Actes de la 8e conférence nationale sur les Applications Pratiques de l'Intelligence Artificielle

Christelle Launois, Céline Rouveirol

AfIA

Association française
pour l'Intelligence Artificielle

APIA

Conférence Nationale
sur les
Applications Pratiques de l'Intelligence Artificielle

PFIA 2022

Table des matières

Éditorial

Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle

L'Intelligence Artificielle poursuit son essor sans précédent dans les laboratoires privés et publics et en entreprise. Les recherches menées ces dernières années ont abouti à des résultats spectaculaires dans certains domaines et des résultats très prometteurs dans d'autres. Aujourd'hui, l'IA se trouve au cœur de nombreuses applications très performantes qui révolutionnent notre vie quotidienne.

Plus que jamais, l'objectif de cette 7ème Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA 2022) est de donner une tribune dans le cadre de PFIA aux applications concrètes de l'IA qui couronnent de succès l'opérationnalisation de l'IA et des travaux de recherche dans ce domaine. APIA cible des contributions décrivant des applications qui s'appuient sur une ou plusieurs méthodes de l'IA dans tous ses domaines. Cette année, dans les 15 articles et le poster retenus et inclus dans ces actes, les domaines abordés, parfois de façon conjointe pour résoudre un problème complexe sont :

- l'Apprentissage Automatique et la Fouille de données,
- l'Ingénierie et le partage des Connaissances
- le Traitement Automatique du Langage Naturel
- l'Explicabilité, l'Ethique et la Confiance
- le Raisonnement à base de modèles, à base de règles, pour l'aide à la décision
- le Raisonnement spatial et temporel dans des environnements physiques
- le Traitement de données complexes (temporelles, mutli-modales, ...), issues de capteurs intelligents, de systèmes physiques, ou de simulations

Qu'elles soient industrielles, sociétales, économiques, politiques, environnementales, artistiques ou autres, cette conférence est l'occasion de présenter des applications concrètes et des travaux dont l'objet d'étude adresse des problèmes et/ou des données opérationnelles. L'objectif est également de comprendre comment ces applications concrètes font remonter des verrous scientifiques que la communauté des chercheurs en IA doit résoudre pour démocratiser encore davantage son utilisation : l'IA est-elle suffisamment expressive et intelligible pour être utilisée ? Est-elle fiable et robuste ? Est-elle capable de passer à l'échelle ? Comment garantir l'interprétabilité ou l'explicabilité de l'IA ? Beaucoup d'articles de ce millésime d'APIA abordent une ou plusieurs de ces questions. Soulignons également que l'article qui a obtenu le prix – remis pour le première fois dans la conférence – du Collège Industriel de l'AFIA : *L'IA au travail : propositions pour outiller la confiance* par Y. Ferguson et C. Pecoste, est centré sur les problèmes d'éthique et de confiance liés à l'utilisattion de l'IA dans les organisations et les métiers.

La conférence invitée de Jean François Puget de NVIDIA France, *Applying Lessons From Kaggle Winning Solutions to Real World Problems* abordera un problème central de l'applicabilité et le transfert des modèles et méthodes d'apprentissage gagnants des compétitions Kaggle – et des compétences acquises acquises dans le cadre de ces compétitions – à la résolution de problèmes réels.

Enfin, afin de favoriser encore davantage l'échange entre chercheurs académiques et industriels et que ces derniers puissent partager leurs expériences et débattre des différents verrous qu'ils rencontrent dans le développement d'applications autour de l'IA, APIA 2022 accueille cette année encore deux présentations invitées de partenaires industriels de l'AFIA, JellySmack et Berger Levrault.

Nous tenons à remercier ici tous ceux qui ont participé de près ou de loin au succès d'APIA 2022, le comité d'organisation de PFIA 2022, les membres du comité de programme, les auteurs des articles, Jean François Puget et les conférenciers invités du Collège Industriel de l'AFIA et enfin tous les participants à la plateforme.

Christelle LAUNOIS, Céline ROUVEIROL

Comité de programme

Présidentes

— Christelle Launois (Société Générale) ;
— Céline Rouveirol (Université Sorbonne Paris Nord).

