y \not\prec x$.

We denote $\mathcal{S} \subseteq \mathcal{P}(C)$ the subsets of C respecting the *non-total redundancy* property. Furthermore we consider $\mathcal{S}^X \subseteq \mathcal{S}$ the set of summaries of a sequence of annotations \hat{X} – each summary respects the properties of being *summarizing* and *faithful* w.r.t. X . Based on these preliminary definitions we formally define by S the function summarizing a sequence of n annotations $\hat{X} \in \mathcal{P}(C)^n$ by a single summary from \mathcal{S}^X :

$$S : \mathcal{P}(C)^n \rightarrow \mathcal{S}, \text{ with } S(\hat{X}) \in \mathcal{S}^X$$

We define the problem of summarizing a sequence of annotations \hat{X} by finding $Y \in \mathcal{S}^X$, the *best* summary for \hat{X} . The following section introduces the model we use for defining function S . Additional properties of interest for characterizing the solution space \mathcal{S}^X are given first.

A summary $Y \in \mathcal{S}^X$ is said to be covering a sequence of annotations \hat{X} if all the concepts mentioned in \hat{X} (i.e. all concepts of X) are abstracted by at least a concept of the summary: $\forall x \in X, \exists y \in Y$ such as $x \preceq y$. We denote the covering summaries of \hat{X} by \mathcal{S}_{cov}^X . We also consider $\mathcal{S}_{p-cov}^X = \mathcal{S}^X \setminus \mathcal{S}_{cov}^X$ the set of summary *partially covering* X , i.e. $\forall Y \in \mathcal{S}_{p-cov}^X, \exists X' \subset X$ such as $Y \in \mathcal{S}_{cov}^{X'}$. Figure 2 illustrates the studied space exposed so far.

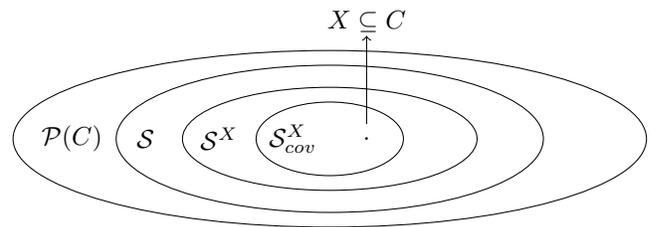


Fig. 2. Graphical representation of the studied space

As we mentioned, the notion of *best* summary is context-dependent. We can however distinguish some notions and quantities that could be used to evaluate the relevance of a summary and define the function S .

III. AUTOMATIC SUMMARY: PROPOSAL

For studying the problem of finding a summary $Y \in \mathcal{S}^X$ for any $\hat{X} \in \mathcal{P}(C)^n$ we define the following objective function:

$$S(\hat{X}) = \arg \max_{Y \in \mathcal{S}^X} \left(\Psi(Y, \hat{X}) - \mathcal{L}(Y, \hat{X}) \right) \quad (4)$$

$$\mathcal{L}(Y, \hat{X}) = \Delta(Y, \hat{X}) + \lambda(Y) + \gamma(Y, \hat{X})$$

$\Psi(Y, \hat{X})$ models the amount of information from \hat{X} covered by Y and $\mathcal{L}(Y, \hat{X})$ the penalty associated to the abstraction, with:

- $\Delta(Y, \hat{X})$ the penalty induced by abstracting by Y the information conveyed by \hat{X} – incorporating penalties regarding *loss*, *addition* and *distortion* of information.
- $\lambda(Y)$ a function evaluating the conciseness of the summary - w.r.t redundancy of information.
- $\gamma(Y, \hat{X})$ a function that can be used to express additional constraints over Y , e.g. to exclude uncovering summaries.

The specific components of the objective function (eq. 4) are detailed hereafter.

 A. $\Psi(Y, \hat{X})$ - Amount of abstracted information

We denote $\Psi(Y, \hat{X})$ the amount of information from \hat{X} covered by Y . It is used to estimate the amount of conceptual information conveyed by \hat{X} which is summarized by Y . We are therefore interested in studying the coverage of abstract notions among those mentioned by \hat{X} and those mentioned by Y . Intuitively this quantity could be defined as follows:

$$\Psi(Y, \hat{X}) = f(\mathcal{A}(Y) \cap \mathcal{A}(X))$$

A specific expression could for instance be:

$$\Psi(Y, \hat{X}) = \sum_{c \in \mathcal{A}(Y) \cap \mathcal{A}(X)} w(c) \times IC(c)$$

with

$$w(c) = \sum_{c' \in X \cap \mathcal{D}(\{c\})} m(c')$$

with $w : C \rightarrow [0, 1]$ the function used to weigh the importance given to any concept w.r.t. to its degree of evocation according to the masses specified by \hat{X} – recall that the elements of X are the focal elements, i.e. the concepts that have non null masses. In this case it corresponds to defining $w(c) = bel(c)$ with bel the classical belief function (Eq. 2) – this is the case since the focal elements are the members of X .³ Since the more abstract a concept will be, the more it will be implicitly mentioned, and therefore easily summarized whatever the selected summary is, we regulate the importance given to any concept $c \in \mathcal{A}(Y) \cap \mathcal{A}(X)$ w.r.t its intrinsic information content.

³Note that we don't want the weighting function w to behave like an extrinsic information content, e.g. Resnik like [4]; since we don't want to give high importance to concepts that are marginal w.r.t \hat{X} but very specific w.r.t any intrinsic IC formula.

Let's consider the following simple example in which $\hat{X} = (\{\text{Violet}, \text{Rose}\})$. In this case, considering the summaries $\{\text{Rose}\}$ (or $\{\text{Violet}\}$) and $\{\text{Floral}\}$ we would observe: $\Psi(\{\text{Floral}\}, \hat{X}) < \Psi(\{\text{Violet}\}, \hat{X})$. Otherwise stated the current model will consider that, based on the quantity of information from \hat{X} which is abstracted, summarising \hat{X} by $\{\text{Rose}\}$ is more relevant than using $\{\text{Floral}\}$. Putting aside the subjective notion of *best* summary, and avoiding arguing on which is the *best* summary, we want to penalize the fact that compared to $\{\text{Floral}\}$, $\{\text{Rose}\}$ only partially covers \hat{X} . In the current formulation we have $\Psi(\{\text{Rose}\}, \hat{X}) = \Psi(\{\text{Floral}\}, \hat{X}) + bel(\text{Rose}) \times IC(\text{Rose})$. However, it can be counter-intuitive to consider that $\Psi(\{\text{Floral}\}, \hat{X})$ will contribute similarly to estimating the quantity of information from \hat{X} abstracted by both $\{\text{Floral}\}$ and $\{\text{Rose}\}$. Otherwise stated, we would like to lower the contribution of $\Psi(\{\text{Floral}\}, \hat{X})$ while computing $\Psi(\{\text{Rose}\}, \hat{X})$ considering that not all the masses coming from specializations of Floral are subsumed by the summary $\{\text{Rose}\}$. Indeed reducing \hat{X} to Rose suppresses the given information that \hat{X} also mentioned Violet. This idea is related to the notion of *distortion* of information we will discuss later while defining the penalties associated to a summary. Nevertheless, we want to integrate this notion while estimating the quantity of information from \hat{X} which is abstracted by a summary. To this end we propose to reconsider the definition of the weighting function as follows:

$$bel_Y(c \in C) = \sum_{c' \in X \cap \mathcal{D}(\{c\}) \cap \mathcal{D}(Y)} m(c')$$

We finally obtain:

$$\Psi(Y, \hat{X}) = \sum_{c \in \mathcal{A}(Y) \cap \mathcal{A}(X)} bel_Y(c) \times IC(c) \quad (5)$$

Note that $\mathcal{D}(X) \cap \mathcal{D}(Y)$ also conveys potentially useful information since $\mathcal{D}(X) \setminus \{X\}$ refers to the information not conveyed by X that are plausible; considering the partial ordering of concepts defined in Figure 1 someone referring to Floral could refer to Violet without being able to refer to this specific odor. It could therefore be interesting to study this quantity as it could be used to characterize the quantity of plausible information captured by the summary. Interesting properties could be achieved analysing this quantity since e.g. $\forall (Y, Y') \in \mathcal{S}_{cov}^X \times \mathcal{S}_{p-cov}^X, \mathcal{D}(Y') \cap \mathcal{D}(X) \subset \mathcal{D}(Y) \cap \mathcal{D}(X)$ could be used to favour covering summaries. Even if this quantity would not be useful for analysing covering summaries, i.e. *per* definition $\forall Y \in \mathcal{S}_{cov}^X, \mathcal{D}(X) \subseteq \mathcal{D}(Y)$, it could be worth considering it for discussing partially covering summaries. Nevertheless, since we consider the rational assumption that analysing exact information is more important than analysing plausible information while criticizing a summary, integrating this quantity in the definition of Ψ is not furthered explored in this paper; a refinement of the proposed approach could be $\Psi(Y, \hat{X}) = f(\mathcal{A}(Y) \cap \mathcal{A}(X), \mathcal{D}(Y) \cap \mathcal{D}(X))$.

$\Psi(Y, \hat{X})$ models the information conveyed by \hat{X} which is conveyed by Y by considering covered masses. We now

introduce how we model the various components of the penalty factor $\mathcal{L}(Y, \hat{X}) = \Delta(Y, \hat{X}) + \lambda(Y) + \gamma(Y, \hat{X})$ (Eq. 4).

B. $\Delta(Y, \hat{X})$ – Penalty of abstraction

In the previous section we have defined a model for estimating the quantity of exact information conveyed by \hat{X} which is conveyed by a summary. For criticizing the relevance of a summary it is also important to discuss penalties regarding *loss*, *addition* and *distortion* of information. We define the penalty of abstraction by:

$$\Delta(Y, \hat{X}) = f(\Delta^{E-}, \Delta^{P+}, \Delta^{P-}, \Delta^D)$$

with

- Δ^{E-} penalty w.r.t to the deletion of exact info
- Δ^{P+} penalty w.r.t to the addition of plausible info
- Δ^{P-} penalty w.r.t to the deletion of plausible info
- Δ^D penalty w.r.t to distortion of information

We define those functions such as for each Δ function we have $\Delta(Y, \hat{X}) \in \mathbb{R}^+$.

Modelling Δ^{E-} and Δ^{P+} , Δ^{P-} :

Δ^{E-} models the amount of exact information conveyed by \hat{X} which is not conveyed by Y – deletion of exact information:

$$\Delta^{E-}(Y, \hat{X}) = f(\mathcal{A}(X) \setminus \mathcal{A}(Y))$$

Δ^{P+} models the amount of plausible information conveyed by Y which is not conveyed by \hat{X} – addition of plausible information; Δ^{P-} models the amount of plausible information conveyed by \hat{X} which is not conveyed by Y – deletion of plausible information:

$$\Delta^{P+}(Y, \hat{X}) = f(\mathcal{D}(Y) \setminus \{\mathcal{D}(X) \cup \mathcal{A}(X)\})$$

$$\Delta^{P-}(Y, \hat{X}) = f(\mathcal{D}(X) \setminus \mathcal{D}(Y))$$

Note that *per* definition, and due to the property of *faithfulness*, a summary cannot (i) add exact information, i.e. provide information which is not conveyed by \hat{X} .⁴ Considering aforementioned operators Δ^{E-} and Δ^{P+} (resp. Δ^{P-}), specific expressions can easily be obtained:

$$\Delta^{E-}(Y, \hat{X}) = \sum_{x \in \mathcal{A}(X) \setminus \mathcal{A}(Y)} (bel(x) \cdot IC(x))$$

$$\Delta^{P+}(Y, \hat{X}) = \sum_{y \in \mathcal{D}(Y) \setminus \{\mathcal{D}(X) \cup \mathcal{A}(X)\}} (pl(y) \cdot IC(y))$$

$$\Delta^{P-}(Y, \hat{X}) = \sum_{y \in \mathcal{D}(X) \setminus \mathcal{D}(Y)} (pl(y) \cdot IC(y))$$

⁴ Due to the definition of a summary we have no addition of exact information $\Delta^{E+}(Y, \hat{X}) = f(\mathcal{A}(Y) \setminus \mathcal{A}(X) = \emptyset)$. It could however be interesting to consider this quantity in specific contexts of use in which the definition of a summary would be less constraining than the one considered.

Modelling Δ^D Penalty - distortion:⁵

The aim of Δ^D is to penalize the distortion which is made considering a specific choice among partially-covering summaries. Considering the previously mentioned simple example in which $\hat{X} = (\{\text{Violet}, \text{Rose}\})$ with the summaries $\{\text{Rose}\}$ (or $\{\text{Violet}\}$) and $\{\text{Floral}\}$ we would have $\Delta^{E-}(\{\text{Rose}\}, \hat{X}) < \Delta^{E-}(\{\text{Floral}\}, \hat{X})$. Even if some adaptation of Ψ have been proposed to penalize the bias induced by the choice of an uncovering summary, an additional penalty has to be modeled for considering additional potential distortion of information that could be made during summarizing. This penalty should be a function of $X \setminus \mathcal{D}(Y)$, i.e., all the elements of X that have not been summarized, and associated masses. We propose the following model to estimate the distortion.

$$\Delta^D(Y, \hat{X}) = \tau \sum_{x \in X \setminus \mathcal{D}(Y)} \sum_{x' \in \mathcal{A}(\{x\}) \setminus \mathcal{A}(Y)} (bel_{\{x\}}(x') \cdot IC(x'))$$

The parameter τ is used to weigh the importance of a specific uncovering and will be introduced later. For each uncovered concept in $x \in X \setminus \mathcal{D}(Y)$, the penalty associated to it is a function of its specificity and the specificity of the concepts abstracting x that are not covered by Y . We however want this penalty to be a function of the amount of masses associated to the concepts mentioned by \hat{X} that are not covered. We want to penalize any distortion that is not motivated by very low masses associated to part that are excluded by the summary. To model this choice, and considering $\beta_{YX} \in [0, 1]$ the amount of masses relative to \hat{X} that are not covered by Y , we use any f function ensuring $f(\beta_{YX}) = 1$ and $f(\beta_{YX}) = \infty$ or any very large value when $\beta_{YX} \rightarrow 1$ and $\beta_{YX} \rightarrow 0$ respectively. Here we adopt the following function: $\tau = -\ln(1 - \beta_{YX}^\alpha)$, with $\alpha \in \mathbb{N}^*$:

$$\beta_{YX} = \frac{\sum_{x \in X \setminus \mathcal{D}(Y)} m(x)}{\sum_{x \in X} m(x)} = \sum_{x \in X \setminus \mathcal{D}(Y)} m(x)$$

Note that we have $\beta_{YX} \in [0, 1]$ for any $Y \in \mathcal{S} \setminus \mathcal{S}^X$, $\beta_{YX} = 0$ for any $Y \in \mathcal{S}_{cov}^X$, as well as $\beta_{YX} \in]0, 1[$ for any $Y \in \mathcal{S}^X \setminus \mathcal{S}_{cov}^X$. The more the distortion will be, the more β_{YX} tends to 1, and the more τ will induce an important penalty. The tuning parameter α is used to define the penalization ratio according to the loss of masses we accept – the lower α is, the more the model will penalize summaries implying mass losses.

We finally simply define:

$$\Delta(Y, \hat{X}) = \delta_{E-} \Delta^{E-} + \delta_{P+} \Delta^{P+} + \delta_{P-} \Delta^{P-} + \delta_D \Delta^D \quad (6)$$

With $\delta_{E-}, \delta_{P+}, \delta_{P-}, \delta_D$ input parameters used to set the importance of each abstraction penalty factor.

⁵Minor corrections have been made in this section (modifying the originally published version) - details are provided in the erratum section.

C. Additional penalties to improve summarizing

Modelling λ - Conciseness and redundancies penalties:

$\lambda(Y) \in \mathbb{R}^+$ is a penalty used to evaluate the conciseness of the summary - w.r.t the number of descriptors, by penalizing redundant information implicitly conveyed by a summary.

$$\lambda(Y) = \epsilon \sum_{y' \in \mathcal{A}(Y)} (|\{y \in Y | y' \in \mathcal{A}(y)\}| - 1) \times IC(y')$$

Using this expression, by avoiding large redundancies we favor abstraction and therefore conciseness, i.e. summaries that do not summarize enough the information carried by \hat{X} will automatically be penalized. The penalization is designed such as each abstracted notions that are repeated more than once will be penalized the number of time the redundant information appears – taking into account of the intrinsic information content of concepts since redundancy cannot be avoided in most cases, and redundancies of very abstract concepts are of minor concern. Tuning ϵ can therefore be used to control the number of descriptors composing a summary.

γ - Additional constraints

$\gamma(Y, \hat{X})$ is a function that can be used to express additional constraints over Y . This constraints can be used to apply specific restrictions on the type of solution we are interested in, e.g. in particular if we relax the definition of a summary:

$$\gamma(Y, \hat{X}) = \begin{cases} 0 & \text{if } \text{valid}(Y, X) \\ +\infty & \text{otherwise} \end{cases}$$

with $\text{valid} : \mathcal{P}(C) \times \mathcal{P}(C) \rightarrow \{\text{true}, \text{false}\}$. As an example, exploring the covering summaries corresponds to defining the following *valid* function:

$$\text{valid}(Y, X) = \forall x \in X, \exists y \in Y \text{ such as } x \preceq y$$

Additional/Other constraints can naturally be defined, for instance, on the size of the summary we would like to generate or on the degree of specificity of the concepts it contains, i.e. by avoiding too abstract concepts.

Note that defining the γ function corresponds to reduce the solution space without explicitly defining a specially designed solution search algorithm. Indeed, by defining γ , restrictions on the solutions can be expressed while using the general search algorithm introduced in the following section.

The model proposed so far (Eq. 4) can be used to evaluate the relevance of a summary and rank several summary alternatives. By providing a summary $Y \in \mathcal{S}^X$ summarizing a sequence of annotation \hat{X} , knowing the importance of each concept of Y , i.e. how many sets of \hat{X} (evaluators) evoke each concept, is of great importance for further data analysis. Therefore, considering the proposed setting, we can define the weight of any concept $y \in Y$ as a function of its belief. The next sections discusses details related to search space construction and analysis.

IV. SUMMARY GENERATION

In the previous section, we have introduced a framework for searching for relevant summaries considering \hat{X} a sequence of conceptual annotations. This section now discusses algorithmic implications and discusses elements of information to practically define a strategy for exploring the search space \mathcal{S}^X . First let's recall some information related to the search space; the number of partitions of C is $2^{|C|}$ and in the worst case, considering that any knowledge representation always has an abstract concept generalizing all the others (a *root*), we have the following theoretical bound $\forall X \subset C, |\mathcal{S}^X| \leq 2^{|C-1|} + 1$.⁶ However in practice, the theoretical bound is always far from the size of real search spaces thanks to the constraints defining a summary, in particular the *faithfulness* and the *non-total redundancy*. The size of \mathcal{S}^X is nevertheless to be taken with high consideration since it largely impacts the computational time of the approach. Indeed, for any sequence of annotations \hat{X} mentioning a large number of concepts (big $|X|$) applying a naive iterative search over \mathcal{S}^X is not feasible. In this section we first propose an algorithm to construct \mathcal{S}^X . We then propose restrictions that can be applied on \mathcal{S}^X to deal with sequence of annotations mentioning a large number of concepts.

A. Building \mathcal{S}^X

Considering a set of concepts $X \subset C$, Algorithm 1 defines how to generate \mathcal{S}^X , the set of summaries for \hat{X} . First recall that $\mathcal{S}^X = \mathcal{S}_{cov}^X \cup \mathcal{S}_{p-cov}^X$. Note also that:

$$\mathcal{S}_{p-cov}^X = \bigcup_{X' \in \mathcal{P}(X) \setminus X} \mathcal{S}_{cov}^{X'}$$

This enables to reformulate the problem of building \mathcal{S}^X as finding the set $\mathcal{S}^X = \bigcup_{X' \in \mathcal{P}(X)} \mathcal{S}_{cov}^{X'}$. Given a sequence of annotations \hat{X} , Algorithm 1 computes \mathcal{S}^X . The algorithm uses a classical directed acyclic graph representation of $O = (\preceq, C)$, named $G = (E, C)$ with $E \subset C \times C$ and $(c_1, c_2) \in E$ means c_2 generalizes c_1 , i.e. which corresponds to $c_1 \prec c_2$. For optimization reasons, G is expected to be reduced according to the transitivity of the relationships defining O (line 1) – otherwise stated there is no relationship in G that can be inferred according to the transitive relationships composing G . Thus for each subset $X' \in \mathcal{P}(X)$ the set \mathcal{S}^X used to compute \mathcal{S}^X is extended by adding the summaries covering X' that have not been found (line 4). Conceptually, this steps corresponds to $\mathcal{S}^X = \mathcal{S}^X \cup \mathcal{S}_{cov}^{X'}$, except that we don't want to cover any subset of $\mathcal{S}_{cov}^{X'}$ that have already been covered. Finally the algorithm returns \mathcal{S}^X as \mathcal{S}^X (line 6).

Given a set of concepts X , and a set of summaries $S_r \subset \mathcal{S}$ respecting the property that $\forall Y \in S_r, \mathcal{S}_{cov}^Y \subset S_r$, Algorithm 2 defines how to compute the subset of \mathcal{S}_{cov}^X that have not been covered considering S_r . The constraint on S_r enables to use a greedy algorithm while computing \mathcal{S}_{cov}^X . Note that for a set

⁶The bound is reached when $O = (\preceq, C)$ is weakly structured and all the pairs composed of the other concepts except the *root* only refer to non-ordered concepts – with $|C| = 100$ it means that the worst case is $|\mathcal{S}^X| = 6.3 \times 10^{29}$.

Algorithm 1 Generate summaries $\mathcal{S}^X \subseteq \mathcal{S}$ for $X \subseteq C$

```

1: transitive reduction  $G$  – graph of  $O = (\preceq, C)$ 
2:  $S^X \leftarrow \emptyset$ 
3: for  $X'$  in  $\mathcal{P}(X)$  do
4:    $S^X = S^X \cup \text{extend}_{cov}(G, X', S^X)$ 
5: end for
6: return  $S^X$  as  $S^X$ 
    
```

of concepts X , defining $S_r = \emptyset$, Algorithm 2 computes \mathcal{S}_{cov}^X , i.e. $\mathcal{S}_{cov}^X = \text{extend}_{cov}(G, X, \emptyset)$. The idea of the algorithm is simple; given a set of concepts X , it will recursively compute the covering summaries for each summary of X that can be obtained by substituting an element $x \in X$ by one of its parents – or any subset of its parents. Line 2 iterates over the element of X . Line 3 computes the parents of the current element c – recall that a transitive reduction has been applied on the graph. Next, the aim is to generate a summary of X by substituting c by one of its abstract representation (subset of parents) – line 3. Line 5 builds a summary of X by adding P'_c the selected subset of parents of x and by removing any concept from X that are generalized by a member of P'_c , e.g. x . This ensures to obtain a set of concepts respecting the *non-redundancy property*. The *faithfulness* property is ensured by the fact that any concept that is replaced is only replaced by one of its abstract representation. The *summarizing* property which ensures that we are processing a summary will next be tested in line 6. Before considering any summary we check if it hasn't been already processed (already in s_{cov}^X) and if it's not in the set of restricted summaries (in S_r). If this is not the case we add the current summary into the set of summaries, and, by applying a recursive procedure, we also add all summaries covering it by excluding any summary already encountered (line 7). Considering that the set $\mathcal{A}(c)$ is finite for each concept $c \in C$ and that in line 5, the set X' obtained is necessarily covering X , it ensures that the algorithm terminates.

Algorithm 2 extend_{cov} : Given a set of concepts X , a set of summaries $S_r \subseteq \mathcal{S}$ such as $\forall Y \in S_r, \mathcal{S}_{cov}^Y \subseteq S_r$ it generates $s_{cov}^X = \mathcal{S}_{cov}^X \setminus S_r$ with $s_{cov}^X \subseteq \mathcal{S}_{cov}^X \subseteq \mathcal{S}^X$. By defining $S_r = \emptyset$, the algorithm generates \mathcal{S}_{cov}^X .

```

1:  $s_{cov}^X \leftarrow \emptyset$ 
2: for  $c$  in  $X$  do
3:    $P_c = \{p \mid \exists (c, p) \in E\}$  // parents of  $c$ 
4:   for  $P'_c$  in  $\mathcal{P}(P_c)$  do
5:      $X' = X \setminus \{x \in X \mid P'_c \cap \mathcal{A}(\{x\}) \neq \emptyset\} \cup P'_c$ 
6:     if  $|X'| < |X| \wedge X' \notin S_r \wedge X' \notin s_{cov}^X$  then
7:        $s_{cov}^X = s_{cov}^X \cup \{X'\} \cup \text{extend}_{cov}(G, X', s_{cov}^X \cup S_r)$ 
8:     end if
9:   end for
10: end for
11: return  $s_{cov}^X$ 
    
```

Considering the ordering of concepts introduced in Figure 3, Figure 4 presents the summaries \mathcal{S}^X for $X = \{a, b, c, e\}$ as well as the partial ordering of $\mathcal{S}^X \cup \{X\}$. This ordering is built such as two groups of concepts $(Y, Y') \in \mathcal{S}^X$, are ordered such as $Y \preceq_{S^X} Y'$ if $\mathcal{D}(Y) \subseteq \mathcal{D}(Y')$.

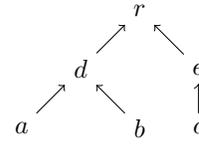


Fig. 3. $O = (\preceq, C)$ example of ordering of concepts

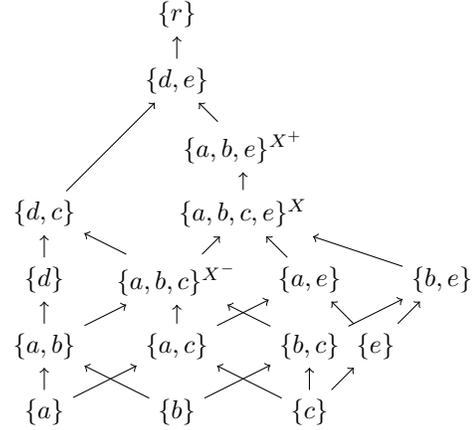


Fig. 4. Considering the set of concepts ordered in Figure 3, ordering of the set of summaries \mathcal{S}^X that could be evaluated to summarize any sequence of annotations \tilde{X} , with $X = \{a, b, c, e\}$, $X^+ = \{a, b, e\}$, $X^- = \{a, b, c\}$; the ordering also includes X .

Indeed, in addition to be the basis for computing \mathcal{S}_{cov}^X and \mathcal{S}^X , note also that Algorithm 2 can be used to compute \preceq_{S^X} the partial order of summaries of any $X \subseteq \mathcal{P}(C)$, i.e. the structure presented in Figure 4 for a specific example. Indeed, in line 5, each X' respecting the summarizing property refers to an ordering $X \preceq_{S^X} X'$.⁷

For convenience, we denote $X^- \subseteq X$ and $X^+ \subseteq X$, the two largest subsets of X respecting the *non-total redundancy* such as $\forall x \in X^-$ (resp. X^+), $\nexists x' \in X, x' \neq x$ such as $x' \prec x$ (resp. $x \prec x'$). Considering Figure 3 and defining $X = \{a, b, c, e\}$ we would have $X^- = \{a, b, c\}$ and $X^+ = \{a, b, e\}$. We also use notations such as $\mathcal{S}^{X'} = \mathcal{S}_{cov}^{X'} \cup \mathcal{S}_{p-cov}^{X'}$, for any subset $X' \subseteq X$, e.g. X^-, X^+ .

Interesting properties helping to better understand \mathcal{S}^X :

Property 1 - If X respects the *non-total redundancy* we have $X^- = X^+ = X$. More particularly we have $X^- \neq X \Leftrightarrow X^+ \neq X$ and $X^- = X \Leftrightarrow X^+ = X$ – this is easily proved using the respect or violation of *non-total redundancy*.

Property 2 - $(Y, Y') \in \mathcal{S}_{cov}^X \times \mathcal{S}_{cov}^Y, Y \preceq_S Y' \implies \mathcal{S}^{Y'} \subseteq \mathcal{S}^Y$. By definition of $\preceq_S, Y \prec_S Y' \implies \mathcal{D}(Y) \subset \mathcal{D}(Y')$. It implies that $\mathcal{A}(Y') \subset \mathcal{A}(Y)$. Knowing that $Y' \in \mathcal{S}$ we have $Y' \in \mathcal{S}^Y$ and therefore $\mathcal{S}^{Y'} \subseteq \mathcal{S}^Y$.

Property 3 - $\mathcal{S}^{X^+} \subseteq \mathcal{S}^{X^-}$ – according to property 2.

Property 4 - $\mathcal{S}_{cov}^{X^+} \subseteq \mathcal{S}_{cov}^{X^-}$ – according to property 2.

Property 5 - $\mathcal{S}_{cov}^X = \mathcal{S}_{cov}^{X^+}$.

Property 6 - $\mathcal{S}_{p-cov}^X = \mathcal{S}^{X^-} \setminus \mathcal{S}_{cov}^{X^+}$.

⁷We denote \preceq_Z the ordering relation over $Z \subseteq \mathcal{P}(C)$ according to the same rule.

These properties are of interest for defining efficient algorithms and heuristics for searching relevant summaries, i.e. distinguishing *best* summaries w.r.t the defined objective function (Eq. 4). Due to space restriction, algorithmic optimization for searching \mathcal{S}^X considering a given set of annotations \hat{X} are not further discussed. Some details regarding the reduction of \mathcal{S}^X that can be applied without introducing additional technical notions are nevertheless introduced in the following subsection.

B. Thought on complexity reduction

Additional properties that can be used to reduce the complexity, e.g. by reducing the set \mathcal{S}^X , are proposed. Computational time reductions can first be obtained while constructing \mathcal{S}^X by applying reduction on X . Two reductions of X are proposed. (1) *Remove redundant concepts from X* . It can easily be proved that, with $X' \preceq_{\mathcal{P}(C)} X$, $\mathcal{S}^X \subseteq \mathcal{S}^{X'}$. X can therefore be substituted by the smaller subset $X' \subseteq X$ such as $\mathcal{A}(X') = \mathcal{A}(X)$ and X' respects the *non-total redundancy property* – this is indeed the more specific covering summary. Thanks to this construction we ensure that $\mathcal{S}^X = \mathcal{S}^{X'}$, with the interesting property $|\mathcal{P}(X')| \leq |\mathcal{P}(X)|$ – with most often in practice $|\mathcal{P}(X')| \ll |\mathcal{P}(X)|$. This approach does not reduce \mathcal{S}^X . (2) *Abstract lower outliers*. Any concept $x \in X$ that has been observed a significantly lower amount of time (e.g. only once) can be substituted by the more specific abstraction of x (element of $\mathcal{A}(x) \setminus \{x\}$) which has the lower mass increasing the one of x . Even if this strategy may reduce \mathcal{S}^X for any cut-off greater than one, considering a cut-off equals to 1 will distinguish all summaries corresponding the abstraction of at least two concepts – these summaries are the ones of interest in most practical applications. This idea can be extended to reduce the number of summaries to evaluate by considering the assumption that any interesting summary should factorize information in order to be meaningful. In accordance to this assumption we can exclude any summary $Y \in \mathcal{S}^X$ for which $\exists Y' \in \mathcal{S}^X$ with $Y' \preceq_{\mathcal{S}^X} Y$, $|Y'| = |Y|$.⁸ This restriction can easily be computed by removing useless concept from G prior to applying Algorithm 1 – this is related to the hypernym closure. Other strategies could be to remove too abstract concepts and to consider specific concepts w.r.t the analysis of specific topological properties – e.g. concepts that are deep but have a large number of descendants are interesting candidate for summarization. Those optimizations techniques are context dependent and must be chosen in agreement with the defined objective function.

V. EVALUATION & DISCUSSION

Evaluating automatic summarization systems is a complex task as well as an open research topic – and no gold-standard dataset exists for evaluating the type of models we are studying. In this section we discuss preliminary results on the evaluation of the performance of the proposed model. This evaluation is based on human defined summaries of

⁸which means that $\forall y \in Y, \exists y' \in Y'$ with $m(y') = m(y)$.

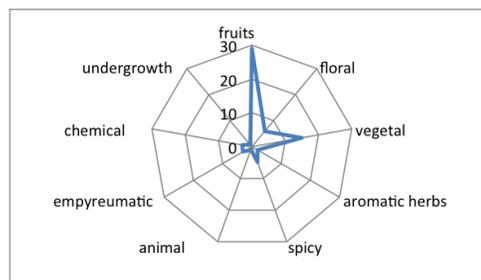


Fig. 5. Example of a radar chart used for sensorial analyses. The weight of each dimension (concept defined in the poset) corresponds to the number of (implicit) occurrences. Here, expected summary is Fruit, Vegetal and Floral.

conceptual annotations provided by domain experts related to odor evaluation. Considering a domain ontology, several experts have, for several products, provided conceptual annotations in the form of sets of concepts. To a specific product, the sequence of annotations \hat{X} has been analysed applying techniques experts use while performing sensorial analyses. Using this protocol we obtained radar charts corresponding to horizontal cuts of the poset, which forms a group of concepts that are homogeneous in term of concept specificity – this is the approach used by domain experts to analyse such data; Figure 5 provides an example. Domain experts finally build a summary by applying a subjective selection of concepts analysing such radar charts. For each product, 3 radar charts of different degrees of specificity have been provided – summaries are further generated by selecting a subset of the most frequently observed concepts. Note that the model studied in this paper can be used to generate more refined summaries composed of concepts having various degrees of specificity. The preliminary analyses we performed have shown that considering specific concepts of interest (those corresponding to the different degrees of specificity) the proposed model was able to find expected summaries, i.e. to rank relevant summaries according to the importance given to each concept in the radar view. Based on the analyses of the model that have been performed, without applying any constraints on the degree of specificity of the concepts to consider, we have also shown that such an approach could be used to automatize the summarizing process by, interestingly, generating more informative summaries.

VI. RELATED WORK

Automatic summarization is a broad research topic related to several domains including Machine Learning, Data Mining, Natural Language Processing, and Information Retrieval. Two main types of fully automatic summarization approaches exist: extraction-based and abstraction-based. Extraction-based methods try to find the most informative elements (e.g. sentences) and to remove repetitive elements (e.g. scenes in videos). Abstraction-based methods first build a representation of the dataset to further analyze this representation to generate a summary – in this case the summary may information that are not explicitly in the original dataset.

Automatic summarization is also intricately linked to clustering and indexing tasks. Literature related to document representation (e.g. vector and probabilistic models), clustering and indexing is of interest [8]–[10]. As an example, LDA [11] clusters the documents and produces a set of topics into which the documents are clustered. Extensions taking account of topic hierarchies and correlation between topics have also been proposed [12], [13]. Approaches for clustering and labeling a collection of resources indexed by concepts of a taxonomy have also been studied [14]. Nevertheless, they remain inspired from information retrieval approaches; the required objective function of the labeling optimization problem is reminiscent of the clustering and diversification processes in information retrieval: the similarity between the inner items of a cluster is maximized while the outer distance between items of two distinct clusters is maximized. The labeling process of clusters is then seen as a continuous optimization problem of distances whereas the feasible solutions are intrinsically discrete since they are related to a specificity level of the taxonomic hierarchy. HSLDA [15] also introduces a hierarchically supervised LDA model to infer hierarchical labels for a document. It assumes an existing label hierarchy in the form of a tree (e.g. multiple inheritance considered in our model is not allowed). The model infers one or more labels such that, if a label is inferred as relevant to a document, then all the labels from to the root of the tree are also inferred as relevant to the document. [16] highlights that applying the proposed inference rule, it is likely that many abstract labels will be classified as relevant without control on the specificity of the labels selected for summarization. To tackle this issue, they introduces a family of submodular functions to identify an appropriate set of topics from a DAG of topics for a group of documents. They characterize topic appropriateness through a set of desirable properties such as coverage, diversity, specificity, clarity, and relevance. Submodular functions are associated to these properties and mixed through a weighted average mean defining the objective function of the optimization problem the best summary results from. The coverage property is central in this approach. Indeed, in [16], unlike Human-like reasoning, no approximate reasoning is allowed over this property since it does not deal with frequency of occurrences, e.g. topics cannot be excluded from a summary even when they are poorly represented, as it is the case in our approach. Their axiomatic approach of the expected properties for summarization is close to our proposal but their related indicators and their management differ from ours: our indicators explicitly integrate masses or beliefs related to concepts; they also allow introducing control rules in the summarization process.

VII. CONCLUSION

Defining mathematical models enabling to automatically abstract and summarize bodies of information in a Human-like manner is a key challenge for Artificial Intelligence. We have proposed a general model to automatically summarize several conceptual annotations by considering knowledge representa-

tions providing *a priori* knowledge in the form of a *poset* formalizing the underlying structure of the concepts composing the annotations to analyze. A rigorous definition of the problem and a formal definition of a summary have been proposed; in addition, several interesting theoretical aspects highlighting the complexity of the challenge, as well as important properties of the search space have been discussed. Applications for data analysis and definition of intelligent agents are numerous considering the growing number of knowledge representations today available for a diversity of domains – e.g., gene analyses, information retrieval, and sensorial analyses. As an example, the evaluation of the proposed model performed in the domain of odor analysis, highlights the benefits of our proposal and shows how it could be used to automatize complex and time-consuming expert summarizing processes. Interestingly for the community, source code implementing the proposed approach as well as datasets are made available.

Additional large-scale experiments in several domains are currently performed to further criticize the model and discuss parameters tuning for specific use cases. Extended theoretical works are also performed to reduce the algorithmic complexity of finding the best summary w.r.t the proposed model. This aspect is of major concern to ensure method efficiency and practicality when applied to large knowledge representations. Interesting results based on the properties of the search space highlighted in this paper are currently studied.

REFERENCES

- [1] L. Saitta and J.-D. Zucker, *Abstraction in artificial intelligence and complex systems*. Springer, 2013, vol. 456.
- [2] J.-D. Zucker, “A grounded theory of abstraction in artificial intelligence,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 358, no. 1435, pp. 1293–1309, 2003.
- [3] S. Staab and R. Studer, *Handbook on ontologies*. Springer Science & Business Media, 2013.
- [4] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, “Semantic similarity from natural language and ontology analysis,” *Synthesis Lectures on Human Language Technologies*, vol. 8, no. 1, pp. 1–254, 2015.
- [5] N. Seco, T. Veale, and J. Hayes, “An Intrinsic Information Content Metric for Semantic Similarity in WordNet,” in *16th European Conference on Artificial Intelligence*. IOS Press, 2004, pp. 1–5.
- [6] S. Harispe, A. Imoussaten, F. Troussset, and J. Montmain, “On the consideration of a bring-to-mind model for computing the information content of concepts defined into ontologies,” in *International Conference on Fuzzy Systems FUZZIEEE*, 2015.
- [7] V. Ranwez, S. Ranwez, and S. Janaqi, “Subontology extraction using hyponym and hypernym closure on is-a directed acyclic graphs,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 12, pp. 2288–2300, 2012.
- [8] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.
- [9] F. Role and M. Nadif, “Beyond cluster labeling: Semantic interpretation of clusters contents using a graph representation,” *Knowledge-Based Systems*, vol. 56, pp. 141–155, 2014.
- [10] A. Bernardini, C. Carpineto, and M. D’Amico, “Full-subtopic retrieval with keyphrase-based search results clustering,” in *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT’09. IEEE/WIC/ACM International Joint Conferences on*, vol. 1. IET, 2009, pp. 206–213.
- [11] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.
- [12] D. Blei and J. Lafferty, “Correlated topic models,” *Advances in neural information processing systems*, vol. 18, p. 147, 2006.
- [13] D. Griffiths and M. Tenenbaum, “Hierarchical topic models and the nested chinese restaurant process,” *Advances in neural information processing systems*, vol. 16, p. 17, 2004.

- [14] N. Fiorini, “Semantic similarities at the core of generic indexing and clustering approaches,” Ph.D. dissertation, University of Montpellier, 2015.
- [15] A. J. Perotte, F. Wood, N. Elhadad, and N. Bartlett, “Hierarchically supervised latent dirichlet allocation,” in *Advances in Neural Information Processing Systems*, 2011, pp. 2609–2617.
- [16] R. B. Bairi, R. Iyer, G. Ramakrishnan, and J. Bilmes, “Summarization of multi-document topic hierarchies using submodular mixtures,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, 2015, pp. 553–563.

ERRATUM

Two minor corrections of the original paper have been made in this version - section III-B of the published paper. These corrections, related to the introduction of β_{YX} and to its bounds, are minor and do not impact the soundness of the proposal. Both Erratum have been proposed by Jeremy Bressand.

Erratum A

In the original version it has been asserted that $\beta_{YX} = 1$ for any $Y \in \mathcal{S} \setminus \mathcal{S}^X$. However, the best we can say is $\beta_{YX} \in [0, 1]$ for any $Y \in \mathcal{S} \setminus \mathcal{S}^X$. Indeed, considering the ordering of concepts introduced in Figure 3, let’s define summaries Y_1, Y_2 for $X = \{a, b\}$:

- $Y_1 = \{d, e\}$

We have $Y_1 \in \mathcal{P}(C)$ and Y_1 respects the *non-total redundancy property* : $d \not\prec e \wedge e \not\prec d$, so we have $Y_1 \in \mathcal{S}$. However Y_1 doesn’t respect the *faithful property* : $e \in Y_1 \wedge (\forall x \in X, x \not\prec e)$, so we have $Y_1 \notin \mathcal{S}^X$. However, despite $Y_1 \in \mathcal{S} \setminus \mathcal{S}^X$, we have β_{Y_1X} :

$$\beta_{Y_1X} = \sum_{x \in X \setminus \{a, b, c, d, e\}} m(x) = 0$$

- $Y_2 = \{e\}$

We have $Y_2 \in \mathcal{P}(C)$ and Y_2 respects the *non-total redundancy property* : $e \not\prec e$, so we have $Y_2 \in \mathcal{S}$. However Y_2 doesn’t respect the *faithful property* : $e \in Y_2 \wedge (\forall x \in X, x \not\prec e)$, so we have $Y_2 \notin \mathcal{S}^X$.

In this case, with $Y_2 \in \mathcal{S} \setminus \mathcal{S}^X$, we do have β_{Y_2X} :

$$\beta_{Y_2X} = \sum_{x \in X \setminus \{c, e\}} m(x) = \sum_{x \in X} m(x) = 1$$

Therefore, it demonstrates that the best we can say is $\beta_{YX} \in [0, 1]$ for any $Y \in \mathcal{S} \setminus \mathcal{S}^X$.

Erratum B

This version also corrects the illustration of expected penalty function behavior w.r.t. to β_{YX} - cf. we use any f function ensuring $f(\beta_{YX}) = 1$ and $f(\beta_{YX}) = \infty$ or any very large value when $\beta_{YX} \rightarrow 1$ and $\beta_{YX} \rightarrow 0$ respectively.



Combining Truth Discovery and RDF Knowledge Bases to Their Mutual Advantage

Valentina Beretta¹ (✉), Sébastien Harispe¹, Sylvie Ranwez¹,
and Isabelle Mougenot²

¹ LGI2P, IMT Mines Ales, Univ Montpellier, Ales, France
valentina.beretta@mines-ales.fr

² UMR 228 Espace Dev UM, Maison de la Télédétection, Montpellier, France

Abstract. This study exploits knowledge expressed in RDF Knowledge Bases (KBs) to enhance Truth Discovery (TD) performances. TD aims to identify *facts* (true claims) when conflicting claims are made by several sources. Based on the assumption that true claims are provided by reliable sources and reliable sources provide true claims, TD models iteratively compute value confidence and source trustworthiness in order to determine which claims are true. We propose a model that exploits the knowledge extracted from an existing RDF KB in the form of rules. These rules are used to quantify the evidence given by the RDF KB to support a claim. This evidence is then integrated into the computation of the confidence value to improve its estimation. Enhancing TD models efficiently obtains a larger set of reliable facts that *vice versa* can populate RDF KBs. Empirical experiments on real-world datasets showed the potential of the proposed approach, which led to an improvement of up to 18% compared to the model we modified.

Keywords: Truth discovery · RDF KBs · Rule mining
Source trustworthiness · Value confidence

1 Introduction

Several popular initiatives, such as DBpedia [2], Yago [17] and Google Knowledge Vault [5], automatically populate Knowledge Bases (KBs) with Web data. The performance of this Knowledge Base Population (KBP) process is critical to ensuring the quality of the KB. In particular, it requires dealing with complex cases in which several conflicting data are extracted from different sources, e.g. different automatic extractors will provide different birth places for Pablo Picasso. Approaches based on voting or naive strategies that only consider the most frequently provided data value are *de facto* limited. Such approaches are unable to deal with spam-based attacks or duplicated errors, which are common on the Web. Dealing with this problem therefore requires distinguishing values

according to their sources. In this study, we propose an approach that serves KBP integrating potentially conflicting data provided by multiple sources; it relies on a general framework that can be used to address conflict resolution problems by exploiting prior knowledge defined in existing KBs.

Several techniques based on Knowledge Fusion have been proposed in order to automatically obtain reliable information. Most of them suppose that information veracity strictly depends on source reliability. Intuitively, the more reliable a source is, the more reliable the information it provides. In turn they also assume that source reliability depends on information veracity, i.e. reliable information is provided by reliable sources. Truth Discovery (TD) methods are unsupervised approaches based on these assumptions aimed at identifying the most reliable of a set of conflicting triples – for *functional* predicates, i.e. when there is a single true value for a property of a real-world entity. This study aims to enhance the TD framework using knowledge extracted from an existing RDF KB to obtain a larger set of correct *facts* that could be used to populate RDF KBs. More precisely, it makes the following contributions:

- A novel approach that can be used to enrich traditional TD models by incorporating additional information given by recurrent patterns extracted from a KB. A state-of-the-art rule mining system is used to extract rules that represent these patterns. A method is proposed for selecting the most useful rules to be used to evaluate veracity of triples. Moreover, since each rule contributes to TD performances according to its quality, a function that aggregates the existing rule quality metrics is also defined. High-quality rules will have a higher weight than low-quality rules;
- An extensive evaluation of the proposed approach; interestingly, it shows that the TD framework can benefit from information derived by rules. As a consequence, we point out how the creation of high quality RDF KBs may benefit from the use of highly reliable TD models. The datasets and source code proposed in this study are open-source and freely accessible online.¹

The paper is organized as follows. Section 2 presents an overview of the TD framework and how it can be applied in the RDF KB context. It also describes the state-of-the-art rule mining techniques that are used in our work to detect interesting recurrent patterns. Section 3 explains how additional information extracted from KBs is integrated into the TD framework. The proposed approach is evaluated and discussed in Sect. 4. Finally, Sect. 5 reports the main findings and discusses perspectives.

2 Related Work and Preliminaries

In this section we introduce the formal aspect of TD, its goal and the key elements required to achieve it. We then formally present rules and their quality metrics. We will then be able to use them to exploit identified recurrent patterns to increase confidence in certain triples.

¹ <https://github.com/lgi2p/TDwithRULES>.

In this study, we assume that sources provide their claims in the form of RDF triples $\langle \text{subject}, \text{predicate}, \text{object} \rangle \in I \times I \times (I \cup L)$ where I is the set of Internationalized Resource Identifiers (IRIs) and L the set of literals.

The following definition introduces all TD components (source, data items and values). Since the TD and Linked Data (LD) fields use different notations, this definition aims at clarifying the correspondence between terms belonging to each field.

Definition 1 (Truth Discovery). *Let $D \subseteq I \times I$ be a set of data items where each $d \in D$ is a pair $(\text{subj}, \text{pred})$ that refers to a functional property ($\text{pred} \in I$) of an entity ($\text{subj} \in I$). Let $V \subseteq I \cup L$ be a set of values that can be assigned to these data items and S be the set of sources. Each source $s \in S$ can associate a value $v \in V$ (corresponding to $\text{obj} \in I \cup L$) to a data item $d \in D$, hence providing a claim v_d that corresponds to the RDF triple $\langle \text{subj}, \text{pred}, \text{obj} \rangle$. **Truth Discovery** associates a value confidence to each claim and a trustworthiness score to each source. It then iteratively estimates these quantities to identify the true value v_d^* for each data item.*

Several TD approaches have been proposed, as detailed in recent surveys [4, 10]. The models differ from one another in the way they compute the value confidence of claims and the trustworthiness of sources. Some of them use no additional information, while others attempt to improve TD performances using external support such as extractor information (i.e. the confidence associated with extracted triples), the temporal dimension, hardness of facts, common sense reasoning or correlations. Models that take correlations into account can be divided according to the kinds of correlations they consider: source correlations, value correlations or data item correlations. To the best of our knowledge, no existing work takes advantage of data item correlations in the form of recurrent patterns to improve TD results. The idea is that the confidence of a certain claim can increase when recurrent patterns occur which are associated with the considered data item. This kind of correlation can be used to enhance existing TD models. In this study, a rule mining procedure is used to identify patterns in data. We specify the major aspects of the rule mining below.

2.1 Recurrent Pattern Detection from RDF KBs

Several techniques can be used to identify regularities in data. For instance, link mining models are often used for that purpose in knowledge base completion [13]. In this study we prefer to use rule mining techniques because they are easily interpretable [1]. Rules generalize patterns in order to identify useful suggestions that can be used to generate new data or correct existing data [6]. We therefore propose to exploit these suggestions in order to solve conflicts among triples provided by different sources. Given our problem setting, where rules are used to reinforce the confidence of a claim, we are particularly interested in Horn rules. Considering Datalog-style, a Horn rule $r : B_1 \wedge B_2 \wedge \dots \wedge B_n \rightarrow H$, i.e. $r : \widehat{B} \rightarrow H$, is an implication from a conjunction of atoms called the body to a single atom

called the head [12]. An atom is usually denoted $pred(subj, obj)$, where $subj$ and obj can be variables or constants. Considering that an instantiation of an atom is a substitution of its variables with IRIs, an atom a holds under an instantiation σ in a KB K if $\sigma(a) \in K$. Moreover, a body \widehat{B} holds under σ in K , if each atom in \widehat{B} holds [7]. Note that in our setting each instantiated atom $pred(subj, obj)$ can also be represented as an RDF triple $\langle subj, pred, obj \rangle$.

Rule extractors rely on the Closed World Assumption (CWA). This means that when a fact is not known (does not belong to the KB) it is considered to be false. This assumption is more often appropriate when KBs are complete. On the contrary, RDF KBs are based on the Open World Assumption (OWA). When dealing with incomplete information the OWA is preferable. If information is missing we need to distinguish between false and unknown information. A triple that does not appear in the KB is not systematically false. In this context, methods have recently been proposed that mine rules from RDF KBs such as DBpedia or Yago, taking the OWA into account [15]. An example of a rule mining system that considers the OWA is AMIE [8]. It is based on the Partial Completeness Assumption (PCA): if a KB contains some object values for a given pair ($subject, predicate$), it is assumed that all object values associated with it are known. This assumption can generate counter-examples, required for rule mining models, but do not appear in RDF KBs, which often contain only positive facts. Alternative assumptions and metrics have been proposed to extract rules under the OWA [9, 13, 18]. In this study, we use AMIE because it is a state-of-the-art system and its source code is freely available online.

2.2 Rule Quality Metrics

Any rule, independently of the system used to extract it, can be evaluated by several quality metrics; among them the most well-recognized measures are *support* and *confidence* [1, 11, 19]. Support represents the frequency of a rule in a KB, while confidence is the percentage of instantiations of a rule in the KB, compared to the instantiations of its body. Based on the formal definition given in [8], for the sake of coherence and clarity, we present how these metrics are computed below. In the rest of the paper we do not make a comparison of the different quality metrics because it is out of the scope of this study. The primary aim here is to evaluate the potential of integrating knowledge extracted from an RDF KB into a TD process. However, since we are aware that robust metrics could have an impact on TD results, we plan to study such a comparison in future studies.

Considering a Horn rule $r : \widehat{B} \rightarrow H$ where H is composed of a single atom $p(x, y)$, its support is defined by:

$$supp(\widehat{B} \rightarrow p(x, y)) := \#(x, y) : \exists z_1, \dots, z_n : \widehat{B} \wedge p(x, y) \quad (1)$$

where z_1, \dots, z_n are the variables contained in the atoms of the rule body \widehat{B} apart from x and y , and $\#(x, y)$ is the number of different pairs x and y .

Its confidence is computed using the following formula:

$$\text{conf}(\widehat{B} \rightarrow p(x, y)) := \frac{\text{supp}(\widehat{B} \rightarrow p(x, y))}{\#(x, y) : \exists z_1, \dots, z_n : \widehat{B}} \quad (2)$$

This formula was introduced to evaluate the quality of rules using the CWA. It is too restrictive when dealing with the OWA. For this reason Galarraga et al. defined a new confidence, called conf_{PCA} [8]. It makes a distinction between false and unknown facts based on PCA. In this setting, if a predicate related to a particular subject, never appears in the KB, then it can neither be considered as true nor false. This new confidence based on PCA is evaluated as follows:

$$\text{conf}_{PCA}(\widehat{B} \rightarrow p(x, y)) := \frac{\text{supp}(\widehat{B} \rightarrow p(x, y))}{\sum_j \text{supp}(\widehat{B} \rightarrow p(x, j))} \quad (3)$$

where j 's are all instantiations of the object variable related to predicate p and having subject x . Using PCA, conf_{PCA} normalizes the support by the set of true and false facts that does not include the unknown ones.

In the next section, we describe how these quality measures are combined into a single measure. Having a more robust metric is important because it is the quality of each rule that will determine its contribution to the computation of the overall evidence that supports a certain claim.

3 Incorporating Rules into the Truth Discovery Framework

This section presents how extracted rules are integrated into truth discovery models. To that end, we define the concepts of *eligible* and *approving* rules, which will be used to identify the most useful rules that need to be taken into account when evaluating the confidence of a claim. Then we describe how information associated with these rules is quantified to further introduce the new confidence estimation formulas used by our TD framework.

3.1 Eligible and Approving Rules

It may not be useful to consider the entire set of extracted rules (denoted R) in order to improve value confidence. For instance, some rules could have a body that is not related to a given data item. Therefore, given a claim $\langle d, v \rangle$, i.e. v_d , where $d = (\text{subj}, \text{pred})$, only *eligible* rules are used as potential evidence to improve its confidence estimation. They are defined in the following way.

Definition 2 (Eligible Rule). *Given a KB K , a set of rules $R = \{r : \widehat{B} \rightarrow H\}$ extracted from K where $H = p(x, y)$ and a claim $\langle d, v \rangle$ where $d = (\text{subj}, \text{pred})$, a rule $r \in R$ is an **eligible rule** when its body holds, i.e. all of its body atoms appear in K when all rule variables are instantiated w.r.t. the data item subject. Moreover, its head predicate has to correspond to the one in the claim under examination, i.e. $(\sigma(\widehat{B}) \in K) \wedge (H = \text{pred}(\text{subj}, y))$.*

In our context, the eligibility of a rule depends on the subject and the predicate that compose a data item d . Thus, all claims related to the same data item $d = (subj, pred)$ have the same set of eligible rules, denoted $R_d = \{r \in R \mid (\sigma(\widehat{B}) \in K) \wedge (H = pred(subj, y))\}$.

Once eligible rules for a claim v_d have been collected, the proposed approach checks how many of these rules endorse (approve) v_d , i.e. how many rules support v_d .

Definition 3 (Approving Rule). *Given a KB K , a set of eligible rules $R_d = \{r : \widehat{B} \rightarrow H\}$ where $H = pred(subj, y)$ and a claim $\langle d, v \rangle$ where $d = (subj, pred)$, a rule $r \in R_d$ is an **approving rule** when the value predicted by r corresponds to the claimed value v , i.e. $(\sigma(\widehat{B}) \in K) \wedge (H = pred(subj, v))$.*

The set of approving rules for v_d is represented by $R_d^v \subseteq R_d$ where d indicates that the rules are eligible for a certain data item d and v indicates that the rules predict/support value v . Formally, we obtain $R_d^v = \{r \in R_d \mid (\sigma(\widehat{B}) \in K) \wedge (H = pred(subj, v))\}$.

Example. Given a KB K , reported in Table 1, and the rules:

- $r_1 : speaks(x, z) \wedge officialLang(y, z) \rightarrow bornIn(x, y)$
- $r_2 : residentIn(x, w) \wedge cityOf(w, y) \rightarrow bornIn(x, y)$

Given the following claims about the birth location of some painters $\langle Picasso, bornIn, Spain \rangle$, $\langle Picasso, bornIn, Málaga \rangle$ and $\langle Monet, bornIn, France \rangle$, the set of eligible rules for data item $d_A = (Picasso, bornIn)$ is $R_{d_A} = \{r_1, r_2\}$. The predicate in the head corresponds to the predicate in the claim and when all occurrences of variable x are replaced by *Picasso* in r_1 's and r_2 's body, they are both verified. However, when $d_B = (Monet, bornIn)$ the set of eligible rules is $R_{d_B} = \{r_2\}$ because, even though the head and claim predicate are the same using both rules, if the x variable is substituted by *Monet* the body of r_1 is not verified.

The set of approving rules for the first, second and third claims are respectively $R_{d_A}^{Spain} = \{r_1\}$, $R_{d_A}^{Málaga} = \emptyset$ and $R_{d_B}^{France} = \{r_2\}$.

Before explaining how additional information related to approving and eligible rules is quantified and then incorporated into the TD framework, we describe a function used to integrate the two quality aspects we are interested in, for each rule. This enables better weighting of each rule's contribution during the evaluation of a claim.

Table 1. Illustrative set of triples.

<i>predicate</i>	<i>subject</i>	<i>object</i>	<i>predicate</i>	<i>subject</i>	<i>object</i>
officialLang	(Spain,	Spanish)	residentIn	(Picasso,	Paris)
speaks	(Picasso,	Spanish)	cityOf	(Paris,	France)
residentIn	(Monet,	Vétheuil)	cityOf	(Vétheuil,	France)

3.2 Combining Rule Quality Measures

Support and conf_{PCA} represent different aspects of a rule, see Sect. 2.2. We propose an aggregate function to combine them into a single quality metric since, in our context, it is important to take both aspects into account. It may happen that two rules r_1 and r_2 have the same confidence, but different supports. For instance, if $\text{conf}_{PCA}(r_1) = \text{conf}_{PCA}(r_2) = 0.8$, $\text{supp}(r_1) = 5$ and $\text{supp}(r_2) = 500$, then r_2 deserves a higher level of *credibility* than r_1 since r_2 has been observed more often than r_1 .

To address this issue, a function $\text{score} : R \rightarrow [0, 1]$ is defined. It is based on Empirical Bayes (EB) methods [16]. EB adjusts estimations resulting from a limited number of examples that may happen by chance. Estimations are modified in function of available examples and prior expectations. When many examples are available, estimation adjustments are small. On the contrary, when there are only few examples, the adjustments are greater. They are corrected w.r.t. the average value that is expected by *a priori* knowledge. Given a family of the prior distribution of available data, EB is able to directly estimate its hyper parameters from the data. Then, it updates the prior belief with new evidence. In other words, the estimation that can be computed from the new examples is modulated w.r.t. prior expectation. The new estimation corresponds to the expected value of a random variable following the updated distribution. In our case, a more robust conf_{PCA} , i.e. the proportion of positive examples among all examples considered, needs to be estimated. The prior expectation on our data can be modelled using a *Beta* distribution that is characterized by parameters α and β . Once the model has estimated them, it uses this distribution as prior to modulate each individual estimate. This estimation will be equal to the expected value of the updated distribution $\text{Beta}(\alpha + X, \beta + (N - X))$, where X is the number of new positive examples and N is the total number of new examples. The new expected value is $(\alpha + X)/(\alpha + \beta + N)$. This value is returned by the aggregation function. In summary, given the hyper parameters α_S and β_S , the value returned by score for a rule $r : \widehat{B} \rightarrow p(x, y)$ is computed as follows:

$$\text{score}(r) = \frac{\alpha_S + \text{supp}(r)}{\alpha_S + \beta_S + \sum_j \text{supp}(\widehat{B} \rightarrow p(x, j))} \quad (4)$$

where $\text{supp}(r)$ is the support of r and $\sum_j \text{supp}(\widehat{B} \rightarrow p(x, j))$ is the number of triples containing data item (x, p) . The returned score appears to be similar to conf_{PCA} , but it takes the cardinality of the examples into account.

Once this score has been estimated for each rule, the proposed approach sums up all this new information and integrates it into the value confidence estimation formula.

3.3 Assessing a Rule's Viewpoint on Claim Confidence

All the evidence provided by rules for a claim v_d is summarized in a *boosting factor* that can be seen as the confidence that is assigned by these rules to

v_d . More precisely, it represents the proportion of eligible rules that confirm a given claim v_d . In other words it evaluates the percentage of approving rules out of the entire set of eligible rules, i.e. $|R_d^v|/|R_d|$. It is returned by a function $boost : D \times V \rightarrow [0, 1]$. As anticipated, the proposed model weights each rule differently w.r.t. its quality *score*. The higher the *score* of a rule, the stronger its impact should be on computing the *boosting factor*. Intuitively, given a claim v_d where $d = (subj, pred)$ and a set of rules R extracted from a KB K , the proposed model evaluates the *boosting factor* in the following way:

$$boost(d, v_d) \approx \frac{\sum_{r \in R_d^v} score(r)}{\sum_{r \in R_d} score(r)} \quad (5)$$

where R_d^v is the set of approving rules, R_d is the set of eligible rules and $score : R \rightarrow [0, 1]$ represents the quality score associated with a rule (as detailed in Sect. 3.2). Since the *boosting factor* consists in evaluating a proportion, EB is used also in this case to obtain a better estimation, less likely to be the result of chance. As explained in Sect. 3.2, when applying EB, initially the parameters α_b and β_b of a *Beta* distribution are estimated from the available data using methods of moments. Then this prior is updated based on evidence associated with a specific v_d . Thus, the *boosting factor*, corresponding to the expected value of the updated prior, is equal to:

$$boost(d, v_d) = \frac{\alpha_b + \sum_{r \in R_d^v} score(r)}{\alpha_b + \beta_b + \sum_{r \in R_d} score(r)} \quad (6)$$

where α_b and β_b are the hyper parameters of the Beta distribution representing the available examples. Since AMIE does not consider any *a priori* knowledge such as the partial order of values to extract rules, we decided to use it to further exploit rule information and compute a more refined boosting factor. More precisely, considering a partial order $\mathcal{V} = (V, \preceq)$, when a rule r explicitly predicts a value v , we assume that it implicitly supports all more general values v' such that $v \preceq v'$. In other words, the evidence provided as support by a rule to a value is propagated to all its generalizations. Therefore, in this case the boosting factor $boost_{PO}(d, v_d)$ indicates the percentage of approving rules out of all eligible rules, for both the value under examination and all of its more specific values. The subscript *PO* in the name of the boosting factor underlines the fact that the Partial Order among values is taken into account. The set R_d^v in Eq. 6 is replaced by the set $R_d^{v+} = \{r \in R_d \mid \hat{B} \wedge H = p(x, v'), v' \preceq v\}$.

3.4 Integrating Rules' Viewpoints into Confidence Computation

All the elements required to integrate information given by recurrent patterns into TD models have been defined. Since the *boosting factor* depends on the

claim, only the confidence formula has been updated. As proof of concept, in this study we modified *Sums* [14] whose estimation formulas are:

$$t^i(s) = \frac{1}{\max_{s' \in S} \sum_{v'_d \in V^{s'}} c^{i-1}(v'_d)} \sum_{v_d \in V^s} c^{i-1}(v_d) \quad (7)$$

$$c^i(v_d) = \frac{1}{\max_{v'_d \in V} \sum_{s' \in S^{v'_d}} t^i(s')} \sum_{s \in S^{v_d}} t^i(s) \quad (8)$$

We modified Eq. 8 proposing *SumsRULES*. This new model integrates the additional information given by rules into the confidence formulas as follows:

$$c_{rules}^i(v_d) = \frac{1}{norm_{v_d}} [(1 - \gamma)c^i(v_d) + \gamma boost(d, v_d)] \quad (9)$$

where $\gamma \in [0, 1]$ is a weight that calibrates the influence assigned to sources and KB for estimating value confidences. For the sake of coherence, when using *boost_{PO}* we considered the partial order also for the computation of the confidence formula, as suggested in a previous study [3]. We refer to the model that uses confidence formula $c_{PO}^i(v_d)$, taking the partial order into account, as *Sums_{PO}*. It computes the confidence of v_d considering all the trustworthiness of sources that provide the value v for the data item d , i.e. the claim v_d under examination, or a more specific value than v . Indeed as highlighted above when claiming a value, we also consider that a source implicitly supports all its generalizations. Similarly, the model that integrates both the *boost_{PO}* and rules is indicated as *SumsRULES&PO* and is defined as follows:

$$c_{RULES\&PO}^i(v_d) = \frac{1}{norm_{v_d}} [(1 - \gamma)c_{PO}^i(v_d) + \gamma boost_{PO}(d, v_d)] \quad (10)$$

Note that, while *Sums* and *SumsRULES* return a true value for each data item selecting the value with the highest confidence, *SumsRULES&PO* and *Sums_{PO}* required a more refined and greedy procedure to select the most informative true value. Indeed, considering the partial order of values, the highest confidence is always assigned to the most general value (it is implicitly supported by all the others). Thus, since systematically returning the most general value each time is not worthwhile, the selection procedure leverages the partial order to identify the expected value. Starting from the root, at each step it selects the closest specialization of the value with the highest confidence. The procedure stops when there are no more specific values, or when the confidence of the selected values is lower than a given threshold θ defining the minimal confidence score required to be considered as a true value. For further details see [3].

4 Experiments and Results

In order to obtain an extended overview of the proposed approach, several experiments were carried out on synthetic and real-world datasets. First of all, experiments were conducted using synthetic datasets to determine the improvement

obtained by $Sums_{RULES}$ (Eq. 9) and $Sums_{RULES\&PO}$ w.r.t. their baseline, i.e. $Sums$ [14] (Eq. 8) and $Sums_{PO}$ (Eq. 10) considering different scenarios. Note that, in both cases, the baseline corresponds to set $\gamma = 0$ in the new confidence formula of the proposed models. A second set of experiments was conducted using a real-world dataset to test the proposed approach in a realistic scenario. A comparison with existing models is also presented.

The rules used in the experiments, as well as their support and $conf_{PCA}$ were extracted from DBpedia by AMIE. To ensure that the rules considered are abstractions of a sufficient number of facts, we selected those with the highest head coverage. We selected 62 rules for the predicate *birthPlace*. Examples of these rules are reported in Table 2.

Table 2. Examples of rules extracted by AMIE from DBpedia for *birthplace* predicate.

@prefix db: <http://dbpedia.org/resource/>.	
@prefix db-owl: <http://dbpedia.org/ontology/>.	
?a db-owl:deathPlace ?b	→?a db-owl:birthPlace ?b
?a db-owl:country ?b	→?a db-owl:birthPlace ?b
?a db-owl:deathPlace ?b ∧ ?b db-owl:language db:English.language	→?a db-owl:birthPlace ?b

4.1 Experiments on Synthetic Data

The synthetic datasets were used to evaluate the proposed model on various scenarios depending on the granularity of the true values provided. Experts usually provide specific true values. Non-expert users provide general values, which remain true. To evaluate the performance in these contexts, we measured the expected value rate/recall (returned values that correspond to expected ones), the true but more general value rate (returned values that are more general than the expected ones) and the erroneous value rate (values that are neither expected nor general) obtained by different model settings.

Generation. The main elements required to generate these datasets are: a ground truth, a partial order and a set of claims provided by several sources on different data items [3]. The ground truth was generated by selecting a subset of 10000 DBpedia instances having the *birthPlace* property, considering the related value as the true one. Also the partial order of values was constructed using the DBpedia ontology. Partial order relationships were added between all classes subsumed, i.e. *rdfs:subClassOf*, by *dbpedia-dbo:Place* class and between those classes and their instances. Moreover, the relationships were added to all instances for which the property *dbpedia-dbo:isPartOf* or *dbpedia-dbo:country* exists. Since *dbpedia-owl:Thing* is the most abstract concept in DBpedia, all the values belonging to the partial order graph were rooted to it. In order to obtain a partial order of values respecting the properties of a Directed Acyclic Graph,

all cycles induced by incorrectness on the part-of property were removed.² For the generation of the claims, 1000 sources and 10000 data items were considered. Table 3 reports all the features regarding the generation of the claim set. The main feature is related to the distribution used to select the granularity of the true values provided. Based on this feature, three types of dataset were generated: EXP, LOW_E and UNI figuring, respectively, the behaviors of experts, a mix of experts and non-experts, and non-expert users. Considering that Picasso was born in Málaga, for example, in the case of EXP datasets, the sources tend to provide true values such as Málaga, Andalusia, Spain, while in the case of UNI datasets they will also provide general values such as Europe or the Continent. For each scenario, 20 synthetic datasets were generated.

Table 3. Features of synthetic datasets.

Feature	Description
Source coverage	Each source provides a number of claims that is exponentially distributed.
Source trustworthiness	The trustworthiness distribution is Gaussian with average 0.6 and standard deviation 0.4. This means that the sources are mostly reliable and only a few of them are always or never correct.
# of true claims per source	Each source provide a true value w.r.t. its trustworthiness level.
# of distinct true values per data item	$1..V_d^{true}$ where $V_d^{true} = \{v \in V : v_d^* \preceq v\}$
Granularity of the true value provided	Each source provides a true value having a granularity that approaches the granularity of the expected true value w.r.t. a high decay-rate exponential distribution (EXP), a low decay-rate exponential distribution (LOW_E) and a uniform distribution (UNI).
# of distinct false values per data item	$1..30$ values belonging to $V_d^{false} = V_d^{true} \setminus \{v v \preceq v_d^*\}$

Results. The results, summarized in Fig. 1, show that the proposed approach enables the definition of TD models that benefit from the use of *a priori* knowledge given by an external RDF KBs. Indeed, the number of correct *facts* identified by the proposed model usually increases w.r.t. the baseline. Intuitively, since

² We assumed that abstract concepts should have higher out-degree than less abstract ones. Thus, for each cycle, the edge whose target is the node with the highest out-degree was removed. Analysing the discarded edges, the heuristic works.

the number of correct *facts* increases, a new KB that is populated with the true claims identified by the improved TD will be of higher quality.

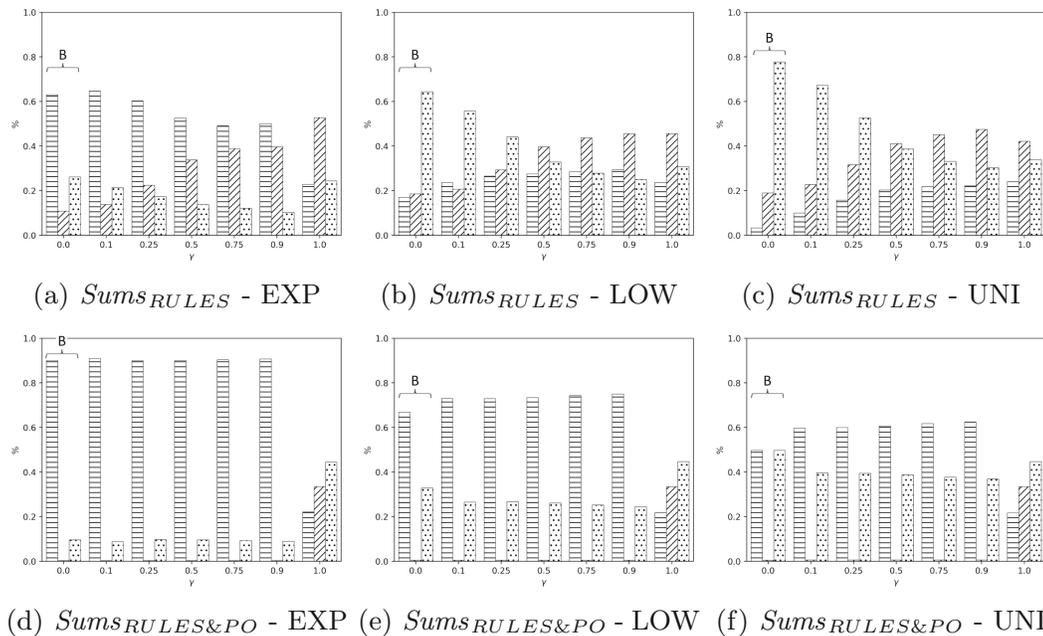


Fig. 1. Expected (horizontal line bars), true but more general (diagonal line bars) and erroneous values (dotted bars) obtained by $SumRules$ and $SumRules\&PO$ on different datasets with several γ . The letter B indicates the baseline model results.

The improvement obtained by considering both $SumRules$ and $SumRules\&PO$ was always greater for UNI datasets than for EXP or LOW_E ones. Since identifying true values in UNI settings was harder than in the other cases (the highest disagreement among sources on the true values was modeled by UNI), the baseline obtained the lowest recall. Using additional information tackles the high level of disagreement among sources and thus enables full exploitation of the higher scope for improvement that was available in the case of the UNI setting.

Considering $SumRules$ the best recall was obtained with different γ values. For UNI datasets, the optimal configuration was when $\gamma = 1$. In such a case, it was considered that no information provided by sources was useful and that only rules should be used to solve conflicts among claims (when rules are available). This was true only for the extreme situation represented by UNI datasets where disagreement among sources was so high that the recall obtained by baseline model remained under 10%. Indeed, in the other cases it was advantageous to take both source trustworthiness and rule information into account. For EXP datasets, the optimal γ value was 0.1, while for LOW_E it was 0.9. Low γ values were preferred in EXP settings because in this case sources that provide true values are quite sure about the expected one, and it is thus less useful to consider

the rules' viewpoints. Moreover, this setting was the only situation where considering external knowledge was damaging in terms of recall. Nevertheless, the error rate obtained by $Sums_{RULES}$ when $0 < \gamma < 1$ was always lower than the error rate achieved when $\gamma = 0$. This is explained by the fact that the average Information Content³ (IC) of values inferred by rules extracted for the *birthPlace* predicate is around 0.53. This means that they often infer values that are general. Many returned values, selected with the highest value confidence criteria, were therefore more general than the expected one but not erroneous. In other words, the rules associated with the *birthPlace* predicate were more effective for discovering the country of birth than the expected location. However using rules were useful, as shown by the results the error rate decreased.

The limitation related to rules that support general values was in part overcome by considering $Sums_{RULES\&PO}$, which also takes the partial order of values into account. In this case rules can improve the selection of the correct value during the first steps of the selection procedure. They were able to handle and dominate the false general values supported by many sources. The selection process was then continued with the fine-grained values evaluated based only on source trustworthiness information since no evidence provided by rules was available. For $Sums_{RULES\&PO}$ tested on EXP datasets, low γ values were preferred, while on LOW_E and UNI datasets high γ values led to the best performance.

The best overall recall was obtained by $Sums_{RULES\&PO}$, which considers both kinds of *a priori* knowledge: extracted rules and partial order of values.

4.2 Experiments on Real-World Data

These experiments were conducted to test the proposed model in a realistic scenario. Since the results of experiments on synthetic data showed that the most interesting results were obtained by considering both extracted rules and the partial order of values, we compared the results obtained in this case with those obtained by existing TD methods⁴ [20]. The evaluation protocol consisted in counting the number of values returned by a model that are equal to the expected values. In this setting, the number of general values returned were not analyzed since the main aim of TD models is to return the expected value, not its generalizations.

Generation. We collected a set of claims related to the predicate *dbo:birthPlace*, i.e. people's birth location. As data item subject, we randomly selected a subset of 480 DBpedia instances of type *dbo:Person* having the property *birthPlace* and having at least one eligible rule. For each data item we collected a set of webpages (up to 50) containing at least one occurrence of the

³ Information Content indicates the degree of abstraction/concreteness of a concept w.r.t. an ontology. It monotonically increases from the most abstract concept (its $IC = 0$) to the most concrete ones discriminating the granularity of different values.

⁴ For these models we used the implementation available at <http://www.github.com/daqcri/DAFNA-EA>.

subject’s full name and the words “was born”, i.e. the natural language expression that is usually used to introduce the birth location of a person. Given a webpage and its data item, we defined two procedures for extracting the provided claim. Procedure A selects, as claimed value, the location (identified by DBpedia-spotlight API) that co-occurs in the same sentence and is nearest to the word “born”. Procedure B adds a constraint to procedure A: a value can be selected only if it appears after the first occurrence of the subject’s full name in the text. Two different datasets were created based on procedures A and B, respectively DataA and DataB. For building our ground truth, we assumed that the values defined in DBpedia as birth location for each data item were the true ones. Since in the collected claims, values that were more specific than the expected one (contained in the ground truth) were provided, we manually checked if these specifications were true. For 20 instances that we manually checked, 10 were found to be true specifications. Note that as partial order we considered the same one as for the experiments on synthetic data. The procedures, source code and datasets obtained are available online at <https://github.com/lgi2p/TDwithRULES>.

Results. We can observe that for both datasets DataA and DataB we improved the performance by 18% and 14% respectively compared to the baseline, i.e. *Sums* – the approach we decided to modify. Table 4 shows the results obtained by the best configuration of parameters where both extracted rules and partial order were considered.

When comparing the proposed approach to existing TD models, it did not outperform the others, see Table 5. Note that our study focused on modifying *Sums* which is considered to be one of the most well studied models, but not necessarily the most effective one. After investigating the errors, we found out that it was mainly due to a limitation of *Sums*: it rewards sources having high coverage and, meanwhile, penalizes those with low coverage. Indeed *Sums* computes the trustworthiness of a source by summing up all the confidence of the claims it provides. Thus the higher the number of claims a source provides, the higher the trustworthiness of the source. The problem is that *Sums* does not distinguish between sources always providing true values, but having different coverage. While Wikipedia.org is correctly considered as a high reliable source, an actor’s fan club website is incorrectly considered as unreliable. Even if the information it provides is correct, because it covers only one data item its trustworthiness will be lower than the one of Wikipedia.org (source having a high coverage). In real-world datasets very few sources have high coverage, and most of them have low coverage – power law phenomenon. In this scenario the sources having high coverage dominate the specialized ones. Therefore, no extraction errors from high coverage sources are allowed. Indeed if an incorrect value is extracted from Wikipedia.org (for instance when the sentence refers to another person), this will be incorrectly considered as the true one. Since this cannot be guaranteed (the extraction procedures we defined are voluntarily naive), we propose a post-processing procedure that alleviates this problem. Before selecting the true value, it sets equal to 0 all the confidence of those values that are

Table 4. Recall obtained using *Sums* and its modifications on DataA and DataB.

Model	DataA	DataB
<i>Sums</i>	0.448	0.473
<i>Sums_{PO}</i> ($\gamma = 0.0, \theta = 0.05$)	0.517	0.566
<i>Sums_{RULES&PO}</i> ($\gamma = 0.3, \theta = 0.0$)	0.527	0.548
<i>Sums_{RULES&PO}</i> ($\gamma = 0.3, \theta = 0.05$)	0.565	0.590
<i>Sums_{RULES&PO}</i> +post-proc. ($\gamma = 0.3, \theta = 0.1$)	0.631	0.614

Table 5. Recall obtained using existing models on DataA and DataB.

Existing model	DataA	DataB
Voting	0.640	0.625
TruthFinder	0.646	0.622
2-Estimates	0.631	0.635
3-Estimates	0.008	0.612
Cosine	0.636	0.635
AccuCopy	0.638	0.640
Accu	0.638	0.660
Depen	0.431	0.494
AccuSim	0.413	0.448
SimpleLCA	0.631	0.660
GuessLCA	0.644	0.646

provided by only a single source. We assume that it is highly improbable that the same extraction error occurs, i.e. the erroneous value should therefore be provided only once. This solution, indicated as *Sums_{RULES&PO}* + post-proc., obtained performances comparable with existing models for DataA and DataB. While it enables to avoid some of the extraction errors (occurring more with the most naive procedure A), it is still not capable of assigning lower trustworthiness levels to specialized sources.

Given these observations, in real-world settings it is very important to consider the power law phenomenon. The results show that *Sums* is not efficient in this kind of situation. Nevertheless, using additional information (partial order and extracted rules) improved the results w.r.t. the baseline approach, and this is promising for the principles introduced in this study. As shown in Table 4, the improvement due to taking this information into account was 18% for DataA and of 14% for DataB. Moreover, through this study we also show that correctness and the granularity of values in DBpedia can be improved using TD models. Claims on data items can easily be collected on the Web. When more specific values than the one contained in DBpedia are found, they can be verified using TD model.

5 Conclusion

Solving information conflicts in an automated fashion is critical for the development of large RDF KBs populated by heterogeneous information extraction systems. In this study, we suggest using TD models as unsupervised techniques to populate RDF KBs. In order to create high quality KBs and exploit current ones, we propose improving an existing TD model (*Sums*) using knowledge

extracted from an external RDF KB in the form of rules. Several experiments that show the validity of the proposed model were conducted. The performances of the proposed model show higher recall than baseline methods (up to 18% of improvement). The datasets, source code and procedures are all available online. We plan to apply the rationale of the proposed model to other TD models in order to outperform them all. In addition, we envisage extending the evaluation methodology in order to consolidate our results by considering other predicates and non-functional ones such as those used in ISWC Semantic Web Challenge 2017. Currently, we do not consider as negative evidence the fact that a rule predicts a different value than the one contained in a claim. In the future, we envisage studying how to incorporate this information, as well as explicit axioms, subjectivity information and contextual dependencies (such as diachronicity).

References

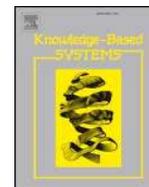
1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD 1993, vol. 22, pp. 207–216. ACM (1993)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
3. Beretta, V., Harispe, S., Ranwez, S., Mougnot, I.: How can ontologies give you clue for truth-discovery? An exploratory study. In: WIMS 2016, p. 15. ACM (2016). <https://doi.org/10.1145/2912845.2912848>
4. Berti-Équille, L., Borge-Holthoefer, J.: Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics. Synthesis Lectures on Data Management. Morgan & Claypool Publishers (2015). <https://doi.org/10.2200/S00676ED1V01Y201509DTM042>
5. Dong, X., et al.: Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: KDD 2014, pp. 601–610. ACM (2014). <https://doi.org/10.1145/2623330.2623623>
6. Galárraga, L.: Interactive rule mining in knowledge bases. In: Actes des 31e Conférence sur la Gestion de Données (BDA 2015), Île de Porquerolles (2015)
7. Galárraga, L., Suchanek, F.M.: Towards a numeric rule mining language. In: Proceedings of Automated Knowledge Base Construction workshop (2014)
8. Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F.M.: Fast rule mining in ontological knowledge bases with AMIE+. VLDB J. **24**(6), 707–730 (2015). <https://doi.org/10.1007/s00778-015-0394-1>
9. Lehmann, J., Völker, J. (eds.): Perspectives On Ontology Learning, vol. 18. IOS Press (2014). <https://doi.org/10.3233/978-1-61499-379-7-i>
10. Li, Y., et al.: A survey on truth discovery. SIGKDD Explor. Newsl. **17**(2), 1–16 (2016). <https://doi.org/10.1145/2897350.2897352>
11. Maimon, O., Rokach, L.: Data Mining and Knowledge Discovery Handbook, vol. 2. Springer, Heidelberg (2005). <https://doi.org/10.1007/b107408>
12. Nebot, V., Berlanga, R.: Finding association rules in semantic web data. Knowl.-Based Syst. **25**(1), 51–62 (2012)
13. Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E.: A review of relational machine learning for knowledge graphs. Proc. IEEE **104**(1), 11–33 (2016)

14. Pasternack, J., Roth, D.: Knowing what to believe (when you already know something). In: COLING 2010, pp. 877–885. Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
15. Quboa, Q.K., Saraee, M.: A state-of-the-art survey on semantic web mining. *Intell. Inf. Manag.* **5**(01), 1–10 (2013)
16. Robbins, H.: An empirical Bayes approach to statistics. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability Volume 1: Contributions to the Theory of Statistics, pp. 157–163. University of California Press, Berkeley, California (1956)
17. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge. In: WWW 2007, pp. 697–706. ACM (2007). <https://doi.org/10.1145/1242572.1242667>
18. Pellissier Tanon, T., Stepanova, D., Razniewski, S., Mirza, P., Weikum, G.: Completeness-aware rule learning from knowledge graphs. In: d’Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10587, pp. 507–525. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68288-4_30
19. Ventura, S., Luna, J.M.: Quality measures in pattern mining. In: Ventura, S., Luna, J.M. (eds.) *Pattern Mining with Evolutionary Algorithms*, pp. 27–44. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-33858-3_2
20. Waguih, D.A., Berti-Equille, L.: Truth discovery algorithms: an experimental evaluation. CoRR abs/1409.6428 (2014)



Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Truth selection for truth discovery models exploiting ordering relationship among values

Valentina Beretta^{*,a}, Sébastien Harispe^a, Sylvie Ranwez^a, Isabelle Mougenot^b^a LGI2P, IMT Mines Ales, Univ Montpellier, Ales, France^b UMR 228 Espace Dev UM, Maison de la Télé-détection, 500 rue JF Breton, Montpellier Cedex 5, 34093, France

ARTICLE INFO

Keywords:

Truth identification
 Truth discovery
 Conflicting values
 Value relationships
 Ontology

ABSTRACT

Data veracity is one of the main issues regarding Web data. Truth Discovery models can be used to assess it by estimating value confidence and source trustworthiness through analysis of claims on the same real-world entities provided by different sources. Many studies have been conducted in this domain. True values selected by most models have the highest confidence estimation. This naive strategy cannot be applied to identify true values when there is a partial order among values that is considered to enhance the final performance. Indeed, in this case, the resulting estimations monotonically increase with respect to the partial order of values. The highest confidence is always assigned to the most general value that is implicitly supported by all the others. Thus, using the highest confidence as criterion to select the true values is not appropriate because it will always return the most general values. To address this problem, we propose a post-processing procedure that, leveraging the partial order among values and their monotonic confidence estimations, is able to identify the expected true value. Experimental results on synthetic datasets show the effectiveness of our approach.

1. Introduction

Developing systems able to automatically evaluate the veracity of the avalanche of data produced by modern information society is of critical importance. Data veracity can be determined comparing information provided by multiple sources on the same subject [1–3]. Numerous scientific communities contribute to studying this complex issue, most notably with respect to (w.r.t.) data integration in information systems and databases. Among the several difficult tasks that data integration addresses and the different approaches that can be used to solve them [4–7], this paper focuses on automatic truth discovery for solving situations in which different sources provide potentially conflicting data about a specific property of an entity of interest, e.g. on the place of birth of a person.

Truth Discovery (TD) consider the accuracy associated with the data sources as an important factor to discriminate data veracity [6,7].

The main aim of the TD models is to identify true information. They intend to automatically solve, in an unsupervised manner, conflicts that may occur among claims. They leverage both the redundancy of the data and the information that is possible to derive from sources (particularly their reliability). More precisely, the backbone of TD is based

on the postulate that reliable sources provide true information and that, conversely, true information is given by reliable sources [3]. To identify reliable sources and true information, TD approaches estimate both source trustworthiness and value confidence; the true value is then considered to be the one with the highest confidence. Note that approaches that leverage information about data sources to check data veracity are currently the focus of a lot of attention in several domains such as social sensing [7] and question-answering [8].

Here we address the problem of selecting the truth for functional predicates when *a priori* knowledge in the form of a partial order of values (e.g. subsumption relationship in an ontology) is considered to improve value confidence and source trustworthiness estimations. A partial order highlights when different values are not conflicting, but they represent the same concepts with different levels of granularity. Indeed, conflict and granularity are two different aspects to consider when identifying the most reliable information. While conflict values produce inconsistency, different granularities only indicate imperfection in data [9]. In formal logic, a predicate p is considered functional if for any *subject* there is a unique value $v \in V$ for which $p(\text{subject}, v)$ is true¹ – birthplace is an example of a functional predicate. Note that this definition does not take subsumption relationships among values into

* Corresponding author.

E-mail addresses: valentina.beretta@mines-ales.fr (V. Beretta), sebastien.harispe@mines-ales.fr (S. Harispe), sylvie.ranwez@mines-ales.fr (S. Ranwez), isabelle.mougenot@umontpellier.fr (I. Mougenot).

¹ Elements of V are here considered to be independent.

<https://doi.org/10.1016/j.knosys.2018.06.023>

Received 19 December 2017; Received in revised form 25 June 2018; Accepted 28 June 2018

Available online 31 July 2018

0950-7051/ © 2018 Elsevier B.V. All rights reserved.

account. This is in accordance to the fact that, for instance, everyone was born in a specific location, but it does not consider that this place can be described using different levels of precision, e.g. district or region. In this case, multiple values can be true given the same *subject*. Thus, considering partial order of values and closed world assumption (our knowledge of the world is complete), a predicate is functional if for any *subject* there is a unique value $v \in V$, for which $p(\text{subject}, v)$ is true, such that there is not another value $v' \in V$ subsumed by v for which $p(\text{subject}, v')$ is true. As far as we know, we are the first to propose to take ontologies (as *a priori* knowledge) into account [10]. In this situation, the traditional final value selection step in the majority of TD approaches cannot be applied. Indeed, in this case, since more abstract values will *de facto* be associated with a higher confidence value in accordance with the partial ordering of values modeling implications among them, the true value cannot be defined as that with the highest confidence. This is due to the hypotheses used by approaches that leverage information related to the structure that may exist among values. Briefly, sources that explicitly claim a value implicitly support all of its generalizations. Therefore, if a source claims that “Pablo Picasso was born in Malaga”, it also implicitly supports the assertion that he was born in Spain, Europe, etc (considering that it is in agreement with the ontology). Thus the most general value is implicitly supported by all the others. Hence, its confidence will always be the highest. However, considering the most general value to be the only truth is not trustworthy (since it is a tautology), e.g. stating that Pablo Picasso was born in a Location is not meaningful. This paper proposes to overcome this problem by studying a solution able to identify more specific true answers (than the most general one) that may exist.

Our contribution consists of:

- proposing a post-processing approach able to identify the truth given the confidence estimations returned by any TD model that considers structured values;
- performing empirical experiments on synthetic datasets – this evaluation uses estimations returned by an adaptation of *Sums* able to take prior knowledge in the form of a partial order among values into account [10] – and comparing the proposed approach with existing ones evaluating identified true values.

The rest of the paper is structured as follows. Section 2 presents an overview of TD approaches taking advantage of potential relationships among sources or claims. This section ends with a discussion about the consequences of using a partial order among values as relationship information. In Section 3, notations are introduced and the problem is formalized. The solution strategy we propose is detailed in Section 4. The model is assessed via several experiments reported in Section 5 and discussed in Section 6. Section 7 summarizes the main findings of the study and the results that have been obtained; while the perspectives opened by our contribution are finally discussed.

2. Related work

Truth discovery aims to solve conflicts among data provided by several sources. The data treated in this domain consists of claims specifying the values that sources associate with certain data items (i.e. a data item represents a particular aspect of a real-world entity). Values can be numerical or categorical/strings. The main assumption of TD is that true information is provided by reliable sources and reliable sources provide true information [3,11]. This rationale can be modeled by defining the value confidence and source trustworthiness.

Many studies have been proposed in this field [3]. The baseline model consists of a voting strategy. For each data item it regards the value which is the most frequently claimed as truth. All sources are therefore implicitly considered similarly in this model. Otherwise TD estimates for each source a different trustworthiness level based on the claims it provides. Some models deal with numerical values [12], others

with categorical claims [7,13–15] and others with both [11,16,17]. While basic TD models limit their complexity to correctly estimate confidence and trustworthiness with different formulas, other approaches incorporate additional information to improve the overall performances. The latter group of approaches is the most relevant for our study and is detailed hereafter.

Among all of these methods, we focus our attention on those that use, as additional knowledge, correlations² that may exist among sources [17–19], data items [7,20,21] or values [11,22].

The first class is related to source interdependences. These models consider source relationships mainly by analyzing the pattern of similar claims with correlated accuracy estimations. They also usually assume that sources sharing common false values are more likely to be dependent than sources sharing common true values. Indeed, it is difficult to identify dependencies between sources stating different false values [23]. Most of these studies only analyze static correlations. To the best of our knowledge, time-course dependency relationship patterns has been only considered in [24]. In this case, dependency among sources is captured by studying the similarity between patterns of updates associated with sources [23]. Several methods take advantage of dependency in terms of the copying relationship [17–19,22]. Other correlations among sources may also occur, such as the common errors made by different extractors that use the same extraction rules or the common values identified by the extractor that use different rules, and so on [17]. Moreover, the dependence relationship is often considered between source pairs, but dependencies may also occur at the group level [18]. The second correlation class is related to data items. The first body of works in this context proposed to deal with the social sensing problem. In crowd sensing, humans coupled with their smartphones become sensors that explicitly or implicitly provide observations about their physical environment. Then it becomes necessary to understand the validity of data sent by sensors. TD models applied to this domain take advantage of both physical [25] and temporal [26] correlations as well as causal relationships [21]. For physical correlations, they assume that co-located data items should have similar values. For instance, gas stations located in the same area should have similar gas prices. For temporal correlations, the assumption is that two temporally close observations cannot have very different values. This kind of correlation is especially useful when analyzed data has a long-tail characteristic, i.e. many data items observed by few sources and few data items observed by many sources. Indeed, in this case the estimations can easily deteriorate if the few sources that provide claims for a data item are also unreliable. Using correlations, information associated with data items having a high number of observations provided by reliable sources can be propagated to data items having only a few claims associated with them. The findings of the two studies [25] and [26] permit to partition data items into small groups without considering any dependency among groups, but the complexity of their solutions is exponential w.r.t. the maximum group size. Alternative models have been proposed to overcome this limitation to be able to deal with a large number of dependencies, e.g. [20,21]. The former classifies the problem as an optimization problem, and the latter, modelling the problem as a Bayesian network, leverages potential conditional independencies among data items. Moreover, in this study they also take into account a second kind of correlation related to the data items: i.e. the category. In this case, a trustworthiness level may be attributed to each source w.r.t. the category a data item belongs to. The main limitation of this approach is that the Bayesian network has to be known or empirically learned from historical data by specific algorithms.

The third type of correlation regards the values. The basic idea is that two correlated values support each other. If one of them is considered true, then the other has a high probability to be true. In order to

²We mean by correlation the interdependences between entities, the relationships that may influence them.

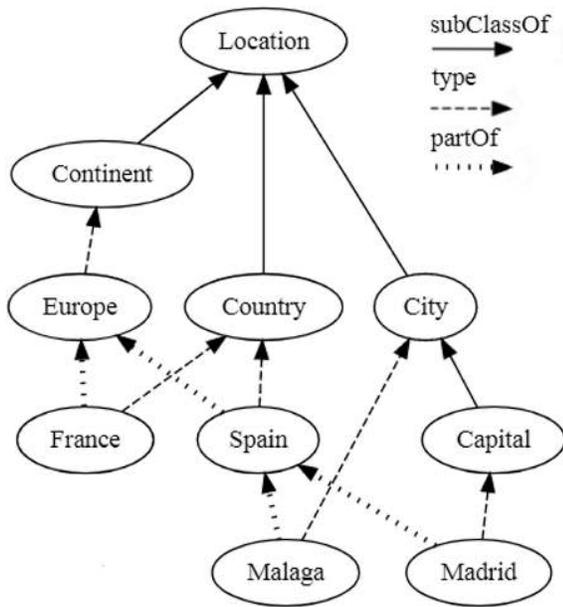


Fig. 1. Example of a partial order that may exist among values.

evaluate value correlations, previous studies such as [11,16,22] use value similarity. For instance, they compute the edit distance of strings, similarity among sets, and difference among numerical values. Otherwise, in our previous study we took advantage of value correlations in the form of partial ordering that may exist among the provided values [10]. An example of partial order is shown in Fig. 1. Given this partial order, we adapted the *Sums* approach to incorporate this new information. The rationale is that if a source associates the value *Spain* to an aspect of a real-world entity, then it also implicitly supports all more general values, e.g. *Europe*. Thus we modified the formula to estimate the confidence of a value changing the set of sources used for the calculus. In *Sums*, confidence is computed by considering all sources that provide an analyzed value, while in the adapted model (*AdaptedSums*) sources that explicitly claim a more specific value than the one being analyzed are also taken into account as well. Indeed, these sources implicitly support all more general values than the one they provide. Using the adapted approach, at the end of the iterative procedure, the value confidence estimations monotonically decrease w.r.t. the partial order among values. In other words, for each value more specific than another, the confidence of the former is lower than or equal to the confidence of the latter. As a result, when considering the partial order of values, the highest confidence score will be always assigned to the most general value. In this context, using the usual strategy adopted by the existing models is not worthy. Indeed, selecting as truth the value having the highest confidence, they will always return as true value the most general one. In the rest of the paper, we describe a refined post-processing strategy able to select the true value leveraging these monotonic estimations.

3. Problem formulation

Let's consider a set of data items D such that each $d \in D$ is composed of a pair (*subject*, *predicate*) where the *subject* represents a real-world entity and the *predicate* represents its aspect of interest, e.g. (*picasso*, *bornIn*). S denotes the set of sources, V the set of values, $V^s \subseteq V$ the set of values provided by $s \in S$ (for each data item for which s provided values), and $V_d \subseteq V$ the set of values associated with data item $d \in D$.

Formally, TD models first aim to identify the set of true values $V_d^* \subseteq V$ for each data item $d \in D$. In the case of a data item d characterized by a functional *predicate*, we have $|V_d^*| = 1$ if elements from V are disjoint, i.e. $\forall (v, v') \in V^2, \neg(v \Rightarrow v') \wedge \neg(v' \Rightarrow v)$. Note that a value v

implies (\Rightarrow) a value v' when v is subsumed by v' . To ease the formal introduction, and in accordance with the literature, we will as often as possible consider the special case of data items associated with the functional predicate.

Dealing with data items composed of functional predicates TD identifies the true values $v_d^* \in V$ for each data item d estimating value confidences $c: V \rightarrow [0, 1]$ – how an information is likely to be true – and source trustworthiness $t: S \rightarrow [0, 1]$ – how reliable is a source. This is done through an iterative procedure that alternatively estimates them. The execution of the model finishes when the stopping criteria is verified, e.g. convergence of estimations, maximum number of iterations, and so on. Hence, each value $v_d \in V$ (w.r.t. a data item) is associated with a confidence level $c(v_d)$ and each source $s \in S$ with a trustworthiness level $t(s)$.