Membres

— Florence Amardeilh (Elzeard)
— Ghislain Atemezing (Mondeca)
— Alain Berger (Ardans)
— Sandra Bringay (LIRMM)
— Stephan Brunessaux (Sensei Consult)
— Davide Buscaldi (LIPN)
— Bruno Carron (Airbus)
— Caroline Chopinaud (Hub France IA)
— Gaël de Chalendar (CEA)
— Yves Demazeau (LIG)
— Sylvie Despres (LIMICS)
— Valentina Dragos (Onera)
— Françoise Fogelman Soulie (Hub France IA)
— Bernard Georges (Société Générale)
— Christophe Guettier (SAFRAN)
— Céline Hudelot (Ecole Centrale Paris)
— Arnaud Lallouet (Huawei)
— Christine Largouët (IRISA)
— Dominique Lenne (Université de Technologie de Compiègne)
— Philippe Leray (Université de Nantes)
— Domitile Lourdeaux (Université de Technologie de Compiègne)
— Sylvain Mahé (EDF Recherche et Développement)
— Juliette Mattioli (Thales)
— Youssef Miloudi (Berger Levrault)
— Marie-Christine Rousset (Université Grenoble Alpes)
— Frédériqe Segond (INRIA)
— Brigitte Trousse (INRIA)

Session 1 : Apprentissage

Sélection d'images satellites pour inversion bathymétrique par réseau de neurones convolutif

Amine Chemchem[1,2], Guillaume Jubelin[1], Grégory Cazanave[1]

[1] IRT Saint Exupery, Toulouse, France [2] Atos, Pôle Data Driven Intelligence, Montpellier, France

lamine.chemchem@atos.net, guillaume.jubelin@irt-saintexupery.com,
gregory.cazanave@irt-saintexupery.com

Résumé

Les satellites optiques Sentinel-2 du programme COPERNICUS de l'ESA permettent d'envisager la production récurrente de cartes bathymétriques à partir des observations acquises. Cette production permettrait un suivi récurrent des fonds océaniques aux abords de la plupart des littoraux du globe mais nécessite pour cela des travaux d'automatisation. La sélection des images candidates à l'inversion bathymétrique reste notamment une tâche manuelle reposant sur l'expertise d'un spécialiste.

Cette sélection est réalisée sur la base de critères environnementaux photo-interprétés : couverture nuageuse, présence de sun glint [1] ou de sillages de bateaux, ou encore hétérogénéité de la masse d'eau. Nous proposons dans ces travaux de substituer cette photo-interprétation de l'expert par une approche d'intelligence artificielle. Elle s'appuie sur un réseau de neurones convolutifs développé à cet effet et entraîné sur une banque d'images labélisées à partir de propriétés optiques de la masse d'eau produites par le logiciel ACOLITE [2].

Le modèle proposé est performant avec une précision moyenne (f-score) de 95%. La mise en oeuvre de méthodes d'explicabilité a démontré que ces performances sont obtenues à partir d'éléments concrets et pertinents.

Mots-clés

Imagerie spatiale optique, paramètres environnementaux, classification, CNN, explicabilité.

Abstract

Sentinel-2 optical satellites of ESA's COPERNICUS programme allow the recurrent production of bathymetric maps based on acquired observations. This would allow recurrent monitoring of the ocean floor near most of the world's coasts, but requires automation work. In particular, the selection of candidate images for bathymetric inversion remains a manual task requiring the expertise of a specialist.

Images selection is based on photo-interpreted environmental criteria : cloud cover, sun glint, boat wakes, or water body heterogeneity. In this work, we propose to replace this expert photo-interpretation by an artificial intelligence approach. It is based on a convolutional neural network developed for this purpose and trained on a set of images labelled using optical properties of the water body produced by the ACOLITE software.

The proposed model performs well with an average accuracy (f-score) of 95%. The implementation of explicability methods has shown that these performances are obtained from concrete and relevant elements.

Keywords

remote sensing, environmental parameters, classification, CNN, explainability.