Existing approaches usually assume that for a specific data item d , elements of V_d are disjoint/independent and they therefore recognize the true value of d is that with the highest confidence score. This straightforward procedure cannot be applied using adapted models that consider ordering among values. Incorporating this information into the model, each value more general than a true value can only be considered as true as well.³ Therefore, considering all values associated with a data item, the estimated confidence scores monotonically increase w.r.t. the partial ordering of values, i.e. $\forall v, v' \in V$: if $(v \Rightarrow v')$, then $c(v) \leq c(v')$. Consequently, the highest confidence score is always assigned to the most general value (that is implicitly supported by all provided claims). To solve this problem, we propose a post-processing procedure able to select the true value for each data item given the estimated confidence scores and the relationships that may exist among values.

We assume that the value dependencies are known *a priori* in the form of a partial order modelled by an ontology $O = (\leq, V)$. Note that even if the domain knowledge is not available, partial order can be automatically constructed [27]. The partial order \leq can be represented by a Directed Acyclic Graph (DAG), $G_O = (V, E)$, where $V = \{v_0, v_1, \dots, v_m\}$ is the set of values representing our vocabulary according to our knowledge of the world (all possible values that can be claimed by sources), and $E = \{(x, y) \in V \times V | x \leq y\}$ is the set of edges specifying the partial ordering that exists between values. Specifically $x \leq y$ when there is a directed path from x to y in the DAG; i.e. when y is reachable from x [28]. Note that a path from x to y is defined as a non-empty sequence of n different nodes $\langle v_0, v_1, \dots, v_{n-1} \rangle$ with $x = v_0$, $y = v_{n-1}$ and for which $\forall i \in [0, n-2] (v_i, v_{i+1}) \in E$. An important characteristic of the graph G_O is that it has to be transitively reduced. This is not a problem because by considering any DAG its transitive reduction can be obtained [29].

Here we introduce several functions that will be useful for manipulating the graphs (with \mathcal{G} a set of graphs):

- *ancestors*: $\mathcal{G} \times V \rightarrow \mathcal{P}(V)$ such that $ancestors(G_O, x) = \{y | x \leq y\}$
- *children*: $\mathcal{G} \times V \rightarrow \mathcal{P}(V)$ designed as $children(G_O, x) = \{z | (z, x) \in E\}$
- *root*: $\mathcal{G} \rightarrow V$ such that as $root(G_O) = \{x | \forall y \in V, y \leq x\}$

These properties enable us to easily explain our procedure to traverse the partial value ordering graph in the next section.

Further important information that can be derived from any ontology is the Information Contents (IC) of its concepts (e.g. [30]). This quantity, related to the concept specificity (see Section 3.3 in [31]) represents the degree of abstraction/concreteness of a concept w.r.t an ontology. One of main IC property is that the IC score monotonically decreases from the root to the leaves, i.e. if $x \leq y$, then $IC(x) \geq IC(y)$ ($IC(root) = 0$). This score will help us to discriminate between different values w.r.t. their granularity.

All of the elements presented in this section will help define the

³ Assuming that the value ordering is consensual.

approach used to select the true values that is described in Section 4.

4. Proposed approach

The entire truth-discovery procedure, from the input consisting of a set of claims to the output consisting of the true values and the degree of reliability associated with each source, is presented in Fig. 2. In this section, we propose a post-processing procedure that selects the true values given the estimations obtained by TD models that relax the assumption related to the disjointness of values.

It involves three steps: (i) selection of the best true value candidates; (ii) ranking of selected values; and (iii) filtering of ranked values w.r.t. defined desirable properties. For instance it may be useful to return a set of solutions that share an ordering relationship or, on the contrary, to return a value set composed only of “alternatives” that are not ordered. The choice related to the appropriate features of the solution set depends mainly on the application scenario.

The first step of the process: (i) permits to retrieve the most specific possible true value(s) and all of its ancestors using available information, such as confidence scores and partial ordering of values. The second step: (ii) orders the selected values based on predefined criteria. The third step: (iii) is required to filter the top k results. For TD, the final aim k should be equal to 1, but in cases where there is uncertainty it may be useful to return a set of values, even if the predicate is functional. Moreover, answers that do not have defined desirable properties (see Section 4.3 for further details) are removed from the result list. Those three steps are detailed hereafter.

4.1. True value selection

The first part of the post-processing procedure concerns the selection of the promising candidate(s) as the most expected value(s) for each data item. We have defined a selection strategy that takes advantage of the partial order of values and step by step refines the granularity of the correct value associated with each data item. Now we will give an overview of the approach followed by Algorithm 1.

Starting from the most general value (implicitly supported by all provided values and surely true), the process aims to detect the most specific expected value(s). A traversing procedure was thus applied on the graph that represents the partial order of values. It starts from the root, it selects the best alternatives among the children of the considered node and moves forward through the selected values. Our

```

1: procedure SELECTIONTRUEVALUES( $d, G_O, c, \theta, \delta$ )
2:    $G \leftarrow transitive\_reduction(G_O)$ 
3:    $V_{visited}^* \leftarrow \{ \}$ 
4:    $queue \leftarrow list(root(G))$ 
5:   while  $!(queue.isEmpty())$  do
6:      $v \leftarrow queue.pop()$ 
7:      $V_{visited}^* \leftarrow V_{visited}^* \cup \{v\}$ 
8:      $V_{ch} \leftarrow children(G, v)$ 
9:      $conf_{max} \leftarrow \max_{child \in V_{ch}} (c(child))$ 
10:     $V_{ch^*} = \{v' \in V_{ch} : c(v') \geq \theta \wedge (conf_{max} - c(v')) \leq \delta\}$ 
11:     $queue.addAll(V_{ch^*} \setminus V_{visited}^*)$ 
12:  return  $\bigcup_{v \in V_{visited}^*} ancestors(G, v)$ 

```

Algorithm 1. True value set computation for any $d \in D$ considering a partial order of values represented as a DAG $G_O = (V, E)$, a threshold $\theta \in (0, 1]$, a threshold $\delta \in [0, 1]$, and a function $c: V \rightarrow [0, 1]$, i.e. confidence of each value.

assumption is that values with the highest confidence locally should be the most likely to be true. Therefore the choice of the best alternative(s) is done by comparing the confidence scores associated with the children of the previously selected node.

In the case of functional predicate, the values can be partially ordered by their granularity, see Fig. 1. Therefore the selection procedure refines, at each step, the level of precision used to describe the single true value associated with a data item. The semantics of each selected node expresses the fact that the node subsumes the correct solution (i.e. the expected true value). The last selected nodes should correspond to the most specific answers that can be identified through the selection process.

The selection process has to handle two main undesirable situations that may occur: (1) selection of values with a confidence score too low to be considered as true, and (2) difficulty in discriminating the best alternative(s) among the children of a node since their confidence scores are not significantly different. As a solution, two thresholds have been defined: θ and δ .

The threshold $\theta \in (0, 1]$ enables us to specify the confidence lower bound required for a value to be part of the set of true values. Note that the value 0 is not included in the θ interval. Indeed, considering claims with confidence scores equal to 0 makes no sense because it would mean considering, as truth, values provided by totally unreliable sources (all with trustworthiness equal to 0). The confidence score that is compared to θ has to be previously normalized w.r.t. each data item, i.e. the confidence score associated with the most general value of each data item always has to be equal to 1. This normalization step is required to avoid the definition of an inconsistent threshold w.r.t. the different data items.

The threshold $\delta \in [0, 1]$ represents the minimum difference that has to exist among values with the highest confidence and all the others so that one prevails over the others. In particular, if the difference between the confidences of two values is less than or equal to δ , then it is hard to make a choice among them. This comparison is done among values that are children of the same father to select the best alternatives.

The definition of different parameter settings produces different behaviours of the selection phase ending in the possibility of obtaining different kinds of solution sets. The main parameter settings are summarized in Table 1.

Configuration 1 ($\theta = \alpha, \delta = 0$) reproduces a naive greedy algorithm that, at each step, selects values with the highest confidence greater than α without performing any other control.

Configuration 2 ($\theta = \alpha, \delta = 1$) is able to return all claimed values with confidence higher than α . It may seem useless, but it is a selection configuration necessary to obtain a particular set of values at the end of

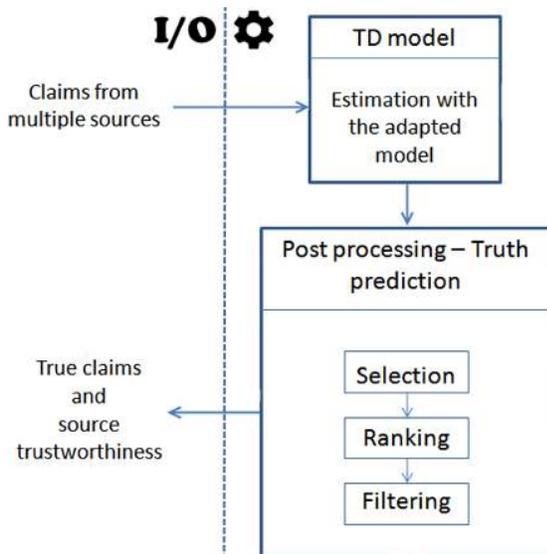


Fig. 2. Diagram of the overall TD procedure.

Table 1
Interesting settings for the selection procedure.

Config	θ	δ	Selection procedure behaviour
1	α	0	Naive greedy procedure that maximizes the confidence score at each step.
2	α	1	Greedy procedure that selects all values greater than α . Since δ is equal to 1, all values with confidence higher than α are selected.
3	α	β	At each iteration a value is collected only if the difference of its confidence and the highest confidence at the current step is lower or equal to β . All values in the returned set have confidence that is greater than α .

the post-processing procedure. A set composed of “promising” alternatives that allow to report values that are, as much as possible, fine-grained and semantically different. In this way, we increase the probability of finding the correct value since we increase the number of different concepts that are considered. Therefore this strategy is useful to deal with cases in which there is a lot of uncertainty. The idea is to return all claimed values and their ancestors and then, using the ranking phase to position in the first places, the most promising alternatives with the properties we have just explained.

Configuration 3 ($\theta = \alpha, \delta = \beta$) is a generalization of the two previous configurations. It selects the set of values that are greater than a threshold θ and they differ, at each step, more than δ from the confidence of the other alternatives.

Algorithm 1 reports the pseudo-code of the selection procedure. The algorithm starts performing a transitive reduction of the graph representation of the partial ordering (line 2). We thus ensure that the choice of the best alternative is done among a set of children that do not share ordered relations. Moreover, this avoids useless comparison of a large number of confidence scores. Then, at each iteration, the algorithm applies a greedy search by maximizing the confidence of the values (lines 5 – 9). It selects all values having confidence higher than or equal to θ whose scores are not significantly different from the highest confidence (line 10). Then, it adds them to the queue (line 11). Note that the confidence scores were computed applying *AdaptedSums*. This model computes the value confidences summing up all trustworthiness of those sources providing the considered value or one of its possible specializations. The procedure stops when the last selected value has no more specific values to be visited. In order to be in accordance with our assumption and problem settings, all values that are more general than that selected will compose the set of true values – due to multiple inheritances some of those values may not have been visited by the greedy procedure (line 12). The fact that confidence score monotonically increases w.r.t. the partial order ensures that the scores related to ancestors of the visited values are higher than or equal to θ .

The termination of Algorithm 1 is ensured by line 6 and line 11. The complexity of the selection of the true value algorithm is related to the number of comparisons required to find the maximum value confidence traversing graph G_O . Therefore, the complexity of the algorithm is $O(E)$ which in turn is $O(V^2)$. At each step, a number of comparisons equal to the number of node children is required. The worst case scenario is verified when the following conditions holds at the same time: (i) graph G_O has depth 2, (ii) its nodes are uniformly distributed between level 2 and 3, (iii) nodes at the same level have the same fathers and the same children and, moreover, (iv) they have equal or not significantly different confidence scores. The conditions related to the morphology of the DAG ensure that the number of comparisons is maximum, and the condition on the confidence score guarantees that the procedure traverses all nodes.

All of the configurations of the algorithm input parameters enable us to select a set of possible true values. Since the aim of TD is to find the most expected solution, a method able to choose it is required. The ranking phase described in the next section is devoted to this.

4.2. True value ranking

Given the true value set selected in the previous step, we have to define a ranking method in order to select the $k \in \mathbb{N}^+$ most expected values where k is a fixed number. In our investigations, k is experimentally set, at the most, at 5. The solution set of most expected true values is indicated as $V^* \subseteq \mathcal{P}(V^*_{candidates})$, where $V^*_{candidates}$ is the value set returned by the selection phase.

Now we propose to rank the values based on rather:

- their IC. This method is useful for situations in which specific answers are expected and when there is not much uncertainty on the data item under consideration. Note that in the following experiments IC is a measure computed according to the definition provided by Seco based on the analysis of the partial ordering topology [30]. In particular, it takes advantage of the number of descendants of a value:

$$IC_{Seco}(v) = 1 - \frac{\log(|descendants(v)| + 1)}{\log(|V|)} \tag{1}$$

where $|V|$ is a non-empty set since an ontology is considered to have at least one concept, i.e. the root value.

IC has been proposed because the user generally expects very precise answers. Often general true values for a data item are well known *a priori*, i.e. it is well known that a person is born in a place. If two or more true values have the same IC, then random selection can be done or, alternatively, another criterion can be used to rank this value subset.

- their source average trustworthiness, denoted $WA_{trust}(v)$. The rationale is that if a lot of unreliable sources support a false value A (increasing its confidence score – *Sums* does not normalize based on the number of sources claiming a value, therefore its confidence estimation is biased), and there are only a few reliable sources that support a true value B, then sources providing B should have higher average trustworthiness scores. This measure is obtained by computing the average trustworthiness associated with sources that explicitly or implicitly claim to have a particular value v and by weighting it by a normalization factor:

$$WA_{trust}(v) = \left(1 - \frac{1}{\eta + |S^{vd+}|}\right) * avg_{trust}(v) \tag{2}$$

where the average of source trustworthiness is represented by $avg_{trust}(v)$, S^{vd+} is the set of sources that implicitly or explicitly provide the value v_d and η is a small number used to avoid that $WA_{trust}(v) = 0$ when v is provided by only one source. The first factor, i.e. normalization, was introduced in order to tune the average w.r.t. the number of sources providing the value. Indeed, inspired by the study presented in [32], the higher the number of sources providing a value, the higher our confidence in the computed average should be.

Moreover in this case, if two values have the same WA_{trust} , then another criterion can be used to rank them.

Once the values are ranked, the next and final step of the post-processing procedure can be performed.

4.3. Filtering of top-k true values

The filtering phase collects the top k values in the rank and returns them to end-users. Before performing selection of the top k values, all the ranked ones have to be controlled. This is necessary because truth discovery models can be applied to different scenarios: high or low uncertainty situations, high or low risky cases in which making an error is, respectively, very dangerous or not. For instance, if truth discovery models are used to populate a medical knowledge base containing,

each symptom, all possible correlated diseases, then the end-users want to be really careful in accepting a value as true. Therefore, based on the possible application contexts, different properties that the solution set V^* has to respect can be defined. In this way various true value sets with different characteristics can be identified:

- the solution set V_{ord}^* contains only values sharing partial ordering relationships; formally $\forall (x, y) \in V_{ord}^* \times V_{ord}^*, x \leq y \vee y \leq x$. The procedure to create a set containing values that respect this property is as follows: it iteratively selects and removes the first element in the ranked list. Then it adds this value to the solution set only if it is an ancestor or descendant of all elements that are already present in it.
- the solution set V_{disj}^* is composed of values that do not share any partial ordering relationships; formally $\forall (x, y) \in V_{disj}^* \times V_{disj}^*, \neg(x \leq y) \wedge \neg(y \leq x) \wedge \nexists w, z \in V_{candidates}^*, w \leq x \vee z \leq y$. This means that all values in the solution set are the most specific among those returned by the selection phase. Indeed, only values that do not have descendants in the returned true value set belong to the solution. The procedure used returns a set of alternatives that are as much as possible very specific and different. In other words, this set of values consists of elements that do not have any of their exclusive descendants in the sorted list. For example, if the values returned by the previous step are *Europe, Continent, Country, City, Location*, then, in accordance with the partial order in Fig. 1, the V_{disj}^* is composed only of *Europe, Country* and *City*.

The first kind of solution can be desirable when there is not much uncertainty (end-users expect to easily find the true answers) or the end-users do not want to deal with potentially different values in a domain where they are not experts. The second property can be adopted when there is a lot of uncertainty and especially when the application context could result in making errors without dangerous consequences. Indeed, when there is uncertainty, to postpone the selection of true values to the end-users, avoiding to automatically select only a specific value and its ancestor, may be useful. In order to support the end-users final choice, returning a set of values composed of the most promising alternatives is important.

Obtaining V_{ord}^* is suitable when $\delta = 0$. Indeed, taking the value with the highest confidence at each step, the process ends with the selection of only one specific true value (and its ancestors). Considering this set of returned values, the first property is often verified without filtering any value out. In any case, very general values often are not returned since only the top- k values are selected after the verification of the property. Otherwise, the second property, i.e. V_{disj}^* , is not useful considering $\delta = 0$. Only the single most specific value contained in the set of returned values is selected when this property must hold. Indeed, all of the others share partial ordering relationships. This corresponds to consider that $\delta = 0, k = 1$ and a solution set V_{ord}^* . Obtaining V_{disj}^* is preferable when $\delta = 1$. Indeed, in those cases all values having confidence higher than θ are returned by the selection phase, but for the final aim of truth discovery (finding the truth) it is suitable to only keep the set of “promising” alternatives that correspond to a set of values that are different and specific as much as possible.

5. Experiments

In this section we describe all experiments performed on synthetic datasets to confirm the validity of our approach. First, we focus on the synthetic datasets: how we generated them and what are the parameter settings we tested. Then, we present the evaluation methodology we used to evaluate the approach and to compare it with existing models.

5.1. Datasets

In order to assess the behaviour of our approach w.r.t. the ontology used to derive the relationship among values, we integrate preliminary experiments carried out using the synthetic *birthPlace* datasets (see [10]

for further details related to their generation) with additional ones performed using different partial ordering structures.

Each synthetic dataset contains a set of claims concerning a specific predicate, a set of sources and the subset of claims provided by each source. We generate 5 different datasets considering the predicates *birthPlace* and *genre* from DBpedia [33], and the predicates *Cellular Component(CC)*, *Molecular Function(MF)* and *Biological Process(BP)* from Gene Ontology [34]. All the datasets are randomly generated based on a ground truth (containing a set of true claims, for each predicate), a partial order among the values contained in the ground truth (an example is shown in Fig. 1) and a set of factitious sources. Note that Table 2 reports the features associated to the different partial ordering structures that we use. Given these elements the generation process can start.

First, a trustworthiness level is associated with each source. We assume that the majority of sources are sufficiently reliable and only a few of them are always or never correct. A Gaussian distribution with an average and standard deviation equal to, respectively, 0.6 and 0.4 was used to model the described behaviour.

Second, we reproduce the long-tail phenomena [35] for which many sources provide values for a few data items and a few sources provide values for many data items. This is modelled using a simple exponential distribution. It associates, with each source, the number of data items on which it has to provide a value. The statistic that confirm that this behaviour is respected by the datasets that were generated are reported in Fig. 3. In Fig. 3(a) we observe that approximately 80% of data items are claimed by less than 500 sources. Fig. 3(b) shows that most of sources have provided at least 1000 data items.

Third, each source claims a true or false value for a specific data item w.r.t. its trustworthiness. In case of true claims, the value is selected among the inclusive ancestors of the value contained in the data item. In the case of false claims, it is selected from the set of values that are neither inclusive ancestors nor descendants of the true one denoted v_d^* . In both cases the values are selected w.r.t. a similarity measure between the values and v_d^* .

For the selection of the true values, three different strategies were adopted: EXP, LOW_E, UNI (Fig. 4). EXP simulates cases in which sources are quite sure about the true values, so they tend to claim values similar to the expected one (contained in the ground truth) when they have to provide a true value. UNI reproduces a world where there is a lot of uncertainty, then the sources tend to indiscriminately select the value from the entire set of possible true values. LOW_E is a trade-off between the previous two types. Sources uniformly select the value from the set of possibilities, but there is a slightly higher probability of choosing values similar to the expected one. For instance, Fig. 4 reports, on the x axis, the values of Fig. 1 ordered according to their similarity measures w.r.t. the true value *Malaga*. Considering that v_d^* is *Malaga* and the EXP law, sources will more often provide values such as *Malaga, Spain* and *City* than values like *Continent* or *Location*. Otherwise, considering the UNI distribution, the probability of claiming these values will be the same.

For the selection of the false values, only a single strategy was considered. A source that has to provide a false claim tends to provide a false value that is similar to the expected true value. For instance, if the true value is *Malaga*, then a source provides the value *Portugal* with an higher probability than the value *Brazil*. Moreover, sources tend to

Table 2
Features of the different partial order structures.

Features	CC	MF	BP	genre	birthPlace
Values	3,984	10,243	28,822	1,838	682,658
≤ depth (max depth)	12	15	16	8	14
Average depth	5.223	5.610	6.906	3.93	5.424
Average # of children	1.451	1.196	1.898	1.041	1.535
Max # of children	466	291	451	824	160,194
Leaves	3,016	8,192	14,797	1,563	663,373

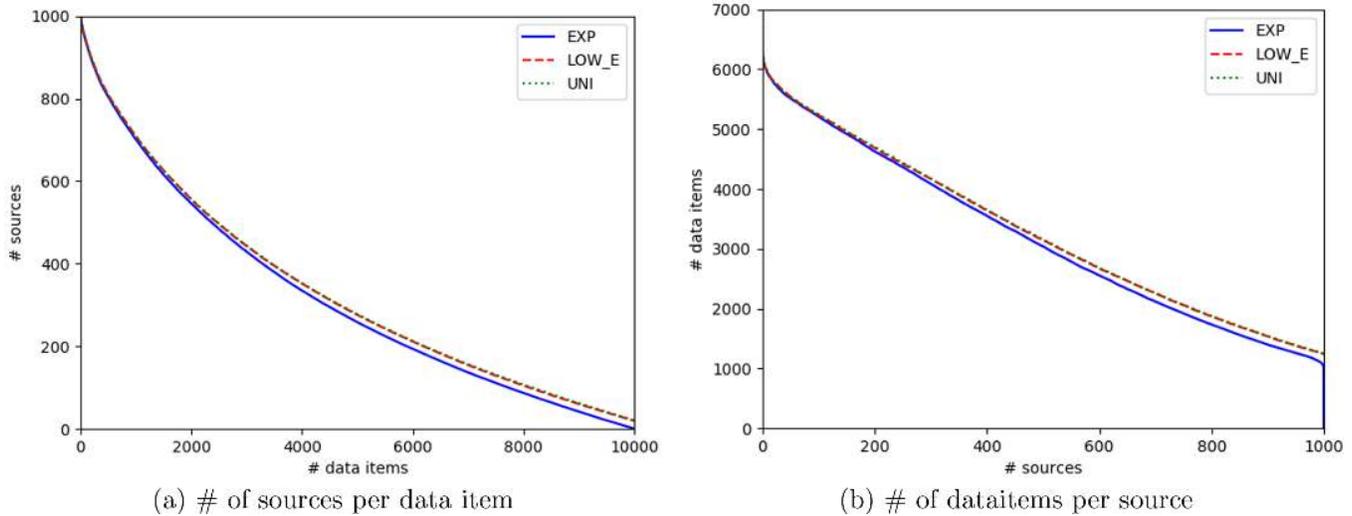


Fig. 3. Statistics of sources-data items for the CC datasets.

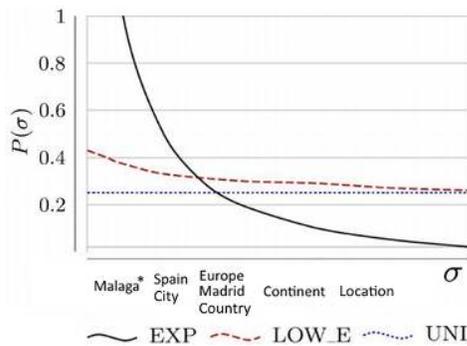


Fig. 4. Distributions used to select true values.

claim the same false values. Therefore, the probability of a value to be selected as false one increases w.r.t. the number of sources that previously claimed it.

For each predicate, 20 datasets were produced w.r.t. the different laws that can be used to select the true values provided by sources. Further details on the generation of the ground truth and how the partial order of values has been derived are provided with the datasets at <http://www.github.com/lgi2p/TDSelection>.

5.2. Experiment settings

In order to provide robust results, considering each predicate, we generated 60 synthetic datasets (20 for each different distribution used to select the true values). Several experiments were conducted on them. Table 3 reports all of the experimental settings in which the datasets were tested. The name associated with each configuration indicates the delta setting. When $\delta = 0$ the approach is called TSbC (Truth Selection of the Best Child). Indeed, the selection algorithm chooses, at each step, the value with the highest confidence. In other words, it selects the best node among the children of the considered one. Otherwise, when $\delta = 1$ the approach is called TSaC (Truth Selection of all Children). Indeed, using this configuration the algorithm selects, at each step, all the children of the considered nodes. Moreover, the subscript specifies the first ranking criteria used, i.e. TSbC_{IC} means that IC is used for the ranking phase as first criteria to order the values. For all the experiments, different threshold θ values were used: 0, 0.1, 0.2, 0.3, 0.4, 0.5. Note that when δ is equal to 0, we test only the property of the solution set indicating that its values share ordering relationships. Indeed, the

Table 3

Set of experiments performed for each predicate.

Conf.	θ	δ	Rank		Filter
			1st	2nd	
TSbC _{trust}	0, ..., 0.5	0	WA _{trust}	IC _{Seco}	V _{ord} [*]
TSbC _{IC}	0, ..., 0.5	0	IC _{Seco}	WA _{trust}	V _{ord} [*]
TSaC _{trust}	0, ..., 0.5	1	WA _{trust}	IC _{Seco}	V _{disj} [*]
TSaC _{IC}	0, ..., 0.5	1	IC _{Seco}	WA _{trust}	V _{disj} [*]

selection procedure in this case chooses, at each step, only values having the highest confidence and therefore only a single most specific value and its ancestors can be returned. No alternatives to the most specific value can be selected. When δ is equal to 1, we test only the property indicating that the values in the solution set do not share a partial order. The procedure may select more than one branch. In this situation, if we force the returned true values to share an ordering relationship, we oblige the algorithm to select only one path. Thus, the main advantage of this configuration, i.e. to propose a set of alternatives, is wasted.

For the confidence and trustworthiness estimations, we initialize the confidence value at 0.5 in order to start the iterative procedure, i.e. AdaptedSums. The stopping criteria used for the iterative procedure is the same as that used in the original paper of Sums [36]: the procedure was stopped after 20 iterations. The algorithms were implemented in Python 3.4. The experiments were performed on a PC with an Intel Core 2 Duo processor (2.93 GHz × 8GB). To give an idea, using the codebase developed for these experiments,⁴ memory consumption varies from 1.6 to 4.3GB depending on the number of values composing the partial order. Using TSbC and TSaC, running times were, respectively, around 0.24 and 1.7 ms per data item. Note that running times may increase significantly when partial orders have specific topological properties. In particular, optimizations have to be studied when dealing with partial orders having values with numerous children (hubs). For instance, the partial order used for the birthPlace predicate contains the value “Settlement” with 160 thousands children; running times using this partial order were 0.02 and 0.8 seconds per data item for TSbC and TSaC respectively. The source code and datasets associated with this study are

⁴ Note that this codebase has been developed for experimental purpose and was not optimized to lower memory consumption and running time.

open-sourced and published on the Web at the following link <http://www.github.com/lgi2p/TDSelection>. Note that for *Sums* and *Adapted-Sums*, we used the code developed by Beretta et al. [10]. Otherwise, for the experiments related to the other existing models, we used the DAFNA-EA⁵ implementation [37]. This API provides the source code for the main existing models.

5.3. Evaluation

The evaluation of the model we proposed to select the true values was carried out using both traditional and hierarchical performance measures of classification problems.

Among traditional metrics, precision and recall were mainly used to compare our approach with the existing models that do not consider the partial order. Our positive class consists of all pairs $(d \in D, v_d^* \in V_d)$ where v_d^* is the value contained in the ground truth for the data item d , and the negative class is composed of all pairs $(d \in D, v_d \in V_d - v_d^*)$. Therefore, the precision is the proportion of pairs (d, v_d^*) returned by the approach among all the pairs it returns. The recall is the proportion of pairs (d, v_d^*) returned by the approach among all pairs contained in the ground truth.

The hierarchical evaluation measures (HEM) were used to analyse the behaviour obtained by different parameter settings of our approach. Indeed, hierarchical metrics distinguish the severity of different errors taking the hierarchy of classes into account. Reasonably if *Malaga* is the true value, then an approach that returns *Portugal* should be less penalized than another that returns *Brazil*. Indeed *Portugal* is in the same continent than *Malaga*, i.e. Europe, while *Brasil* is in a different continent, i.e. America. A detailed study related to hierarchical measures was presented in [38]. They distinguish the main dimensions that characterize hierarchical classification problems and suggest, for each possible combination, which are the best evaluation metrics to use. They recommend F_{LCA} , P_{LCA} , R_{LCA} and $MGIA$ when dealing with single-label problem and DAG hierarchy. This situation corresponds to our initial problem settings: for each data item there is a single expected true value and our partial order among values is represented using a DAG. F_{LCA} , P_{LCA} and R_{LCA} are set-based measures. They use hierarchical relations to augment the sets of returned and true values and to compute precision and recall. Since adding ancestors over-penalize errors that occur to nodes with many of them, F_{LCA} , P_{LCA} , R_{LCA} use the notion of the Lowest Common Ancestor to limit this undesirable effect. $MGIA$ is a pair-based metric that uses graph distance measures to compare returned and true values. Its limitation is that it does not change with depth. For further details related to the computation of these measure please refer to Kosmopoulos et al. [38]. Now, we briefly describe the main characteristics of these hierarchical measures through an illustrative example. This enable the reader to better understand the result discussion in the next section. Considering the DAG in Fig. 1 and *Malaga* as the true value, the HEMs related to several returned values are reported in Table 4. As shown, if the returned value is more general than the expected one, then P_{LCA} is not affected, while R_{LCA} decreases when increasing the distance from the expected value. Otherwise, if the returned value is an error (neither the expected value nor more general one), then P_{LCA} and R_{LCA} decrease w.r.t. the position of the returned value in the partial order. $MGIA$ indicates the distance among the returned value and the expected one without considering if one value is more general or specific than the other.

6. Results

All of the experimental settings presented in Section 5 were tested. Here, the results are presented and discussed. Note that a robust analysis were conducted given the artificial nature of the synthetic

Table 4

Example of HEM considering the DAG in Fig. 1 and *Malaga* as the true value.

Returned value	P_{LCA}	R_{LCA}	F_{LCA}	$MGIA$
Malaga	1	1	1	1
Spain	1	0.5	0.7	0.9
Country	1	0.3	0.5	0.8
Madrid	0.5	0.5	0.5	0.8
France	0.5	0.3	0.4	0.7

datasets.

Results show that our approach enables successfully addressing the problem of selecting true values. Recall that our study considers a setting where value confidence estimations w.r.t. the partial order of values monotonically increases. The most effective configuration settings of our selection procedure were $Tsac_{trust}$ and $TsbC_{IC}$ as shown in Fig. 6. These settings coupled with the *AdaptedSums* model were able to outperform, in terms of recall, existing truth discovery methods on the different datasets and predicates that were used for the experiments, see Fig. 5. Note that in these experiments we compared our post-processing strategies considering $k = 1$ with the other models. Indeed, the general aim of TD is to return a single answer for each data item.

In the rest of this section we detail the comparison of the proposed approach with existing truth discovery models and we study different configuration settings of the post-processing procedure analysing its behaviour considering different k , δ and θ values.

Both $Tsac_{trust}$ and $TsbC_{IC}$ obtained good performance, but $Tsac_{trust}$ was the most robust approach independently of the predicate and dataset type, as shown in Fig. 5. It resulted to be only slightly influenced by source disagreement increase (UNI dataset case). Indeed, $Tsac_{trust}$ aimed to analyse and compare the trustworthiness of sources providing the most specific values that do not share partial order relationships. This was done selecting and returning all provided values higher than θ , i.e. $\delta = 1$. Then ranking the values according to the weighted average trustworthiness of sources claiming them. Finally, filtering the first k values that did not share ordering relationships. Following this post-processing procedure, $Tsac_{trust}$ performance was not affected when the number of sources providing true general values increased (UNI dataset). Precisely, analysing the recall obtained by the different models from EXP to UNI dataset types, we observed that, when increasing sources that provided general true values, $Tsac_{trust}$ had a recall drop equal to 0.073 against a recall drop around 0.528 obtained by existing truth discovery models. Indeed, the average recall, over the different predicates, obtained by $Tsac_{trust}$ was 0.954, 0.912 and 0.881 respectively for EXP, LOW_E and UNI dataset types. The average recall achieved by existing truth discovery models was 0.595, 0.243 and 0.067 respectively for EXP, LOW_E and UNI dataset types.

On the contrary $TsbC_{IC}$ performance was more influenced by source disagreement increase than $Tsac_{trust}$ performance. It is the post-processing strategy that employed the greedy algorithm to select the true value, i.e. at each step the selection phase chooses the values with the highest confidence. Then it ordered them w.r.t. their IC. Finally, it kept only values that shared a partial order. Therefore, it used as selection criterion, at each step, the value confidence. When sources provided more general true values the information associated to these claims were propagated to less values. Thus the confidence estimations were less informative in the last steps of the procedure. Anyway, also $TsbC_{IC}$ outperformed existing methods obtaining recall levels that were equal to 0.889, 0.670 and 0.531 for EXP, LOW_E and UNI dataset respectively (thus with a recall drop of 0.358).

Observing Fig. 5 we analysed for which predicates our approaches, $Tsac_{trust}$ and $TsbC_{IC}$, obtained slightly lower performances. Even in these cases our models stills outperformed existing ones.

Considering $Tsac_{trust}$, the worst recall performance were achieved for *birthPlace* and *BP* predicate. Analysing the features shown in Table 2 related to the different predicate partial order, it is clear that this

⁵ <http://www.github.com/daqcri/DAFNA-EA>.

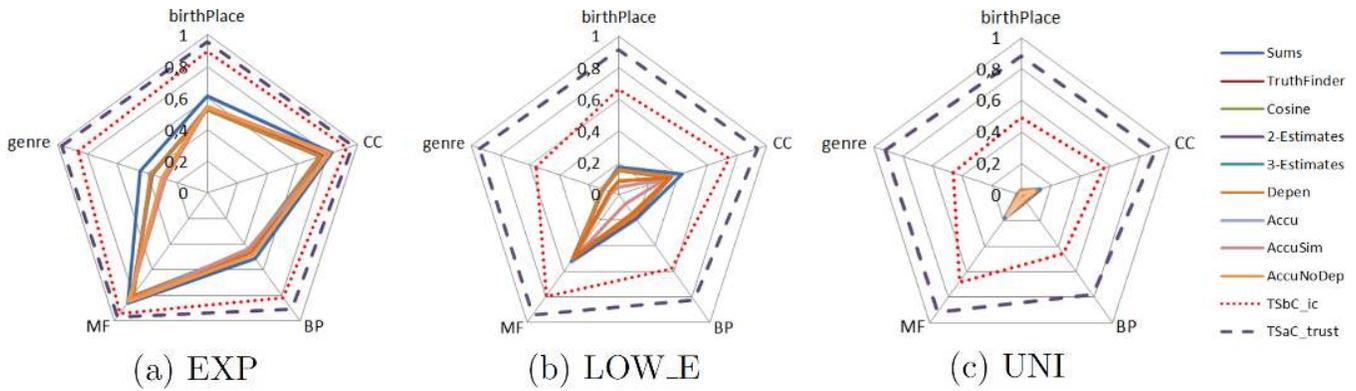


Fig. 5. Recall obtained by applying our approaches $TSbC_{IC}$ (dotted line) and $TSaC_{trust}$ (dashed line), both with $k = 1$ and $\theta = 0$, and the models provided by DAFNA API (solid lines) on the synthetic datasets.

configuration setting was influenced by the average number of children in the partial order. Indeed *birthPlace* and *BP* were the two predicates with the highest children average number. Moreover, the ranking of predicates w.r.t. their recall corresponded to the predicate ranking w.r.t. the children average number in decreasing order.

Otherwise, when considering $TSbC_{IC}$ approach the worst performance in terms of recall were obtained considering *genre* and *BP* predicate. We found out that $TSbC_{IC}$ performance depended both on the children average number and the average depth of expected solutions w.r.t. the maximum depth. Indeed, at each step of $TSbC_{IC}$ the probability of error is related to the number of alternatives among which the procedure can select a value. Moreover, it also related to the percentage of the partial order that the selection procedure has to traverse in order to reach the expected solutions w.r.t. the maximum depth. The probability of error increased when the part of the graph to traverse augmented. For instance *genre* predicate had the lowest children average number, but it obtained performance lower than *MF*, *CC* and *birthPlace* predicate. This because its expected values had a depth that required to traverse a bigger part of the partial order than in the other cases. To better understand the best parametrization for the post-processing procedure several experiments were conducted w.r.t. the different settings reported in Table 3.

First of all, we compared the different post-processing strategies we proposed, evaluating the recall at different levels of k . The results are reported in Fig. 6. Note that we show the results for the predicates *genre* and *MF*, but a similar behaviour was obtained with all the others.

We observe that the best results were obtained by the $TSaC_{trust}$ for any k value. It took advantage of the fact that it returned a set of

alternatives as different as possible from each other and, at the same time, as specific as possible. Usually $TSbC_{IC}$ also outperformed the baseline model (*Sums*), but for higher values of k it was worse than *Sums*. This is because we forced the result of $TSbC_{IC}$ to share ordered relationships, while in the case of *Sums*, k values with the highest confidence were returned (no additional filter was applied on these values). Note that the recall of $TSbC_{IC}$ did not improve when increasing the value of k . This means that a situation in which a returned value is more specific than the expected one never occurs. This is in accordance with the policy we adopted to generate the synthetic datasets. Given the expected value, we cannot say anything about its descendants. Each of them may be a true specification of the expected truth or not. Consequently, we removed all of the descendants from the set of possible true and false values. In other words, no sources provide a claim that contains one of the descendants of the expected value associated with the considered data item. Otherwise, in all the other configurations, increasing the number of values returned (k) enhanced the recall.

The $TSaC_{IC}$ and $TSbC_{trust}$ configurations were for the majority of cases worse than those of the baseline approaches. $TSaC_{IC}$ consists of the selection strategy with $\delta = 1$, i.e. all provided values having confidence higher than θ are selected, and the use of IC as first ranking criterion. It obtained low performance because IC_{Seco} was not a good discriminator among values that did not share ordering relationships. Indeed it is based on the number of descendant values and it may happen in situations in which x is the expected value and y has the same father as x . If x has descendants, while y has none, y will be preferred by the ranking based on the IC_{Seco} even if it is not a true value. Thus, the WA_{trust} ranking is more suitable in these cases.

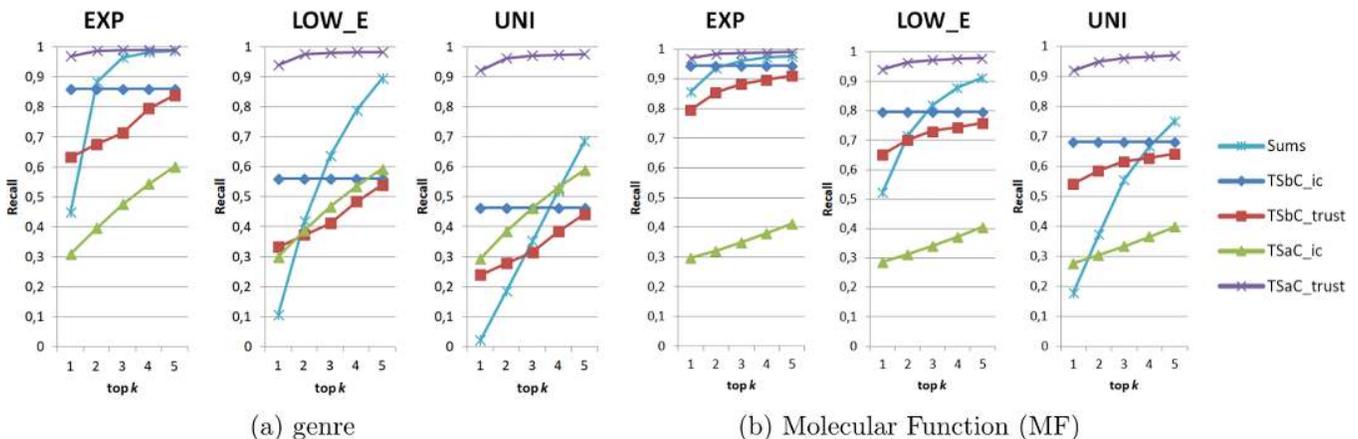


Fig. 6. Recall obtained by applying our approach and the proposed models (with $\theta = 0$) on the synthetic datasets w.r.t. the dataset type and number of returned values.

Table 5
HEM obtained for the different predicates w.r.t. the model and the threshold θ considered.

Predicate	HEM	Model											
		TSbC _{IC}						TSaC _{TRUST}					
		θ						θ					
	0	0.1	0.2	0.3	0.4	0.5	0	0.1	0.2	0.3	0.4	0.5	
CC	F _{LCA}	0.836	0.826	0.770	0.694	0.617	0.561	0.958	0.890	0.783	0.690	0.613	0.560
	P _{LCA}	0.824	0.874	0.943	0.986	0.991	0.988	0.959	0.954	0.967	0.985	0.989	0.989
	R _{LCA}	0.862	0.812	0.693	0.568	0.469	0.407	0.959	0.861	0.701	0.563	0.465	0.406
	MGIA	0.879	0.910	0.907	0.890	0.855	0.818	0.963	0.945	0.917	0.887	0.851	0.816
MF	F _{LCA}	0.878	0.865	0.800	0.697	0.637	0.572	0.962	0.914	0.807	0.695	0.636	0.572
	P _{LCA}	0.870	0.907	0.960	0.990	0.994	0.994	0.964	0.965	0.971	0.989	0.994	0.994
	R _{LCA}	0.898	0.850	0.729	0.568	0.492	0.414	0.963	0.893	0.734	0.567	0.491	0.414
	MGIA	0.909	0.937	0.926	0.892	0.862	0.824	0.966	0.958	0.928	0.890	0.860	0.824
BP	F _{LCA}	0.745	0.689	0.620	0.540	0.484	0.438	0.881	0.725	0.607	0.527	0.477	0.436
	P _{LCA}	0.732	0.859	0.957	0.979	0.976	0.968	0.886	0.935	0.963	0.976	0.974	0.967
	R _{LCA}	0.783	0.624	0.494	0.391	0.335	0.293	0.882	0.642	0.481	0.379	0.329	0.291
	MGIA	0.792	0.853	0.836	0.774	0.707	0.641	0.881	0.865	0.815	0.754	0.696	0.635
birthPlace	F _{LCA}	0.791	0.773	0.709	0.640	0.587	0.532	0.946	0.855	0.713	0.627	0.576	0.530
	P _{LCA}	0.788	0.841	0.936	0.988	0.993	0.990	0.948	0.941	0.953	0.988	0.991	0.989
	R _{LCA}	0.800	0.744	0.601	0.483	0.424	0.372	0.946	0.813	0.602	0.469	0.414	0.369
	MGIA	0.909	0.912	0.897	0.877	0.845	0.807	0.968	0.948	0.900	0.869	0.838	0.805
genre	F _{LCA}	0.784	0.775	0.708	0.657	0.617	0.571	0.963	0.930	0.729	0.657	0.617	0.571
	P _{LCA}	0.781	0.791	0.855	0.979	0.995	0.997	0.966	0.952	0.878	0.980	0.994	0.997
	R _{LCA}	0.793	0.774	0.641	0.505	0.454	0.409	0.962	0.920	0.660	0.505	0.454	0.409
	MGIA	0.903	0.904	0.889	0.887	0.867	0.833	0.974	0.967	0.897	0.887	0.867	0.833

Otherwise TSbC_{trust} is a post-processing strategy with $\delta = 0$, i.e. at each step of the selection process only one value is selected, with the use of source average as ranking criterion. Obtaining low recall for this model means that WA_{trust} was not a good discriminator to rank the values sharing partial order relationships returned by the selection phase.

Moreover, Fig. 6 shows that when disagreement among sources providing true values increased these two latter approaches (TSaC_{IC} and TSbC_{trust}) could be useful anyway. The recall they obtained for $k = 1$ was higher than the recall of Sums model. Therefore in case of high level of disagreement also a not optimal procedure can be advantageous.

As expected, in all the cases, the precision always decrease when increasing k . Moreover, comparing the different settings of the proposed approach, we observed that the ranking based on their precision performances was the same that the one obtained w.r.t. their recall. Therefore, we omit these repetitive results.

Our further analysis focused on models TSaC_{trust} and TSbC_{IC} since they were the models among the proposed ones that achieved the best performances. We examined the impact of different threshold values, setting $k = 1$, w.r.t. the hierarchical evaluation metrics: F_{LCA}, P_{LCA}, R_{LCA} and MGIA.

The results are reported in Table 5. Considering TSbC_{IC}, we noticed that, when slightly increasing θ , MGIA increased in the majority of the cases. This occurred because there are expected values (supported by few reliable sources) with a confidence lower than false ones (supported by many unreliable sources), even though the former have a higher WA_{trust} than the latter. Thus, using TSbC_{IC} and $\theta = 0$, these values were selected as true values. Increasing θ allows the procedure to avoid a part of these errors. Indeed, eliminating the values with confidence score very low enables the procedure to return, with high probability,

the father of the expected value. Anyway, further increasing the threshold caused a loss of MGIA because the returned values result to be very general. This does not happen with TSaC_{trust} since this kind of errors are already overcome considering WA_{trust} as first ranking criterion.

Moreover, we observed that, in the majority of cases, when increasing θ the R_{LCA} always decreased, while the P_{LCA} always increased. Precisely, the highest R_{LCA} for both TSaC_{trust} and TSbC_{IC} was obtained with $\theta = 0$. The highest P_{LCA} was obtained for both approaches with different θ values depending on the predicate as shown in Table 5.

Summarising, the most effective configuration settings were TSaC_{trust} and TSbC_{IC}. They were both able to obtain better performance than existing truth discovery models. We noted that increasing the number of values returned for each data item allow increasing the performance. Nevertheless this can be applied only in the case where a group of experts can select the true values among the ones proposed by the proposed approach for each data item. Otherwise, we have to force the parametrization $k = 1$. Regarding the threshold θ , a high θ value is recommended when the application scenario does not permit to assume many risks. In this case it is important to have a high precision. In other words, obtaining a general true value than the specific false one is preferred. Therefore, the different parameter settings of the proposed post-processing procedure allow dealing with different application scenarios taking their requirements into account.

7. Conclusion

In this paper, we have presented a post-processing procedure able to select true values after estimation of the value confidences using the AdaptedSums approach we proposed in our previous work. This general

procedure can be used with any TD approach when partial order of values is taken into account as *a priori* knowledge. The post-processing process involves three main steps. The first one consists of the selection procedure. It aims to identify the set of possible true values using relationships among them and includes two parameters (δ and θ). Based on their tuning, different behaviours of the selection process can be obtained. The second step ranks the returned values of the selection phase. Finally, the third step permits to filter the top k values and ensure desirable properties (values that share or not relationships). The results confirmed our preliminary finding: using partial ordering of values helps to improve both source trustworthiness estimation, as already demonstrated by our preliminary study [10], and the true value identification. More precisely, the best results are obtained with the configuration of the algorithm that selects a set of alternatives, not sharing ordering relationships, and ranks them through the average trustworthiness of sources claiming those values. The results showed a similar behaviour on the datasets obtained by the two different ontologies (DBpedia and Gene Ontology).

As prospects, we envisage to incorporate our framework by adapting it to another existing model. Indeed we would like to show the flexibility of our approach and further enhance the results. Moreover, we intend to explore other kinds of additional information such as correlations among data items and values. For instance, usually the birth location of people is correlated with the language they speak. Therefore, if we know that a person speaks Italian, we can increase the confidence in those claims that contain Italy as value for the bornIn predicate. More precisely, we plan to design a model that integrates, into existing approaches, information extracted from external knowledge bases in the form of rules. The idea is to add, in the confidence formula, a boosting factor indicating the confidence level of each claim according to the external knowledge base.

References

- [1] J. Bleiholder, F. Naumann, Data fusion, *ACM Comput. Surv. (CSUR)* 41 (1) (2009) 1.
- [2] C. Li, V.S. Sheng, L. Jiang, H. Li, Noise filtering to improve data and model quality for crowdsourcing, *Knowl. Based Syst.* 107 (Supplement C) (2016) 96–103. [10.1016/j.knsys.2016.06.003](https://doi.org/10.1016/j.knsys.2016.06.003).
- [3] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, J. Han, A survey on truth discovery, *SIGKDD Explorations* 17 (2) (2015) 1–16.
- [4] T. Knap, J. Michelfeit, M. Necaský, Linked open data aggregation: conflict resolution and aggregate quality, *Computer Software and Applications Conference Workshops (COMPSACW)*, 2012 IEEE 36th Annual, (2012), pp. 106–111.
- [5] A. Guzman-Arenas, A.-D. Cuevas, A. Jimenez, The centroid or consensus of a set of objects with qualitative attributes, *Expert Syst. Appl.* 38 (5) (2011) 4908–4919, <https://doi.org/10.1016/j.eswa.2010.09.169>.
- [6] L. Berti-Equille, J. Borge-Holthoefer, Veracity of data: from truth discovery computation algorithms to models of misinformation dynamics, *Synth. Lect. Data Manage.* 7 (3) (2015) 1–155.
- [7] D. Wang, T. Abdelzaher, L. Kaplan, *Social Sensing: Building Reliable Systems on Unreliable Data*, Morgan Kaufmann, 2015.
- [8] G. Zhou, J. Zhao, T. He, W. Wu, An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities, *Knowl. Based Syst.* 66 (2014) 136–145, <https://doi.org/10.1016/j.knsys.2014.04.032>.
- [9] B. Khaleghi, A. Khamis, F.O. Karray, S.N. Razavi, Multisensor data fusion: a review of the state-of-the-art, *Inf. Fusion* 14 (1) (2013) 28–44, <https://doi.org/10.1016/j.inffus.2011.08.001>.
- [10] V. Beretta, S. Harispe, S. Ranwez, I. Mougenot, How can ontologies give you clue for truth-discovery? an exploratory study, *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, WIMS 2016, Nîmes, France, June 13–15, (2016)*.
- [11] X. Yin, J. Han, S.Y. Philip, Truth discovery with multiple conflicting information providers on the web, *IEEE Trans. Knowl. Data Eng.* 20 (6) (2008) 796–808.
- [12] R.W. Ouyang, L.M. Kaplan, A. Toniolo, M. Srivastava, T.J. Norman, Aggregating crowdsourced quantitative claims: additive and multiplicative models, *IEEE Trans. Knowl. Data Eng.* 28 (7) (2016) 1621–1634.
- [13] R.W. Ouyang, M. Srivastava, A. Toniolo, T.J. Norman, Truth discovery in crowd-sourced detection of spatial events, *IEEE Trans. Knowl. Data Eng.* 28 (4) (2016) 1047–1060.
- [14] J. Pasternack, D. Roth, Latent credibility analysis, *Proceedings of the 22nd International Conference on World Wide Web, ACM, 2013*, pp. 1009–1020.
- [15] B. Zhao, B.I. Rubinstein, J. Gemmell, J. Han, A bayesian approach to discovering truth from conflicting sources for data integration, *Proc. VLDB Endowment* 5 (6) (2012) 550–561.
- [16] X. Yin, W. Tan, Semi-supervised truth discovery, *Proceedings of the 20th International Conference on World wide web, (2011)*, pp. 217–226.
- [17] R. Pochampally, A. Das Sarma, X.L. Dong, A. Meliou, D. Srivastava, Fusing data with correlations, *Proceedings of the 2014 ACM SIGMOD International Conference on Management of data - SIGMOD '14, (2014)*, pp. 433–444. [arXiv:1503.00306](https://arxiv.org/abs/1503.00306), [doi:10.1145/2588555.2593674](https://doi.org/10.1145/2588555.2593674).
- [18] G.-J. Qi, C.C. Aggarwal, J. Han, T. Huang, Mining collective intelligence in diverse groups, *Proceedings of the 22nd International Conference on World Wide Web, WWW '13, (2013)*, pp. 1041–1052.
- [19] X. Wang, Q.Z. Sheng, X.S. Fang, L. Yao, X. Xu, X. Li, An integrated bayesian approach for effective multi-truth discovery, *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM '15, (2015)*, pp. 493–502.
- [20] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, Y. Cheng, Truth discovery on crowd sensing of correlated entities, *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, SenSys '15, (2015)*, pp. 169–182.
- [21] S. Wang, L. Su, S. Li, S. Hu, T. Amin, H. Wang, S. Yao, L. Kaplan, T. Abdelzaher, Scalable social sensing of interdependent phenomena, *Proceedings of the 14th International Conference on Information Processing in Sensor Networks, IPSN '15, (2015)*, pp. 202–213.
- [22] X.L. Dong, L. Berti-Equille, D. Srivastava, Integrating conflicting data: the role of source dependence, *Proc. VLDB Endowment* 2 (1) (2009) 550–561.
- [23] L. Berti-Equille, A.D. Sarma, X. Dong, A. Marian, D. Srivastava, Sailing the information ocean with awareness of currents: discovery and application of source dependence, *CIDR* (2009). [arXiv:0909.1776](https://arxiv.org/abs/0909.1776).
- [24] X.L. Dong, L. Berti-Equille, D. Srivastava, Truth discovery and copying detection in a dynamic world, *Vldb* 2 (1) (2009) 562–573.
- [25] D. Wang, T. Abdelzaher, L. Kaplan, R. Ganti, S. Hu, H. Liu, Exploitation of physical constraints for reliable social sensing, *Proceedings of the 2013 IEEE 34th Real-Time Systems Symposium, RTSS '13, (2013)*, pp. 212–223.
- [26] S. Wang, D. Wang, L. Su, L. Kaplan, T.F. Abdelzaher, Towards cyber-physical systems in social spaces: the data reliability challenge, *Real-Time Systems Symposium (RTSS)*, 2014 IEEE, (2014), pp. 74–85.
- [27] A. Bronselaer, M. Szymczak, S. Zadrozny, G.D. Trè, Dynamical order construction in data fusion, *Inf. Fusion* 27 (Supplement C) (2016) 1–18, <https://doi.org/10.1016/j.inffus.2015.05.001>.
- [28] D.C. Kozen, *The Design and Analysis of Algorithms*, Springer-Verlag New York, Inc., 1992.
- [29] A.V. Aho, M.R. Garey, J.D. Ullman, The transitive reduction of a directed graph, *SIAM J. Comput.* 1 (2) (1972) 131–137.
- [30] N. Seco, T. Veale, J. Hayes, An intrinsic information content metric for semantic similarity in wordnet, *Proceedings of the 16th European Conference on Artificial Intelligence, (2004)*, pp. 1089–1090.
- [31] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, Semantic similarity from natural language and ontology analysis, *Synth. Lect. Human Lang. Technol.* 8 (1) (2015) 1–254.
- [32] P.-A. Jean, S. Harispe, S. Ranwez, P. Bellot, J. Montmain, Uncertainty detection in natural language: a probabilistic model, *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, WIMS 2016, Nîmes, France, June 13–15, (2016)*.
- [33] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, ISWC'07/ ASWC'07, Springer-Verlag, Berlin, Heidelberg, 2007*.
- [34] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., Gene ontology: tool for the unification of biology, *Nat. Genet.* 25 (1) (2000) 25–29.
- [35] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, J. Han, A confidence-aware approach for truth discovery on long-tail data, *Proc. VLDB Endowment* 8 (4) (2014) 425–436.
- [36] J. Pasternack, D. Roth, Knowing what to believe (when you already know something), *Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10, (2010)*, pp. 877–885.
- [37] D.A. Waguih, L. Berti-Equille, Truth discovery algorithms: an experimental evaluation, *CoRR* (2014). [abs/1409.6428](https://arxiv.org/abs/1409.6428)
- [38] A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras, I. Androutsopoulos, Evaluation measures for hierarchical classification: a unified view and novel approaches, *Data Min. Knowl. Discov.* 29 (3) (2015) 820–865.

Selecting Relevant Association Rules From Imperfect Data

Cécile L'Héritier^{1,2}, Sébastien Harispe¹, Abdelhak Imoussaten¹,
Gilles Dusserre¹, and Benoît Roig²

¹ LGI2P, IMT Mines Ales, Univ Montpellier, Alès, France
{firstname.name}@mines-ales.fr

² EA7352 CHROME, Université de Nîmes, France
{firstname.name}@unimes.fr

Abstract. Association Rule Mining (ARM) in the context of imperfect data (*e.g.* imprecise data) has received little attention so far despite the prevalence of such data in a wide range of real-world applications. In this work, we present an ARM approach that can be used to handle imprecise data and derive imprecise rules. Based on evidence theory and Multiple Criteria Decision Analysis, the proposed approach relies on a selection procedure for identifying the most relevant rules while considering information characterizing their interestingness. The several measures of interestingness defined for comparing the rules as well as the selection procedure are presented. We also show how *a priori* knowledge about attribute values defined into domain taxonomies can be used to (i) ease the mining process, and to (ii) help identifying relevant rules for a domain of interest. Our approach is illustrated using a concrete simplified case study related to humanitarian projects analysis.

Keywords: Association rules · Imperfect data · Evidence theory · Multiple Criteria Decision Analysis (MCDA).

1 Introduction

Association rule mining (ARM) is a well-known data mining technique designed to extract interesting patterns in databases. It has been introduced in the context of market basket analysis [1], and has received a lot of attention since then [15]. An association rule is usually formally defined as an implication between an *antecedent* and a *consequent*, being conjunctions of attributes in a database, *e.g.* “People who have age-group between 20 and 30 and a monthly income greater than \$2k are likely to buy product X”. Such rules are interesting for extracting simple intelligible knowledge from a database; they can also further be used in several applications, *e.g.* recommendation, customer or patient analysis. A large literature is dedicated to the study of ARM, and numerous algorithms have been defined for efficiently extracting rules handling a large range of data types, *e.g.*, nominal, ordinal, quantitative, sequential [15]. Nevertheless, only a few contributions of the literature study the case of ARM with imperfect data, *e.g.* [13, 24], even if such data is central in numerous real-world applications.

In order to extend the body of work related to ARM with imperfect data, and to answer some of the limitations of existing contributions, this paper presents a novel ARM approach that can be used to handle imprecise data and derive imprecise rules. In this study, to simplify, the proposed approach focuses on a specific case where the *antecedent* and the *consequent* are composed of predefined disjoint sets of attributes forming a partition of the whole set of attributes. This particular case is relevant, for example in classification tasks in which the label value to predict can be defined as consequent of the rules of interest. To sum up, our goal is threefold: (i) to enrich the expressivity of existing proposed frameworks, (ii) to complement them with a richer procedure for selecting relevant rules, and (iii) to present simple way to incorporate domain knowledge to ease the mining process, and to help identifying relevant rules for a domain of interest. Based on the evidence theory framework and Multiple Criteria Decision Analysis, a selection procedure for identifying the most relevant rules while considering information characterizing their interestingness is proposed. The several measures of interestingness defined for comparing the rules, as well as the selection procedure, are presented. We also show how *a priori* knowledge in the form of taxonomies about consequent and antecedent (i.e. attribute values) can be used to focus on rules of interest for a domain. We also present an illustration using a simplified case study related to humanitarian projects analysis.

The paper is structured as follows: Section 2 formally introduces traditional ARM, the theoretical notions on which our approach is based, and formally defines the problem we are considering. It also introduces related work focusing on rule selection and ARM with imperfect data. The proposed approach is detailed in Section 3, and Section 4 presents the illustration. Finally, perspectives and concluding remarks are provided in Section 5.

2 Theoretical background and related work

This section briefly presents some of the theoretical notions required to introduce our work. We next provide the problem statement of ARM with imperfect data, and our positioning w.r.t. existing contributions.

2.1 Theoretical background

Association Rule Mining (ARM): In classical ARM [1], a database $\mathcal{D} = \{d_1, \dots, d_m\}$ to be mined consists of m observations of a set of n attributes. The set of attribute indices is denoted by $N = \{1, \dots, n\}$. Each attribute i takes its values in a discrete -boolean, nominal or numerical- finite scale denoted Θ_i . An association rule r denoted $r : X \rightarrow Y$ links an antecedent X with a consequent Y where $X \in \prod_{i \in I} \Theta_i$, $I \subset N$ and $Y \in \prod_{j \in J} \Theta_j$, $J \subseteq N \setminus I$.

The main challenge in ARM is to extract *interesting* rules from a large search space, *e.g.*, n and m are large. In this context, defining the *interestingness* of a rule is central.

Interestingness of rules. Numerous works have studied notions related to the *interestingness* of a rule, [16, 22, 23]. No formal and widely accepted definition arose from those works, and discussing the numerous existing formulations is out of the scope of this paper. However, interestingness is generally regarded as a general concept covering several features of interest for a rule, e.g. *reliability* (how reliable is the rule?) and *conciseness* (is the rule complex?, i.e. based on numerous attribute-value pairs). Other aspects of a rule are also considered, e.g. *peculiarity*, *surprisingness*, or *actionability*, to name a few - the reader can refer to [12] for details. The literature also distinguishes objective and subjective measures, the latter being defined based on domain-dependent considerations. The two main (objective) measures used in the literature are *Support* and *Confidence* [2]. The *support* of a rule $r : X \rightarrow Y$ denoted $supp(X \rightarrow Y)$ is traditionally defined as the proportion of the realization of X and Y in \mathcal{D} , and the *confidence* denoted $conf(X \rightarrow Y)$ is defined as the proportion of the realization of Y when X is observed in \mathcal{D} . Given support and confidence thresholds, ARM usually aims at identifying rules exceeding those thresholds [2]. In classical ARM, support and confidence are quantified using probability theory framework. When ARM involves imperfect data, this quantification requires reformulating the problem in a theoretical framework suited for handling data imperfection. In this work, we focus on contributions based on evidence theory.

Evidence theory has been introduced to represent imprecision and uncertainty [21]. We briefly introduce its main concepts. Let Θ be a finite set of elements being the most precise available information, referred to as the *frame of discernment*. A *mass function* $m : 2^\Theta \rightarrow [0, 1]$ is a set function such that $\sum_{A \subseteq \Theta} m(A) = 1$. The quantity $m(A)$, $A \subseteq \Theta$ is interpreted as the portion of belief that is exactly committed to A and to nothing smaller. The subsets of Θ having a strictly positive mass are called *focal elements*, their set is denoted \mathcal{F} . The total belief committed to any $A \subseteq \Theta$ is measured by the *belief function*: $Bel : 2^\Theta \rightarrow [0, 1]$ with $Bel(A) = \sum_{B \subseteq \Theta, B \subseteq A} m(B)$. In evidence theory, $Bel(\bar{A})$, where \bar{A} denotes the complement of A in Θ , is characterized through the notion of *plausibility*: $Pl : 2^\Theta \rightarrow [0, 1]$, with $Pl(A) = 1 - Bel(\bar{A}) = \sum_{B \subseteq \Theta, B \cap A \neq \emptyset} m(B)$.

In order to provide a complete generalization of the probability framework, conditioning has also been defined in evidence theory. Several expressions have been proposed, none of them leading to a full consensus [7, 10]. In this paper, we will adopt the definition corresponding to the conditioning process stated by Fagin et al. [10], a natural extension of the Bayesian conditioning. We do not consider the definition proposed in Dempster [7] based on Dempster-Shafer combination rule, where a new information is interpreted as a modification of the initial belief function and used in a revision process [9]. Thus, for $A, B \subseteq \Theta$, such that $Bel(A) > 0$, we will further consider:

$$Bel(B|A) = \frac{Bel(A \cap B)}{Bel(A \cap B) + Pl(A \cap \bar{B})}, \quad Pl(B|A) = \frac{Pl(A \cap B)}{Pl(A \cap B) + Bel(A \cap \bar{B})}$$

4 C. L'Héritier et al.

2.2 Problem statement and related work

Problem statement. In classical ARM, where only precise information is considered, *e.g.*, the value of attribute i is $X_i \in \Theta_i, i \in N$. In this paper, we consider observations as “the value of attribute i is in $A_i \subseteq \Theta_i$ ”. The case $A_i \subset \Theta_i$ with $|A_i| > 1$ corresponds to imprecision, while $A_i = \Theta_i$ is considered when information is missing, *i.e.* it corresponds to the ignorance about the value of attribute i . In this setting, a rule r is defined as:

$$r : A \rightarrow B \text{ where } A = \prod_{i \in I} A_i, A_i \subseteq \Theta_i \text{ and } B = \prod_{j \in J} B_j, B_j \subseteq \Theta_j \\ \text{for all } I \subset N \text{ and } J \subseteq N \setminus I$$

As mentioned previously, in this paper we consider the case where antecedent A concerns only a subset $I_1 \subset N$ of attributes and consequent B concerns a subset $I_2 \subset N$ where I_1 and I_2 form partition of N , and $I_1 \neq \emptyset$. Thus:

$$r : A \rightarrow B \text{ where } A = \prod_{i \in I_1} A_i, A_i \subseteq \Theta_i \text{ and } B = \prod_{j \in I_2} B_j, B_j \subseteq \Theta_j \quad (1)$$

We denote by \mathcal{R} the set of rules defined by Formula (1). The problem addressed here is to reduce \mathcal{R} by selecting only the relevant rules.

Related work and positioning. As stated in the introduction, our goal is threefold: (i) to enrich the expressivity of existing proposed frameworks dedicated to ARM with imperfect data, (ii) to complement them with a richer procedure for selecting relevant rules (rule pruning), and (iii) to present a simple way to incorporate domain knowledge to ease the mining process, and to help identifying relevant rules for a domain of interest.³

Rule pruning. Most of the approaches use thresholds to select rules - only using support and confidence most often allows drastically reducing the number of rules in traditional ARM [1]. A post-mining step is generally performed to rank the remaining rules according to one specific interestingness measure -the measure used is generally selected according to the application domain and context-specific measure properties [23, 27]. Nevertheless, processing this way does not enable selecting rules when conflicting interestingness measures are used, *e.g.* maximizing both support and specificity of rules. This is the purpose of MCDA methods. Some works propose to take advantage of MCDA methods [3–6, 17] in the context of ARM. Those works can be divided into two categories: 1) those incorporating the end-user’s preferences using Analytic Hierarchy Process (AHP) and Electre II [6], or using Electre tri [3]; and 2) those that do not incorporate such information and use Data Envelopment Analysis (DEA) [5, 26], or Choquet

³ Note that the simplification of the mining process here refers to a reduction of complexity in terms of the number of rules analysed, *i.e.* search space size. Algorithmic contributions and therefore complexity analyses regarding efficient implementations of the proposed approach are left for future work.

integral [17]. Our approach is hybrid and falls within the two categories. First, selection is made based only on database information as in Bouker et al. [4]. Second, if the set of selected rules is large, a trade-off based on end-user's preferences is used within an appropriate MCDA method. As our aim is to select a subset of interesting rules, Electre I [18] seems to be the most appropriate.

ARM and imperfect data. Several frameworks have been studied to deal with imperfect data in ARM. The assumptions entailed in the approaches based on probabilistic models do not preserve imprecision and might lead to unreliable inferences [13]. Uncertainty theories have also been investigated for imperfect data in ARM using fuzzy logic [14], or using possibility theory [8]. In the case of missing and incomplete data, evidential theory seems the appropriate setting to handle ARM problem [13, 19, 24, 25]. Our approach is adopting this setting. In addition to studying a richer modelling that enables incorporating more information, we propose to combine it with a selection process taking advantage of an MCDA method, namely Electre I, to assess rules interestingness considering different viewpoints. Although some works previously mentioned tackle rule selection using MCDA, and few approaches have been addressing ARM problem using evidence theory, none of them is addressing both issues simultaneously.

We also present how to benefit from *a priori* knowledge about attribute values -organised into taxonomies- for improving the rule selection process, and reducing the increase of complexity induced by the proposed extension of modellings used so far in existing ARM approaches suited for imperfect data.

3 Proposed approach

This section presents our ARM approach for imperfect data. We first introduce how rule interestingness is evaluated by presenting the selected measures and their formalization in the evidence theory framework. Then, the main steps of the proposed approach for selecting rules based on these measures are detailed.

3.1 Assessing rule interestingness from imprecise data

In this study, we focus on important objective measures of interestingness - subjective ones, involving further interactions with final user, are most often considered context-dependent and will not be considered in this paper. We propose to evaluate rules according to (i) their support, (ii) their confidence, as well as (iii) indirect evaluations used to criticize their potential relevance. In addition, since in our context rules are imprecise, and since very imprecise rules are most often considered useless, the (iv) degree of imprecision embedded in the mined rules is also evaluated. These four notions of interest considered in the study are defined below. For convenience, we consider that we are computing measures to evaluate a rule $r : A \rightarrow B$ where $A = \prod_{i \in I_1} A_i, A_i \subseteq \Theta_i$ and $B = \prod_{j \in I_2} B_j, B_j \subseteq \Theta_j$ with $I_1 \cup I_2 = N$. In our context, since we consider $n = |N|$ attributes, the

6 C. L'Héritier et al.

set functions mass m , belief Bel and plausibility Pl are defined on subsets of $\Theta = \prod_{i \in N} \Theta_i$.

Support. A rule is said to be supported if observations of its realization are frequent [2]. In our context, the support of a rule relates to the masses of evidence associated to observations supporting the rule, either explicitly or implicitly. The belief function is thus used to express support:

$$supp(r : A \rightarrow B) = Bel(A \times B) \quad (2)$$

Note that the belief function is monotone, then, the rules composed of the most imprecise attribute values will necessarily be the most supported.

Confidence. A rule is said to be reliable if the relationship described by the rule is verified in a sufficiently great number of applicable cases [12]. The *Confidence* measure is traditionally evaluated as a conditional probability [1]. Its natural counterpart in evidence theory is given by the conditional belief, leading to the following expression:

$$conf(r : A \rightarrow B) = Bel(B | A) = \frac{Bel(A \times B)}{Bel(A \times B) + Pl(A \times \bar{B})} \quad (3)$$

The elements defining the consequent are conditioned to the elements composing the antecedent. Note that the belief and conditional belief functions have also been adopted to express support and confidence for ARM with imprecise data [13, 24]. In those cases the modelling and domain definition were different, i.e. restricted to the cartesian products of the power-sets of attribute domains.

Indirect measures of potential relevance. These measures will be introduced through an illustration. Consider humanitarian projects described by two attributes: the *transport means* with $\Theta_1 = \{truck, motorbike, helicopter\}$, and the final *coverage reached* in the project (proportion of beneficiaries), with $\Theta_2 = \{low, moderate, high\}$. To criticize the relevance of a rule $r : A \rightarrow B$, e.g. $r : \{truck\} \rightarrow \{high\}$, we propose to evaluate the following relations:

- $A \rightarrow \bar{B}$. In the example, if the rule $\{truck\} \rightarrow \{\overline{high}\}$ holds, it means that most often using *trucks* also leads to a *coverage* that is *not high*. Hence we consider that validating $A \rightarrow \bar{B}$ conveys a contradictory information w.r.t. to the rule $A \rightarrow B$ and tends to invalidate it.
- $\bar{A} \rightarrow B$. If the rule $\{\overline{truck}\} \rightarrow \{high\}$ holds, it means that in some cases, some of the *other means of transport* also allow to reach a *high coverage*. Such an information tends to decrease the interest of the rule $r : A \rightarrow B$ if we assume that B is not explained by multiple causes.
- $\bar{A} \rightarrow \bar{B}$. The rule $\{\overline{truck}\} \rightarrow \{\overline{high}\}$ means that when *trucks* are not used, a *low or moderate coverage* (not high) is obtained. We assume that most commonly, if $\{truck\} \rightarrow \{high\}$ is somehow assumed to be considered as valid, supporting $\{truck\} \rightarrow \{high\}$ will reinforce our interest over $\{truck\} \rightarrow \{high\}$.

In a probabilistic framework, only the relationship $\bar{A} \rightarrow \bar{B}$ would have to be studied, since the other ones do not provide additional information, i.e. $P(\bar{B}|A) = 1 - P(B|A)$, $P(B|\bar{A}) = 1 - P(\bar{B}|\bar{A})$, $P(A \times \bar{B}) = P(A)P(\bar{B}|A)$ and $P(\bar{A} \times B) = (1 - P(A))P(B|\bar{A})$. Thus, the potential relevance of a rule takes into consideration the confidence of the rule composed of the complements of the antecedent and the consequent, given by: $P(\bar{B}|\bar{A})$. Note that, in the literature, this measure is also referred to as *specificity*. When considering evidence theory, the information about the complement is provided by the plausibility function, such as $Bel(A) = 1 - Pl(\bar{A})$ and then $Bel(B|A) = 1 - Pl(\bar{B}|A)$. In this context, Table 1 introduces the relationships between the confidence of a rule (conditional belief) and the ones involving the complement of its antecedent and/or consequent.

Note that to criticize the relevance of a rule using the three rules involving its complements, we propose to consider their respective *support* and *confidence*: criticizing a rule on the basis of weakly supported rules would not be appropriate.

Table 1. Relationships between support and confidence of a rule $r : A \rightarrow B$ and rules involving its complements.

Rule	Support	Confidence	depends on quantities:
$A \rightarrow B$	$Bel(A \times B)$	$Bel(B A)$	$Bel(A \times B)$ and $Pl(A \times \bar{B})$
$A \rightarrow \bar{B}$	$Bel(A \times \bar{B})$	$Bel(\bar{B} A) = 1 - Pl(B A)$	$Bel(A \times \bar{B})$ and $Pl(A \times B)$
$\bar{A} \rightarrow B$	$Bel(\bar{A} \times B)$	$Bel(B \bar{A}) = 1 - Pl(\bar{B} \bar{A})$	$Bel(\bar{A} \times B)$ and $Pl(\bar{A} \times \bar{B})$
$\bar{A} \rightarrow \bar{B}$	$Bel(\bar{A} \times \bar{B})$	$Bel(\bar{B} \bar{A})$	$Bel(\bar{A} \times \bar{B})$ and $Pl(\bar{A} \times B)$

Specificity using Information Content. Finally, we propose to incorporate the specificity of a rule. Let's consider the information "the value of attribute i is in the subset A_i ". This information is more specific than the information "the value of attribute i is in the subset A'_i " where $A_i \subset A'_i$. Based on the notion of Information Content (IC) defined for comparing concept specificities in ontologies [20], we propose to quantify the specificity of a rule r by:

$$IC(r : A \rightarrow B) = 1 - \frac{\log |\{X : X \subseteq A \times B\}|}{|\Theta|} \quad (4)$$

$|X|$ denotes the number of elements in the set X and $\Theta = \prod_{i \in N} \Theta_i$.

3.2 Search space reduction

Let us remind the starting set \mathcal{R} -see Formula (1)- of rules from which a small subset \mathcal{R}^* of interesting rules should be selected:

$$\mathcal{R} = \{r : A \rightarrow B \mid A = \prod_{i \in I_1} A_i, A_i \subseteq \Theta_i, B = \prod_{j \in I_2} B_j, B_j \subseteq \Theta_j\}$$

We assume that I_1 and I_2 are fixed before starting the ARM process.

To simplify notations in the rest of the paper, we will denote by $r_{A,B}$ the rule $r : A \rightarrow B$ where A and B are as in the Formula (1). Two restrictions are proposed below:

8 C. L'Héritier et al.

1. All rules being supported are generalizations (supersets) of focal elements \mathcal{F} , i.e. $\mathcal{F} = \{X : X \subseteq \Theta, m(X) > 0\}$. Since support is a prerequisite for assessing rule validity, we further consider that the evaluation will be restricted to the set:

$$\mathcal{R}_r = \{r_{A,B} \in \mathcal{R} \mid \exists X \in \mathcal{F} \text{ st. } X \subseteq A \times B\}$$

2. The search space can also be reduced using prior knowledge defined into ontologies expressing taxonomies of attribute values. Since the ontology defines the concepts of interest for a domain, a restriction can be performed only considering the attribute values defined into taxonomies. Thus, for each $i \in N$, only a subset \mathcal{O}_i of 2^{Θ_i} of the information of interest for a domain is considered. We can then define the following restriction:

$$\mathcal{R}_{r,t} = \{r_{A,B} \in \mathcal{R}_r \mid A = \prod_{i \in I_1} A_i, A_i \in \mathcal{O}_i, B = \prod_{j \in I_2} B_j, B_j \in \mathcal{O}_j\}$$

3.3 Rules selection process

The proposed approach aims at selecting the most relevant rules \mathcal{R}^* according to their evaluations on a set of interestingness measures listed in Table 2. We here consider that the evaluated rules are members of the restriction $\mathcal{R}_{r,t} \subseteq \mathcal{R}$, even if that condition could further be relaxed. We denote the set of interestingness measures by K ($|K| = 9$), and $g_k(r)$ the score of rule r for the measure $k \in K$. To simplify notations, we consider that $g_k(r)$ is to maximize⁴ for all $k \in K$. A two-step pruning strategy is proposed.

Table 2. Summary of interestingness measures considered in the selection process

$k \in K$	Measures	Formulae $\forall r \in \mathcal{R}_{r,t} r : A \rightarrow B$	variation	weight
1	Rule Support	$supp(r) = Bel(A \times B)$	maximize	w_1
2	Rule Confidence	$conf(r) = Bel(B A)$	maximize	w_2
3	Rule Specificity	$IC(r)$	maximize	w_3
4	$A \rightarrow \bar{B}$	$Bel(A \times \bar{B})$	minimize	w_4
5		$Bel(\bar{B} A)$	minimize	w_5
6	$\bar{A} \rightarrow B$	$Bel(\bar{A} \times B)$	minimize	w_6
7		$Bel(B \bar{A})$	minimize	w_7
8	$\bar{A} \rightarrow \bar{B}$	$Bel(\bar{A} \times \bar{B})$	maximize	w_8
9		$Bel(\bar{B} \bar{A})$	maximize	w_9

⁴ Indeed all the measures used in our approach take values in the interval $[0, 1]$, then a measure k to minimize can be changed to a measure to maximize by considering $1 - g_k(r)$ instead of $g_k(r)$.

Step 1: Dominance-based pruning. A reduction of the concurrent rules in $\mathcal{R}_{r,t}$ is carried out by focusing on non-dominated rules on the basis of the considered measures. A rule r_1 dominates a rule r_2 , we write $r_2 \prec r_1$, iff r_1 is at least equal to r_2 on all measures and it exists a measure where r_1 is strictly superior to r_2 . More formally,

$$r_2 \prec r_1 \text{ iff } g_k(r_2) \leq g_k(r_1), \forall k \in K \text{ and } \exists j \in K \text{ such that } g_j(r_2) < g_j(r_1).$$

The reduced set of rules can be stated as:

$$\mathcal{R}_{r,t,d} = \{r \in \mathcal{R}_{r,t} \mid \nexists r' \in \mathcal{R}_{r,t} : r \prec r'\}$$

Step 2: Pruning using Electre I. When $\mathcal{R}_{r,t,d}$ remains too large to be manually analyzed, a subjective pruning procedure based on the selection procedure Electre I is applied. This MCDA method enables expressing subjectivity through parameters that can be given by decision makers [18]. We use it for finding the final set of rules $\mathcal{R}^* \subseteq \mathcal{R}_{r,t,d}$. Electre I builds an outranking relation between pairs of rules allowing to select a subset of the best rules: \mathcal{R}^* . This subset is such that (i) any rules excluded from $\mathcal{R}_{r,t,d}$ is outranked by at least one rule from \mathcal{R}^* , (ii) rules from \mathcal{R}^* do not outrank each other. To do so, Electre I procedure (a) constructs outranking relationships through pairwise comparisons of rules, to further (b) exploit those relationships to build \mathcal{R}^* .

a) Outranking relations: the relationship “ r outranks r' ” (rSr') means that r is at least as good as r' on the set of measures K . The outranking assertion rSr' holds if: (i) a sufficient coalition of measures supports it, and (ii) none of the measures is too strongly opposed to it. These conditions are respectively referred to as concordance $c(rSr')$ and discordance indices $d(rSr')$, such that:

$$c(rSr') = \frac{\sum_{\{k: g_k(r) \geq g_k(r')\}} w_k}{\sum_{\{k: g_k(r) \geq g_k(r')\}} w_k} \text{ and } d(rSr') = \max_{\{k: g_k(r) < g_k(r')\}} [g_k(r') - g_k(r)],$$

with w_k the relative importance of measure k .

From these notations, we consider rSr' if $c(rSr') \geq \hat{c}$ and $d(rSr') \leq \hat{d}$; with \hat{c} and \hat{d} , two thresholds defining when the outranking should be considered or not.

b) Relations exploitation: a graph of outranking relationships is obtained from these pairwise comparisons. The kernel of this graph is our final reduced set of rules \mathcal{R}^* to be considered, such that:

$$\begin{aligned} & - \forall r' \in \mathcal{R}_{r,t,d} \setminus \mathcal{R}^*, \exists r \in \mathcal{R}^* : rSr', \text{ and} \\ & - \forall (r, r') \in \mathcal{R}^* \times \mathcal{R}^*, \neg(rSr'). \end{aligned} \quad (5)$$

The set of model parameters that have to be defined for applying the subjective reduction based on Electre I are: weights $w_k, \forall k \in K$, and the concordance and discordance thresholds, \hat{c}, \hat{d} .⁵ The choice of parameter values will be further discussed in the illustration Section 4.

⁵ Evaluating support and confidence of $\bar{A} \rightarrow B$ and $\bar{A} \rightarrow \bar{B}$ can lead to undefined values, e.g. evaluating $\bar{A} \rightarrow B$, we have $Bel(\bar{A} \times B) = 0$ when \bar{A} has never been

4 Illustration

As an illustration, we consider the context of humanitarian projects carried out for answering to emergency situations. A dataset of observations describes these emergency situations according to four attributes: 1) the *type of disaster* faced, 2) the *season*, 3) the *environment* in which it occurred, and 4) an evaluation of the situation w.r.t. the *human cost*. We further refer to these attributes using their number, considering that they respectively take discrete values in: $\Theta_1 = \{\textit{tsunami}, \textit{earthquake}, \textit{epidemic}, \textit{conflict}, \textit{pop.displacement}\}$, $\Theta_2 = \{\textit{spring}, \textit{summer}, \textit{autumn}, \textit{winter}\}$, $\Theta_3 = \{\textit{urban}, \textit{rural}\}$, $\Theta_4 = \{\textit{low}, \textit{medium}, \textit{high}, \textit{veryHigh}\}$. Besides, for each attribute, prior knowledge is defined into ontologies determining the values of interest. In this specific case study, the purpose of association rules is to highlight the influence of a situation contextual features on its evaluation according to the *Human Cost*, a useful information for project planning. Thus the searched rules $r : A \rightarrow B$ will imply the attributes in the following set $I_1 = \{1, 2, 3\}$ in the *antecedent* and in $I_2 = \{4\}$ for the *consequent*.

Table 3. Database of observations expressed using precise, imprecise or missing values.

	Disaster Type	Season	Environment	Human Cost
d_1	{ <i>earthquake</i> }	{ <i>autumn</i> }	{ <i>rural</i> }	{ <i>medium</i> }
d_2	{ <i>tsunami</i> }	{ <i>autumn</i> }	{ <i>urban</i> }	{ <i>medium</i> }
d_3	{ <i>epidemic</i> }	-	{ <i>urban</i> }	{ <i>veryHigh</i> }
d_4	{ <i>earthquake, epidemic, tsunami</i> }	{ <i>spring</i> }	-	{ <i>high, veryHigh</i> }
d_5	{ <i>epidemic</i> }	{ <i>spring</i> }	{ <i>urban</i> }	{ <i>high</i> }
d_6	{ <i>epidemic</i> }	{ <i>spring, summer</i> }	-	{ <i>high, veryHigh</i> }
d_7	{ <i>epidemic</i> }	{ <i>spring, summer</i> }	{ <i>urban</i> }	{ <i>high, veryHigh</i> }
d_8	{ <i>epidemic</i> }	{ <i>spring, summer</i> }	{ <i>urban</i> }	{ <i>veryHigh</i> }
d_9	{ <i>earthquake, epidemic, tsunami</i> }	{ <i>summer</i> }	{ <i>rural</i> }	{ <i>high</i> }
d_{10}	{ <i>epidemic</i> }	{ <i>summer</i> }	{ <i>urban</i> }	{ <i>high</i> }
d_{11}	{ <i>epidemic</i> }	{ <i>summer</i> }	{ <i>urban</i> }	{ <i>veryHigh</i> }
d_{12}	{ <i>earthquake</i> }	{ <i>winter</i> }	{ <i>rural</i> }	{ <i>high, medium, veryHigh</i> }
d_{13}	{ <i>earthquake</i> }	{ <i>winter</i> }	{ <i>rural</i> }	{ <i>low</i> }
d_{14}	{ <i>earthquake, epidemic, tsunami</i> }	{ <i>winter</i> }	{ <i>rural</i> }	{ <i>high</i> }

Among the observations of 14 projects given in Table 3, some attribute values are expressed with imprecision, e.g. *Human cost* values may be unclear such that “*human Cost is High or VeryHigh*”. When values are missing the total ignorance is considered. In this setting, the size of the initial studied space \mathcal{R} is $\prod_{i=1}^4 2^{|\Theta_i \setminus \emptyset|} = 20925$. Using the restrictions focusing on rules with non-null support, and involving attribute values of interest defined into ontologies (cf. Section 3), we obtain a reduced search space $\mathcal{R}_{r,t}$ composed of 484 rules.

The rule evaluation and selection process is further applied to $\mathcal{R}_{r,t}$ using the 9 interestingness measures proposed in Table 2. Using dominance-based pruning

observed, leading to $Bel(B|\bar{A})$ being undefined. However, pruning using dominance and Electre I requires the same measures to be defined. Undefined values are thus substituted by an arbitrary value that neither favor nor penalize the evaluation of the rule $A \rightarrow B$. The median of $Bel(\bar{A} \times B)$ (resp. $Bel(\bar{A} \times \bar{B})$) has been chosen. Note that $A \rightarrow \bar{B}$ is not concerned since evaluating $A \rightarrow B$ implies evidence on A .

(Step 1/2), a set of 18 non-dominated rules $\mathcal{R}_{r,t,d}$ is identified among the 484 rules initially considered. These rules are listed in Table 4, and indexed from r_0 to r_{17} . Pruning using Electre I is then applied over the set of non-dominated rules $\mathcal{R}_{r,t,d}$ (Step 2/2). Different sets of selected rules -i.e. \mathcal{R}^* - are given in Table 5 for different sets of model parameters. The results being sensitive to parameter values, we propose to discuss different parameter settings. We remind that these parameters are: $\forall k \in K$, w_k the weights of interestingness measures, and \hat{c} and \hat{d} the concordance and discordance thresholds. They represent end-user's preferences. They can be given directly; the weights w_k can also be elicited using Simos, a well-known weighting procedure [11].

Table 4. Set of non-dominated rules, $\mathcal{R}_{r,t,d}$.

	Disaster Type	Season	Environment	Human Cost
r_0	{ <i>earthquake</i> }	\wedge { <i>autumn</i> }	\wedge { <i>rural</i> }	\rightarrow { <i>medium</i> }
r_1	{ <i>earthquake, tsunami</i> }	\wedge { <i>autumn</i> }	\wedge Θ_3	\rightarrow { <i>medium</i> }
r_2	{ <i>tsunami</i> }	\wedge { <i>autumn</i> }	\wedge { <i>urban</i> }	\rightarrow { <i>medium</i> }
r_3	{ <i>earthquake, epidemic, tsunami</i> }	\wedge Θ_2	\wedge Θ_3	\rightarrow Θ_4
r_4	{ <i>earthquake, epidemic, tsunami</i> }	\wedge Θ_2	\wedge Θ_3	\rightarrow { <i>high, medium, veryHigh</i> }
r_5	{ <i>earthquake, epidemic, tsunami</i> }	\wedge Θ_2	\wedge Θ_3	\rightarrow { <i>high, veryHigh</i> }
r_6	{ <i>epidemic</i> }	\wedge Θ_2	\wedge Θ_3	\rightarrow { <i>high, veryHigh</i> }
r_7	{ <i>epidemic</i> }	\wedge Θ_2	\wedge { <i>urban</i> }	\rightarrow { <i>veryHigh</i> }
r_8	{ <i>earthquake</i> }	\wedge { <i>autumn, winter</i> }	\wedge { <i>rural</i> }	\rightarrow { <i>medium</i> }
r_9	{ <i>earthquake, tsunami</i> }	\wedge { <i>autumn, winter</i> }	\wedge Θ_3	\rightarrow { <i>low, medium</i> }
r_{10}	{ <i>earthquake, tsunami</i> }	\wedge { <i>autumn, winter</i> }	\wedge Θ_3	\rightarrow { <i>medium</i> }
r_{11}	{ <i>earthquake, epidemic, tsunami</i> }	\wedge { <i>spring, summer</i> }	\wedge Θ_3	\rightarrow { <i>high, veryHigh</i> }
r_{12}	{ <i>epidemic</i> }	\wedge { <i>spring, summer</i> }	\wedge Θ_3	\rightarrow { <i>high, veryHigh</i> }
r_{13}	{ <i>epidemic</i> }	\wedge { <i>spring, summer</i> }	\wedge { <i>urban</i> }	\rightarrow { <i>high, veryHigh</i> }
r_{14}	{ <i>epidemic</i> }	\wedge { <i>spring, summer</i> }	\wedge { <i>urban</i> }	\rightarrow { <i>veryHigh</i> }
r_{15}	{ <i>epidemic</i> }	\wedge { <i>summer</i> }	\wedge { <i>urban</i> }	\rightarrow { <i>high, veryHigh</i> }
r_{16}	{ <i>epidemic</i> }	\wedge { <i>summer</i> }	\wedge { <i>urban</i> }	\rightarrow { <i>veryHigh</i> }
r_{17}	{ <i>earthquake</i> }	\wedge { <i>winter</i> }	\wedge { <i>rural</i> }	\rightarrow { <i>low</i> }

Table 5. Final sets of rules (\mathcal{R}^*) obtained with Electre I pruning using four parameter settings (a to e).

Different sets of parameters, with $\hat{c} = 0.7$											
	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	\hat{d}	\mathcal{R}^*
a	0.27	0.15	0.1	0.08	0.08	0.08	0.08	0.08	0.08	0.3	{ $r_1, r_3, r_6, r_9, r_{11}$ }
b	0.18	0.18	0.18	0.1	0.1	0.1	0.1	0.03	0.03	0.3	{ r_1, r_3, r_6 }
										0.2	{ $r_0, r_1, r_2, r_3, r_6, r_{13}, r_{16}, r_{17}$ }
c	0.12	0.2	0.2	0.08	0.08	0.08	0.08	0.08	0.08	0.3	{ r_1, r_3, r_6 }
										0.2	{ $r_0, r_1, r_2, r_3, r_6, r_{13}, r_{16}, r_{17}$ }
d	0.15	0.25	0.25	0.05	0.05	0.05	0.05	0.075	0.075	0.3	{ r_1, r_3, r_6, r_{17} }
										0.2	{ $r_0, r_1, r_2, r_3, r_6, r_{13}, r_{16}, r_{17}$ }
e	0.33	0.33	0.34	0	0	0	0	0	0	0.3	$\mathcal{R}_{r,t,d} \setminus \{r_8, r_{10}, r_{16}, r_{17}\}$

Among the considered interestingness measures, according to the literature, we assume that *support*, *confidence* and *IC* are the most significant ones w.r.t. rule interest. They have to be associated to the most important weights. Conversely, we assume that the other measures -about rule complements- are secondary and will provide additional information for comparing and criticizing the

relevance of rules. In the first set of parameters (a) (cf. Table 5), the weight given to *support* and *confidence* is maximized to represent 60% of the votes required for the outranking (to exceed $\hat{c} = 0.7$). This setting will tend to favor the rules having a high degree of imprecision, being well supported and then reliable, since $Bel(B|A) \geq Bel(A \times B)$. For example, in this setting the rules r_3, r_6, r_{11} , see Tables 5 and 4, are among the selected rules in \mathcal{R}^* ; e.g. with r_3 involving the total imprecision on three attributes.

When restricting \hat{d} to 0.2 with the parameter settings (b), (c), (d), it increases the size of the kernel, while still discarding more than half of the rules among the set of non-dominated ones. With parameters (d) and $\hat{d} = 0.3$, highest importance is given to *confidence* and *IC*, providing these 2 measures with 71% of the voting power to reach the outranking condition $\hat{c} = 0.7$. Thus, a rule with a better score on *confidence*, *IC* and on some of the other measures -except *support*- can be selected while having a low support. This is illustrated with the selection of r_{17} for example. Lastly, the parameter setting (e) is equivalent to considering only the three main measures with equal importance. Here, it enables to discard only 4 extra rules in comparison to dominance relationships. This is explained by the fact that the absence of dominance between rules is more frequent.

Finally, the parameter settings (b), (c) or (d) with $\hat{d} = 0.2$, favoring the *support*, *confidence* and *IC* over the other measures tend to provide interesting results. This setting enables the selection of both precise and imprecise rules of interest w.r.t. the initial set of observations, such as r_{16} and r_{13} . In the initial dataset -see Table 3- the imprecise information $\{spring, summer\}$ for the *season* or $\{high, veryHigh\}$ for the *Human Cost* are frequently observed. Indeed, selecting the imprecise rule $r_{13} : \{epidemic\} \wedge \{spring, summer\} \wedge \{urban\} \rightarrow \{high, veryHigh\}$ in \mathcal{R}^* is not surprising. As an interpretation of this rule, we say that the analysis of the database tends to relate the occurrence of epidemics in urban areas to a specific season, spring or summer, and human cost. In particular, the rule seems valid at least for one the conjunction “summer and high human cost”, “summer and a very High human cost”, “spring and high” or “spring and veryHigh”. In this illustration, different sets of parameters and their results on rule selection have been presented. However, these parameters have to be set by the end-user.

To further discuss these results, it is interesting to note that all the selected measures for rules comparison, except the *IC*, are based on observations frequency. In order to counterbalance the preponderance of this factor, it might be relevant to add subjective measures and not only data-driven ones. Subjective interestingness measures have been studied in the literature. Relying on these works, we could include here measures based for example on user expected rules or expected conjunction of attribute values. Furthermore, investigating the dependencies among frequency based measures, and considering them in the selection process will be valuable. Nevertheless, considering additional measures (especially data-driven), as the ones proposed for classical ARM, is not necessarily straightforward within the evidence theory framework. It indeed implies to define their right expression and meaning in this framework.

5 Conclusion and perspectives

Mining association rules from imperfect data is a key challenge for real-world applications dealing with imperfect data, e.g., imprecise, missing data, etc. The ARM approach introduced in this paper enables to deal with imprecise data and derive imprecise rules under specific conditions (e.g. fixing both antecedent and consequent). Relying on evidence theory and Multiple Criteria Decision Analysis, this new framework enriches expressivity of existing works while providing a novel selection procedure for identifying most interesting rules according to several viewpoints. To this aim, several interestingness measures have been proposed, and used in a two-step selection procedure based on dominance relationships and Electre I. A restriction using *a priori* knowledge has also been proposed to focus and ease the mining process by incorporating symbolic knowledge defined into domain ontologies. To further improve the approach, additional measures of interestingness could be added. Future work related to subjective measures (e.g., user-oriented) would be particularly relevant to enrich the set of frequency-based measures that are currently involved in the approach. Studying the interactions between the measures would also be of interest. Finally, only an illustration using a simplified case study related to humanitarian projects analysis has been presented in this paper. Thorough algorithmic complexity and performance evaluations of the approach have to be discussed. Difficult challenges related to algorithmic complexity and efficiency issues of the procedure also have to be addressed in order to mine rules involving numerous attributes.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: *Acm sigmod record*. vol. 22, pp. 207–216. ACM (1993)
2. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: *Proc. 20th int. conf. very large data bases, VLDB*. vol. 1215, pp. 487–499 (1994)
3. Ait-Mlouk, A., Gharnati, F., Agouti, T.: Multi-agent-based modeling for extracting relevant association rules using a multi-criteria analysis approach. *Vietnam Journal of Computer Science* **3**(4), 235–245 (2016)
4. Bouker, S., Saidi, R., Yahia, S.B., Nguifo, E.M.: Ranking and selecting association rules based on dominance relationship. In: *2012 IEEE 24th international conference on tools with artificial intelligence*. vol. 1, pp. 658–665. IEEE (2012)
5. Chen, M.C.: Ranking discovered rules from data mining with multiple criteria by data envelopment analysis. *Expert Systems with Applications* **33**(4), 1110–1116 (2007)
6. Choi, D.H., Ahn, B.S., Kim, S.H.: Prioritization of association rules in data mining: Multiple criteria decision approach. *Expert Systems with Applications* **29**(4), 867–878 (2005)
7. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics* **38**, 325–339 (1967)
8. Djouadi, Y., Redaoui, S., Amroun, K.: Mining association rules under imprecision and vagueness: towards a possibilistic approach. In: *2007 IEEE International Fuzzy Systems Conference*. pp. 1–6. IEEE (2007)

14 C. L'Héritier et al.

9. Dubois, D., Denoeux, T.: Conditioning in dempster-shafer theory: prediction vs. revision. In: *Belief Functions: Theory and Applications*, pp. 385–392. Springer (2012)
10. Fagin, R., Halpern, J.Y.: A new approach to updating beliefs. In: *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*. pp. 347–374. UAI '90, Elsevier Science Inc., New York, NY, USA (1991), <http://dl.acm.org/citation.cfm?id=647233.760137>
11. Figueira, J., Roy, B.: Determining the weights of criteria in the electre type methods with a revised simos' procedure. *European Journal of Operational Research* **139**(2), 317–326 (2002)
12. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: a survey. *ACM Computing Surveys* **38**(3), 9–es (2006)
13. Hewawasam, K., Premaratne, K., Subasingha, S., Shyu, M.L.: Rule mining and classification in imperfect databases. In: *2005 7th International Conference on Information Fusion*. vol. 1, pp. 8–pp. IEEE (2005)
14. Hong, T.P., Lin, K.Y., Wang, S.L.: Fuzzy data mining for interesting generalized association rules. *Fuzzy sets and systems* **138**(2), 255–269 (2003)
15. Kotsiantis, S., Kanellopoulos, D.: Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering* **32**(1), 71–82 (2006)
16. Liu, B., Hsu, W., Chen, S., Ma, Y.: Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems* **15**(5), 47–55 (2000). <https://doi.org/10.1109/5254.889106>
17. Nguyen Le, T.T., Huynh, H.X., Guillet, F.: Finding the most interesting association rules by aggregating objective interestingness measures. In: Richards, D., Kang, B.H. (eds.) *Knowledge Acquisition: Approaches, Algorithms and Applications*. pp. 40–49. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)
18. Roy, B.: Classement et choix en présence de points de vue multiples. *Revue française d'informatique et de recherche opérationnelle* **2**(8), 57–75 (1968)
19. Samet, A., Lefèvre, E., Yahia, S.B.: Evidential data mining: precise support and confidence. *Journal of Intelligent Information Systems* **47**(1), 135–163 (2016)
20. Seco, N., Veale, T., Hayes, J.: An intrinsic information content metric for semantic similarity in wordnet. In: *Ecai*. vol. 16, p. 1089 (2004)
21. Shafer, G.: *A mathematical theory of evidence*, vol. 42. Princeton university press (1976)
22. Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and data engineering* **8**(6), 970–974 (1996)
23. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 32–41. ACM (2002)
24. Tobji, M.B., Yaghlane, B.B., Mellouli, K.: A new algorithm for mining frequent itemsets from evidential databases. In: *Proceedings of IPMU*. vol. 8, pp. 1535–1542 (2008)
25. Tobji, M.A.B., Yaghlane, B.B., Mellouli, K.: Frequent itemset mining from databases including one evidential attribute. In: *International Conference on Scalable Uncertainty Management*. pp. 19–32. Springer (2008)
26. Toloo, M., Sohrabi, B., Nalchigar, S.: A new method for ranking discovered rules from data mining by dea. *Expert Systems with Applications* **36**(4), 8503–8508 (2009)
27. Vaillant, B., Lenca, P., Lallich, S.: A clustering of interestingness measures. In: *International Conference on Discovery Science*. pp. 290–297. Springer (2004)

Fair and Efficient Alternatives to Shapley-based Attribution Methods

Charles Condevaux¹, Sébastien Harispe², and Stéphane Mussard¹

¹ Univ. Nimes CHROME, France

{charles.condevaux, stephane.mussard}@unimes.fr

² EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, France
sebastien.harispe@mines-ales.fr

Abstract. Interpretability of predictive machine learning models is critical for numerous application contexts that require decisions to be understood by end-users. It can be studied through the lens of local explainability and attribution methods that focus on explaining a specific decision made by a model for a given input, by evaluating the contribution of input features to the results, e.g. probability assigned to a class. Many attribution methods rely on a game-theoretic formulation of the attribution problem based on an approximation of the popular Shapley value, even if the underlying rationale motivating the use of this specific value is today questioned. In this paper we introduce the FESP - Fair-Efficient-Symmetric-Perturbation - attribution method as an alternative approach sharing relevant axiomatic properties with the Shapley value, and the Equal Surplus value (ES) commonly applied in cooperative games. Our results show that FESP and ES produce better attribution maps compared to state-of-the-art approaches in image and text classification settings.

Keywords: Machine learning interpretability · XAI · local interpretability · Attribution method

1 Introduction

Deep learning models are today state-of-the-art to tackle a large variety of machine learning problems in image or natural language processing (NLP) to cite a few. The use of these efficient models is however still limited due to their intrinsic black-box nature, i.e. deciphering the complex input-output mapping performed by trained deep learning models -sometimes involving billions of parameters- is still an open problem [21]. Indeed, many application contexts require not only models with good average performance, but also significant explanations allowing to fully understand and interpret predictor outputs. This is not only true for obvious critical use cases, e.g. in the medical field, in which sensitive decisions have to be supported by evidence [10,21,29]. More generally, legitimate concerns about potential harmful bias of inscrutable models are more expressed. Due to those issues, regulators introduce more and more legal requirements imposing life-impacting automated decision making to be explainable [15,28].

In this context, numerous works analyze approaches contributing to deep learning model explainability, in particular through the notions of global and local interpretability while dealing with predictive tasks [16,4,21]. Global interpretability sheds some light on the general model behavior, e.g. global decision rules, while local interpretability focuses on explaining a specific decision (output) for a given input instance. This paper is concerned with local interpretability, and in particular with *attribution methods* (AMs). These methods aim at explaining the prediction made by a predictor for an input instance by assigning a scalar *attribution value* to each input feature. The purpose is therefore, for a given instance, to distinguish the features best explaining a model output prediction. The core problem is thus to define an AM that assigns a relevant attribution value to each feature in a specific predictive context.

Assigning attribution values in a meaningful way is an open question that has been studied through different angles. A large body of works focuses on AMs backed on axiomatic motivations defining supposedly intuitive properties that these methods should respect [36,38,24,22,2]. A wide range of contributions are in particular considering the Shapley value [32] as ground truth since it defines the unique way to solve the game-theoretic formulation of the attribution problem considering admitted axioms in coalition games - attribution is made considering a cooperative game between model features; attribution among the features is then made based on the Shapley value [2,38,22]. In that context, several approaches have been proposed to approximate the prohibitive computation of the Shapley value which requires evaluating 2^N feature subsets considering inputs of N features (NP-hard). Nevertheless, even if contributions stress that AMs based on the Shapley value seem to agree with human intuitive expectations [2], no clear agreement on that matter has been reached and the ground truth status of the Shapley value is today questioned [20,12]. Axiomatically grounded and algorithmically efficient, AMs have still to be investigated.

The contributions proposed in this paper are threefold:

1. We introduce FESP (Fair-Efficient-Symmetric-Perturbation value), an axiomatically grounded and algorithmically efficient AM that shares some properties with the Shapley value.
2. We propose the use of the equal surplus (ES) value, an $O(N)$ AM employed in cooperative games, which is linear, efficient and symmetric.
3. We show that FESP and ES achieve good accuracy on image and text classification compared with usual AMs.³⁴ The results outline their benefits with different benchmarks, e.g. issued from SHAP [22] and gradients [33,38].

The paper is structured as follows: Section 2 introduces existing AMs and discussions about the Shapley value. Section 3 presents FESP and ES. Section 4 evaluates ES, FESP and existing AMs on image and text classification tasks, with discussions on performances with respect to different protocols. Section 5 discusses our findings before mentioning perspectives they open.

³ Our experiments: <https://github.com/ccdv-ai/fesp.es.git>. This work used HPC resources of IDRIS (allocation 2022-AD011011309R2) made by GENCI.

⁴ This work has benefited from LAWBOT (ANR-20-CE38-0013) grant.

2 State of the art

In this section we present the attribution problem focusing on a multiclass classification setting, as well as state-of-the art AMs proposed to solve it.

2.1 The attribution problem

Considering a predictor, the attribution problem consists in attributing a scalar value to each input feature characterizing an instance with respect to (w.r.t.) a predicted value (e.g. class probability in a classification setting or real value in a regression setting). This value represents the contribution of a specific feature to the prediction, e.g. in a classification setting, this value may be useful to understand which input features support a given class.

Without loss of generality, a multiclass classification setting is considered with a set of classes $\mathcal{C} := \llbracket 1, C \rrbracket$, with $\llbracket a, b \rrbracket$ denoting the interval of all integers between a and b included. In that context, a predictor f takes an N -dimensional feature input $\mathbf{x} := [x_1, \dots, x_N] \in \mathbb{R}^N$ and produces a probability distribution $f(\mathbf{x}) := [f_1(\mathbf{x}), \dots, f_C(\mathbf{x})] \in [0, 1]^C$, with $f_i(\mathbf{x})$ the probability assigned to class $i \in \mathcal{C}$ by f for \mathbf{x} . $\mathcal{N} := \llbracket 1, N \rrbracket$ is the set of feature indices.

Considering this setting, given predictor f and an input $\mathbf{x} \in \mathbb{R}^N$, an AM φ aims at computing a contribution vector $\varphi(\mathbf{x}, f_i)$ for any class $i \in \mathcal{C}$ such as $\varphi(\mathbf{x}, f_i) = [\varphi_1(\mathbf{x}, f_i), \dots, \varphi_N(\mathbf{x}, f_i)] \in \mathbb{R}^N$, with $\varphi_j(\mathbf{x}, f_i)$ the attribution value of feature $j \in \mathcal{N}$ w.r.t. $f_i(\mathbf{x})$. Otherwise stated, considering the AM φ , $\varphi_j(\mathbf{x}, f_i)$ is the contribution of feature j to the probability assigned by the predictor f to class i for the input \mathbf{x} .

The two main classes of approaches studied in the literature to solve the attribution problem are introduced hereafter. They are both based on the evaluation of a perturbation of the input features on the predictive value under study.

2.2 Attribution using feature coalisation analysis

In the local interpretability setting, numerous perturbation-based approaches define an AM φ by evaluating the contribution $\varphi_j(\mathbf{x}, f_i)$ of a specific feature $j \in \mathcal{N}$ (to $f_i(\mathbf{x})$) as its contribution to coalitions of features. Considering a coalition including all features except j (i.e. $\mathcal{N} \setminus \{j\}$), the contribution of j to that coalition is assessed by evaluating the impact of a perturbation of x_j on $f_i(\mathbf{x})$. Such a perturbation aims at mimicking the removal of the studied feature, e.g. by naively setting its value to zero or a baseline value.

For any $\mathcal{S} \subseteq \mathcal{N}$, $\mathbf{x}(\mathcal{S})$ refers to the vector \mathbf{x} in which all feature values x_k , $k \in \mathcal{N} \setminus \mathcal{S}$ have been substituted by a baseline value. As the input \mathbf{x} is implicitly fixed in our discussions, $f_i(\mathcal{S})$ is used to denote $f_i(\mathbf{x}(\mathcal{S}))$, which is the probability assigned by f to class $i \in \mathcal{C}$ w.r.t. $\mathbf{x}(\mathcal{S})$.

The *marginal contribution* of a feature $j \in \mathcal{N}$ to a coalition \mathcal{S} ($j \notin \mathcal{S}$) is thus defined by $f_i(\mathcal{S} \cup \{j\}) - f_i(\mathcal{S})$. Numerous AMs based on this notion of marginal contribution have been studied [13,6,41,39]. Game theory allows us to obtain such contributions, through the (least) core [18] but also through the Shapley

4 C. Condevaux et al.

value often considered as the ground truth value to explain the role of a given variable [2].

Attribution value as the Shapley value: The Shapley value averages marginal contributions over all possible feature coalitions:

$$\varphi_j^{Sh}(\mathbf{x}, f_i) := \sum_{S \subseteq \mathcal{N} \setminus \{j\}} P(S) (f_i(S \cup \{j\}) - f_i(S)),$$

for all $j \in \mathcal{N}$; $f_i(\emptyset) := 0$ for all $i \in \mathcal{C}$ by convention, and $P(S) := (N - S - 1)!S!/N!$ ($S := |S|$).

The Shapley value implies (and is implied by) four axioms: *efficiency*, *additivity*, *symmetry* and the *null player axiom*, see [32].⁵ These axioms make the Shapley value appealing from a theoretical point of view, and have motivated the *de facto* ground truth status given to this value.

However, considering N features, 2^N coalitions have to be evaluated which makes the Shapley value prohibitively expensive to compute. A natural way to reduce computation complexity is to rely either on coalition sampling to compute the marginal contributions [8], on local coalitions [9] or on Boolean circuits [3]. The first approach can however be slow to converge when the number of features is large. Instead of directly modifying original inputs, DASP [2] relies on distribution propagation using an auxiliary network based on Lightweight Probabilistic Deep Networks [14]. This model sequentially produces an estimate for each coalition size, thus allowing to greatly reduce the complexity from $O(2^N)$ to $O(N^2)$. Although this approximation is accurate, building an additional network is cumbersome, especially when fine tuning a pretrained model (as it requires rewriting each layer and activation function).

Attribution based on Occlusion: In order to determine whether a feature or a group of features impacts a prediction, occlusion models measure the effect of removing them from the input (marginal contribution). In computer vision, these feature coalitions generally take the form of a sliding block [42], of a predefined size, inside which pixels are disturbed or replaced by a specific value (e.g. 0). Although such perturbation and occlusion models can accurately measure the marginal contribution of a variable, they tend to be slower than other AMs since they require multiple forward passes to fully cover the input and are thus dependent on the number of features. The size of the block is also an additional hyperparameter which can have a significant impact on overall performances.

2.3 Attribution based on gradient analysis

Gradient-based approaches rely on various gradient computations through back-propagation evaluations. They compute the attribution value of a feature evaluating the partial derivative of the studied predicted value with regard to the

⁵ It is noteworthy that *additivity* implies *linearity* but the converse does not hold. Invoking *linearity* enlarges the class of admissible AMs, see Theorem 1 below.

feature value, e.g., $\varphi_j(\mathbf{x}, f_i)$ is defined as a function of $\partial f_i(\mathbf{x})/\partial x_j$. In this context $\varphi_j(\mathbf{x}, f_i)$ is then evaluated based on the impacts on $f_i(\mathbf{x})$ induced by a local change of x_j . The function φ_j should be carefully chosen to respect some properties or specific behaviors. For instance, multiplying the gradient by the input [34] increases the sharpness of the attribution map but fails to handle specific functions like ReLU, which can produce zero values. More sophisticated models like DeepLift [33] and Integrated Gradient [38] satisfy a desirable axiom called completeness which is closely related to the efficiency axiom in cooperative game theory: for a baseline \mathbf{x}' we have $\sum_{j \in \mathcal{N}} \varphi_j(\mathbf{x}, f) = f(\mathbf{x}) - f(\mathbf{x}')$.

To compute the contribution map, DeepLift takes all neurons and compares their activations after feeding a true sample and a reference input which can depend on the task and on the dataset. This model is inspired by Layer-wise Relevance Propagation which relies on a similar idea without the use of a reference [5]. Integrated Gradient averages different gradients: the input is modified multiple times along a linear path between itself and a baseline often set to zero. This continuous setting has been connected to another branch of the literature based on coalisation analysis, such as the Aumann-Shapley value [37].

3 Fair-Efficient-Symmetric Perturbations-based AMs

3.1 The Equal Surplus Value

It is well established that the Shapley value is easy to interpret since it displays the average of all marginal contributions of each feature; in this respect, it is a marginalist value. It shares some common properties with other marginalist values which form the Linear-Efficient-Symmetric values family (LES values) [31]. To our knowledge, this family has not been studied in the context of the attribution problem. The axioms respected by LES values are introduced hereafter.

Axiom 1 Linearity: For all predictors f, g , an AM φ satisfies linearity if, $\varphi(\mathbf{x}, \alpha_1 f_i + \alpha_2 g_i) = \alpha_1 \varphi(\mathbf{x}, f_i) + \alpha_2 \varphi(\mathbf{x}, g_i)$, for all $\alpha_1, \alpha_2 \in \mathbb{R}$ and for all classes $i \in \mathcal{C}$.

Axiom 2 Efficiency: For all predictors f , an AM φ satisfies efficiency if, $\sum_{j \in \mathcal{N}} \varphi_j(\mathbf{x}, f_i) = f_i(\mathcal{N})$, for all classes $i \in \mathcal{C}$.

Axiom 3 Symmetry: For all predictors f , an AM φ satisfies symmetry if, for all features $j \in \mathcal{N}$, $\varphi_j(\mathbf{x}, f_i) = \varphi_{\pi(j)}(\mathbf{x}_\pi, f_i)$ for all permutations π over the set of $N!$ permutations on \mathcal{N} and for all classes $i \in \mathcal{C}$.

LES values have been extensively characterized outside the machine learning literature first by [31], then by [17,25,27] through the following theorem:

Theorem 1. For all predictors f and all classes $i \in \mathcal{C}$, an AM φ satisfies linearity, efficiency and symmetry if and only if there exists a unique sequence of $N - 1$ real numbers $\{b_s\}_{s=1}^{N-1}$ such that for each $j \in \mathcal{N}$ with $b_0 = 0$ and $b_N = 1$:

$$\varphi_j(\mathbf{x}, f_i) = \sum_{S \subseteq \mathcal{N} \setminus \{j\}} P(S) \left(b_{s+1} f_i(S \cup \{j\}) - b_s f_i(S) \right).$$

6 C. Condevaux et al.

LES values are all based on marginal contributions, therefore they provide feature contributions and interpretations very close to the usual Shapley value. The Shapley value φ^{Sh} is indeed a particular case of the LES family considering all marginal contributions equally weighted ($b_s = 1$ for all $s = 1, \dots, N - 1$). Other well-known LES values, studied in the cooperative game literature are: the Equal Surplus value (ES) [11], the Solidarity value [26], the Prenucleolus [30], and the Consensus value [19]. The ES value φ_j^{ES} ($b_s = 0$ if $1 < s < N$, $b_s = 1$ if $s = N$, $b_s = N - 1$ if $s = 1$) is a peculiar member of the LES family since it is of complexity $O(N)$ whereas the others are $O(2^N)$:

$$\varphi_j^{ES}(\mathbf{x}, f_i) = f_i(\{j\}) + \frac{f_i(\mathcal{N}) - \sum_{k=1}^N f_i(\{k\})}{N}. \quad (1)$$

The first term of the right-hand side of Equation (1) is the contribution of feature x_j alone: its individual marginal contribution compared to a model composed of all features with baseline values $f_i(\{j\}) - f_i(\emptyset)$. The second term is the equal surplus: $f_i(\mathcal{N}) - \sum_{j=1}^N f_i(\{j\})$, i.e. the additional gain produced by the grand coalition in excess of the sum of the individual marginal contributions of features x_j , which evolve independently of the others.⁶

3.2 FESP

An AM grounded on the individual marginal contributions of each feature $f_i(\{j\}) - f_i(\emptyset)$ as in the ES is welcome since it outlines the role of each feature independently of the others. However, the equal surplus term is a constant for all features, consequently it cannot display the interaction of each feature with the grand coalition. In order to capture this specific effect, the exclusion of one feature from the whole set of features is employed, which consists in the occlusion technique. Occlusion related to feature x_j over class i may be simply characterized by $f_i(\mathcal{N} \setminus \{j\})$ instead of the equal surplus $f_i(\mathcal{N}) - \sum_{j=1}^N f_i(\{j\})$.

Then, two extreme feature coalitions could be considered for an AM: the one with the feature itself (such as Fig. 1 on the right-hand side - considering features as superpixels), and the one associated with occlusion, i.e. the entire image minus a given feature (center of Fig. 1). On this basis, we propose the following family of AMs based on extreme feature coalitions:

Definition 1. Family of AMs based on extreme feature coalitions:

$$\varphi_j(\mathbf{x}, f_i) = w_i f_i(\{j\}) + (1 - w_i)(-f_i(\mathcal{N} \setminus \{j\})), \quad (2)$$

with $w_i \in [0, 1]$ a weight associated with class $i \in \mathcal{C}$.

The first component of the family, $w_i f_i(\{j\})$, is grounded on the individual marginal feature contribution, which is always positive. Then, as far as the feature is discriminant, its contribution to the classification in class i increases. The

⁶ The study of the independence is of importance for the tractability of the Shapley value, this is the case with fully factorized data distributions [7].



Fig. 1. Extreme feature coalitions

second component, $(1 - w_i)(-f_i(\mathcal{N} \setminus \{j\}))$, is the contribution of occlusion, it is always negative. Occlusion of a discriminant feature x_j for class i entails that the probability f_i collapses, implying that the second component tends to zero. If an AM does not lie in the family of extreme feature coalitions, anything guarantees that bad features would be penalized by occlusion. Indeed, whenever a feature x_j is not discriminant for the classification in class i , the second component becomes negative, and the attribution value $\varphi(\mathbf{x}, f_i)$ can also become negative so that feature x_j is considered non-explanatory for the task. Furthermore, in order to gauge whether a feature is more *relevant* than another, the *fair treatment* axiom must be respected.

A feature x_k is said to be more relevant compared to feature x_ℓ when the association of x_k with all feature coalitions $\mathcal{S} \setminus \{k, \ell\}$ provides a greater attribution value compared to that of x_ℓ [27]. This property is welcome for all classification tasks such as image and text classifications. For instance, in an image classification setting, if a pixel x_k is more relevant compared to another one, because it allows some important shapes to be outlined, then the AM provides a higher contribution for x_k .

Axiom 4 Fair treatment: For all models f , and two given features x_k, x_ℓ , an AM φ satisfies fair treatment if, whenever feature x_k is more relevant compared to feature x_ℓ , i.e. $f_i(\mathcal{S} \cup \{k\}) \geq f_i(\mathcal{S} \cup \{\ell\})$ for all $\mathcal{S} \setminus \{k, \ell\}$, then $\varphi_k(\mathbf{x}, f_i) \geq \varphi_\ell(\mathbf{x}, f_i)$, for any given class $i \in \mathcal{C}$.

FESP is an $O(N)$ complexity AM that shares a common structure with members of the LES family: it respects *efficiency*, *symmetry* and *fair treatment* (see Appendix A and B).

Proposition 1. If an AM φ lies in the family of AMs based on extreme feature coalitions, and if it satisfies efficiency, then it is the FESP (Fair-Efficient-Symmetric-Perturbation) value given by, for $j \in \mathcal{N}$ and for $i \in \mathcal{C}$,

$$\varphi_j^{FESP}(\mathbf{x}, f_i) = w_i f_i(\{j\}) + (1 - w_i)(-f_i(\mathcal{N} \setminus \{j\})), \quad (3)$$

$$w_i = \frac{f_i(\mathcal{N}) + \sum_{k=1}^N f_i(\mathcal{N} \setminus \{k\})}{\sum_{k=1}^N f_i(\{k\}) + \sum_{k=1}^N f_i(\mathcal{N} \setminus \{k\})}. \quad (4)$$

4 Experiments

This section presents results and evaluation protocols defined for comparing AMs on image and text classification tasks. We report experiments running ES and FESP along-side Integrated Gradients [38], DeepLIFT rescale [33], GradientShap [22] and Occlusion model. We also compare to the SHAP library using the Deep-Explainer model [22] for vision tasks and the NLP pipeline named ShapExplainer for language tasks.

Local explainability. A model is first trained to solve the predictive task under consideration. In order to focus on AM evaluation and to avoid any interpretative bias, we consider *simple* predictive tasks for which good performances are today easily achieved. Based on the predictor obtained, an AM is then evaluated regarding the features it brings out as important to explain the prediction obtained for a given input (only predictions are performed, no training phase is involved while evaluating AMs).

Top- k model accuracy. This metric consists in evaluating how the predictor accuracy evolves only using top- k input-dependent contributing features according to an AM φ . If φ identifies the features that best explain an input classification, the predictor should keep achieving good performances only considering those features, i.e., the more φ performs correctly, the better should be the predictor accuracy only considering a subset of features provided by φ . Unselected features for a given input are simply masked during prediction; the shape of the predictor input is not modified. For a given task, the same predictor is therefore employed independently of the features considered during prediction.

4.1 Image classification: protocols and results

A pretrained VGG16 model [35] is fine tuned on a binary classification task related to Oxford-IIIT Pet Dataset⁷ (dog vs cat, fine-tuning: 3 epochs over 6325 images). It achieves 99% accuracy (1024 images, features are pixels).

Masking strategy. An image segmentation dataset gives for an image, the pixels of the shape of interest (segmentation mask), as well as its label, e.g. for an image labeled *dog*, the pixels of the dog are known. As focus is put on a simple classification task, it is assumed that pixels inside the mask should be relevant and have high attribution values (Appendix C presents a similar experiment modifying the pixels outside the segmentation mask with a random value). Considering an AM φ and a given image \mathbf{x} , the top- k contributing pixels of \mathbf{x} are computed w.r.t. φ , with k a fixed number of pixels set based on the size of the segmentation mask of \mathbf{x} . *AM-Precision@ k* is computed: *AM-Precision@20*(φ, \mathbf{x}, i) is the precision of φ on \mathbf{x} over class i , only considering the top- k pixels, k being here equal to 20% of the size of the segmentation mask for \mathbf{x} on class i . The precision of an AM is set to the average of the precision obtained for each image.

⁷ <https://www.robots.ox.ac.uk/~vis/vgg/data/pets/>

Averaging ES and FESP. Despite their $O(N)$ complexity, computing ES and FESP is slower than gradient based attribution methods since a forward pass is required for each feature. For a 224×224 RGB image, 50,176 passes would be in theory necessary to compute attribution values. In practice, removing or inferring a class by modifying a single pixel has little to no impact on the prediction of the VGG16. 56×56 superpixels are considered for ES, FESP and Occlusion, so that the image becomes a grid of 16 superpixels. These methods are run on the superpixels, and the process is repeated by moving the grid with a stride of 8. All pixels inside a given superpixel get the same attribution score φ_j for the current pass; these scores are then averaged resulting in an overlapping process (each pixel gets masked the same number of times in order to get a balanced average).⁸ This approach is similar to Occlusion and DeepExplain implementation.⁹

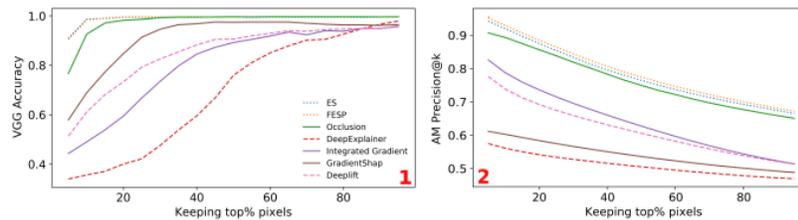


Fig. 2. Effect of feature selection on the predictor accuracy/precision.

Plot 1 (Figure 2) shows the accuracy of the pretrained model (VGG16) while only considering the top- k features (pixels) evaluated as important by each AM. Considering the top 10% of the features, ES and FESP provide good predictive performances (90% accuracy). Except Occlusion, the other AMs must consider almost 90% of the selected features in order to reach the full input predictor accuracy (99%). This accuracy is reached only using 15% to 20% of the features identified as important by FESP and ES.

Plot 2 (Figure 2) presents the *AM-Precision@k*, i.e. the capacity of each AM to outline expected informative pixels (pixels of the segmentation mask). AMs generally tend to consider the most important pixels to be inside the segmentation mask at first. FESP, ES and Occlusion achieves very good performances compared to other methods according to that test.

Figure 3 shows which image parts are recognized as relevant by the different AMs to explain the network prediction (top-10%).

We observe very different behaviors. AMs based on backpropagation independently treat pixels and therefore may return a noisy representation that is

⁸ Good tradeoff between performance and time complexity since large superpixels lead to higher performances while small ones tend to be noisier.

⁹ <https://github.com/slundberg/shap>

10 C. Condevaux et al.

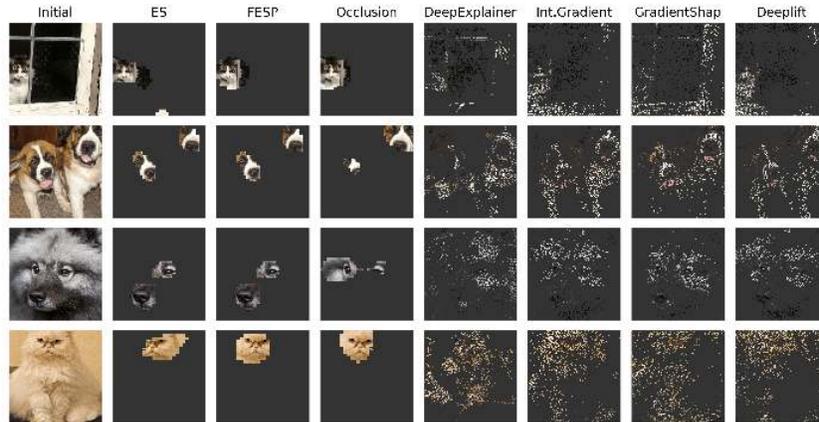


Fig. 3. Top 10% highest contributing pixels.

difficult to understand even for a relatively simple task where the discriminating criteria are fairly high level. The other AMs choose very localized areas, ignoring the rest of the image since they take benefit from convolutional layers that rely on local information. FESP and ES have a similar behavior but we observe in practice a noisier selection using ES, sometimes resulting in small artifacts. This behavior partly explains the performances obtained in the prediction task on partially masked inputs. Indeed FESP and ES tend to quickly identify very discriminant groups of features enabling to achieve good predictive performances even with a very limited set of features (Figure 2). Thus, normalizing and merging the best performing AMs can be a good solution to improve the overall selection as shown in Appendix D.

Robustness of ES and FESP. An additional evaluation protocol is conducted based on recent contributions on AM evaluations [40] (refining [1]). It relies on a binary classification setting involving fictive composite images, each one being composed of 2×2 images from the Oxford-IIIT Pet Dataset. Each composite image is labeled *cat* or *dog* and only contains a single image among four corresponding to its label. Considering a specific composite image and a good predictive model, we assume that an efficient AM should make it easy to distinguish the single image corresponding to its label. Each composite image is a random mix of: (i) a labeled image (*cat* or *dog*), (ii) 3 unrelated additional images, (iii) the locations of the 4 images in the 2×2 grid. The train and test sets are generated using the same approach compared to disjoint subsets of images. The same pretrained VGG16 architecture is fine-tuned on 6325 images over 4 epochs (96% accuracy on the test set).

Figure 4 shows that ES and FESP display the highest top- k model accuracy, indeed an accuracy greater than 90% is reached with 3% of top pixels (plot 1).

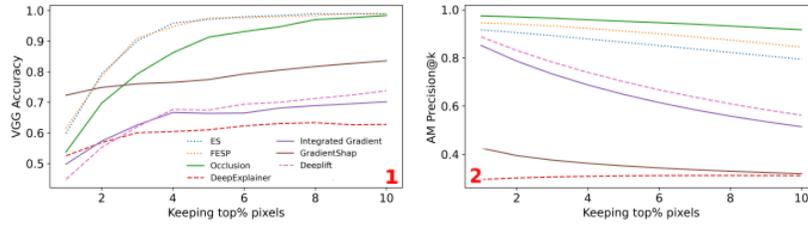


Fig. 4. Effect of feature selection on accuracy and precision.

The $AM-Precision@k$ (plot 2) gives the percentage of pixels located within the labeled subimage (among the four) given the selection of the top- k contributing pixels. According to this metric, FESP is between Occlusion and ES with 95% for only 2% of top pixels.

As shown in Figure 5 with 2×2 images, Occlusion, FESP and ES bring out relevant areas of the classes `dog` and `cat` and tend to be less noisy than gradient-based techniques (more images are available in Appendix D).

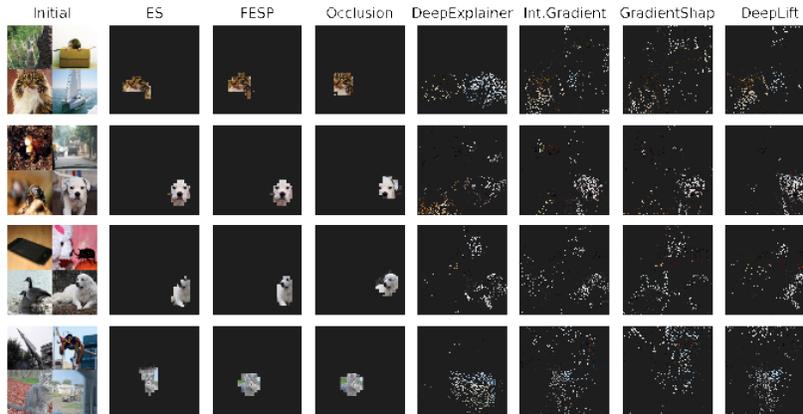


Fig. 5. Top 5% highest contributing pixels.

4.2 Text classification: protocols and results

A binary text classification task is performed using IMDB dataset [23]. The model is a pretrained RoBERTa fine tuned on IMDB dataset, for which features

12 C. Condevaux et al.

are words (95.5% accuracy).¹⁰ Testing is made on a subset of 1024 samples of the official testing set.

Masking strategy. The masking strategy is task dependent. Transformers can take benefit from the softmax function inside the self-attention mechanism to fully mask a token and avoid all connections. This is not possible with convolution layers used in vision tasks.

Averaging ES and FESP. A block of size 1 with a stride of 1 is used, consequently ES and FESP are directly estimated without an averaging strategy.

Compared to the image classification task, the same accuracy performances are obtained (Figure 6). On the one hand, the top 5% of words yield 95% accuracy for ES and FESP, and 90% for Occlusion.

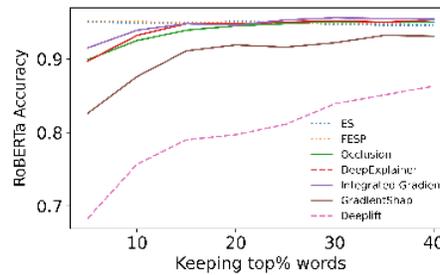


Fig. 6. Effect of feature selection on the predictor accuracy.

Finally, Figure 7 depicts words selected by the seven AMs over one example of the IMDB testing set. Additional examples are provided in Appendix D. For words chunked into several subwords by the tokenizer, the maximum score is used. AMs are normalized in such a way that each feature contribution takes value between 0 and 1, with 1 the highest contributing tokens being colored red. As expected, all AMs easily capture positive words such as "love" and "sexy", but these are not necessarily associated to the highest contribution. For instance Occlusion assigns the most important contribution to "this" and "I".

4.3 Discussions

Occlusion, ES and FESP. FESP and ES behave similarly most of the time although ES being slightly noisier since each feature evolves independently (of the grand coalition). In order to remedy this problem, FESP straddles the line between ES and Occlusion since it can be considered as a weighted mean of the two methods. In terms of interpretability, these models differ greatly. Occlusion is unable to determine the sign of the feature contributions unlike FESP & ES. This difference makes interpretability difficult in many cases, especially for word

¹⁰ <https://huggingface.co/textattack/roberta-base-imdb>



Fig. 7. Word importance normalized scores.

importance tasks as can be shown in Figure 7. In the case of image classifications a feature with a high contribution does not mean that it contributes positively to the prediction in the case of Occlusion.

From local to global explainability. Although we focused on local explainability, ES and FESP can also be used in a global explainability context. This can be achieved by using specific metrics (e.g accuracy, coefficient of determination) to measure the average impact of a feature on the predictions of a given predictor (see Appendix E for examples).

5 Conclusion

We have presented Equal Surplus (ES) and FESP (Fair-Efficient-Symmetric-Perturbation), two AMs based on marginalist values that can be used for local explainability of deep supervised learning models. These AMs compute attribution values that share relevant axiomatic properties with the Shapley value while ensuring an $O(N)$ time complexity for N -dimensional inputs.

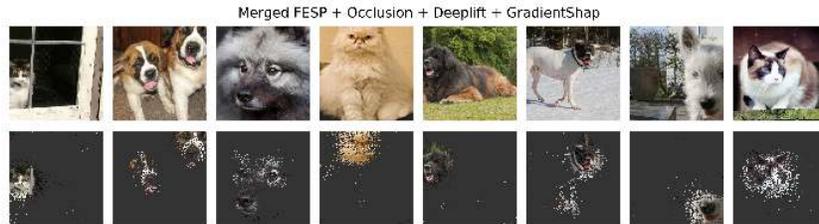


Fig. 8. Merged top 30% highest contributing pixels relative to mask size.

According to the proposed evaluations based on two image and text classification tasks, FESP, ES and Occlusion seem to be more suited for tasks with spatial or temporal dependencies such as computer vision and NLP. Indeed, in these contexts, backpropagation and gradient-based approaches tend to be noisy and generally more difficult to interpret for humans. Additionally, our results also corroborate literature findings highlighting that backpropagation gradient-based approaches tend to act like shape detectors, and therefore achieve good results in distinguishing the global shape of an object of interest in a local attribution setting [1]. This paves the way to the study of various AMs mixing both approaches highlighting different but often complementary features (see Figure 8).

Finally, the quantitative and qualitative results achieved by ES and FESP motivate the study of fast and axiomatically grounded AMs derived from LES values, which could reveal more, for example with the employ of AMs issued from the least square prenucleolus [30,31], or in the global attribution setting.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. arXiv preprint arXiv:1810.03292 (2018)
2. Ancona, M., Öztireli, C., Gross, M.: Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In: International Conference on Machine Learning. pp. 272–281. PMLR (2019)
3. Arenas, M., Barceló, P., Bertossi, L.E., Monet, M.: The tractability of shap-score-based explanations for classification over deterministic and decomposable boolean circuits. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. pp. 6670–6678. AAAI Press (2021)
4. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion* **58**, 82–115 (2020)
5. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wiserelevance propagation. *PLoS one* **10** (2015)

6. Brink, R., Funaki, Y., Ju, Y.: Reconciling marginalism with egalitarianism: consistency, monotonicity, and implementation of egalitarian shapley values. *Social Choice and Welfare* **40**, 693–714 (2013)
7. den Broeck, G.V., Lykov, A., Schleich, M., Suci, D.: On the tractability of shap explanations. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**(07), 6505–6513 (2021)
8. Castro, J., Gómez, D., Tejada, J.: Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research* **36**(5), 1726–1730 (2009)
9. Chen, J., Song, L., Wainwright, M.J., Jordan, M.I.: L-shapley and c-shapley: Efficient model interpretation for structured data. In: *International Conference on Learning Representations* (2019), <https://openreview.net/forum?id=S1E3Ko09F7>
10. Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Zietz, M., Hoffman, M.M., et al.: Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface* **15**(141), 20170387 (2018)
11. Driessen, T.S.H., Funaki, Y.: Coincidence of and collinearity between game theoretic solutions. *Operations-Research-Spektrum* **13**(1), 15–30 (1991)
12. Frye, C., Rowat, C., Feige, I.: Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems* **33** (2020)
13. Funaki, Y., Hoede, K., Aarts, H.: A marginalistic value for monotonic set games. *International Journal of Game Theory* **26**, 97–111 (1997)
14. Gast, J., Roth, S.: Lightweight probabilistic deep networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018)
15. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine* **38**(3), 50–57 (2017)
16. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* **51**(5), 1–42 (2018)
17. Hernández-Lamóneda, L., Juárez, R., Sánchez-Sánchez, F.: Dissection of solutions in cooperative game theory using representation techniques. *International Journal of Game Theory* **35**, 395–426 (2007)
18. Heskes, T., Sijben, E., Bucur, I.G., Claassen, T.: Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In: *Advances in Neural Information Processing Systems*. vol. 33, pp. 4778–4789. Curran Associates, Inc. (2020)
19. Ju, Y., Borm, P., Ruys, P.: The consensus value: a new solution concept for cooperative games. *Social Choice and Welfare* **28**, 685–703 (2007)
20. Kumar, I.E., Venkatasubramanian, S., Scheidegger, C., Friedler, S.: Problems with shapley-value-based explanations as feature importance measures. In: *International Conference on Machine Learning*. pp. 5491–5500. PMLR (2020)
21. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: A review of machine learning interpretability methods. *Entropy* **23**(1), 18 (2021)
22. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems* 30, pp. 4765–4774. Curran Associates, Inc. (2017)
23. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

16 C. Condevaux et al.

- pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA (June 2011), <http://www.aclweb.org/anthology/P11-1015>
24. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition* **65**, 211–222 (2017)
 25. Nembua, C.C., Andjiga, N.G.: Linear, efficient and symmetric values for TU-games. *Economics Bulletin* **3**, 1–10 (2008)
 26. Nowak, A.S., Radzik, T.: A solidarity value for n-person transferable utility games. *International Journal of Game Theory* **23**, 43–48 (1994)
 27. Radzik, T., Driessen, T.: On a family of values for tu-games generalizing the shapley value. *Mathematical Social Sciences* **65**, 105–111 (2013)
 28. Ras, G., van Gerven, M., Haselager, P.: Explanation methods in deep learning: Users, values, concerns and challenges. In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pp. 19–36. Springer (2018)
 29. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**(5), 206–215 (2019)
 30. Ruiz, L.M., Valenciano, F., Zarzuelo, J.M.: The least square prenucleolus and the least square nucleolus. two values for tu games based on the excess vector. *International Journal of Game Theory* **25**, 113–34 (1996)
 31. Ruiz, L.M., Valenciano, F., Zarzuelo, J.M.: The family of least square values for transferable utility games. *Games and Economic Behavior* **24**, 109–130 (1998)
 32. Shapley, L.S.: A value for n-person games. *Contributions to the Theory of Games* **2**(28), 307–317 (1953)
 33. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Precup, D., Teh, Y.W. (eds.) *Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 70, pp. 3145–3153. PMLR (06–11 Aug 2017)
 34. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. *CoRR abs/1605.01713* (2016)
 35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
 36. Sun, Y., Sundararajan, M.: Axiomatic attribution for multilinear functions. In: *Proceedings of the 12th ACM conference on Electronic commerce*. pp. 177–178 (2011)
 37. Sundararajan, M., Najmi, A.: The many shapley values for model explanation. In: III, H.D., Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 119, pp. 9269–9278. PMLR (13–18 Jul 2020)
 38. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: *International Conference on Machine Learning*. pp. 3319–3328. PMLR (2017)
 39. Wang, J., Zhang, Y., Kim, T.K., Gu, Y.: Shapley q-value: A local reward approach to solve global reward games. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(05), 7285–7292 (Apr 2020)
 40. Yona, G., Greenfeld, D.: Revisiting sanity checks for saliency maps (2021). <https://doi.org/10.48550/ARXIV.2110.14297>, <https://arxiv.org/abs/2110.14297>
 41. Young, P.: Monotonic solutions of cooperative games. *International Journal of Game Theory* **29**, 65–72 (1985)
 42. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. *CoRR abs/1311.2901* (2013)

Neural Order-First Split-Second Algorithm for the Capacitated Vehicle Routing Problem

Ali Yaddaden¹, Sébastien Harispe^{1*}, and Michel Vasquez¹

EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France
 {firstname.lastname}@mines-ales.fr

Abstract. Modern machine learning, including deep learning models and reinforcement learning techniques, have proven effective for solving difficult combinatorial optimization problems without relying on handcrafted heuristics. In this work, we present NOFSS, a Neural Order-First Split-Second deep reinforcement learning approach for the Capacity Constrained Vehicle Routing Problem (CVRP). NOFSS consists of a hybridization between a deep neural network model and a dynamic programming shortest path algorithm (Split). Our results, based on intensive experiments with several neural network model architectures, show that such a two-step hybridization enables learning of implicit algorithms (i.e. policies) producing competitive solutions for the CVRP.

Keywords. Neural Combinatorial Optimization Capacitated Vehicle Routing Problem
 Order-first Split-second Deep Reinforcement Learning

1 Introduction

Modern machine learning, including deep learning models and reinforcement learning techniques, have proven effective for solving difficult combinatorial optimization problems without relying on handcrafted heuristics [1]. The framework known as Neural Combinatorial Optimization (NCO), which proposes to solve combinatorial optimization problems using recent neural networks architectures, is in this context widely studied for routing problems such as the traveling salesman problem (TSP) [2–5] and the capacitated vehicle routing problem (CVRP) [6, 5].

Current NCO approaches implement a construction-based strategy. For the CVRP, such approaches build (i.e. construct) candidate solutions step by step, by selecting at each time step either to visit a client or to go back to the depot to refill, until each client is served. The action to perform at each construction step is chosen based on a probability distribution that will be estimated by a deep neural network, either using supervised or reinforcement learning. This discrete probability distribution defines the probability that an extension of the partial solution under construction, considering each available choices (unsatisfied clients and depot), will lead to the optimal solution. Considering such construction-based NCO approaches, solving the CVRP is therefore reframed as a learning goal aiming to obtain a good estimate of the probability distribution, such as step decisions based on this estimate minimize solution costs.

Using such an approach, the models handle both clients routing and returns to depot. In this context, choices of when to return to the depot are critical. Indeed, more returns to the depot can *de facto* lead to candidate solutions with a number of tours¹ greater than the optimal one. This will result in models failing to efficiently learn interesting resolution strategies, i.e. routing *policies*, due to poor quality candidate solutions, and/or large computational costs inducing prohibitive learning process (millions of learning steps). Handcrafted heuristics and metaheuristics may nevertheless be used to handle return to depot by using an exact tour splitting algorithm - solving a shortest path problem in an auxiliary graph that represents the clients’ visit order [7, 8]. Inspired by this problem decomposition, this paper presents NOFSS, Neural Order-First Split-Second, a novel two-step learning-based approach proposing to:

1. Learn how to order clients into a giant tour, using a deep neural network.
2. Optimally split the giant tour into a feasible solution using an exact split algorithm.

* This work used HPC resources of IDRIS (allocation 2022-AD011011309R2) made by GENCI.

¹ A tour is the ordering of clients the vehicle will visit before returning back to the depot. The optimal number of tours will therefore depend on client’s demands and vehicle capacity.

NOFSS is a generic approach that will be introduced and tested in the context of CVRP, even if it may be used for a larger class of routing problems. NOFSS relies on a deep neural network that learns a giant tour policy and a dynamic programming algorithm, called Split [8]. Split modifies the giant tour into a feasible solution with respect to vehicle capacity and clients demands. It acts as an oracle that provides feedback on the quality (the total travelled distance) of the giant tour generated from our neural network. This makes it possible to train the NOFSS model through REINFORCE algorithm.

Alongside NOFSS introduction, we present an extensive comparison of various NOFSS and NCO models with state-of-the-art CVRP (meta)heuristics. Results show that, by exploring the search space of giant tours, NOFSS allows to implicitly learn competitive routing policies.²

The paper is organized as follows: Section 2 formally introduces the CVRP and notations; Section 3 introduces related work focusing on approaches based on machine learning; Section 4 presents NOFSS; Section 5 presents the experimental protocol as well as results. Discussions and perspectives conclude the paper.

2 Problem Statement

The Capacitated Vehicle Routing Problem (CVRP) is one of the basic types of routing problems where information associated with the clients, the depot and the vehicles are deterministic and known in advance. We consider a set of n clients dispatched on the Euclidean plan and a single depot. In the depot, there is a fleet of homogeneous vehicles with identical transport capacity C . We associate to the clients their coordinates (x_i, y_i) and their demands of goods to deliver $0 \leq d_i \leq C$ ($i \in \{1, \dots, n\}$). We associate to the depot its coordinates (x_0, y_0) . The demands cannot be split, meaning that a vehicle must satisfy the demand at once. The objective is to minimize the total travelled distance when serving all the clients.

The problem can also be formulated using graph theory [9]. We consider a complete graph $G(V, E)$, where $V = \{0, \dots, n\}$ is the vertex set (the vertex 0 represents the depot) and $E = \{(u, v) \in V \times V, u \neq v\}$ is the edge set. We associate with each edge a cost defined as the distance between two vertices. We can represent it as a cost matrix D where $D_{uv} = \sqrt{(x_u - x_v)^2 + (y_u - y_v)^2}$, $(u, v) \in E$. The goal is in this case to find simple circuits called tours such that all clients are served without transgressing the vehicles' capacity and the total travelled distance is as minimum as possible.

3 Related Work

3.1 Neural Combinatorial Optimization (NCO) for the CVRP

We refer to the use of end-to-end deep neural network approaches for solving difficult combinatorial optimization as the Neural Combinatorial Optimization (NCO) framework [3]. In this section, we review the use of this framework to learn construction-based policies for routing problems.

Although the use of neural networks for solving combinatorial optimization problems dates back longer than the appearance of modern deep learning architectures [10], their use has faded away in favor of more efficient metaheuristics. The success of deep learning and reinforcement learning has revived the interest in studying deep neural networks for solving this class of problems. More precisely, with the appearance of the sequence-to-sequence type approaches and the attention mechanism. The general framework (Figure 1) considers two neural networks called respectively encoder and decoder, which can be of different types. The encoder generates the *embeddings* of each element of a problem instance (clients and depot). Embeddings can be viewed as an alternative representation of the element in a higher dimension vector space (\mathbb{R}^d with generally $d = 64$ or $d = 128$). This representation is intended to encompass meaningful features that will be used during the decoding phase. The decoder uses the history of the already visited elements (clients or depot) to compute a query vector that summarizes the solution under construction through a single vector. The query along with the embeddings are used to compute a probability distribution of selecting

² Our implementation and results will be available on the following repository <https://github.com/AYaddaden/NOFSS>

the next element via an attention module. To do so, the attention module confronts the query $q \in \mathbb{R}^d$ to the elements embeddings $e_i \in \mathbb{R}^d$ in order to give attention scores s_i either via a scaled dot-product (i.e. $s_i = \frac{q \cdot e_i^T}{\sqrt{d}}$) or via an additive attention defined as $s_i = v^T \cdot \tanh(W_q \cdot q + W_e \cdot e_i)$ with $W_q, W_e \in \mathbb{R}^{d \times d}$, $v \in \mathbb{R}^d$ being learnable parameters. The scores s_i will be converted into a probability distribution by a softmax function³.

Pointer Networks [2] was the seminal work that considered training LSTM-based encoder and decoder along with an additive attention module via supervised learning on a dataset of TSP instances. The approach successfully solved instances of sizes between 10 and 50 cities. It was next improved by using a policy-based reinforcement learning algorithm for training, namely REINFORCE with critic baseline, thus avoiding the need of a supervision, i.e. to have ground truth optimal solutions for the TSP dataset's instances [3]. Reinforcement learning proved to be more effective for training models on instances of size between 20 and 100 cities, thus achieving better results than Pointer Networks.

Nazari et al. [6] applied the NCO approach to CVRP. Their model considered 1D convolutions instead of an LSTM encoder in order not to bias the model on the inputs' order – LSTM are indeed better suited for modeling sequences where input's order matters. Comparison with classic CVRP algorithms (Clarke and Wright *savings* heuristic and the Sweep algorithm) shows that the deep neural network model performs better on training and test instances' sizes ranging from 10 to 100 clients. It appears also, that the choices of the encoder and decoder are of extreme importance in order to improve the learned policy. The Attention Model (AM) improves the results on the TSP and the CVRP by introducing a model entirely based on the attention mechanism [5]. It uses a Transformer encoder and computes the query vector using a Multi-head attention [11]. The Transformer encoder allows taking into account the graph structure of the TSP and the CVRP in the same way Graph Neural Networks do, thus giving a better representation of the instances. Also, they introduce a new baseline for the REINFORCE algorithm; a greedy rollout baseline that is a copy of AM that gets updated less often.

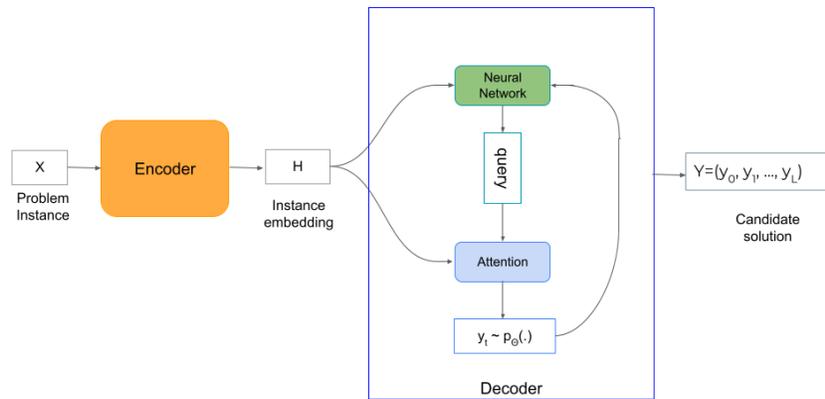


Fig. 1. The general encoder-decoder framework used to solve routing problems. The encoder takes as input a problem instance X and outputs an alternative representation H in an embedding space. The decoder iteratively constructs the candidate solution Y by adding a client or a depot y_t at each step t until all clients are visited.

3.2 Two-step algorithms for the vehicle routing problem

Classical two-step construction approaches for solving the CVRP involve (i) partitioning the clients into feasible clusters with regard to vehicle capacity and (ii) ordering them into routes of minimum length. Based on how the two operations are orchestrated, we can distinguish two types of two-step algorithms: Cluster-first Route-second and Order-first Split-Second.

In Cluster-first Route-second algorithms, the clients are first grouped together following the vehicle capacity constraint, then a traveling salesman problem is solved for each cluster using an

³ $\text{softmax}(s_i) = \frac{\exp(s_i)}{\sum_{j=1}^K \exp(s_j)}$

exact solver or heuristics. The Sweep algorithm is the most common algorithm of this type [12]. Feasible clusters are constructed by considering the polar angle between the clients and the depot, then for each cluster a TSP is solved. An extension of this algorithm called the petal algorithm considers generating several routes and selects the final routes of the solution by solving a set partitioning problem [13]. Another work considers obtaining the clusters by solving a generalized assignment problem [14]. One major drawback of this approach is that it is not computationally efficient due to the clustering algorithms [15].

On the other hand, Order-first Split-second algorithms consider first ordering the customers into a sequence called a giant tour, to then, decompose it into a set of feasible tours considering the vehicle capacity. Traveling salesman problem heuristics are used to get giant tours, and the CVRP tours can be obtained optimally from the giant tours by solving a shortest path problem, as we will detail later. The first documented approach of this type generates the giant tour by random permutation of clients' visit order, followed by a 2-opt improvement, and then builds the routes using Floyd's algorithm [7]. Prins proposed the first genetic algorithm for the CVRP that relies on the Order-first Split-second approach, which was competitive with the best metaheuristic at that time (Tabu Search) [8]. In their approach, the authors proposed a representation of the chromosomes as giant tours and introduced the *Split* procedure based on an auxiliary acyclic graph generated on top of a giant tour. Bellman's algorithm is used in order to extract the feasible routes. HGS, today's state of the art metaheuristic for the CVRP, also uses a giant tour representation and the Split algorithm [16].

The Order-first Split-second approach is appealing. A recent review of this approach surveys more than 70 research papers that build heuristics and metaheuristics to successfully solve vehicle routing problems [17]. Computationally, it is less expensive to build a giant tour and then to split it than building clusters of clients. Also, the search space is reduced to the space of giant tours instead of the direct solution representations with depot placement. As highlighted in the survey, this search space reduction does not make the optimal solution unattainable, since there is an *optimal* giant tour which corresponds to the optimal solution. In addition, for a given giant tour, only its optimal split is retained. This ensures to prevent too many poor quality solutions from appearing often.

3.3 Graph Neural Networks

Since CVRP instances can be modelled as a graph, it is interesting to use neural networks that takes advantage of this structure. This makes Graph Neural Networks (GNNs) an ideal choice to compute a representation of an instance that captures useful information for the resolution process. We define a GNN by stacking K GNN blocks. Each block k relies on message passing in order to compute the node embeddings $h_u^k, \forall u \in V$. This mechanism can be viewed as a differentiable function that computes node embeddings as follows: $h_u^k = F(h_u^{k-1}, \{h_v^{k-1}\}_{v \in \mathcal{N}(u)}, \{e(u, v)\}_{v \in \mathcal{N}(u)})$, with $\mathcal{N}(u)$ being the set of the neighbor nodes of a node $u \in V$ and $\{e(u, v)\}_{v \in \mathcal{N}(u)}$ the set of edges that link the node u to its neighbors $v \in \mathcal{N}(u)$. We use the instance features as an initial input of the first GNN block. The function F itself relies on two mechanisms: neighborhood message aggregation and node embedding update, defined as:

$$\begin{aligned} m_u^{(k)} &= \text{AGGREGATE}(\{h_v^{(k-1)}\}_{v \in \mathcal{N}(u)}, \{e(u, v)\}_{v \in \mathcal{N}(u)}) \\ h_u^{(k)} &= \text{UPDATE}(h_u^{(k-1)}, m_u^{(k)}) \end{aligned}$$

Aggregation can either be the mean, the maximum or the sum of neighbors' node embeddings. It can also be a weighted sum with weights computed using an attention mechanism [18]. It can take into consideration the edge weights of the neighboring nodes $e(u, v)$. The update function is a deep neural network that computes a new node embedding by using the message from the aggregation and the node embedding from the preceding block. Graph neural network models differ depending on the choice of the AGGREGATE and the UPDATE functions.

We can distinguish two families of GNNs: spectral and spatial GNNs. Spectral GNNs rely on spectral graph representations based on graph signal processing theory, such as GCN [19]. Spatial GNNs, such as GAT [18], exploit the graph topology. Refer to Zhou et al. for a GNN review [20].

In the next section, we describe how we use the *Split* algorithm along with the NCO framework to train GNN models for solving the CVRP.

4 The Neural Order-first Split-second algorithm

As mentioned in the previous section, actual NCO construction-based policies for the CVRP produce a sequence by routing the clients and choosing when to return to the depot iteratively until all clients are served. These policies may lead to more returns to depot than necessary and produce poor quality solutions. For example, a policy can decide to refill in the depot after serving each client even if the vehicle capacity allows for serving more than one client at once. Learning from poor quality solutions can slow down and hamper the learning process and produce suboptimal policies. Instead of this, we propose to let the deep neural network build an indirect solution representation via the construction of the giant tour and to delay the routes construction to the Split algorithm. Thus, our neural network implicitly learns to solve vehicle routing problem instances by exploring the space of giant tours. Alternatively, we can view the neural network's output as a permutation of the clients' visit order, which is close to what is done in works for the TSP [3, 4]. This also simplifies the masking procedure used to avoid the appearance of a client twice in the solution. Another advantage of this approach is that the neural network can learn different policies depending on the variant of the vehicle routing problem (e.g. Capacitated VRP, VRP with Time Windows) without additional adaptation. The Split algorithm will handle the additional constraints, and the neural network learns the policy accordingly. Unlike other learning-based construction approaches that build a solution in a variable number of steps due to the return to the depot to refill, our neural network builds the giant tour in a fixed number of steps equal to the number of clients in the instance. Algorithm 1 presents the general approach that will be detailed afterwards.

For a given instance X of the CVRP, our neural network defines a stochastic policy that outputs the probability of generating a giant tour as a sequence Y . Using the probability chain rule, and with θ the parameters of the neural network, this policy is defined as follows:

$$P_{\theta}(Y|X) = \prod_{t=0}^{n-1} p_{\theta}(y_t|y_0, \dots, y_{t-1}, X)$$

After sampling a sequence Y from P_{θ} , Y is then transformed into feasible routes using the Split algorithm with regard to the vehicle's capacity constraint. The Split algorithm can be viewed as an oracle that evaluates the goodness of a giant tour by returning the associated solution's total travelled distance. This evaluation makes it possible to train our deep neural network via reinforcement learning. We define the loss as the expected tour lengths of the Y sequences evaluated by the Split algorithm, i.e. $\mathcal{L}(\theta) = \mathbb{E}_{X \sim \mathcal{D}, Y \sim P_{\theta}(\cdot|X)} [\text{Split}(Y, X)]$. The objective is to find the best parameters θ that will output good quality sequences Y that would result on short tour lengths. For this, we rely on ADAMW as a gradient descent optimizer during training. In order to compute the gradient of the loss, we use REINFORCE with Rollout baseline [5]:

$$\nabla_{\theta} \mathcal{L}(\theta) = \mathbb{E}_{X \sim \mathcal{D}, Y \sim P_{\theta}(\cdot|X)} \left[\left(\text{Split}(Y, X) - b(X) \right) \nabla_{\theta} \log P_{\theta}(Y|X) \right]$$

The gradient $\nabla_{\theta} \mathcal{L}(\theta)$ is approximated using Monte Carlo sampling over a batch of B i.i.d CVRP instances as follows:

$$\nabla_{\theta} \mathcal{L}(\theta) \approx \frac{1}{B} \sum_{i=1}^B \left[\left(\text{Split}(Y_i, X_i) - b(X_i) \right) \nabla_{\theta} \log P_{\theta}(Y_i|X_i) \right]$$

The baseline $b(X)$ is used to reduce the gradient variance, leading to an acceleration of the learning process. We use the greedy rollout baseline $b(X) = \text{Split}(Y^{BL}, X)$ which is an evaluation of the optimal Split of the giant tour Y^{BL} resulting from a copy of the learning neural network with parameters θ^{BL} that acts greedily, i.e. it chooses the next client with the highest probability of appearance at each time step. This baseline proved to be more efficient than actor-critic or REINFORCE with an exponential moving average baseline [5]. During validation, if the performance of θ is significantly better than that of θ^{BL} according to a t-test ($\alpha = 5\%$), the baseline is updated with the parameters of P_{θ} , i.e. θ^{BL} is set to θ .

4.1 Instance features

For each instance X , we define the nodes and edges features as follows:

Algorithm 1.1: NOFSS REINFORCE with Rollout Baseline

```

1 Inputs:  $\theta$ , Number of epochs  $E$ , batch size  $B$ , number of instances  $K$ , number of clients  $n$ ,
   vehicle capacity  $C$ , t-test threshold  $\alpha$ 
2  $T \leftarrow \frac{K}{B}$ 
3  $\theta^{BL} \leftarrow \theta$ 
4 for  $e \leftarrow 1$  to  $E$  do // train for  $E$  epochs
5   for  $t \leftarrow 1$  to  $T$  do // loop over the  $T$  instance batches
6     // Get a batch of  $B$  CVRP instances with  $n$  clients
7      $X_i \leftarrow \text{getInstance}(n, C), \quad \forall i \in \{1, \dots, B\}$ 
8     // Sample a giant tour according to the learning policy  $P_\theta$ 
9      $Y_i \leftarrow \text{SampleGiantTour}(X_i, P_\theta), \quad \forall i \in \{1, \dots, B\}$ 
10    // Generate a giant tour greedily according to the policy  $P_{\theta^{BL}}$ 
11     $Y_i^{BL} \leftarrow \text{GreedyGiantTour}(X_i, P_{\theta^{BL}}), \quad \forall i \in \{1, \dots, B\}$ 
12    // Evaluate giant tours total travel cost
13     $L_i \leftarrow \text{Split}(X_i, Y_i, C) \quad \forall i \in \{1, \dots, B\}$ 
14     $L_i^{BL} \leftarrow \text{Split}(X_i, Y_i^{BL}, C) \quad \forall i \in \{1, \dots, B\}$ 
15    // Compute the loss and update the neural network parameters
16     $\nabla_\theta \mathcal{L} \leftarrow \frac{1}{B} \sum_{i=1}^B (L_i - L_i^{BL}) \nabla_\theta \log P_\theta(Y_i | X_i)$ 
17     $\theta \leftarrow \text{AdamW}(\theta, \nabla_\theta \mathcal{L})$ 
18  end
19  if  $t\text{-test}(P_\theta, P_{\theta^{BL}}) < \alpha$  then
20     $\theta^{BL} \leftarrow \theta$ 
21  end

```

Node features. Each node $u \in V$ is represented as a quadruplet $(x_u, y_u, \hat{d}_u, a_u)$ where (x_u, y_u) are the node coordinates sampled from a uniform distribution $\mathcal{U}([0, 1] \times [0, 1])$, $\hat{d}_u = d_u/C \in [0, 1]$ is the normalized demand and $a_u = \text{atan}((y_u - y_0)/(x_u - x_0)) \in]-\pi/2, \pi/2[$ is the polar angle between the node u and the depot node 0.

Edge features. For each edge $(u, v) \in E$, we define the edge features as the Euclidean distance between the nodes u and v (i.e. $d(u, v) := \|u - v\|, \forall (u, v) \in E$). The distance between two nodes in the instance is an interesting feature in the case of vehicle routing problems, since it is information that characterizes the problem well, and it appears in the objective function.

4.2 NOFSS Encoding-Decoding architectures

The NOFSS approach is agnostic to the choice of the encoding and decoding model architectures. Thus, we propose to train various encoder-decoder models that rely on different graph neural networks (GNNs) and a GRU recurrent cell for decoding. The decoded sequence is passed to the Split algorithm in order to retrieve a candidate solution for the instance (Figure 2).

Encoding. We experiment three GNN Encoders for our approach: GCN (a spectral GNN), GAT (a spatial GNN) and TransformerConv (a spatial GNN) [21]. Each encoder have K similar blocks. The GNN outputs an embedding for each node (clients and depot) $h_u^{(K)} \in \mathbb{R}^d, \forall u \in V$ and a graph representation computed using an average pooling $\bar{h} = 1/|V| \sum_{u \in V} h_u^{(K)}$. Finally, to distinguish

the clients embeddings from the depot embedding $h_0^{(K)}$, we pass them into a feedforward layer $h_u = W_c \cdot h_u^{(K)} + b_c, \forall u \in V - \{0\}$, with $W_c \in \mathbb{R}^{d \times d}, b_c \in \mathbb{R}^d$ being respectively the weights and the bias of the layer.

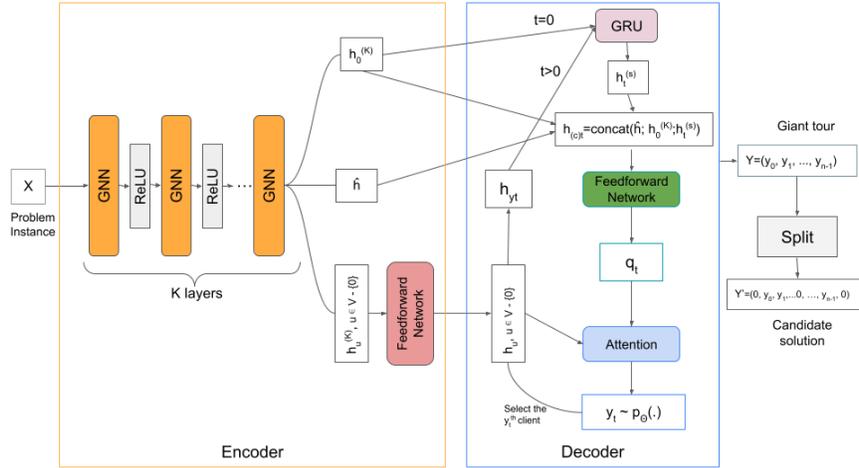


Fig. 2. Our proposed NOFSS model for solving CVRP instances.

Neighborhood definition. As highlighted in Section 2, we can define a CVRP instance as a complete graph. We define the neighborhood $\mathcal{N}(u)$ of a client node $u \in V - \{0\}$ as the κ nearest nodes in terms of Euclidean distance and the depot 0, since it is important for the client’s representation to be aware of the depot’s existence (i.e. $\mathcal{N}(u) = \{v_1, v_2, \dots, v_\kappa \in V; \|v_1 - u\| \leq \|v_2 - u\| \leq \dots \leq \|v_\kappa - u\|\} \cup \{0\}$). For the depot, we consider that it is connected to every client. An example of an instance neighborhood definition is depicted in Figure 3. The central node (red square) represents the depot, while the other nodes (blue circles) represent the clients. An edge exists between nodes u and v if $v \in \mathcal{N}(u)$. The number of nearest neighbors κ is determined per instance. We set it to be the average number of clients per route as if they were uniformly distributed on the routes, i.e. $\kappa = \frac{n}{m}$ with n being the number of clients and m being the lower bound of the number of routes. m is determined as the sum of all clients’ demands divided by the vehicle’s capacity rounded to the next integer ($m = \lceil \frac{\sum_{i=1}^n d_i}{C} \rceil$). The advantage of such a definition of κ is that it takes into account the characteristics of the instance in terms of the number of clients, their demands, and the capacity of the vehicles instead of selecting an arbitrary number of neighbors.

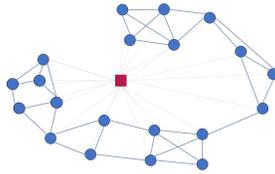


Fig. 3. CVRP instance with relationships between neighboring nodes (central square node is the depot).

Decoding. Since we are decoding a sequence of clients’ order, we use a GRU recurrent cell [22]. GRU is relevant as it enables capturing the sequence representation while taking into account the order of its elements. It takes as input the previously selected client representation at step $t - 1$ concatenated with the depot representation $h_0^{(K)}$ and incorporates it in the global representation

of the partial giant tour. At $t = 0$, we only use the depot representation $h_0^{(K)}$ as input to the GRU.

$$h_t^{(s)} = \begin{cases} \text{GRU}(h_0^{(K)}), & t = 0 \\ \text{GRU}([h_{y_{t-1}}; h_0^{(K)}]), & t > 0 \end{cases}$$

The graph embedding \hat{h} , the depot embedding $h_0^{(K)}$ and the sequence embedding $h_t^{(s)}$ are then concatenated together to form a context vector $h_c \in \mathbb{R}^{3d}$. The context vector is then passed to a feedforward layer made of two linear layers with *ReLU* activation function in between to output a query vector $q_t \in \mathbb{R}^d$ i.e. $q = W_2 \cdot \text{ReLU}(W_1 \cdot h_c + b_1) + b_2$ with $W_2 \in \mathbb{R}^{d \times 3d}$, $W_1 \in \mathbb{R}^{d \times d}$, $b_1, b_2 \in \mathbb{R}^d$ being the parameters of the feedforward layer.

To compute the probability of selecting the next client $p_\theta(y_t | y_0, \dots, y_{t-1}, X)$, we compute attention scores s_u ($\forall u \in V - \{0\}$) using a scaled dot-product with a masking mechanism in order to avoid selecting the same client twice. These scores are then clipped within $[-10, 10]$ using *tanh* [5].

$$s_u = \begin{cases} c \cdot \tanh\left(\frac{q_t h_u^\top}{\sqrt{d}}\right), & u \neq y_{t'} \quad t' < t, c = 10 \\ -\infty & \text{otherwise} \end{cases}$$

The attention scores are converted into a probability distribution using the softmax function $p_i = p_\theta(y_t = i | y_0, \dots, y_{t-1}, X) = \text{softmax}(s_i)$. By setting the value of the attention score to $-\infty$, we can perform the masking of already visited clients. Thus, when passed to the softmax function, its associated probability will be 0.

The Split procedure. The algorithm works on the basis of the giant tour output by the neural network augmented with the depot, i.e. $\mathcal{Y} = (y_0, y_1, \dots, y_n)$ with $y_0 = 0$ being the depot. Using the giant tour, we define an auxiliary graph $H(V^H, E^H)$ with $|V^H| = n + 1$. The nodes in V^H indicate the depot (either for return or departure). The edge set indicates all possible sub-sequences that starts from y_i to y_j (y_i, y_{i+1}, \dots, y_j) that do not transgress the vehicle’s capacity constraint. We formulate it as follows: $E^H = \{(i, j) \in V^H \times V^H; i < j, \sum_{k=i+1}^j d_{y_k} \leq C\}$. The edges are weighted as follows: for an edge $(i, j) \in E^H$ we associate the total travelled distance starting from the depot to the client y_{i+1} , visiting the tour (y_{i+1}, \dots, y_j) and going back to the depot from y_j :

$$D^H = \{d_{ij} = \text{dist}(0, y_{i+1}) + \sum_{\substack{k=i+1 \\ j-i > 1}}^{j-1} \text{dist}(y_k, y_{k+1}) + \text{dist}(y_j, 0), \quad \forall (i, j) \in E^H\}$$

This gives us a direct acyclic graph where we solve a shortest path problem using Bellman’s algorithm. The associated shortest path cost represents the best solution length (total travelled distance) for the CVRP instance with regard to the given giant tour.

5 Experiments

Data generation. We follow the data generation protocol of Nazari et al. [6] to consider 3 types of CVRP instances with number of clients $n = 20, 50$ and 100 . For each problem size, we have generated $100k$ instances for training, and two sets of $10k$ instances for validation and test. Clients and depot locations are generated from a uniform distribution $\mathcal{U}(\{[0, 1] \times [0, 1]\})$. The clients’ demands are also uniformly drawn from the interval $[1, 9]$. Vehicles’ capacities are set to $30, 40$ and 50 respectively for $n = 20, 50, 100$.

Hyperparameters. We use an embedding dimension $d = 128$ and a uniform parameter initialization for our deep neural networks $\mathcal{U}(-1/\sqrt{d}, 1/\sqrt{d})$ and set the learning rate to $\eta = 10^{-3}$. The models are trained with a time limit of 100 hours and batch size $B = 128$ on a single NVIDIA V100 GPU with 16 GB of VRAM. For each encoder type, we use $K = 3$ GNN blocks. Implementations use PyTorch and PyTorch Geometric for graph neural networks [23] (Python), while the Split algorithm is implemented in C.

Baselines. We use HGS⁴ [16] as baseline as it is one of the state of the art metaheuristics for the CVRP. We also use classical CVRP heuristics⁵: (i) RFCS [7] as a two-step order-first split-second heuristic, (ii) Sweep [12] as a two-step cluster-first route-second approach, and (iii) Nearest Neighbor heuristic as a single-step construction approach [24]. We also trained the model with TransformerConv encoder in an end-to-end manner for depot and clients choice (Full-learning). We first note that NOFSS models are faster to train, completing $E = 1000$ of learning epochs in the 100 hours time budget, while the Full-learning models perform 1000, 500 and 200 training epochs for instance sizes of 20, 50 and 100 respectively. For the exploitation of the learned policies, we use a greedy decoding which considers the highest probability at each decoding step and a sampling strategy which samples 1280 candidate solutions for each test instance from the probability distributions given by the models. Table 1 reports the results of each approach on the test specifying: average solution lengths (obj.), the average gap (in percentage) to the best average solution lengths and the running time (in seconds) to output a candidate solution for a single instance.

Table 1. NOFSS vs. other algorithms. FL for Full-Learning; exploitation, greedy (G), sampling (S).

Method	$n = 20$			$n = 50$			$n = 100$		
	obj.	gap (%)	time (s)	obj.	gap (%)	time (s)	obj.	gap (%)	time (s)
HGS	6.13	0.00	0.003	10.34	0.00	0.09	15.57	0.00	0.69
RFCS	6.30	2.76	0.02	10.90	5.39	0.57	16.62	6.73	7.53
Sweep	7.55	23.16	0.01	15.60	50.93	0.06	28.56	83.37	0.23
Nearest neighbor	7.39	20.57	0.0004	12.63	22.19	0.001	18.95	21.68	0.01
NOFSS-GCN (G)	6.83	11.41	0.0008	12.31	19.05	0.003	19.41	24.66	0.007
NOFSS-GAT (G)	6.59	7.50	0.006	11.74	13.53	0.02	18.34	17.80	0.05
NOFSS-Transformer (G)	6.50	6.03	0.006	11.57	11.89	0.02	18.13	16.44	0.06
FL-Transformer (G)	6.49	5.87	0.006	11.34	9.67	0.02	17.69	13.61	0.06
NOFSS-Transformer (S)	6.24	1.79	1.37	11.03	6.67	1.56	17.45	12.07	2.43
FL-Transformer (S)	6.18	0.81	2.09	10.79	4.35	2.35	17.32	11.23	8.29

5.1 Comparison with a Full-learning setting

Figure 4 presents the evolution of the average solution length per epoch during training and validation on CVRP instances with 20 clients (left) and 50 clients (right). During training, candidate solutions are sampled from the model and their total lengths are averaged over the training set. Let us note that the models' parameters are updated each time a batch is processed via gradient descent, thus the performance of the models changes every batch during training, while validation is performed using the model resulting from the processing of the last batch in the training set, which is theoretically the best model achieved at the end of the epoch. Also, in validation, we use a greedy decoding instead of sampling. The evolution of the average solution lengths shows that the NOFSS model is able to learn an implicit policy for solving the CVRP by learning to output an indirect representation of the solution. On instances with 20 clients, we can observe that during training, the NOFSS model achieves better average solution lengths than the Full-learning model. On validation, we observe the same trend as in training, but starting from the 600th epoch, the Full-learning model slightly outperforms the NOFSS model. The equivalent performance of the two models is confirmed on the test set with average solution lengths of 6.50 and 6.49 on greedy decoding for NOFSS and Full-learning respectively with similar execution times. On sampling decoding, similar performances are observed, with 0.9 % difference in performance between the two models, but with an advantage in execution time in favor of NOFSS. On CVRP with 50 clients, we observe that NOFSS has a better jump start performance on training and a better final performance for the Full-learning model. We observe 2 % difference in performance for greedy and sampling decoding on the test set. We also note similar sampling times for the two types of models in greedy decoding, while NOFSS being 52 %, 50% and 241% faster in sampling respectively for $n = 20, 50$ and 100.

⁴ <https://github.com/vidalt/HGS-CVRP>

⁵ <https://github.com/yorak/VeRyPy>

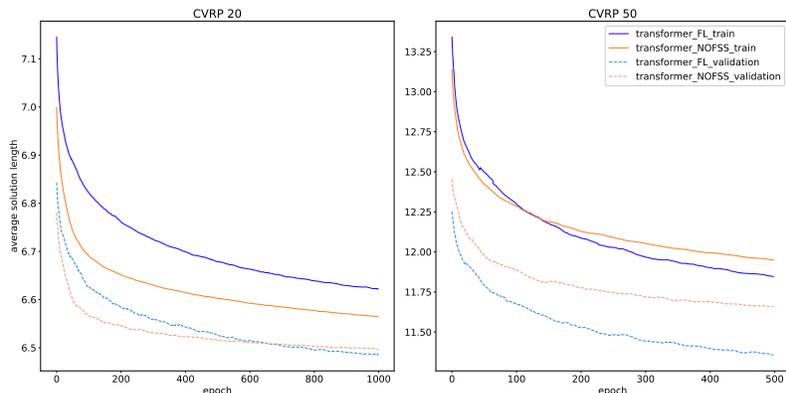


Fig. 4. Learning curves in training and validation for Full-learning (blue) and NOFSS models (orange) on CVRP instances with 20 (CVRP20) and 50 clients (CVRP50); lower is better.

5.2 Comparison to handcrafted heuristics

When compared to handcrafted heuristics, we can observe from Table 1 that either with greedy or sampling exploitation, NOFSS models outperform the Sweep and Nearest neighbor algorithms. NOFSS model seems to output better solution lengths, on average, than RFCS on CVRP with 20 clients when using the sampling strategy but seems to fail scaling to CVRP with 50 and 100 clients. Let us note that while RFCS and NOFSS belong to the same type of two-step strategy, there is a difference in the two approaches in that RFCS explicitly solves a Traveling Salesman Problem, while NOFSS directly evaluates the giant tour using the Split algorithm. The difference in average solution lengths may suggest that NOFSS learned policy is different from a policy that learns to solve a Traveling Salesman Problem.

5.3 Influence of the type of encoder

We investigate the influence of the choice of GNN encoder on models' performance. Figure 5 shows the evolution of the average solutions lengths per epoch in training and validation phases for the 3 types of GNN encoders: GCN, GAT and TransformerConv on CVRP with 20 and 50 clients. We observe the same trends for both training and validation phases, with TransformerConv having the best convergence, followed by GAT encoder and finally by GCN encoder. The instances' representation plays an important role in the resolution process, because a good representation leads to the exploitation of meaningful features and, thus, gives a better solution. The choice of the encoder seems to be a critical part of the model's architecture. It appears from these results that spatial GNNs better perform than spectral GNNs in our evaluation setting. Exploiting the graph topology in the spatial domain seems to benefit more in the context of vehicle routing problems than exploiting the graph structure in the spectral domain. While TransformerConv and GAT are both spatial GNNs, it seems that the way they exploit the node and edges information has an impact on the overall performance of the models.

5.4 On models generalization

We propose to study the generalization of the models trained on a set of instances with a specific size to instances of different size. For this, we evaluate the different test sets on instances of different sizes. For example, we evaluate the NOFSS Transformer model trained on CVRP with 20 clients instances (Transformer-20) on instances with 20, 50 and 100 clients. Table 2 sums up our results. We report the average solution lengths for both greedy and sampling exploitation strategies. For greedy decoding, we report the results for the models trained on the different instance sizes while for sampling, we focus on the model trained on instances sizes which seems more promising based on our findings on the greedy decoding. We observe that for the Transformer-20, the NOFSS model

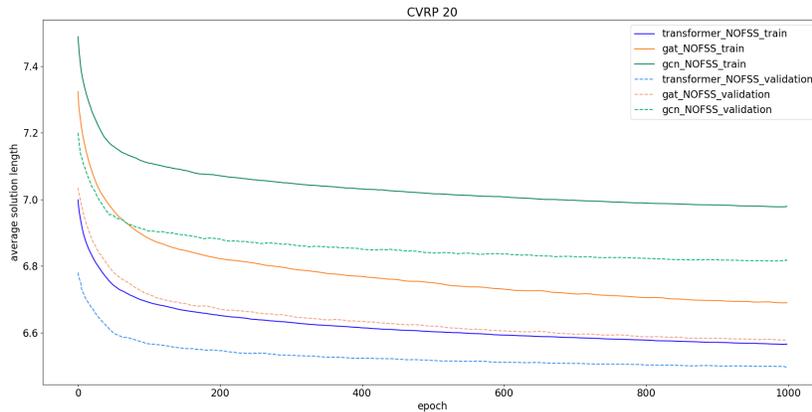


Fig. 5. Comparison of Graph Neural Network encoders on models’ performance (training and validation).

has a better generalization property than the Full-learning model, with performance similar for $n = 20$ and $n = 100$ and better for $n = 50$. Since training models on instances with 20 clients is faster, it is relevant to identify that the NOFSS model is a better choice.

For Transformer-50 and Transformer-100, it appears that, for $n = 20$ NOFSS models have better performances than their Full-learning counterparts while staying competitive for $n = 50$ and $n = 100$. An interesting result observed on Transformer-50 is its good generalization to CVRP instances with 100 clients, as it appears that it achieves better performance than the models trained on instances with 100 clients. This may suggest that relevant invariants that are beyond the instance size are learned while training on instances with 50 clients. We push further our investigations on Transformer-50 by analyzing its performance with a sampling exploitation strategy. While for the instances with 20 clients, the models stay competitive with the ones trained on that size, they achieve the best performances on the sets with instances with 50 and 100 clients. Transformer-50 appears to be a good trade-off between learning speed (it is faster to train than Transformer-100) and performance.

Table 2. Comparison of average solution lengths achieved by the NOFSS and Full-learning models on different instance sizes of the test set.

Trained model	NOFSS (G)			Full-learning (G)		
	20	50	100	20	50	100
Transformer-20	6.50	11.62	18.34	6.49	12.01	18.33
Transformer-50	6.64	11.57	17.97	6.76	11.34	17.52
Transformer-100	6.94	11.79	18.13	6.98	11.65	17.69
	NOFSS (S)			Full-learning (S)		
	20	50	100	20	50	100
Transformer-50	6.31	11.03	17.40	6.25	10.79	17.22

6 Conclusion

In this work, we proposed NOFSS, a two-step algorithm hybridizing a deep neural network model and an exact tour splitting procedure for the Capacitated Vehicle Routing Problem. To the best of our knowledge, this is the first model that proposes a hybridization between a deep neural network and a dynamic programming algorithm to successfully learn an implicit policy based on giant tour generation to solve the CVRP. We conducted extensive experiments on the proposed models with various Graph Neural Network encoders and compared them against classic CVRP heuristics and an end-to-end Full-learning model. Our results show that NOFSS is very competitive, even if it

currently does not surpass end-to-end full-learning approaches. NOFSS is however faster than end-to-end approaches in both training and evaluation. It also shows good generalization properties when trained on instances with a specific size and applied to solve instances of different sizes. The NOFSS model is easier to implement than an end-to-end learning-based policy and does not rely on sophisticated handcrafted search strategies to find good quality solutions.

Future work should investigate more on the generalization of the method to instances of bigger sizes. Also, while we tested only greedy and sampling strategies for exploiting the trained models, other relevant strategies may be interesting such as beam search, or using bigger sample sizes than the one we used since NOFSS has a faster execution time. The solution given by NOFSS can also be a good warm start for further improvement by local search algorithms. Finally, since our approach is generic, it would be interesting to evaluate it on other problems, such as the Vehicle Routing Problem with Time Windows.

References

1. Bengio, Y., Lodi, A., Prouvost, A.: Machine learning for combinatorial optimization: a methodological tour d’horizon. *European Journal of Operational Research* **290** (2021) 405–421
2. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. [arXiv:1506.03134](https://arxiv.org/abs/1506.03134) (2015)
3. Bello, I., Pham, H., Le, Q.V., Norouzi, M., Bengio, S.: Neural combinatorial optimization with reinforcement learning. [arXiv:1611.09940](https://arxiv.org/abs/1611.09940) (2016)
4. Deudon, M., Cournut, P., Lacoste, A., Adulyasak, Y., Rousseau, L.M.: Learning heuristics for the tsp by policy gradient. In: *International conference on the integration of constraint programming, artificial intelligence, and operations research*, Springer (2018) 170–181
5. Kool, W., Van Hoof, H., Welling, M.: Attention, learn to solve routing problems! [arXiv:1803.08475](https://arxiv.org/abs/1803.08475) (2018)
6. Nazari, M., Oroojlooy, A., Snyder, L.V., Takáč, M.: Reinforcement learning for solving the vehicle routing problem. [arXiv:1802.04240](https://arxiv.org/abs/1802.04240) (2018)
7. Beasley, J.E.: Route first—cluster second methods for vehicle routing. *Omega* **11** (1983) 403–408
8. Prins, C.: A simple and effective evolutionary algorithm for the vehicle routing problem. *Computers & operations research* **31** (2004) 1985–2002
9. Toth, P., Vigo, D.: *The vehicle routing problem*. SIAM (2002)
10. Smith, K.A.: Neural networks for combinatorial optimization: a review of more than a decade of research. *INFORMS Journal on Computing* **11** (1999) 15–34
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: *Advances in neural information processing systems*. (2017) 5998–6008
12. Gillett, B.E., Miller, L.R.: A heuristic algorithm for the vehicle-dispatch problem. *Operations research* **22** (1974) 340–349
13. Ryan, D.M., Hjorring, C., Glover, F.: Extensions of the petal method for vehicle routeing. *Journal of the Operational Research Society* **44** (1993) 289–296
14. Fisher, M.L., Jaikumar, R.: A generalized assignment heuristic for vehicle routing. *Networks* **11** (1981) 109–124
15. Hiquebran, D., Alfa, A., Shapiro, J., Gittoes, D.: A revised simulated annealing and cluster-first route-second algorithm applied to the vehicle routing problem. *Engineering Optimization* **22** (1993) 77–107
16. Vidal, T.: Hybrid genetic search for the cvrp: Open-source implementation and swap* neighborhood. *Computers & Operations Research* **140** (2022) 105643
17. Prins, C., Lacomme, P., Prodhon, C.: Order-first split-second methods for vehicle routing problems: A review. *Transportation Research Part C: Emerging Technologies* **40** (2014) 179–200
18. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. [arXiv:1710.10903](https://arxiv.org/abs/1710.10903) (2017)
19. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. [arXiv:1609.02907](https://arxiv.org/abs/1609.02907) (2016)
20. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. *AI Open* **1** (2020) 57–81
21. Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., Sun, Y.: Masked label prediction: Unified message passing model for semi-supervised classification. [arXiv:2009.03509](https://arxiv.org/abs/2009.03509) (2020)
22. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. preprint [arXiv:1409.1259](https://arxiv.org/abs/1409.1259) (2014)
23. Fey, M., Lenssen, J.E.: Fast graph representation learning with PyTorch Geometric. In: *ICLR Workshop on Representation Learning on Graphs and Manifolds*. (2019)
24. Rasku, J., Kärkkäinen, T., Musliu, N.: Meta-survey and implementations of classical capacitated vehicle routing heuristics with reproduced results. *Toward Automatic Customization of Vehicle Routing Systems* (2019)

LSG Attention: Extrapolation of pretrained Transformers to long sequences

Charles Condevaux¹[0000-0002-0819-9056] and
Sébastien Harispe²[0000-0001-5630-2743]

¹ CHROME, University of Nîmes, France
charles.condevaux@unimes.fr

² EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, France
sebastien.harispe@mines-ales.fr

Abstract. Transformer models achieve state-of-the-art performance on a wide range of NLP tasks. They however suffer from a prohibitive limitation due to the self-attention mechanism, inducing $O(n^2)$ complexity with regard to sequence length. To answer this limitation we introduce the LSG architecture which relies on Local, Sparse and Global attention. We show that LSG attention is fast, efficient and competitive in classification and summarization tasks on long documents. Interestingly, it can also be used to adapt existing pretrained models to efficiently extrapolate to longer sequences without additional training. Along with the introduction of the LSG attention mechanism, we propose a PyPI package to train new models and adapt existing ones based on this mechanism.

Keywords: Attention mechanism · Long sequences · Extrapolation

1 Introduction

Transformer models [33] are nowadays state-of-the-art in numerous domains, and in particular in NLP where they are used in general language models, and to successfully tackle several specific tasks such as document summarization, machine translation and speech processing to cite a few [13,26]. The cornerstone of Transformer models is the Attention mechanism used to iteratively build complex context-dependent representations of sequence elements, e.g. tokens, by dynamically aggregating prior representations of these elements. Using self-attention, a popular Attention flavour, this is made by computing full attention scores defining how each prior element representation will contribute to building the new representation of an element. Considering a sequence of n elements, the computation of the attention scores is therefore of complexity $O(n^2)$ which is prohibitive dealing with long sequences. Since a large number of models based on full attention have been trained on various datasets and tasks, we are therefore interested in extrapolating those models to long sequences by simply, post training, substituting the full attention trained on shorter input sequences by new attention mechanisms adapted to longer sequences. Common pretrained

models (e.g. RoBERTa) are indeed known to underperform when extrapolated to sequences of length exceeding the 512 tokens considered during training. This is due to the nature of the attention mechanism which largely impacts extrapolation capabilities: full attention usually fails to extrapolate, even considering post hoc adaptations, e.g. using a relative positional embedding [30] or duplicating the positional embedding [3]. Defining new attention mechanisms that can efficiently substitute full attention in pretrained models that are not originally capable of handling long sequences would avoid the costs induced by training large language models from scratch. The main contributions of this paper are:

1. LSG (Local Sparse Global) attention, an efficient $O(n)$ approach to approximate self-attention for processing long sequences.³
2. Results demonstrating that LSG is fast, efficient and competitive on classification and summarization tasks applied to long documents. It is also shown that LSG can adapt and extrapolate existing pretrained models not based on LSG, with minimal to no additional training.
3. A procedure and a PyPI package to convert existing models and checkpoints (e.g. RoBERTa, DistilBERT, BART) to their LSG variant.⁴

Compared to several contributions aiming at reducing the complexity of self-attention introduced hereafter, a specific focus is given in our work on the extrapolation of existing Transformer models, i.e. reuse, to longer sequences.

2 Related works

Several contributions have been devoted to the optimization of the Attention mechanism. Four categories of approaches can be distinguished in the literature: (i) recurrent models such as Transformers-XL [12] and Compressive Transformers [25] which maintain a memory of past activation at each layer to preserve long-range contextual information; (ii) factorization or kernels aiming at compressing attention score matrices, such as Linformer [34] or Performer [9]; (iii) models based on clustering such as Reformer [21] that dynamically define eligible attention patterns (i.e. where attention may be made); and (iv) models based on fixed or adaptive attention patterns, e.g. Longformer [3] or Big Bird [37].

Recurrent approaches iteratively process the sequence by maintaining a memory to enable long-range dependencies. They generally suffer limitations induced by specific, slow, and difficult to implement forward and back propagation procedures. Alternatively, one of the main line of study for reducing the complexity of Attention is thus to perform sparsity by limiting the number of elements on which new representations will be based, i.e. reducing the number of elements with non-null attention scores. This approach is motivated by the observation of global or data-dependent positional patterns of non-null attention scores depending on the task [7]. The sparsity of attention scores in the traditional Attention mechanism is indeed documented in the literature. It has for instance been shown that in practice, full attention tends to overweight close elements

³ Checkpoints and datasets are available at <https://huggingface.co/ccdv>

⁴ https://github.com/ccdv-ai/convert_checkpoint_to_lsg

in average, in particular for MLM, machine translation, and seq-to-seq tasks in general [10]. Moreover, according to analyses on the use of multi-head full attention on specific tasks, e.g. machine translation, numerous heads learn similar simple patterns [27]. Such redundant patterns may be hardcoded implementing fixed-positional patterns, eventually in a task-dependent manner.

Two main approaches are discussed in the literature for implementing sparsity: fixed or adaptative patterns based on whether attention scores are computed considering (1) predefined fixed elements based on their location in the sequence, or (2) elements selected from a given procedure. As an example, [35] have shown that fixed $O(n)$ convolutions can perform competitively on machine translation. Longformer proposes an alternative $O(n)$ approach based on sliding and global patterns [3]. In the context of image, audio, and text processing, [7] propose sparse Transformer, an $O(n\sqrt{n})$ model based on sparse factorization of the attention matrix relying on specific 2D factorized attention schemes. Those approaches however prevent the use of task-dependent dynamic patterns. Considering adaptative patterns, [35] also introduced dynamic convolutions as an $O(n)$ complexity substitute to self-attention. Kernels defining the importance of context elements are specified at inference time rather than fixed after training. Another example is Reformer [21], an $O(n \log n)$ approach based on locality-sensitive hashing (LSH) based on random projections.

In a transverse manner, several authors, explicitly or implicitly motivated by the compositional nature of language have studied structured approaches in which subsequences (i.e. blocks) are processed independently and then aggregated. This aims at implementing a local or global dynamic memory for considering close to long-range dependencies. Some approaches use a blockwise approach to reduce the quadratic complexity induced by large sequences in encoder-decoder architectures [4]. Other propose a chunkwise attention in which attention is performed in a blockwise manner adaptively splitting the sequence into small chunks over which soft attention is computed [8]. This idea is also used in Transformer-XL [12]. Masked block self-attention mechanism in which the entire sequence is divided into blocks, to further 1) apply self-attention intra-block for modeling local contexts, to further 2) apply self-attention inter-block for capturing long-range dependencies, as also been proposed [31]. Such an approach enables implementing some forms of connectivity between all positions over several steps without being restricted by full attention limitations. This can also be achieved by factorization techniques, e.g. [7]. More recently authors have proposed global attention mechanisms encoding information related to blocks on which attention is based [1,39,16].

This paper presents LSG (Local, Sparse and Global) attention based on block local attention to capture local context, sparse attention to capture extended context, and global attention to improve information flow. Contrary to prior work mostly focusing on defining new models, the proposed LSG Attention mechanism is model agnostic and aims to facilitate adapting existing (pretrained) models for them to be used on long sequences.

3 LSG: mixing Local, Sparse and Global attentions

LSG assumes (1) that locally, a token needs to capture precise low level information using dense attention, (2) as the context grows, higher level information is sufficient, i.e. a limited number of tokens specifically selected are sufficient. LSG therefore relies on block local attention to capture local context, sparse attention to capture extended context, and global attention to improve information flow.

Local Attention. LSG takes advantage of a block-based processing of the input. The sequence is split into n_b non-overlapping chunks of size b_t . For a given block, each token attends to the tokens inside the block, as well as to those in the previous and next blocks. The local attention window is asymmetrical since a token can connect up to $2 \times b_t - 1$ tokens on the left or on the right.

Sparse Attention. Sparse connections are used to expand the local context by selecting additional tokens. These tokens can be directly selected based on a specific metric or using some computation such as a pooling method. In the proposed approach, each attention head can process different sparse tokens independently. Sparse attention also relies on a block structure where the sparse selection is done inside each block. Five alternative criteria can be used in LSG.

1. *Head-wise strided*: Each attention head attend to a set of tokens following a specific stride defined as the sparsify factor. Figure 1 shows the selection pattern.
2. *Head-wise block strided* selects consecutive tokens, see Figure 2.

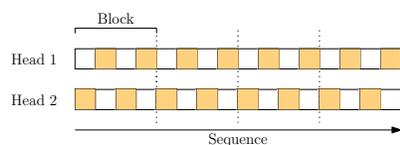


Fig. 1. Head-wise selection (stride 2).

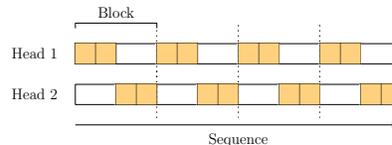


Fig. 2. Block selection (stride 2).

3. *Average pooling*: sparse tokens are computed using average pooling on blocks. For a block of size b_t and a sparsify factor f , pooling is applied to each block with a window of f and a stride of f to produce b_t/f tokens.

4. *Max norm*: selects tokens that are most likely to have high scores. Finding those keys efficiently is difficult in practice so we use a simple and deterministic heuristic selecting inside each block and each head b_t/f tokens with the highest key norm. Indeed, note that for a query and a key $\mathbf{q}, \mathbf{k} \in \mathbb{R}^d$, $\mathbf{q}\mathbf{k}^\top = \cos(\theta)\|\mathbf{q}\|\|\mathbf{k}\|$. If $\cos(\theta)$ is positive and $\|\mathbf{k}\|$ is high, the key will likely dominate the softmax regardless of the query.

5. *LSH Clustering*: non deterministic approach relying on the LSH algorithm [2]. For each block, b_t/f clusters are built using a single round LSH. To get $c = b_t/f$ hashes and for an input $\mathbf{x} \in \mathbb{R}^d$, a random matrix $\mathbf{R} \in \mathbb{R}^{d \times c/2}$ is generated, such

that $h(\mathbf{x}) = \arg \max([\mathbf{x}\mathbf{R}; -\mathbf{x}\mathbf{R}])$ with $[\mathbf{a}; \mathbf{b}]$ the concatenation of two vectors. Using the key matrix as input, each token inside the block gets a cluster index from $h(\mathbf{x})$. Tokens inside a cluster are averaged.

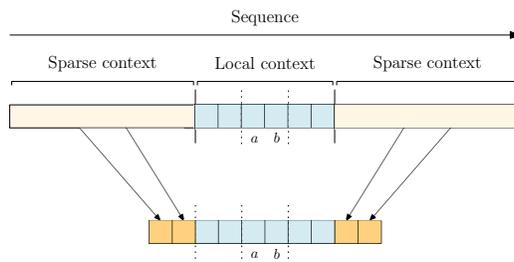


Fig. 3. Local and sparse contexts with a block size of 2 and a sparsity factor of 4. Queries a and b will attend to 6 local keys and 4 sparse keys.

Global Attention. Global tokens improve the flow of information inside the model. They attend to every tokens across the sequence and all tokens attend to them. Rather than picking a subset of tokens, additional tokens are prepended to the sequence and trained using their own embedding matrix (their number is an hyperparameter). When a model is converted to its LSG version, the first global token is initialized as the sum of the [CLS] token and the first position from the positional embedding. The other global tokens are initialized as the sum of [MASK] token and the other positions from the positional embedding. Thus, they can be trained and fine-tuned independently.

Positional Embedding. It is necessary to modify the positional embedding matrix to reuse existing models to process long sequences. In LSG, instead of randomly initializing the new positions, the original matrix is duplicated and concatenated until the desired max sequence length is reached.

4 Experiments

We evaluate LSG in the context of model extrapolation by replacing full attention by LSG attention in various architectures. The official RoBERTa-base checkpoint for classification tasks and BART-base checkpoint for summarization tasks are extrapolated using LSG attention. All metrics are reported for the test set except in the case where only the validation set is available – datasets are all available on the HuggingFace hub. We use a batch size of 32, a linear decaying learning rate, a dropout rate of 0.10 and Adam (0.9, 0.999) optimizer [20] for classification and summarization experiments. An experiment comparing several attention approximations to extrapolate RoBERTa in an MLM task is first discussed; it is used to limit the number of tested alternatives, and therefore reduce the cost of the proposed evaluations. All experiments are conducted on NVIDIA Quadro RTX 8000 48Gb GPUs.

6 C. Condevaux and S. Harispe

4.1 RoBERTa extrapolation on MLM

A test on a MLM task is performed to question the ability of an attention mechanism to extrapolate a model to longer sequences without additional training. A RoBERTa-base model is here considered and two experiments are conducted. First, the full attention is substituted by different kinds of attention (kernel, factorization, local, fixed pattern) and each model is evaluated on sequences of the same length as those considered during RoBERTa initial training (512 tokens). In the second experiment, their ability to extrapolate to 4096 tokens sequences without additional training is tested (positional embedding duplicated 8 times).

A random sample from Wikipedia + BookCorpus + CC_News is used; BPC and MLM accuracy are in Table 1. RoBERTa’s author report a 1.880 BPC loss; we obtain a comparable loss of 1.881 on this random sample.

Only Longformer, Big Bird and LSG obtain competing BPC while processing sequences of the same length as those considered during the original RoBERTa training. Other approaches such as Linformer, Performer or Reformer require additional MLM fine-tuning to leverage an existing checkpoint. It can be seen that RoBERTa fails to extrapolate to longer sequences (+2.454 BPC), which highlights that full attention is not suitable for extrapolation. Longformer and Big Bird are able to perform some form of extrapolation. Therefore, we restrict our comparison to these two approaches in order to limit experimentation costs.

Attention	512 length		4,096 length	
	BPC	Accuracy	BPC	Accuracy
RoBERTa (full) [23]	1.881	0.732	4.335	0.359
Linear Attn. [19]	11.324	0.061	11.474	0.058
Efficient Attn. [32]	21.022	0.102	20.574	0.097
Performer [9]	10.382	0.107	10.556	0.102
Linformer (128 proj.) [34]	22.176	0.098	20.386	0.032
Reformer [21]	17.602	0.003	18.608	0.002
Longformer (512) [3]	1.929	0.726	2.051	0.708
Big Bird (64) [37]	1.881	0.732	2.439	0.659
LSG-Norm (128/2) (block size / sparsity)	1.919	0.727	2.032	0.712
LSG-Stride (128/2)	1.938	0.724	2.046	0.710
LSG-BlockStride (128/2)	1.940	0.724	2.048	0.709
LSG-Pooling (128/2)	1.968	0.720	2.064	0.706
LSG-LSH (128/2)	1.969	0.719	2.065	0.705

Table 1. BPC and MLM accuracy of RoBERTa-base with various Attention.

4.2 Classification Tasks

We compare LSG to Longformer [3] and Big Bird [37], two approaches able to process long sequences with a similar number of parameters. Tests are performed considering sparse attentions with a block size of 128 and a sparsity factor of 4.

Datasets. Standard NLP datasets are used. *IMDb* [24]: binary sentiment analysis classification task from movie reviews. *ArXiv* [17]: set of documents from ArXiv where the objective is to predict a topic from 11 available classes. Because there is no official split, a random one is made of 28K, 2.5K and 2.5K documents for train, validation and test. *Patent* [29]: subset of the Big Patent summarization dataset. The task is redefined as a classification task where the objective is to predict the patent category using the full document (9 classes, random split of 25K, 5K and 5K documents for train, validation and test). Some specific domains are highly dependent on processing long sequences, e.g. legal domain in which sentences tend to be long and complex. To demonstrate the ability of LSG to leverage pretrained models in such cases, the following three datasets are chosen from LexGlue [6], a benchmark focused on legal documents. Tasks where the input is on average significantly longer than 512 tokens have been selected. *Scotus*: given a court opinion, the task is to predict the relevant issue area among 14 choices. *ECtHRa* and *ECtHRb*: the objective is to predict which articles of the European Court of Human Rights (ECHR) have been violated (if any) from case description: multi-label task (10 + 1 labels).

Training setup and architecture. To make a fair comparison between models and architectures, fine-tuning is done with the same learning rate, number of steps and batch size. To show that LSG is compatible with different architectures, the LexGlue tasks are also run with an LSG version of LEGAL-BERT [5].

Results. Micro and Macro F-1 (Table 2) show that LSG outperforms most of the time Longformer and Big Bird models with input sequences up to 4096 tokens long. A major difference lies in the implementation itself since the LSG variant is twice as fast to train on these lengths with no additional memory cost.⁵

	IMDb	Arxiv	Patent	Scotus	ECtHRa	ECtHRb
Epochs	3	3	3	7	5	5
Learning rate	2e-5	5e-5	2e-5	1e-4	1e-4	1e-4
RoBERTa (512-length)	95.5	87.2/86.8	66.6/61.8	69.4/60.8	62.9/58.2	72.0/65.9
Longformer	95.9	88.2/87.9	69.8/63.8	72.9/62.6	68.3/59.7	78.9/72.2
Big Bird ETC	95.4	85.9/85.5	69.4/63.9	69.4/58.2	68.3/60.3	80.0/70.6
LSG-Local (256/0)	96.0	87.5/87.1	69.9/64.8	73.3/63.7	68.8/63.7	79.9/73.4
LSG-Stride (128/4)	95.6	88.2/87.9	69.2/64.0	70.5/60.0	69.5/62.3	79.3/71.6
LSG-BlockStride (128/4)	95.7	87.7/87.4	69.6/64.1	72.5/63.1	69.1/58.6	79.5/71.8
LSG-Norm (128/4)	95.7	87.0/86.6	70.0/64.4	71.3/60.8	70.1/61.9	79.4/72.1
LSG-Pooling (128/4)	95.9	87.5/87.3	69.4/64.1	72.6/60.9	70.2/61.4	79.0/73.1
LSG-LSH (128/4)	95.8	88.2/87.9	69.5/64.2	70.3/54.6	71.0/60.3	78.9/71.0
Legal-BERT (512-length)	-	-	-	73.5/60.5	64.2/58.2	73.2/65.9
LSG-Legal-BERT (256/0)	-	-	-	74.5/62.6	71.7/63.9	81.0/75.1

Table 2. Micro/Macro F-1 on classification datasets.

⁵ See https://github.com/ccdv-ai/convert_checkpoint_to_lsg for a benchmark.

On Patent, ECtHRa and ECtHRb tasks, the ability to process longer sequences improves significantly the F-measures compared to a vanilla (full attention) RoBERTa model. We also observe that Big Bird model is in general slightly under its counterpart except for the ECtHRb dataset. This probably comes from the random attention mechanism which may require additional training steps. LSG-LSH and Big Bird models are affected by randomness during inference, thus their performance can differ between runs.

Extrapolating LEGAL-BERT with LSG to handle longer sequences improves predictions. The choice of the sparse attention is likely task specific. Using local attention only with a large block size is also a viable option. The role of global tokens is not discussed here since we only use one for all experiments. We show in the next section with summarization tasks the utility of such tokens.

4.3 Summarization Tasks

We evaluate our models on summarization tasks where the input is significantly longer than 1k tokens only. The models have been fine-tuned on each dataset.⁶

Datasets. In both *ArXiv and Pubmed* [11], the goal is to generate an abstract using a document as input. *MultiNews* [14] involves generating human-written summaries from sets of news documents. *MediaSum* [40] consists of using interview transcripts from CNN and NPR media to generate a summary.

Models	Params.
PRIMERA [36]	447M
LED [3]	460M
HAT-BART [28]	471M
Pegasus [38]	577M
Big Bird-Peg. [37]	577M
Hepos [18]	406M
LongT5-Base [15]	220M
LongT5-L	770M
LongT5-XL	3B
Ours, LSG-BART-base (256/0)	145M

Table 3. Parameters count of summarization models.

Training setup and architecture. The BART-base model [22] is converted to its LSG version by replacing the full attention in the encoder part and adding global tokens. The model is then fine-tuned on 4096-length inputs and evaluated. To reduce computational costs, experiments on 16384-length inputs are warm started from the 4096-length experiments. The model is then fine-tuned during a single epoch if necessary using the same training parameters. We propose 3

⁶ All summarization experiments are run using a 8e-5 learning rate, a 10% warmup, a length penalty of 2.0 and a beam size of 5 for beam search.

setups for the 16384-length. First we evaluate the model with pure extrapolation from 4096-length (no additional training). In the second setup, we extrapolate by adding 64 global tokens we choose to fine-tune. In the last setup, we extrapolate by adding 64 global tokens and by fine-tuning the full model. Extrapolation is done by concatenating 4 copies of the positional embedding matrix (4×4096).

The tested model - LSG-BART-base - is significantly smaller than common models from the existing literature (Table 3). An input sequence of 16384 tokens can fit on a 32Gb GPU (without attention dropout) during training without a specific memory reduction tool (i.e gradient checkpointing).

Results. LSG-BART is compared to state-of-the-art models by reporting the results from their respective papers. We use ROUGE-1, ROUGE-2 and ROUGE-L evaluation metrics as comparison points.

Models	R1	R2	RL
Pegasus (1K)	45.49	19.90	27.69
Big Bird-Peg. (4K)	46.32	20.65	42.33
HAT-BART (4K)	48.36	21.43	37.00
Hepos-LSH (7.2K)	48.12	21.06	42.72
Hepos-SKN (10.2K)	47.93	20.74	42.58
LongT5-Base (4K)	47.77	22.58	44.38
LongT5-L (16K)	49.98	24.69	46.46
LongT5-XL (16K)	50.23	24.76	46.67
Ours (4K)	47.37	21.74	43.67
Ours (16K)	48.03	22.42	44.32
+ global tuning	48.12	20.46	44.40
+ full tuning	48.32	22.52	44.57

Table 4. ROUGE on PubMed dataset.

Models	R1	R2	RL
TG-MultiSum	47.10	17.55	20.73
PRIMERA (4K)	49.90	21.10	25.9
LongT5-Base (4K)	46.01	17.37	23.50
LongT5-L (4K)	46.99	18.21	24.08
LongT5-L (8K)	47.18	18.44	24.18
LongT5-XL (8K)	48.17	19.43	24.90
Ours (4K)	47.10	18.94	25.22
Ours (16K)	47.30	19.19	25.38
+ global tuning	47.23	19.18	25.29
+ full tuning	47.07	19.04	25.35

Table 5. ROUGE on MultiNews.

Models	R1	R2	RL
Pegasus (1K)	44.70	17.27	25.80
Big Bird-Peg. (4K)	46.63	19.02	41.77
LED (16K)	46.63	19.62	41.83
PRIMERA (4K)	47.58	20.75	42.57
HAT-BART (4K)	46.68	19.07	42.17
Hepos-LSH (7.2K)	48.24	20.26	41.78
Hepos-SKN (10.2K)	47.87	20.00	41.50
LongT5-Base (4K)	44.87	18.54	40.97
LongT5-L (16K)	48.28	21.63	44.11
LongT5-XL (16K)	48.35	21.92	44.27
Ours (4K)	46.65	18.91	42.18
Ours (16K)	47.03	20.19	42.69
+ global tuning	48.08	20.42	43.65
+ full tuning	48.74	20.88	44.23

Table 6. ROUGE on ArXiv dataset.

Models	R1	R2	RL
BART-Large (1K)	35.09	18.05	31.44
T5-large (1K)	30.68	14.88	27.88
LongT5-Base (4K)	35.09	18.35	31.87
LongT5-L (4K)	35.54	19.04	32.20
LongT5-XL (4K)	36.15	19.66	32.80
Ours (4K)	35.16	18.13	32.20
Ours (16K)	35.17	18.13	32.21
+ global tuning	35.22	18.08	32.22
+ full tuning	35.31	18.35	32.47

Table 7. ROUGE on MediaSum.

As shown in Tables 4, 5, 6 and 7, LSG achieves very competitive results by enabling adapting existing pretrained models to longer sequences. On the ArXiv dataset (Table 6), LSG is competitive with every size of the LongT5 model, despite the limited number of model parameters. On the PubMed dataset

(Table 4), LSG also outperforms Pegasus and Big Bird Pegasus, and is close to Hepos models. On the MultiNews dataset (Table 5), LSG is close to the large L and XL LongT5 models. We note that while extrapolation improves metrics, additional fine-tuning has a negative impact in this case. Since this dataset is rather small (45K examples, $\sim 1.4k$ steps), fine-tuning a single epoch is not enough for the model to converge properly; longer training is required. On the MediaSum dataset (Table 7), LSG is close to the LongT5-base model again. This dataset has the shortest inputs, thus processing a maximum of 16384 tokens has a marginal impact on performances. These results underline the ability of LSG to efficiently substitute full-attention mechanisms to process long sequences.

The second surprising and important finding is the ability of LSG to improve metrics from 4096 to 16384-length inputs without additional fine-tuning. This is especially true on ArXiv and PubMed datasets which have the longest input sequences. Fine tuning additional global tokens further improves metrics while limiting cost and training time compared to a fully tuned model.

5 Conclusion

We have presented LSG attention, a novel efficient $O(n)$ alternative to the full attention mechanism relying on local, sparse and global attentions. Our results on MLM, classification and summarization tasks show that LSG is a fast and very competitive full attention substitute for pretrained Transformers to efficiently extrapolate to long input sequences. We also proposed an optimized implementation of the LSG attention mechanism on HuggingFace, improving training speed by a factor of 2 without additional memory cost compared to Longformer and Big Bird models. By providing a PyPI package conversion tool to leverage existing models and checkpoints (BERT, RoBERTa, DistilBERT, BART), the proposed approach removes the need of a costly re-training of existing models to handle long sequences.⁷

References

1. Ainslie, J., Ontanon, S., Alberti, C., Cvícek, V., Fisher, Z., Pham, P., Ravula, A., Sanghai, S., Wang, Q., Yang, L.: Etc: Encoding long and structured inputs in transformers. arXiv preprint arXiv:2004.08483 (2020)
2. Andoni, A., Indyk, P., Laarhoven, T., Razenshteyn, I.P., Schmidt, L.: Practical and optimal LSH for angular distance. CoRR **abs/1509.02897** (2015)
3. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv:2004.05150 (2020)
4. Britz, D., Guan, M.Y., Luong, M.T.: Efficient attention using a fixed-size memory representation. arXiv preprint arXiv:1707.00110 (2017)
5. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: The muppets straight out of law school. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 2898–2904 (Nov 2020)

⁷ This work has benefited from LAWBOT (ANR-20-CE38-0013) grant and HPC resources from GENCI-IDRIS (Grant 2023-AD011011309R3).

6. Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D.M., Aletras, N.: Lexglue: A benchmark dataset for legal language understanding in english. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland (2022)
7. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509 (2019)
8. Chiu, C.C., Raffel, C.: Monotonic chunkwise attention. arXiv preprint arXiv:1712.05382 (2017)
9. Choromanski, K., Likhoshesterov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., Weller, A.: Rethinking attention with performers. arXiv:2009.14794 (2021)
10. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does bert look at? an analysis of bert's attention. arXiv preprint arXiv:1906.04341 (2019)
11. Cohan, A., Dernoncourt, F., Kim, D.S., Bui, T., Kim, S., Chang, W., Goharian, N.: A discourse-aware attention model for abstractive summarization of long documents. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (2018)
12. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
14. Fabbri, A.R., Li, I., She, T., Li, S., Radev, D.R.: Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model (2019)
15. Guo, M., Ainslie, J., Uthus, D.C., Ontañón, S., Ni, J., Sung, Y., Yang, Y.: Longt5: Efficient text-to-text transformer for long sequences. CoRR **abs/2112.07916** (2021)
16. Guo, Q., Qiu, X., Liu, P., Shao, Y., Xue, X., Zhang, Z.: Star-transformer. arXiv preprint arXiv:1902.09113 (2019)
17. He, J., Wang, L., Liu, L., Feng, J., Wu, H.: Long document classification from local word glimpses via recurrent attention learning. IEEE Access **7**, 40707–40718 (2019)
18. Huang, L., Cao, S., Parulian, N., Ji, H., Wang, L.: Efficient attentions for long document summarization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online (Jun 2021)
19. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. CoRR **abs/2006.16236** (2020)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014)
21. Kitaev, N., Kaiser, L., Levskaya, A.: Reformer: The efficient transformer. CoRR **abs/2001.04451** (2020)
22. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880. Association for Computational Linguistics, Online (Jul 2020)
23. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019)

12 C. Condevaux and S. Harispe

24. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 142–150. Association for Computational Linguistics (Jun 2011)
25. Rae, J.W., Potapenko, A., Jayakumar, S.M., Lillicrap, T.P.: Compressive transformers for long-range sequence modelling. arXiv preprint arXiv:1911.05507 (2019)
26. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**(140), 1–67 (2020)
27. Raganato, A., Scherrer, Y., Tiedemann, J.: Fixed encoder self-attention patterns in transformer-based machine translation. arXiv preprint arXiv:2002.10260 (2020)
28. Rohde, T., Wu, X., Liu, Y.: Hierarchical learning for generation with long source sequences. *CoRR* **abs/2104.07545** (2021)
29. Sharma, E., Li, C., Wang, L.: BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2204–2213. Association for Computational Linguistics, Florence, Italy (Jul 2019)
30. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 464–468. New Orleans, Louisiana (Jun 2018)
31. Shen, T., Zhou, T., Long, G., Jiang, J., Zhang, C.: Bi-directional block self-attention for fast and memory-efficient sequence modeling. arXiv preprint arXiv:1804.00857 (2018)
32. Shen, Z., Zhang, M., Yi, S., Yan, J., Zhao, H.: Factorized attention: Self-attention with linear complexities. *CoRR* **abs/1812.01243** (2018)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
34. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. *CoRR* **abs/2006.04768** (2020)
35. Wu, F., Fan, A., Baeviski, A., Dauphin, Y.N., Auli, M.: Pay less attention with lightweight and dynamic convolutions. arXiv preprint arXiv:1901.10430 (2019)
36. Xiao, W., Beltagy, I., Carenini, G., Cohan, A.: PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5245–5263. Dublin, Ireland (May 2022)
37. Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al.: Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems* **33** (2020)
38. Zhang, J., Zhao, Y., Saleh, M., Liu, P.J.: Pegasus: Pre-training with extracted gap-sentences for abstractive summarization (2019)
39. Zhang, X., Wei, F., Zhou, M.: Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. arXiv preprint arXiv:1905.06566 (2019)
40. Zhu, C., Liu, Y., Mei, J., Zeng, M.: Mediasum: A large-scale media interview dataset for dialogue summarization. arXiv preprint arXiv:2103.06410 (2021)

B.1 Décision de justice

Exemple de décision de justice française [57] :

RÉPUBLIQUE FRANCAISE AU NOM DU PEUPLE FRANCAIS

6ème Chambre A

ORDONNANCE No 283

R. G : 17/ 04448

Mme Céline X...

C/

M. Jean-Marie Michel Lucien Y...

Déclare l'acte de saisine caduc

Copie exécutoire délivrée le :

à :

RÉPUBLIQUE FRANÇAISE AU NOM DU PEUPLE FRANÇAIS

COUR D'APPEL DE RENNES ORDONNANCE DE MISE EN ETAT DU 04 DECEMBRE 2017

Le quatre Décembre deux mille dix sept, par mise à disposition au Greffe,

Monsieur Yves LE NOAN, Magistrat de la mise en état de la 6ème Chambre A, assisté de Xavier LE COLLEN, faisant fonction de Greffier,

Statuant dans la procédure opposant :

Madame Céline X... née le 29 Avril 1974 à SAINT RENAN (29290) ... Représentée par Me Françoise NAUDY-ORTAIS de la SCP KERDILES-KAYA & NAUDY-ORTAIS, Plaidant/ Postulant, avocat au barreau de BREST (bénéficie d'une aide juridictionnelle Totale numéro 2017/004995 du 09/ 06/ 2017 accordée par le bureau d'aide juridictionnelle de RENNES)

APPELANTE

à

Monsieur Jean-Marie Michel Lucien Y... né le 22 Avril 1964 à PALAISEAU (91120) ...
Représenté par Me Luc BOURGES de la SELARL LUC BOURGES, Plaidant/ Postulant,
avocat au barreau de RENNES

INTIME

A rendu l'ordonnance suivante :

Vu la demande d'observations sur la caducité de la déclaration d'appel adressée aux parties le 2 novembre 2017 ;

Vu l'absence d'observation des parties ;

Vu les dispositions des articles 908, 911 et 911-1 du code de procédure civile, dans leur rédaction, applicable au présent incident, antérieure au décret no 2017-891 du 6 mai 2017 ;

Selon l'article 908 du code de procédure civile, à peine de caducité de la déclaration d'appel, relevée d'office, l'appelant dispose d'un délai de trois mois à compter de la déclaration d'appel pour conclure ;

Selon l'article 911 du code de procédure civile, sous les sanctions prévues aux articles 908 à 910, les conclusions sont notifiées aux avocats des parties dans le délai de leur remise au greffe de la cour. Sous les mêmes sanctions, elles sont signifiées dans le mois suivant l'expiration de ce délai aux parties qui n'ont pas constitué avocat ; cependant, si, entre-temps, celles-ci ont constitué avocat avant la signification des conclusions, il est procédé par voie de notification à leur avocat ;

En l'espèce, la déclaration d'appel de madame Céline X...a été effectuée le 20 juin 2017. L'appelante a déposé ses conclusions au greffe le 7 septembre 2017, soit dans le délai prévu à l'article 908. A cette date, monsieur Jean-Marie Y..., intimé, n'avait pas constitué avocat, ce qu'il n'a fait que le 19 septembre 2017. Il appartenait en conséquence à l'appelante, en application de l'article 911, de notifier ses conclusions à l'avocat de l'intimé dans le mois suivant l'expiration du délai prévu à l'article 908, soit au plus tard le 20 octobre 2017, à défaut d'avoir signifié ces conclusions à l'intimé avant sa constitution d'avocat ;

Si l'appelante a signifié sa déclaration d'appel à l'intimé le 28 septembre 2017, conformément à l'article 902, elle ne lui a en revanche pas signifié ses conclusions, ni notifié celles-ci à son conseil avant le 20 octobre 2017 ; PAR CES MOTIFS

Prononce la caducité de la déclaration d'appel,

Condamne l'appelante aux dépens.

Le Greffier, Le Conseiller de la mise en état,

TBox	ABox
Woman \equiv Person \sqcap Female	Woman(Lucie)
Man \equiv Person $\sqcap \neg$ Female	Man(Marc)
Mother \equiv Woman $\sqcap \exists$ hasChild.Person	Mother(Lea)
Father \equiv Man $\sqcap \exists$ hasChild.Person	hasChild(Marc, Tom)
Parent \equiv Mother \sqcup Father	hasChild(Lucie, Tom)
GrandMother \equiv Mother $\sqcap \exists$ hasChild.Parent	Person(Tom)
...	...

TABLE B.1 Exemple d'ontologie (TBox, ABox) adapté de [111].

B.2 Représentations des connaissances et représentations opérationnelles

Nous utiliserons dans nos travaux des représentations formelles des connaissances de type ontologies. Celles-ci reposent pour la plupart sur des formalismes qui correspondent à des logiques descriptives. Nous utiliserons cependant des termes variés pour faire référence à ces représentations en fonction des contextes d'étude, e.g. ontologies, bases de connaissances, graphes de connaissances, graphes sémantiques, taxonomies. Cette annexe fournit quelques précisions sur l'usage que nous ferons de ces termes.

Nous utiliserons les termes « ontologie » et « base de connaissances » de manière interchangeable. Nous ferons alors référence dans ce cas aux notions de TBox et de ABox classiquement considérées en Représentation des Connaissances [111]. La Tbox définit le niveau terminologique de la représentation des connaissances ; elle précise en particulier la définition des concepts et des rôles (propriétés ou relations) qui seront utilisés. Nous ne précisons pas ici formellement la notion d'interprétation qui permet de formaliser la sémantique d'une TBox. La Abox définit le niveau factuel sous la forme d'assertions sur les individus (instances) telles que des appartenances à des classes, ou des assertions de rôle. Une exemple de TBox et de Abox d'une ontologie est fourni dans la table B.1. En fonction de l'expressivité de la logique engagée, des expressions de connaissances plus ou moins riches pourront être retrouvées dans ces deux composantes. Des procédures de raisonnement pourront être appliquées sur nos bases de connaissances, notamment pour inférer les connaissances implicites qu'elles encodent, e.g. taxonomie de classes ou des connaissances sur les individus. A partir de la table B.1 nous pourrions par exemple inférer que Marc est un père car c'est un homme qui a un enfant (qui est une personne). Ces bases de connaissances pourront être exprimées à l'aide de différents langages, en particulier OWL [206].

Nous utiliserons généralement les termes « graphe de connaissances » ou « graphe sémantique » pour faire référence à une représentation, via le modèle de graphe RDF [227], des connaissances relatives à une base de connaissances (avant ou après une procédure d'inférence amenée par un raisonneur), e.g. taxonomie des classes et les faits sur les individus. Nous retrouverons dans ce cas des relations de différentes natures dans le graphe (en fonction des rôles définis, exemple page 64).

Les représentations des connaissances auxquelles nous ferons référence dans nos travaux seront très souvent des ontologies dites légères, e.g. *lightweight ontology*. Bien que celles-ci soient fréquemment utilisées dans la littérature, elles ne sont très généralement pas formellement définies par les auteurs qui les considèrent (dans notre littérature) et peuvent varier en fonction des travaux, e.g. sur les

mesures sémantiques. Elles correspondront très généralement à un ensemble de connaissances simples relatives à une taxonomie : ses classes et leur structuration. Elles incluront dans certains cas les instances des classes et les relation de typage afférentes. Ces connaissances seront la plupart du temps manipulées sous la forme de graphes orientés acycliques (ou DAG pour *Directed Acyclic Graph*) réduits transitivement. Elles pourront aussi dans certains cas être manipulées à l'aide d'autres cadres théoriques que celui proposé par la théorie des graphes. Nous pourrons ainsi mobiliser des représentations de type POSET (*Partially Ordered Set*) de la théorie des ensembles que nous manipulerons parfois à l'aide des fonctions de croyance (théorie de Dempster-Shafer). Plusieurs cadres d'analyse seront par ailleurs quelquefois considérés conjointement dans nos travaux.

Il est important de souligner que la représentation formelle des connaissances – ici au sens donné par le domaine de la Représentation des Connaissances, e.g. exprimée à l'aide d'une logique descriptive – sera souvent restreinte par l'expressivité des représentations permises par les cadres d'étude que nous considérerons. A titre d'exemple, la relation de sous-classe d'une taxonomie, telle qu'exprimée en logique descriptive entre deux concepts atomiques sera exprimée dans un DAG par un arc entre deux nœuds du graphe, e.g. $Person \sqsubseteq Animal$ sera exprimée par une relation orientée entre les nœuds *Person* et *Animal* associés respectivement aux concepts *Person* et *Animal*. Notez que nous dissociions le concept *Person* défini dans la représentation des connaissances, de sa représentation dans la structure que nous manipulerons, e.g. nœud d'un graphe. Cette précision est à notre connaissance éludée dans la plupart des travaux liés par exemple aux mesures sémantiques. Elle est cependant selon nous importante car seule une partie de la sémantique exprimée dans certaines ontologies ne pourra parfois trouver correspondance dans les représentations que nous manipulerons. Ce type de DAG ne pourra en aucune façon exprimer la richesse de la sémantique formelle définie par des langages de représentation des connaissances tels que OWL, e.g. un exemple simple étant la relation `disjointWith` qui peut être exprimée entre deux classes en OWL. Le DAG sera alors dans ce cas vu comme une *représentation opérationnelle* de notre représentation des connaissances initiale. Ainsi, les représentations que nous manipulerons, e.g. sous forme de graphes, seront parfois le produit d'un traitement réalisé à partir d'une représentation des connaissances plus riche – entendez par là la construction d'une vue de la représentation des connaissances, ou une simple conversion voire transpilation partielle de connaissances.

Une approche de construction simple et classiquement considérée pour l'obtention des DAG représentant une taxonomie à partir d'une ontologie OWL exprimée en RDF(S) est la suivante :

1. Utilisation d'un raisonneur sur l'ontologie.
2. Définition d'un DAG à partir des triplets RDF qui engagent le prédicat `rdfs:subClassOf`.
3. Application d'un traitement de type réduction transitive sur le DAG.

La construction de ce type de représentations opérationnelles des bases de connaissances est détaillée dans le manuscrit de thèse de l'auteur [112] - cf. annexe A.2 pour plus de détails. Des discussions sur ce sujet sont aussi proposées dans la documentation de la SML [107].

Les plus simples représentations des connaissances que nous considérerons seront donc de véritables taxonomies manipulées sous la forme de DAG (au sens explicité ci-dessus et avec les limites de représentation que nous venons de mentionner). Des instances de classes pourront aussi parfois enrichir le DAG en considérant qu'un arc entre les nœuds associés à une instance et à sa classe existe pour chaque triplet (`?instance`, `rdf:type`, `?Classe`). Par abus de langage, nous considérerons aussi souvent des entités annotées par des concepts, e.g. gènes annotés par des concepts, documents annotés conceptuellement. Une entité sera alors une instance et ses annotations seront les classes auxquelles elle appartient.

Bibliographie

- [1] *A proposal for the Dartmouth summer research project on Artificial Intelligence*. <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>. Accessed : 2023-09-22.
- [2] Bellandi A., Furletti B., Grossi V. et Romei A. « Ontology-Driven Association Rule Extraction : A Case Study ». Dans : *International Workshop on Contexts and Ontologies : Representation and Reasoning*, pp. 10–19, Roskilde University, Denmark, 21 August 2007. CEUR-WS.org, Aachen, DEU, 2007.
- [3] Alan AKBIK, Tanja BERGMANN, Duncan BLYTHE, Kashif RASUL, Stefan SCHWETER et Roland VOLLGRAF. « FLAIR : An easy-to-use framework for state-of-the-art NLP ». Dans : *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*. 2019.
- [4] Md Shafiqul ALAM et Ying XIE. « Appley : Approximate shapley value for model explainability in linear time ». Dans : *2022 IEEE International Conference on Big Data (Big Data)*. IEEE. 2022.
- [5] Yaddaden ALI. « Deep reinforcement learning for the Vehicle Routing Problem ». Theses. Université Montpellier, nov. 2017.
- [6] Ebtesam ALMAZROUEI et al. « Falcon-40B : an open large language model with state-of-the-art performance ». Dans : <https://huggingface.co/tiiuae/falcon-40b> (2023).
- [7] Marco ANCONA, Cengiz OZTIRELI et Markus GROSS. « Explaining deep neural networks with a polynomial time algorithm for shapley value approximation ». Dans : *International Conference on Machine Learning*. PMLR. 2019.
- [8] Rohan ANIL et al. *PaLM 2 Technical Report*. 2023. arXiv : 2305.10403 [cs.CL].
- [9] Sören AUER, Christian BIZER, Georgi KOBILAROV, Jens LEHMANN, Richard CYGANIAK et Zachary IVES. « Dbpedia : A nucleus for a web of open data ». Dans : *international semantic web conference*. Springer. 2007.
- [10] Dzmitry BAHDANAU, Kyunghyun CHO et Yoshua BENGIO. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv : 1409.0473 [cs.CL].
- [11] Yuntao BAI et al. *Constitutional AI : Harmlessness from AI Feedback*. 2022. arXiv : 2212.08073 [cs.CL].
- [12] Collin F BAKER, Charles J FILLMORE et John B LOWE. « The Berkeley Framenet project ». Dans : *COLING 1998 Volume 1 : The 17th International Conference on Computational Linguistics*. 1998.
- [13] Anton BAKHTIN, Noam BROWN, Emily DINAN, Gabriele FARINA, Colin FLAHERTY, Daniel FRIED, Andrew GOFF, Jonathan GRAY, Hengyuan HU et al. « Human-level play in the game of Diplomacy by combining language models with strategic reasoning ». Dans : *Science* 378.6624 (2022).
- [14] Montserrat BATET, Sébastien HARISPE, Sylvie RANWEZ, David SANCHEZ et Vincent RANWEZ. « An information theoretic approach to improve semantic similarity assessments across multiple ontologies ». Dans : *Information Sciences* 283 (2014). DOI : 10.1016/j.ins.2014.06.039.

- [15] John E BEASLEY. « Route first—cluster second methods for vehicle routing ». Dans : *Omega* 11.4 (1983).
- [16] Irwan BELLO, Hieu PHAM, Quoc V. LE, Mohammad NOROUZI et Samy BENGIO. « Neural Combinatorial Optimization with Reinforcement Learning ». Dans : *CoRR* abs/1611.09940 (2016). arXiv : 1611.09940.
- [17] Iz BELTAGY, Matthew E. PETERS et Arman COHAN. *Longformer : The Long-Document Transformer*. 2020. arXiv : 2004.05150 [cs.CL].
- [18] Yoshua BENGIO, Aaron COURVILLE et Pascal VINCENT. *Representation Learning : A Review and New Perspectives*. 2014. arXiv : 1206.5538 [cs.LG].
- [19] Yoshua BENGIO, Réjean DUCHARME et Pascal VINCENT. « A Neural Probabilistic Language Model ». Dans : *Advances in Neural Information Processing Systems*. T. 13. MIT Press, 2000.
- [20] Yoshua BENGIO, Andrea LODI et Antoine PROUVOST. « Machine Learning for Combinatorial Optimization : a Methodological Tour d’Horizon ». Dans : *CoRR* abs/1811.06128 (2018). arXiv : 1811.06128.
- [21] Valentina BERETTA. « Data veracity assessment : enhancing Truth Discovery using a priori knowledge ». Theses. IMT Mines Alès, oct. 2018.
- [22] Valentina BERETTA, Katarina BOLAND, Luke LO SEEN, Sébastien HARISPE, Konstantin TODOROV et Andon TCHECHMEDJIEV. « Can Knowledge Graph Embeddings Tell Us What Fact-checked Claims Are About ? » Dans : *Workshop on Insights from Negative Results in NLP*. Online, Dominican Republic : Association for Computational Linguistics, nov. 2020. DOI : 10.18653/v1/2020.insights-1.11.
- [23] Valentina BERETTA, Sébastien HARISPE, Sylvie RANWEZ et Isabelle MOUGENOT. « Combining Truth Discovery and RDF Knowledge Bases to Their Mutual Advantage ». Dans : *The Semantic Web – ISWC 2018*. T. 11136. Monterey, California, United States, oct. 2018. DOI : 10.1007/978-3-030-00671-6_38.
- [24] Valentina BERETTA, Sébastien HARISPE, Sylvie RANWEZ et Isabelle MOUGENOT. « Truth selection for truth discovery models exploiting ordering relationship among values ». Dans : *Knowledge-Based Systems* 159 (nov. 2018). DOI : 10.1016/j.knsys.2018.06.023.
- [25] Valentina BERETTA, Sébastien HARISPE, Sylvie RANWEZ et Isabelle MOUGENOT. « Utilisation d’ontologies pour la quête de vérité : une étude expérimentale ». Dans : *IC 2016, Journées francophones d’Ingénierie des Connaissances*. Montpellier, France, juin 2016.
- [26] Valentina BERETTA, Sylvie RANWEZ, Sébastien HARISPE et Isabelle MOUGENOT. « Améliorer la recherche de vérité en exploitant la connaissance de domaines via les ontologies et les règles d’association ». Dans : *Revue des Sciences et Technologies de l’Information - Série RIA : Revue d’Intelligence Artificielle* 32.3 (2018). DOI : 10.3166/ria.32.373-405.
- [27] Valentina BERETTA, Sylvie RANWEZ, Sébastien HARISPE et Isabelle MOUGENOT. « Contribution à la recherche de vérité : modèles exploitant des règles d’association extraites d’une base de connaissances ». Dans : *IC 2017, Journées francophones d’Ingénierie des Connaissances*. Caen, France, juill. 2017.
- [28] Laure BERTI-EQUILLE et Javier BORGE-HOLTHOEFER. *Veracity of Data*. Springer Nature, 2022.
- [29] Laure BERTI-EQUILLE et Javier BORGE-HOLTHOEFER. *Veracity of Data : From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics*. Morgan & Claypool Publishers, 2015.
- [30] Jeffrey R BINDER, Rutvik H DESAI, William W GRAVES et Lisa L CONANT. « Where is the semantic system ? A critical review and meta-analysis of 120 functional neuroimaging studies ». Dans : *Cerebral cortex* 19.12 (2009).
- [31] William BLACOE et Mirella LAPATA. « A Comparison of Vector-based Representations for Semantic Composition ». Dans : *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea, juill. 2012.
- [32] Emmanuel BLANCHARD. « Exploitation d’une hiérarchie de subsomption par le biais de mesures sémantiques ». Theses. Université de Nantes, mai 2008.

- [33] Emmanuel BLANCHARD, Mounira HARZALLAH et Pascale KUNTZ. « A generic framework for comparing semantic similarities on a subsumption hierarchy. » Dans : *ECAI 2018 - European Conference on Artificial Intelligence*. T. 2008. 2008.
- [34] David M BLEI, Andrew Y NG et Michael I JORDAN. « Latent dirichlet allocation ». Dans : *Journal of machine Learning research* 3.Jan (2003).
- [35] Piotr BOJANOWSKI, Edouard GRAVE, Armand JOULIN et Tomas MIKOLOV. « Enriching Word Vectors with Subword Information ». Dans : *arXiv preprint arXiv :1607.04606* (2016).
- [36] Danushka BOLLEGALA et James O' NEILL. « A Survey on Word Meta-Embedding Learning ». Dans : *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. Survey Track. International Joint Conferences on Artificial Intelligence Organization, juill. 2022. DOI : 10.24963/ijcai.2022/758.
- [37] Tolga BOLUKBASI, Kai-Wei CHANG, James Y ZOU, Venkatesh SALIGRAMA et Adam T KALAI. « Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings ». Dans : *Advances in Neural Information Processing Systems*. T. 29. Curran Associates, Inc., 2016.
- [38] Antoine BORDES, Nicolas USUNIER, Alberto GARCIA-DURAN, Jason WESTON et Oksana YAKHNENKO. « Translating Embeddings for Modeling Multi-relational Data ». Dans : *Advances in Neural Information Processing Systems*. T. 26. Curran Associates, Inc., 2013.
- [39] Nick BOSTROM. *Superintelligence : Paths, Dangers, Strategies*. 1st. USA : Oxford University Press, Inc., 2014.
- [40] Greg BROCKMAN, Vicki CHEUNG, Ludwig PETTERSSON, Jonas SCHNEIDER, John SCHULMAN, Jie TANG et Wojciech ZAREMBA. *OpenAI Gym*. 2016. arXiv : 1606.01540 [cs.LG].
- [41] Jane BROMLEY, Isabelle GUYON, Yann LECUN, Eduard SÄCKINGER et Roopak SHAH. « Signature Verification Using a "Siamese" Time Delay Neural Network ». Dans : *Proceedings of the 6th International Conference on Neural Information Processing Systems*. NIPS'93. Denver, Colorado : Morgan Kaufmann Publishers Inc., 1993.
- [42] Tom B. BROWN et al. *Language Models are Few-Shot Learners*. 2020. arXiv : 2005.14165 [cs.CL].
- [43] Bryan CAI, Sihang ZENG, YUCONG LIN, Zheng YUAN, Doudou ZHOU et Lu TIAN. *Hierarchical Pretraining for Biomedical Term Embeddings*. 2023. arXiv : 2307.00266 [cs.CL].
- [44] Jeremy J. CARROLL, Ian DICKINSON, Chris DOLLIN, Dave REYNOLDS, Andy SEABORNE et Kevin WILKINSON. *Jena : Implementing the Semantic Web Recommendations*. Rapp. tech. HPL-2003-146. Hewlett Packard Laboratories, 2003.
- [45] Valentina CEASU et Sylvie DESPRÈS. « A Semantic Case-Based Reasoning Framework for Text Categorization ». Dans : *The Semantic Web*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2007.
- [46] Dhivya CHANDRASEKARAN et Vijay MAGO. « Evolution of semantic similarity—a survey ». Dans : *ACM Computing Surveys (CSUR)* 54.2 (2021).
- [47] *ChatGPT sets record for fastest-growing user base - analyst note (Reuters)*. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>. Accessed : 2023-09-22.
- [48] Sneha CHAUDHARI, Varun MITHAL, Gungor POLATKAN et Rohan RAMANATH. *An Attentive Survey of Attention Models*. 2021. arXiv : 1904.02874 [cs.LG].
- [49] D. CHAVALARIAS. *Toxic Data*. Flammarion, 2022.
- [50] Hugh CHEN, Ian C COVERT, Scott M LUNDBERG et Su-In LEE. « Algorithms to estimate Shapley value feature attributions ». Dans : *Nature Machine Intelligence* (2023).
- [51] Jiaoyan CHEN, Pan HU, Ernesto JIMENEZ-RUIZ, Ole Magnus HOLTER, Denver ANTONYRAJAH et Ian HORROCKS. « OWL2Vec* : Embedding of OWL ontologies ». Dans : *Machine Learning* 110.7 (2021).
- [52] Shouyuan CHEN, Sherman WONG, Liangjian CHEN et Yuandong TIAN. *Extending Context Window of Large Language Models via Positional Interpolation*. 2023. arXiv : 2306.15595 [cs.CL].
- [53] Xiushi CHEN, Jyun-Yu JIANG, Kun JIN, Yichao ZHOU, Mingyan LIU, P Jeffrey BRANTINGHAM et Wei WANG. « ReLiable : Offline Reinforcement Learning for Tactical Strategies in Professional Basketball Games ». Dans : *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2022.

- [54] Jaekeol CHOI, Euna JUNG, Sungjun LIM et Wonjong RHEE. *Finding Inverse Document Frequency Information in BERT*. 2022. arXiv : 2202.12191 [cs.IR].
- [55] Kevin CLARK, Urvashi KHANDELWAL, Omer LEVY et Christopher D. MANNING. « What Does BERT Look at? An Analysis of BERT's Attention ». Dans : *Proceedings of the 2019 ACL Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy : Association for Computational Linguistics, août 2019. DOI : 10.18653/v1/W19-4828.
- [56] Ronan COLLOBERT, Jason WESTON, Leon BOTTOU, Michael KARLEN, Koray KAVUKCUOGLU et Pavel KUKSA. *Natural Language Processing (almost) from Scratch*. 2011. arXiv : 1103.0398 [cs.LG].
- [57] Charles CONDEVAUX. « Méthodes d'apprentissage automatique pour l'analyse de corpus jurisprudentiels. (Machine learning methods for the analysis of jurisprudential corpus) ». Thèse de doct. University of Nîmes, France, 2021.
- [58] Charles CONDEVAUX et Sébastien HARISPE. « LSG Attention : Extrapolation of pretrained Transformers to long sequences ». Dans : *PAKDD 2023 - The 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Osaka, Japan, mai 2023. DOI : 10.1007/978-3-031-33374-3_35.
- [59] Charles CONDEVAUX, Sébastien HARISPE et Stéphane MUSSARD. « Fair and Efficient Alternatives to Shapley-based Attribution Methods ». Dans : *ECMLPKDD 2022 - The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Grenoble, France, sept. 2022. DOI : 10.1007/978-3-031-26387-3_19.
- [60] Charles CONDEVAUX, Sébastien HARISPE, Stéphane MUSSARD et Guillaume ZAMBRANO. « Classification One-Shot faiblement supervisée par réseaux de neurones récurrents avec attention : application à la détection de résultat juridique. » Dans : *EGC 2020, Extraction et Gestion des Connaissances*. Bruxelles, Belgium, jan. 2020.
- [61] Charles CONDEVAUX, Sébastien HARISPE, Stéphane MUSSARD et Guillaume ZAMBRANO. « Weakly Supervised One-shot Classification using Recurrent Neural Networks with Attention : Application to Claim Acceptance Detection ». Dans : *JURIX 2019 32nd International Conference on Legal Knowledge and Information Systems*. Madrid, Spain, déc. 2019. DOI : 10.3233/FAIA190303.
- [62] *Conference Acceptance Rate*. <https://github.com/lixin4ever/Conference-Acceptance-Rate>. Accessed : 2023-09-22.
- [63] Gene Ontology CONSORTIUM. « The Gene Ontology (GO) database and informatics resource ». Dans : *Nucleic acids research* 32.suppl_1 (2004).
- [64] The UniProt CONSORTIUM. « UniProt : the universal protein knowledgebase ». Dans : *Nucleic Acids Research* 45.D1 (nov. 2016). DOI : 10.1093/nar/gkw1099. eprint : <https://academic.oup.com/nar/article-pdf/45/D1/D158/23819877/gkw1099.pdf>.
- [65] William J COOK, David L APPLGATE, Robert E BIXBY et Vasek CHVATAL. *The traveling salesman problem : a computational study*. Princeton university press, 2011.
- [66] Romain COUILLET, Yagmur Gizem CINAR, Eric GAUSSIER et Muhammad IMRAN. « Word Representations Concentrate and This is Good News ! » Dans : *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online : Association for Computational Linguistics, nov. 2020. DOI : 10.18653/v1/2020.conll-1.25.
- [67] Nick CRASWELL, Bhaskar MITRA, Emine YILMAZ, Daniel CAMPOS, Jimmy LIN, Ellen M. VOORHEES et Ian SOBOROFF. « Overview of the TREC 2022 deep learning track ». Dans : *Text REtrieval Conference (TREC)*. NIST. TREC, mars 2023.
- [68] Tri DAO, Daniel Y. Fu, Stefano ERMON, Atri RUDRA et Christopher Ré. *FlashAttention : Fast and Memory-Efficient Exact Attention with IO-Awareness*. 2022. arXiv : 2205.14135 [cs.LG].
- [69] Cedric DE BOOM, Steven VAN CANNEYT, Steven BOHEZ, Thomas DEMEESTER et Bart DHOEDT. « Learning Semantic Similarity for Very Short Texts ». Dans : *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. 2015. DOI : 10.1109/ICDMW.2015.86.
- [70] *Décret n° 2007-468 du 28 mars 2007 portant statut particulier des enseignants de l'Institut Mines-Télécom*. <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000647115>. Accessed : 2023-07-11.

- [71] Jia DENG, Wei DONG, Richard SOCHER, Li-Jia LI, Kai LI et Li FEI-FEI. « Imagenet : A large-scale hierarchical image database ». Dans : *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.
- [72] Michel DEUDON. « Learning semantic similarity in a continuous space ». Dans : *Advances in Neural Information Processing Systems*. T. 31. Curran Associates, Inc., 2018.
- [73] Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE et Kristina TOUTANOVA. *BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018.
- [74] Michel Marie DEZA et Elena DEZA. *Encyclopedia of Distances*. Springer Berlin Heidelberg, 2009. DOI : 10.1007/978-3-642-00234-2_1.
- [75] Ning DING, Yankai LIN, Zhiyuan LIU et Maosong SUN. « Representation Learning for Compositional Semantics ». Dans : *Representation Learning for Natural Language Processing*. Springer Nature Singapore Singapore, 2023.
- [76] Xin DONG, Evgeniy GABRILOVICH, Jeremy HEITZ, Wilko HORN, Ni LAO, Kevin MURPHY, Thomas STROHMANN, Shaohua SUN et Wei ZHANG. « Knowledge vault : a web-scale approach to probabilistic knowledge fusion ». Dans : *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. ACM, 2014. DOI : 10.1145/2623330.2623623.
- [77] David DUBIN. « The most influential paper Gerard Salton never wrote ». Dans : *Graduate School of Library and Information Science. University of Illinois*. (2004).
- [78] Phan Minh DUNG. « On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games ». Dans : *Artificial intelligence 77.2* (1995).
- [79] Cheikh Brahim EL VAIGH, François GOASDOUÉ, Guillaume GRAVIER et Pascale SÉBILLOT. « A Novel Path-Based Entity Relatedness Measure for Efficient Collective Entity Linking ». Dans : *The Semantic Web – ISWC 2020*. Cham : Springer International Publishing, 2020.
- [80] Ronen ELKAN et Yuanzhi LI. *TinyStories : How Small Can Language Models Be and Still Speak Coherent English?* 2023. arXiv : 2305.07759 [cs.CL].
- [81] Orri ERLING et Ivan MIKHAILOV. « Virtuoso : RDF Support in a Native RDBMS ». Dans : *Semantic Web Information Management*. 2009.
- [82] *Évaluation du HCERES sur l'unité : Institut de Recherche en Informatique de Toulouse IRIT, campagne d'évaluation 2014-2015*. https://www.hceres.fr/sites/default/files/media/publications/rapports_evaluations/pdf/A2016-EV-0311384L-S2PUR160009711-010079-RF.pdf. Accessed : 2023-09-22.
- [83] *Évaluation du HCERES sur l'unité : Institut de Recherche en Informatique de Toulouse IRIT, campagne d'évaluation 2019-2020*. <https://www.hceres.fr/sites/default/files/media/downloads/a2021-ev-0311384l-der-pur210019163-031247-rf.pdf>. Accessed : 2023-09-22.
- [84] *Exemple de roue des odeurs (aromaster.com)*. <https://aromaster.com/fr/produit/roue-des-aromes-du-vin/>. Accessed : 2023-10-15.
- [85] Anthony FADER, Stephen SODERLAND et Oren ETZIONI. « Identifying Relations for Open Information Extraction ». Dans : *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*. Edinburgh, Scotland, UK, juill. 2011.
- [86] Ronald FAGIN et Joseph Y. HALPERN. « A New Approach to Updating Beliefs ». Dans : *CoRR abs/1304.1119* (2013). arXiv : 1304.1119.
- [87] Richárd FARKAS, Veronika VINCZE, György MÓRA, János CSIRIK et György SZARVAS. « The CoNLL-2010 Shared Task : Learning to Detect Hedges and their Scope in Natural Language Text ». Dans : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*. Uppsala, Sweden : Association for Computational Linguistics, juill. 2010.
- [88] Manaal FARUQUI, Jesse DODGE, Sujay Kumar JAUHAR, Chris DYER, Eduard HOVY et Noah A. SMITH. « Retrofitting Word Vectors to Semantic Lexicons ». Dans : *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Denver, Colorado : Association for Computational Linguistics, juin 2015. DOI : 10.3115/v1/N15-1184.

- [89] Alhussein FAWZI, Matej BALOG, Aja HUANG, Thomas HUBERT, Bernardino ROMERA-PAREDES, Mohammadamin BAREKATAIN, Alexander NOVIKOV, Francisco J R RUIZ, Julian SCHRITTWIESER, Grzegorz SWIRSZCZ et al. « Discovering faster matrix multiplication algorithms with reinforcement learning ». Dans : *Nature* 610.7930 (2022).
- [90] Qinrong FENG, Duoqian MIAO et Yi CHENG. « Hierarchical decision rules mining ». Dans : *Expert Systems with Applications* 37.3 (2010). DOI : <https://doi.org/10.1016/j.eswa.2009.06.065>.
- [91] José Rui FIGUEIRA, Vincent MOUSSEAU et Bernard ROY. « ELECTRE methods ». Dans : *Multiple criteria decision analysis : State of the art surveys* (2016).
- [92] Anthony FINCH, Alexander CROWELL, Mamta BHATIA, Pooja PARAMESHWARAPPA, Yung-Chieh CHANG, Jose MARTINEZ et Michael HORBERG. « Exploiting hierarchy in medical concept embedding* ». Dans : *JAMIA Open* 4.1 (mars 2021). DOI : [10.1093/jamiaopen/oaab022](https://doi.org/10.1093/jamiaopen/oaab022).
- [93] Nicolas FIORINI, Sébastien HARISPE, Sylvie RANWEZ, Jacky MONTMAIN et Vincent RANWEZ. « Annotation sémantique de clusters ». Dans : *ROADEF 2015, Congrès annuel de la société Française de Recherche Opérationnelle et d'Aide à la Décision*. Marseille, France, fév. 2015.
- [94] Nicolas FIORINI, Sébastien HARISPE, Sylvie RANWEZ, Jacky MONTMAIN et Vincent RANWEZ. « Fast and reliable inference of semantic clusters ». Dans : *Knowledge-Based Systems* 111 (2016). DOI : [10.1016/j.knsys.2016.08.008](https://doi.org/10.1016/j.knsys.2016.08.008).
- [95] Nicolas FIORINI, Sylvie RANWEZ, Sébastien HARISPE, Jacky MONTMAIN et Vincent RANWEZ. « USI at BioASQ 2015 : a semantic similarity-based approach for semantic indexing ». Dans : *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF), Toulouse, France*. 2015.
- [96] LOIC FOSSE, Duc-Hau NGUYEN, Pascale SÉBILLOT et Guillaume GRAVIER. « Une étude statistique des plongements dans les modèles transformers pour le français (An empirical statistical study of embeddings in French transformers) ». French. Dans : *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*. Avignon, France : ATALA, juin 2022.
- [97] *Frederick Jelinek Wikipedia*. https://en.wikipedia.org/wiki/Frederick_Jelinek. Accessed : 2023-09-20.
- [98] Drew FUDENBERG et Jean TIROLE. *Game theory*. MIT press, 1991.
- [99] Luis GALÁRRAGA, Christina TEFLIoudI, Katja HOSE et Fabian SUCHANEK. « Fast Rule Mining in Ontological Knowledge Bases with AMIE+ ». Dans : *The VLDB Journal* (2015).
- [100] Juri GANITKEVITCH, Benjamin VAN DURME et Chris CALLISON-BURCH. « PPDB : The Paraphrase Database ». Dans : *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Atlanta, Georgia : Association for Computational Linguistics, juin 2013.
- [101] Liqiang GENG et Howard J. HAMILTON. « Interestingness Measures for Data Mining : A Survey ». Dans : *ACM Comput. Surv.* 38.3 (sept. 2006). DOI : [10.1145/1132960.1132963](https://doi.org/10.1145/1132960.1132963).
- [102] Kaoutar GHAZI, Sébastien MARCHAL, Andon TCHECHMEDJIEV, Pierre-Antoine JEAN, Nicolas SUTTON-CHARANI et Sébastien HARISPE. « Classification de phrases courtes : des approches non-supervisées aux approches faiblement supervisées ». Dans : *EGC 2020, Extraction et Gestion des Connaissances (TextMine - Atelier sur la fouille de textes)*. Bruxelles, Belgium, jan. 2020.
- [103] Kaoutar GHAZI, Andon TCHECHMEDJIEV, Pierre-Antoine JEAN, Sébastien HARISPE et Nicolas SUTTON-CHARANI. « On the evaluation of retrofitting for supervised short-text classification. » Dans : *1st International Workshop, Part of The Joint Ontology Workshops JOWO 2020 at FOIS 2020*. Bozen-Bolzano, Italy, juin 2020.
- [104] Amelia GLAESE et al. *Improving alignment of dialogue agents via targeted human judgements*. 2022. arXiv : 2209.14375 [cs.LG].
- [105] Josu GOIKOETXEA, Eneko AGIRRE et Aitor SOROA. « Single or Multiple? Combining Word Representations Independently Learned from Text and WordNet ». Dans : *Proceedings of the AAAI Conference on Artificial Intelligence* 30.1 (mars 2016). DOI : [10.1609/aaai.v30i1.10321](https://doi.org/10.1609/aaai.v30i1.10321).
- [106] Palash GOYAL et Emilio FERRARA. « Graph embedding techniques, applications, and performance : A survey ». Dans : *Knowledge-Based Systems* 151 (2018).

- [107] *Graph data model used in the SML library*. https://factory.euromov.eu/sml/index.php?q=doc_graph_based. Accessed : 2023-07-12.
- [108] Gregory GREFENSTETTE. « INRIASAC : Simple Hypernym Extraction Methods ». Dans : *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado : Association for Computational Linguistics, juin 2015. DOI : 10.18653/v1/S15-2152.
- [109] Suriya GUNASEKAR et al. *Textbooks Are All You Need*. 2023. arXiv : 2306.11644 [cs.CL].
- [110] Ulrike HAHN. « Similarity ». Dans : *Wiley Interdisciplinary Reviews : Cognitive Science* 5.3 (2014).
- [111] *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, 2004.
- [112] Sébastien HARISPE. « Knowledge-based Semantic Measures : from Theory to Applications ». Thèse de doct. Université Montpellier 2, 2014.
- [113] Sébastien HARISPE, Abdelhak IMOUSSATEN, François TROUSSET et Jacky MONTMAIN. « On the consideration of a bring-to-mind model for computing the Information Content of concepts defined into ontologies ». Dans : *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2015. DOI : 10.1109/FUZZ-IEEE.2015.7337964.
- [114] Sébastien HARISPE, Massissilia MEDJKOUNE et Jacky MONTMAIN. « Eliciting Implicit Evocations Using Word Embeddings and Knowledge Representation ». Dans : *Scalable Uncertainty Management*. Cham : Springer International Publishing, 2017. DOI : 10.1007/978-3-319-67582-4_6.
- [115] Sébastien HARISPE, Massissilia MEDJKOUNE et Jacky MONTMAIN. « Summarizing conceptual descriptions using knowledge representations ». Dans : *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. Déc. 2016. DOI : 10.1109/SSCI.2016.7850024.
- [116] Sébastien HARISPE, Sylvie RANWEZ, Stefan JANAQI et Jacky MONTMAIN. « Semantic Measures Based on RDF Projections : Application to Content-Based Recommendation Systems ». Dans : *On the Move to Meaningful Internet Systems : OTM 2013 Conferences*. Graz (Austria) : Springer Berlin Heidelberg, 2013. DOI : 10.1007/978-3-642-41030-7_44.
- [117] Sébastien HARISPE, Sylvie RANWEZ, Stefan JANAQI et Jacky MONTMAIN. *Semantic Similarity from Natural Language and Ontology Analysis*. T. 8. Synthesis Lectures on Human Language Technologies 1. preprint sur ArXiv : <https://arxiv.org/abs/1704.05295>. Morgan & Claypool publishers, mai 2015. DOI : 10.2200/S00639ED1V01Y201504HLT027.
- [118] Sébastien HARISPE, Sylvie RANWEZ, Stefan JANAQI et Jacky MONTMAIN. « Semantic Similarity from Natural Language and Ontology Analysis ». Dans : *CoRR abs/1704.05295* (2017). arXiv : 1704.05295.
- [119] Sébastien HARISPE, David SÁNCHEZ, Sylvie RANWEZ, Stefan JANAQI et Jacky MONTMAIN. « A Framework for Unifying Ontology-based Semantic Similarity Measures : a Study in the Biomedical Domain ». Dans : *Journal of Biomedical Informatics* 48 (2014). DOI : 10.1016/j.jbi.2013.11.006.
- [120] Zellig S HARRIS. « Distributional structure ». Dans : *Word* 10.2-3 (1954).
- [121] Kazuma HASHIMOTO, Caiming XIONG, Yoshimasa TSURUOKA et Richard SOCHER. « A Joint Many-Task Model : Growing a Neural Network for Multiple NLP Tasks ». Dans : *CoRR abs/1611.01587* (2016). arXiv : 1611.01587.
- [122] James HENDLER, Fabien GANDON et Dean ALLEMANG. *Semantic web for the working ontologist : Effective modeling for linked data, RDFS, and OWL*. Morgan & Claypool, 2020.
- [123] Geoffrey E HINTON et al. « Learning distributed representations of concepts ». Dans : *Proceedings of the eighth annual conference of the cognitive science society*. T. 1. Amherst, MA. 1986.
- [124] Pascal HITZLER. « A review of the semantic web field ». Dans : *Communications of the ACM* 64.2 (2021).
- [125] Jordan HOFFMANN et al. *Training Compute-Optimal Large Language Models*. 2022. arXiv : 2203.15556 [cs.CL].
- [126] Edward J. HU, Yelong SHEN, Phillip WALLIS, Zeyuan ALLEN-ZHU, Yuanzhi LI, Shean WANG, Lu WANG et Weizhu CHEN. *LoRA : Low-Rank Adaptation of Large Language Models*. 2021. arXiv : 2106.09685 [cs.CL].
- [127] E. A. HUERTA et al. « FAIR for AI : An interdisciplinary and international community building perspective ». Dans : *Scientific Data* 10.1 (juill. 2023). DOI : 10.1038/s41597-023-02298-6.
- [128] *Hugging Face datasets*. <https://huggingface.co/docs/datasets>. Accessed : 2023-09-21.

- [129] Muhammad Jawad HUSSAIN, Heming BAI, Shahbaz Hassan WASTI, Guangjian HUANG et Yuncheng JIANG. « Evaluating semantic similarity and relatedness between concepts by combining taxonomic and non-taxonomic semantic features of WordNet and Wikipedia ». Dans : *Information Sciences* 625 (2023).
- [130] Stefan JANAQI, Sébastien HARISPE, Jacky MONTMAIN et Sylvie RANWEZ. « Sélection Robuste de Mesures de Similarité Sémantique à partir de Données Incertaines d'Expertise ». Dans : *LFA 2014, Rencontres francophones sur la logique floue et ses applications*. Cargèse, France, oct. 2014.
- [131] Stefan JANAQI, Sébastien HARISPE, Sylvie RANWEZ et Jacky MONTMAIN. « Robust Selection of Domain-Specific Semantic Similarity Measures from Uncertain Expertise ». Dans : *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Cham : Springer International Publishing, 2014. DOI : 10.1007/978-3-319-08852-5_1.
- [132] Pierre-Antoine JEAN. « Gestion de l'incertitude et de l'imprécision dans un processus d'extraction de connaissances à partir des textes ». Theses. Université Montpellier, nov. 2017.
- [133] Pierre-Antoine JEAN, Sébastien HARISPE, Sylvie RANWEZ, Patrice BELLOT et Jacky MONTMAIN. « Étude d'un modèle d'inférence de connaissances à partir de textes ». Dans : *CORIA 2017, Conférence en Recherche d'Information et Applications*). Marseille, France, mars 2017.
- [134] Pierre-Antoine JEAN, Sébastien HARISPE, Sylvie RANWEZ, Patrice BELLOT et Jacky MONTMAIN. « Un modèle probabiliste pour la détection de l'incertitude dans le langage naturel ». Dans : *Document numérique - Revue des sciences et technologies de l'information. Série Document numérique* 19-2016.2-3 (déc. 2016). DOI : 10.3166/dn.19.2-3.53-73.
- [135] Pierre-Antoine JEAN, Sébastien HARISPE, Sylvie RANWEZ, Patrice BELLOT et Jacky MONTMAIN. « Un modèle probabiliste pour la détection de l'incertitude dans le langage naturel ». Dans : *CORIA 2016, Conférence en Recherche d'Information et Applications*. Toulouse, France, 2016.
- [136] Pierre-Antoine JEAN, Sébastien HARISPE, Sylvie RANWEZ, Patrice BELLOT et Jacky MONTMAIN. « Uncertainty detection in natural language : a probabilistic model ». Dans : *WIMS'16, Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*. Nîmes, France, juin 2016. DOI : 10.1145/2912845.2912873.
- [137] *jeu Pyramide Wikipedia*. [https://fr.wikipedia.org/wiki/Pyramide_\(jeu_t%C3%A9%91%C3%A9vis%C3%A9\)](https://fr.wikipedia.org/wiki/Pyramide_(jeu_t%C3%A9%91%C3%A9vis%C3%A9)). Accessed : 2023-09-20.
- [138] Adulam JEYASOTHY, Thibault LAUGEL, Marie-Jeanne LESOT, Christophe MARSALA et Marcin DETYNECKI. « Integrating Prior Knowledge in Post-hoc Explanations ». Dans : *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Cham : Springer International Publishing, 2022.
- [139] Armand JOULIN, Edouard GRAVE, Piotr BOJANOWSKI et Tomas MIKOLOV. *Bag of Tricks for Efficient Text Classification*. 2016. arXiv : 1607.01759 [cs.CL].
- [140] Luc JULIA. *L'Intelligence artificielle n'existe pas*. First Edition, 2019.
- [141] *Kaggle*. <https://www.kaggle.com/>. Accessed : 2023-09-21.
- [142] Jaz KANDOLA, Nello CRISTIANINI et John SHAWE-TAYLOR. « Learning Semantic Similarity ». Dans : *Advances in Neural Information Processing Systems*. T. 15. MIT Press, 2002.
- [143] Wafaa S. EL-KASSAS, Cherif R. SALAMA, Ahmed A. RAFEA et Hoda K. MOHAMED. « Automatic text summarization : A comprehensive survey ». Dans : *Expert Systems with Applications* 165 (2021). DOI : <https://doi.org/10.1016/j.eswa.2020.113679>.
- [144] Yevgeny KAZAKOV, Markus KRÖTZSCH et František SIMANČIK. « The Incredible ELK : From Polynomial Procedures to Efficient Reasoning with Ontologies ». Dans : *Journal of automated reasoning* 53.1 (2014).
- [145] Nikita KITAEV, Łukasz KAISER et Anselm LEVSKAYA. *Reformer : The Efficient Transformer*. 2020. arXiv : 2001.04451 [cs.LG].
- [146] Mika KLEMETTINEN, Heikki MANNILA, Pirjo RONKAINEN, Hannu TOIVONEN et A. Inkeri VERKAMO. « Finding Interesting Rules from Large Sets of Discovered Association Rules ». Dans : *Proceedings of the Third International Conference on Information and Knowledge Management. CIKM '94*. Gaithersburg, Maryland, USA : Association for Computing Machinery, 1994. DOI : 10.1145/191246.191314.

- [147] Natalia KONSTANTINOVA, Sheila C.M. de SOUSA, Noa P. CRUZ, Manuel J. MAÑA, Maite TABOADA et Ruslan MITKOV. « A review corpus annotated for negation, speculation and their scope ». Dans : *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey : European Language Resources Association (ELRA), mai 2012.
- [148] Wouter KOOL, Herke van HOOF et Max WELLING. « Attention, Learn to Solve Routing Problems ! » Dans : *International Conference on Learning Representations*. 2019.
- [149] Alex KRIZHEVSKY, Ilya SUTSKEVER et Geoffrey E. HINTON. « ImageNet Classification with Deep Convolutional Neural Networks ». Dans : *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012.
- [150] Nicholas D KULLMAN, Martin COUSINEAU, Justin C GOODSON et Jorge E. MENDOZA. « Dynamic Ridehailing with Electric Vehicles ». Dans : *Transportation Science* (juill. 2021).
- [151] Maxat KULMANOV, Fatima Zohra SMAILI, Xin GAO et Robert HOEHNDOERF. « Semantic similarity and machine learning with ontologies ». Dans : *Briefings in bioinformatics 22.4* (2021).
- [152] Ajay KUMAR, Shashank Sheshar SINGH, Kuldeep SINGH et Bhaskar BISWAS. « Link prediction techniques, applications, and performance : A survey ». Dans : *Physica A : Statistical Mechanics and its Applications 553* (2020).
- [153] Matt KUSNER, Yu SUN, Nicholas KOLKIN et Kilian WEINBERGER. « From Word Embeddings To Document Distances ». Dans : *Proceedings of the 32nd International Conference on Machine Learning*. T. 37. Proceedings of Machine Learning Research. Lille, France : PMLR, juill. 2015.
- [154] Cécile L 'HÉRITIER, Sébastien HARISPE, Abdelhak IMOUSSATEN, Gilles DUSSEYRE et Benoit ROIG. « Etude d'une approche de Retour d'Expérience pour la découverte d'enseignements génériques dans le domaine humanitaire ». Dans : *IC 2018, Journées Francophones d'Ingénierie des Connaissances*. Nancy, France, juill. 2018.
- [155] Cécile L 'HÉRITIER, Sébastien HARISPE, Abdelhak IMOUSSATEN, Gilles DUSSEYRE et Benoit ROIG. « Selecting Relevant Association Rules From Imperfect Data ». Dans : *13th international conference on Scalable Uncertainty Management (SUM 2019)*. Compiègne, France, déc. 2019. DOI : 10.1007/978-3-030-35514-2_9.
- [156] Cécile L 'HÉRITIER, Abdelhak IMOUSSATEN, Sébastien HARISPE et Gilles DUSSEYRE. « Identifying criteria most influencing strategy performance : Application to humanitarian logistical strategy planning ». Dans : *19th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), 2018*. Cádiz, Spain, juin 2018. DOI : 10.1007/978-3-319-91479-4_10.
- [157] Cécile L 'HÉRITIER, Abdelhak IMOUSSATEN, Sébastien HARISPE, Gilles DUSSEYRE et Benoit ROIG. « Identification de l'information pertinente pour la prise de décision : Application à la logistique humanitaire ». Dans : *LFA 2018, Rencontres francophones sur la logique floue et ses applications*. Arras, France, nov. 2018.
- [158] Cécile L'HÉRITIER. « Une approche de retour d'expérience basée sur l'analyse multicritère et l'extraction de connaissances : Application au domaine humanitaire ». Theses. Université de Nîmes, juin 2020.
- [159] Cécile L'HÉRITIER, Gilles DUSSEYRE, Sébastien HARISPE et Abdelhak IMOUSSATEN. « Contribution à l'aide à la décision en situations d'urgence humanitaire par analyse des retours d'expérience. » Dans : *ACTIS*. 2017.
- [160] Jonathan LAJUS. « Minage de règles rapide, exact et exhaustif dans de larges bases de connaissances ». Theses. Institut Polytechnique de Paris, fév. 2021.
- [161] Thomas K LANDAUER, Danielle S McNAMARA, Simon DENNIS et Walter KINTSCH. *Handbook of latent semantic analysis*. Psychology Press, 2013.
- [162] Levi B. LARKEY et Arthur B. MARKMAN. « Processes of Similarity Judgment ». Dans : *Cognitive Science 29.6* (2005). DOI : https://doi.org/10.1207/s15516709cog00000_30.
- [163] Thibault LAUGEL. « Local post-hoc interpretability for black-box classifiers ». Thèse de doct. Sorbonne Université, 2020.

- [164] Thibault LAUGEL, Marie-Jeanne LESOT, Christophe MARSALA, Xavier RENARD et Marcin DETYNIĘCKI. « The dangers of post-hoc interpretability : Unjustified counterfactual explanations ». Dans : *arXiv preprint arXiv :1907.09294* (2019).
- [165] John LAWRENCE et Chris REED. « Argument Mining : A Survey ». Dans : *Computational Linguistics* 45.4 (déc. 2019). DOI : 10.1162/coli_a_00364.
- [166] Yann LECUN, Yoshua BENGIO et Geoffrey HINTON. « Deep learning ». Dans : *nature* 521.7553 (2015).
- [167] Guang-He LEE et Yun-Nung CHEN. *MUSE : Modularizing Unsupervised Sense Embeddings*. 2018. arXiv : 1704.04601 [cs.CL].
- [168] John Boaz LEE, Ryan A ROSSI, Sungchul KIM, Nesreen K AHMED et Eunye KOH. « Attention models in graphs : A survey ». Dans : *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13.6 (2019).
- [169] Alessandro LENCI et Magnus SAHLGREN. *Distributional semantics*. Cambridge University Press, 2023.
- [170] Mike LEWIS, Yinhan LIU, Naman GOYAL, Marjan GHAZVININEJAD, Abdelrahman MOHAMED, Omer LEVY, Ves STOYANOV et Luke ZETTLEMOYER. *BART : Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. arXiv : 1910.13461 [cs.CL].
- [171] Raymond LI et al. *StarCoder : may the source be with you!* 2023. arXiv : 2305.06161 [cs.CL].
- [172] Yaliang LI, Jing GAO, Chuishi MENG, Qi LI, Lu SU, Bo ZHAO, Wei FAN et Jiawei HAN. *A Survey on Truth Discovery*. 2015. arXiv : 1505.02463 [cs.DB].
- [173] Yuxi LI. *Deep Reinforcement Learning : An Overview*. 2018. arXiv : 1701.07274 [cs.LG].
- [174] Percy LIANG et Christopher POTTS. « Bringing machine learning and compositional semantics together ». Dans : *Annu. Rev. Linguist.* 1.1 (2015).
- [175] Dekang LIN. « An Information-Theoretic Definition of Similarity ». Dans : *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML '98. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1998.
- [176] Tianyang LIN, Yuxin WANG, Xiangyang LIU et Xipeng QIU. *A Survey of Transformers*. 2021. arXiv : 2106.04554 [cs.LG].
- [177] Grace W LINDSAY. « Convolutional neural networks as a model of the visual system : Past, present, and future ». Dans : *Journal of cognitive neuroscience* 33.10 (2021).
- [178] Carolyn E LIPSCOMB. « Medical subject headings (MeSH) ». Dans : *Bulletin of the Medical Library Association* 88.3 (2000).
- [179] Shuming LIU, Christopher BREWSTER et Duncan SHAW. « Ontologies for crisis management : A review of state of the art in ontology design and usability. » Dans : *ISCRAM* (2013).
- [180] Yinhan LIU, Myle OTT, Naman GOYAL, Jingfei DU, Mandar JOSHI, Danqi CHEN, Omer LEVY, Mike LEWIS, Luke ZETTLEMOYER et Veselin STOYANOV. « Roberta : A robustly optimized bert pretraining approach ». Dans : *arXiv preprint arXiv :1907.11692* (2019).
- [181] Cédric LOPEZ, Molka TOUNSI DHOUB, Elena CABRIO, Catherine FARON-ZUCKER, Fabien GANDON et Frédéric SEGOND. « SMILK, trait d'union entre langue naturelle et données sur le web ». Dans : *Revue d'intelligence artificielle* 32.3 (juin 2018). DOI : 10.3166/ria.32.287-312.
- [182] Scott M LUNDBERG et Su-In LEE. « A Unified Approach to Interpreting Model Predictions ». Dans : *Advances in Neural Information Processing Systems*. T. 30. Curran Associates, Inc., 2017.
- [183] Bill MACCARTNEY et Christopher D. MANNING. « Natural Logic for Textual Inference ». Dans : *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Prague : Association for Computational Linguistics, juin 2007.
- [184] Louis MARTIN, Benjamin MULLER, Pedro Javier ORTIZ SUÁREZ, Yoann DUPONT, Laurent ROMARY, Éric de la CLERGERIE, Djamé SEDDAH et Benoît SAGOT. « CamemBERT : a Tasty French Language Model ». Dans : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. DOI : 10.18653/v1/2020.acl-main.645.
- [185] Massissilia MEDJKOUNE. « Vers une approche non orientée pour l'évaluation de la qualité des odeurs ». Theses. Université Montpellier, mars 2018.

- [186] Pablo N MENDES, Max JAKOB, Andrés GARCIA-SILVA et Christian BIZER. « DBpedia spotlight : shedding light on the web of documents ». Dans : *Proceedings of the 7th international conference on semantic systems*. 2011.
- [187] Jean-Christophe MENSONIDES, Sébastien HARISPE, Jacky MONTMAIN et Véronique THIREAU. « Analyse automatique d'arguments et apprentissage multi-tâches : un cas d'étude ». Dans : *Revue Ouverte d'Intelligence Artificielle* 3.3-4 (2022). DOI : 10.5802/roia.29.
- [188] Jean-Christophe MENSONIDES, Sébastien HARISPE, Jacky MONTMAIN et Véronique THIREAU. « Automatic Detection and Classification of Argument Components using Multi-task Deep Neural Network ». Dans : *3rd International Conference on Natural Language and Speech Processing*. Trento, Italy, sept. 2019.
- [189] Jean-Christophe MENSONIDES, Sébastien HARISPE, Jacky MONTMAIN et Véronique THIREAU. « Réseaux de neurones récurrents multi-tâches pour l'analyse automatique d'arguments ». Dans : *CNIA 2018, Conférence Nationale d'Intelligence Artificielle*. Nancy, France, juill. 2018.
- [190] Jean-Christophe MENSONIDES, Pierre-Antoine JEAN, Andon TCHECHMEDJIEV et Sébastien HARISPE. « DÉfi Fouille de Textes 2019 : indexation par extraction et appariement textuel ». Dans : *TALN-RECITAL 2019, Défi Fouille de Textes (atelier)*. Toulouse, France, juill. 2019.
- [191] Jean-Christophe MENSONIDES, Pierre-Antoine JEAN, Andon TCHECHMEDJIEV et Sébastien HARISPE. « IMT Mines Ales at HASOC 2019 : Automatic Hate Speech Detection ». Dans : *FIRE 2019 - 11th Forum for Information Retrieval Evaluation*. T. 2517. CEUR Workshop proceedings - Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation. Kolkata, India, déc. 2019.
- [192] Timothee MICKUS. « On the Status of Word Embeddings as Implementations of the Distributional Hypothesis ». Thèse de doct. Université de Lorraine, 2022.
- [193] Tomas MIKOLOV, Ilya SUTSKEVER, Kai CHEN, Greg S CORRADO et Jeff DEAN. « Distributed representations of words and phrases and their compositionality ». Dans : *Advances in neural information processing systems* 26 (2013).
- [194] George A MILLER. « WordNet : a lexical database for English ». Dans : *Communications of the ACM* 38.11 (1995).
- [195] Jeff MITCHELL et Mirella LAPATA. « Vector-based Models of Semantic Composition ». Dans : *Proceedings of ACL-08 : HLT*. Columbus, Ohio : Association for Computational Linguistics, juin 2008.
- [196] Nikola MRKŠIĆ, Diarmuid Ó SÉAGHDHA, Blaise THOMSON, Milica GAŠIĆ, Lina M. ROJAS-BARAHONA, Pei-Hao SU, David VANDYKE, Tsung-Hsien WEN et Steve YOUNG. « Counter-fitting Word Vectors to Linguistic Constraints ». Dans : *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. San Diego, California : Association for Computational Linguistics, juin 2016. DOI : 10.18653/v1/N16-1018.
- [197] *Munk Debate 2023 : Artificial Intelligence*. <https://munkdebates.com/debates/artificial-intelligence>. Accessed : 2023-09-18.
- [198] Mark A. MUSEN. « The Protégé Project : A Look Back and a Look Forward ». Dans : *AI Matters* 1.4 (juin 2015). DOI : 10.1145/2757001.2757003.
- [199] Stéphane MUSSARD et Fattouma SOUSSI-BENREJAB. « Gini-PLS Regressions ». Dans : *Journal of Quantitative Economics* 17.3 (sept. 2019). DOI : 10.1007/s40953-018-0132-9.
- [200] Richard NGO, Lawrence CHAN et Sören MINDERMAN. *The alignment problem from a deep learning perspective*. 2023. arXiv : 2209.00626 [cs.AI].
- [201] Gildas Tagny NGOMPE, Sébastien HARISPE, Jacky MONTMAIN, Stéphane MUSSARD et Guillaume ZAMBRANO. « Reconnaissance de sections et d'entités dans les décisions de justice : application des modèles probabilistes HMM et CRF. » Dans : *EGC 2017, Extraction et Gestion des Connaissances*. Grenoble, France, 2017.
- [202] Gildas Tagny NGOMPÉ, Sébastien HARISPE, Guillaume ZAMBRANO, Jacky MONTMAIN et Stéphane MUSSARD. « Detecting Sections and Entities in Court Decisions Using HMM and CRF Graphical Models ». Dans : *Advances in Knowledge Discovery and Management : Volume 8*. Cham : Springer International Publishing, 2019. DOI : 10.1007/978-3-030-18129-1_4.
- [203] OPENAI. « GPT-4 Technical Report ». Dans : *ArXiv abs/2303.08774* (2023).

- [204] Martin J OSBORNE et Ariel RUBINSTEIN. *A course in game theory*. MIT press, 1994.
- [205] Long OUYANG et al. *Training language models to follow instructions with human feedback*. 2022. arXiv : 2203.02155 [cs.CL].
- [206] *OWL 2 Web Ontology Language Document Overview (Second Edition)*. <https://www.w3.org/TR/owl2-overview/#Semantics>. Accessed : 2023-07-11.
- [207] *Papers with code (website)*. <https://paperswithcode.com/>. Accessed : 2023-09-22.
- [208] Omkar M PARKHI, Andrea VEDALDI, CV JAWAHAR et Andrew ZISSERMAN. « The truth about cats and dogs ». Dans : *2011 International Conference on Computer Vision*. IEEE, 2011.
- [209] Jeff PASTERNAK et Dan ROTH. « Knowing what to believe (when you already know something) ». Dans : *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 2010.
- [210] Adam PASZKE et al. « PyTorch : An Imperative Style, High-Performance Deep Learning Library ». Dans : *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.
- [211] *Pause Giant AI Experiments : An Open Letter*. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>. Accessed : 2023-09-21.
- [212] F. PEDREGOSA et al. « Scikit-learn : Machine Learning in Python ». Dans : *Journal of Machine Learning Research* 12 (2011).
- [213] Nicole PEINELT, Dong NGUYEN et Maria LIAKATA. « tBERT : Topic models and BERT joining forces for semantic similarity detection ». Dans : *Proceedings of the 58th annual meeting of the association for computational linguistics*. 2020.
- [214] Jeffrey PENNINGTON, Richard SOCHER et Christopher D MANNING. « Glove : Global vectors for word representation ». Dans : *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
- [215] Catia PESQUITA, Daniel FARIA, Andre O FALCAO, Phillip LORD et Francisco M COUTO. « Semantic similarity in biomedical ontologies ». Dans : *PLoS computational biology* 5.7 (2009).
- [216] Matthew E. PETERS, Mark NEUMANN, Mohit IYYER, Matt GARDNER, Christopher CLARK, Kenton LEE et Luke ZETTLEMOYER. *Deep contextualized word representations*. 2018. arXiv : 1802.05365 [cs.CL].
- [217] Victor PILLAC, Michel GENDREAU, Christelle GUÉRET et Andrés L MEDAGLIA. « A review of dynamic vehicle routing problems ». Dans : *European Journal of Operational Research* 225.1 (2013).
- [218] Jocelyn PONCELET, Pierre-Antoine JEAN, Jacky MONTMAIN, François TROUSSET, Sébastien HARISPE et Nicolas PECHEUR. « Impact des mesures de similarité sémantique dans un algorithme de partitionnement : d'un cas biomédical à la détection de comportements de consommation ». Dans : *SFC 2019, Rencontres de la Société Francophone de Classification*. Nancy, France, sept. 2019.
- [219] Céline POU DAT et Frédéric LANDRAGIN. « Explorer un corpus textuel ». Dans : *Explorer un corpus textuel* (2017). DOI : 10.3917/DBU.POU DA.2017.01.
- [220] Christian PRINS. « A simple and effective evolutionary algorithm for the vehicle routing problem ». Dans : *Computers & Operations Research* 31.12 (2004). DOI : [https://doi.org/10.1016/S0305-0548\(03\)00158-8](https://doi.org/10.1016/S0305-0548(03)00158-8).
- [221] Christian PRINS, Philippe LACOMME et Caroline PRODHON. « Order-first split-second methods for vehicle routing problems : A review ». Dans : *Transportation Research Part C : Emerging Technologies* 40 (2014). DOI : <https://doi.org/10.1016/j.trc.2014.01.011>.
- [222] Julien RABATEL, Madalina CROITORU, Dino IENCO et Pascal PONCELET. « Contextual Itemset Mining in DBpedia ». Dans : *LD4KD : Linked Data for Knowledge Discovery*. T. CEUR Workshop Proceedings. 1232. Nancy, France : CEUR Workshop Proceedings, sept. 2014.
- [223] Karim RADOUANE, Andon TCHECHMEDJIEV, Binbin XU et Sebastien HARISPE. « Comparison of Deep Learning Approaches for Protective Behaviour Detection Under Class Imbalance from MoCap and EMG data ». Dans : *ACIIW 2021 - 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*. Nara, Japan : IEEE, sept. 2021. DOI : 10.1109/ACIIW52867.2021.9666417.

- [224] Tadeusz RADZIK et Theo DRIESSEN. « On a family of values for TU-games generalizing the Shapley value ». Dans : *Mathematical Social Sciences* 65.2 (2013). DOI : <https://doi.org/10.1016/j.mathsocsci.2012.10.002>.
- [225] *Rapport de Cédric Villani : donner un sens à l'intelligence artificielle*. <https://www.enseignementsup-recherche.gouv.fr/fr/rapport-de-cedric-villani-donner-un-sens-l-intelligence-artificielle-ia-49194>. Accessed : 2023-09-21.
- [226] Philip RESNIK. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. 1995. arXiv : [cmp-lg/9511007](https://arxiv.org/abs/cmp-lg/9511007) [cmp-lg].
- [227] *Resource Description Framework (RDF)*. <https://www.w3.org/RDF/>. Accessed : 2023-10-15.
- [228] Petar RISTOSKI et Heiko PAULHEIM. « RDF2Vec : RDF Graph Embeddings for Data Mining ». Dans : *The Semantic Web – ISWC 2016*. Cham : Springer International Publishing, 2016.
- [229] Herbert E. ROBBINS. « An Empirical Bayes Approach to Statistics ». Dans : *Breakthroughs in Statistics : Foundations and Basic Theory*. New York, NY : Springer New York, 1992. DOI : [10.1007/978-1-4612-0919-5_26](https://doi.org/10.1007/978-1-4612-0919-5_26).
- [230] Robin ROMBACH, Andreas BLATTMANN, Dominik LORENZ, Patrick ESSER et Björn OMMER. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. arXiv : [2112.10752](https://arxiv.org/abs/2112.10752) [cs.CV].
- [231] Benedek ROZEMBERCZKI, Lauren WATSON, Péter BAYER, Hao-Tsung YANG, Olivér KISS, Sebastian NILSSON et Rik SARKAR. « The shapley value in machine learning ». Dans : *arXiv preprint arXiv :2202.05594* (2022).
- [232] Magnus SAHLGREN. « The distributional hypothesis ». Dans : *Italian Journal of Disability Studies* 20 (2008).
- [233] Hinrich SCHÜTZE, Christopher D MANNING et Prabhakar RAGHAVAN. *Introduction to information retrieval*. T. 39. Cambridge University Press Cambridge, 2008.
- [234] Glenn SHAFER. *A mathematical theory of evidence*. T. 42. Princeton university press, 1976.
- [235] Sapan SHAH, Sreedhar REDDY et Pushpak BHATTACHARYYA. « Emotion Enriched Retrofitted Word Embeddings ». Dans : *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea : International Committee on Computational Linguistics, oct. 2022.
- [236] Avanti SHRIKUMAR, Peyton GREENSIDE et Anshul KUNDAJE. « Learning important features through propagating activation differences ». Dans : *International conference on machine learning*. PMLR. 2017.
- [237] David SILVER et al. « Mastering the Game of Go with Deep Neural Networks and Tree Search ». Dans : *Nature* 529.7587 (jan. 2016). DOI : [10.1038/nature16961](https://doi.org/10.1038/nature16961).
- [238] *Site ANR JCJC ModPuls : Modélisation de données temporelles, rythmiques et de synchronisations sociales par réseaux de neurones impulsifs*. <https://modpuls.wp.imt.fr/>. Accessed : 2023-10-19.
- [239] Pinky SRIKISHU, Kritish PAHI, Pujan THAPA et Subarna SHAKYA. « A comparison of semantic similarity methods for maximum human interpretability ». Dans : *2019 artificial intelligence for transforming business and society (AITB)*. T. 1. IEEE. 2019.
- [240] Fatima Zohra SMAILI, Xin GAO et Robert HOEHNDOERF. « Onto2vec : joint vector-based representation of biological entities and their ontology-based annotations ». Dans : *Bioinformatics* 34.13 (2018).
- [241] Philippe SMETS. « Belief functions on real numbers ». Dans : *International Journal of Approximate Reasoning* 40.3 (2005). DOI : <https://doi.org/10.1016/j.ijar.2005.04.001>.
- [242] Surbhi K. SOLANKI et Jalpa T. PATEL. « A Survey on Association Rule Mining ». Dans : *2015 Fifth International Conference on Advanced Computing Communication Technologies*. 2015. DOI : [10.1109/ACCT.2015.69](https://doi.org/10.1109/ACCT.2015.69).
- [243] Rita T SOUSA, Sara SILVA et Catia PESQUITA. « The Supervised Semantic Similarity Toolkit ». Dans : *European Semantic Web Conference*. Springer. 2022.
- [244] Rita T. SOUSA, Sara SILVA et Catia PESQUITA. « Supervised Biomedical Semantic Similarity ». Dans : *IEEE Access* 11 (2023). DOI : [10.1109/ACCESS.2023.3285406](https://doi.org/10.1109/ACCESS.2023.3285406).
- [245] Fabian M SUCHANEK, Gjergji KASNECI et Gerhard WEIKUM. « Yago : a core of semantic knowledge ». Dans : *Proceedings of the 16th international conference on World Wide Web*. 2007.

- [246] Yutao SUN, Li DONG, Shaohan HUANG, Shuming MA, Yuqing XIA, Jilong XUE, Jianyong WANG et Furu WEI. *Retentive Network : A Successor to Transformer for Large Language Models*. 2023. arXiv : 2307.08621 [cs.CL].
- [247] Richard S SUTTON et Andrew G BARTO. *Reinforcement learning : An introduction*. MIT press, 2018.
- [248] György SZARVAS, Veronika VINCZE, Richárd FARKAS et János CSIRIK. « The BioScope corpus : annotation for negation, uncertainty and their scope in biomedical texts ». Dans : *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Columbus, Ohio : Association for Computational Linguistics, juin 2008.
- [249] György SZARVAS, Veronika VINCZE, Richárd FARKAS, György MÓRA et Iryna GUREVYCH. « Cross-Genre and Cross-Domain Detection of Semantic Uncertainty ». Dans : *Computational Linguistics* 38.2 (juin 2012). DOI : 10.1162/COLI_a_00098.
- [250] Gildas TAGNY NGOMPE. « Méthodes D'Analyse Sémantique De Corpus De Décisions Jurisprudentielles ». Theses. IMT - MINES ALES - IMT - Mines Alès Ecole Mines - Télécom, jan. 2020.
- [251] Gildas TAGNY-NGOMPÉ, Stéphane MUSSARD, Guillaume ZAMBRANO, Sébastien HARISPE et Jacky MONTMAIN. « Identification of Judicial Outcomes in Judgments : A Generalized Gini-PLS Approach ». Dans : *Stats* 3.4 (déc. 2020). DOI : 10.3390/stats3040027.
- [252] Yi TAY, Mostafa DEGHANI, Samira ABNAR, Yikang SHEN, Dara BAHRI, Philip PHAM, Jinfeng RAO, Liu YANG, Sebastian RUDER et Donald METZLER. *Long Range Arena : A Benchmark for Efficient Transformers*. 2020. arXiv : 2011.04006 [cs.LG].
- [253] Andon TCHECHMEDJIEV, Pavlos FAFALIOS, Katarina BOLAND, Malo GASQUET, Matthaus ZLOCH, Benjamin ZAPILKO, Stefan DIETZE et Konstantin TODOROV. « ClaimsKG : A Knowledge Graph of Fact-Checked Claims ». Dans : *ISWC 2019 - 18th International Semantic Web Conference*. Auckland, New Zealand, oct. 2019. DOI : 10.1007/978-3-030-30796-7_20.
- [254] GraphIK TEAM. *Brief Overview of the Existential Rule Framework*. Rapp. tech. LIRMM-Inria, Montpellier, 2012.
- [255] *The Oxford-IIIT Pet Dataset (Website)*. <https://www.robots.ox.ac.uk/~vgg/data/pets/>. Accessed : 2023-10-15.
- [256] Emanuel TODOROV, Tom EREZ et Yuval TASSA. « MuJoCo : A physics engine for model-based control ». Dans : *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2012. DOI : 10.1109/IRoS.2012.6386109.
- [257] Julian TOGELIUS et Georgios N. YANNAKAKIS. *Choose Your Weapon : Survival Strategies for Depressed AI Academics*. 2023. arXiv : 2304.06035 [cs.OH].
- [258] François TORREGROSSA, Robin ALLESIARDO, Vincent CLAVEAU, Nihel KOOLI et Guillaume GRAVIER. « A survey on training and evaluation of word embeddings ». Dans : *International Journal of Data Science and Analytics* 11 (2021).
- [259] Hugo TOUVRON et al. *LLaMA : Open and Efficient Foundation Language Models*. 2023. arXiv : 2302.13971 [cs.CL].
- [260] Théo TROUILLON, Johannes WELBL, Sebastian RIEDEL, Eric GAUSSIER et Guillaume BOUCHARD. « Complex Embeddings for Simple Link Prediction ». Dans : *Proceedings of The 33rd International Conference on Machine Learning*. T. 48. Proceedings of Machine Learning Research. New York, New York, USA : PMLR, 20–22 Jun 2016.
- [261] P. D. TURNEY et P. PANTEL. « From Frequency to Meaning : Vector Space Models of Semantics ». Dans : *Journal of Artificial Intelligence Research* 37 (fév. 2010). DOI : 10.1613/jair.2934.
- [262] Amos TVERSKY. « Features of similarity ». Dans : *Psychological Review* 84 (4 juill. 1977). DOI : 10.1037/0033-295X.84.4.327.
- [263] Eduardo UCHOA, Diego PECIN, Artur PESSOA, Marcus POGGI, Thibaut VIDAL et Anand SUBRAMANIAN. « New benchmark instances for the capacitated vehicle routing problem ». Dans : *European Journal of Operational Research* 257.3 (2017).
- [264] Maaïke VAN ROY, Pieter ROBBERECHTS, Wen-Chi YANG, Luc DE RAEDT et Jesse DAVIS. « A Markov Framework for Learning and Reasoning About Strategies in Professional Soccer ». Dans : *Journal of Artificial Intelligence Research* 77 (2023).

- [265] Ashish VASWANI, Noam SHAZEER, Niki PARMAR, Jakob USZKOREIT, Llion JONES, Aidan N GOMEZ, Łukasz KAISER et Illia POLOSUKHIN. « Attention is all you need ». Dans : *Advances in neural information processing systems* 30 (2017).
- [266] Lucia VICENTE et Helena MATUTE. « Humans inherit artificial intelligence biases ». Dans : *Scientific Reports* 13.1 (2023).
- [267] Bernard VICTORRI. « Espaces sémantiques et représentation du sens ». Dans : *éc/artS* 3 (2002).
- [268] Oriol VINYALS, Igor BABUSCHKIN, Wojciech M CZARNECKI, Michaël MATHIEU, Andrew DUDZIK, Junyoung CHUNG, David H CHOI, Richard POWELL, Timo EWALDS, Petko GEORGIEV et al. « Grandmaster level in StarCraft II using multi-agent reinforcement learning ». Dans : *Nature* 575.7782 (2019).
- [269] Denny VRANDEČIĆ et Markus KRÖTZSCH. « Wikidata : a free collaborative knowledgebase ». Dans : *Communications of the ACM* 57.10 (2014).
- [270] Zhiguo WANG, Haitao MI et Abraham ITTYCHERIAH. *Sentence Similarity Learning by Lexical Decomposition and Composition*. 2017. arXiv : 1602.07019 [cs.CL].
- [271] Ronald J. WILLIAMS. « Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning ». Dans : *Mach. Learn.* 8.3–4 (mai 1992). DOI : 10.1007/BF00992696.
- [272] Eyal WINTER. « The shapley value ». Dans : *Handbook of game theory with economic applications* 3 (2002).
- [273] Thomas WOLF et al. « Transformers : State-of-the-Art Natural Language Processing ». Dans : *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*. Online : Association for Computational Linguistics, oct. 2020.
- [274] Deyi XIONG. « Book Reviews : Semantic Similarity from Natural Language and Ontology Analysis by Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain ». Dans : *Computational Linguistics* 42.4 (déc. 2016). DOI : 10.1162/COLI_r_00269.
- [275] Wenyi XU et Christophe MARSALA. « Personalized search in smart indoor environments : Combining a formal location model, user preferences and semantic similarity ». Dans : *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE. 2016.
- [276] Ali YADDADEN, Sébastien HARISPE et Michel VASQUEZ. « Apprentissage par transfert : du TSP au VRP ». Dans : *ROADEF 2021, Congrès annuel de la société Française de Recherche Opérationnelle et d'Aide à la Décision*. Mulhouse (en ligne), France, avr. 2021.
- [277] Ali YADDADEN, Sébastien HARISPE et Michel VASQUEZ. « Is Transfer Learning Helpful for Neural Combinatorial Optimization Applied to Vehicle Routing Problems? ». Dans : *Computing and Informatics* 41.1 (2022). DOI : 10.31577/cai_2022_1_172.
- [278] Ali YADDADEN, Sébastien HARISPE et Michel VASQUEZ. « Neural Order-First Split-Second Algorithm for the Capacitated Vehicle Routing Problem ». Dans : *OLA 2022 - International Conference on Optimization and Learning*. T. 1684. Communications in Computer and Information Science. Syracuse, Italy : Springer International Publishing, juill. 2022. DOI : 10.1007/978-3-031-22039-5_14.
- [279] Ali YADDADEN, Sébastien HARISPE et Michel VASQUEZ. « Une méthode à base d'apprentissage par renforcement pour le problème de tournées de véhicules avec contrainte de capacité ». Dans : *ROADEF 2023, Congrès annuel de la Société Française de Recherche Opérationnelle et d'Aide à la Décision*. Rennes, France, fév. 2023.
- [280] Ali YADDADEN, Sébastien HARISPE, Michel VASQUEZ et Mirsad BULJUBASIC. « Apprentissage automatique pour l'optimisation combinatoire : Étude du problème du voyageur de commerce ». Dans : *CNIA 2020, Conférence Nationale en Intelligence Artificielle*. Angers, France, juin 2020.
- [281] Bishan YANG, Wen-tau YIH, Xiaodong HE, Jianfeng GAO et Li DENG. *Embedding Entities and Relations for Learning and Inference in Knowledge Bases*. 2015. arXiv : 1412.6575 [cs.CL].
- [282] Amel YESSAD, Catherine FARON-ZUCKER, Rose DIENG-KUNTZ et Med Tayeb LASKRI. « Ontology-based semantic relatedness for detecting the relevance of learning resources ». Dans : *Interactive Learning Environments* 19.1 (2011).
- [283] Zhiqi YIN, Zeshi YANG, Michiel VAN DE PANNE et KangKang YIN. « Discovering diverse athletic jumping strategies ». Dans : *ACM Transactions on Graphics (TOG)* 40.4 (2021).
- [284] Manzil ZAHEER et al. *Big Bird : Transformers for Longer Sequences*. 2021. arXiv : 2007.14062 [cs.LG].

- [285] Jun ZHANG, Shuyang JIANG, Jiangtao FENG, Lin ZHENG et Lingpeng KONG. *CAB : Comprehensive Attention Benchmarking on Long Sequence Modeling*. 2023. arXiv : 2210.07661 [cs.LG].
- [286] Yizhen ZHANG, Kuan HAN, Robert WORTH et Zhongming LIU. « Connecting concepts in the brain by mapping cortical representations of semantic relations ». Dans : *Nature Communications* 2020 11 :1 11 (1 avr. 2020). DOI : 10.1038/s41467-020-15804-w.
- [287] Qiankun ZHAO et Sourav S BHOWMICK. « Association rule mining : A survey ». Dans : *Nanyang Technological University, Singapore* 135 (2003).
- [288] Jiangbin ZHENG, Yile WANG, Ge WANG, Jun XIA, Yufei HUANG, Guojiang ZHAO, Yue ZHANG et Stan Z. LI. *Using Context-to-Vector with Graph Retrofitting to Improve Word Embeddings*. 2023. arXiv : 2210.16848 [cs.CL].