1 Introduction

La classification d'images satellites comme exploitable ou non pour l'inversion bathymétrique est un problème complexe nécessitant de prendre en compte aussi bien le contenu atmosphérique (couvert nuageux, aérosols) que l'état de surface de la masse d'eau (déferlante, sillage, réflexion spéculaire) ou encore de sa composition (panache turbide, remise en suspension, efflorescence algale) [1].

Alors qu'un expert est capable de qualifier la contribution dans l'image des phénomènes précités, des algorithmes de quantification de ceux-ci sont disponibles : définition d'un masque de nuage, correction du glint, estimation de la turbidité de la masse d'eau, etc.

Plutôt que d'assembler et de paramétrer ces algorithmes dans un système expert pour réaliser la classification, l'idée des travaux présentés est d'exploiter ces algorithmes pour constituer un jeu de données pour l'entraînement d'un réseau de neurones profond. En procédant de la sorte, la tâche chronophage et complexe de définition d'un système expert est supprimée. Nous évitons également de mettre en œuvre de nombreux algorithmes sans savoir si l'image traitée sera exploitée par la suite. Enfin, nous nous appuyons sur les performances éprouvées des réseaux de neurones profonds pour cette tâche complexe de classification [2].

La création du jeu de données d'entraînement est supervisée par l'expert avec une interprétation visuelle et la mise en oeuvre de règles de décisions simples prises sur des paramètres environnementaux extraits automatiquement des

1. sun glint : réflexion spéculaire de l'éclairement solaire en direction du capteur.
2. ACOLITE : processeur d'images satellites du Royal Belgian Institute of Natural Sciences.

images. Il est ainsi possible de constituer rapidement un jeu de données assez important pour l'entraînement d'un réseau de neurones. L'effort de calcul est concentré en amont et pendant la phase d'entraînement du réseau, la simple inférence de ce dernier est réalisée au moment de la classification. La capacité à généraliser ce problème de classification ne dépend plus du paramétrage des algorithmes en fonction des littoraux traités mais de la représentativité de la diversité de cas introduite dans le jeu de données.

La substitution d'un modèle déterministe comme un système expert par un modèle probabiliste tel qu'un réseau de neurone soulève toutefois la problématique du contrôle de la prise de décision. Pour comprendre celle-ci et contrôler le fait que le réseau de neurones s'appuie sur des éléments cohérents de l'image, des méthodes d'explicabilité sont utilisées.

Cet article est organisé comme suit : la section 2 rapporte la littérature des travaux connexes, la section 3 détaille la méthodologie de cette étude, et la section 4 analyse les résultats obtenus avant de conclure.

2 État de l'art

La tache de classification, au sens association d'un label (exploitable ou non-exploitable dans notre cas) à une image, n'est pas répandue dans le domaine de la télédétection. L'image de télédétection est surtout un support pour la cartographie et la classification dans ce domaine s'entend pixel à pixel correspondant ainsi à la tache de segmentation dans le domaine de l'apprentissage profond. Chen et al. [3] présente tout de même une revue détaillée des méthodes de classification d'image basées sur les réseaux de neurones convolutifs tandis que [4, 5] se focalisent sur l'utilisation des réseaux de neurones sur des données de télédétection.

D'autres travaux d'apprentissage machine et d'apprentissage profond essaient de traiter des problématiques connexes sans passer par les méthodes de classification, comme par exemple le papier[6] qui propose un réseau de neurones profond capable d'évaluer des cartes de bathymétrie, en modélisant le problème sous forme de classification pixel par pixel. Le papier [7] essaie d'exploiter la corrélation spatiale locale entre les pixels avec un modèle CNN pour la prédiction de la profondeur de l'eau, en tenant compte de la relation non linéaire entre la valeur de radiance et la valeur de la profondeur de l'eau des pixels adjacents et centraux. Des données de terrain telles que des données de profondeurs d'eau mesurées et des données de sondages lidar ont été utilisées comme entrées pour construire ce modèle, les résultats démontrent l'efficacité du modèle obtenu. L'inconvénient majeur de cette approche reste la partie recueil de données terrain (avec des relevés Lidar) qui est généralement coûteuse et qui nécessite souvent un travail laborieux d'analyse et de correction d'erreurs de récupération de données avant de lancer l'apprentissage du modèle.

Nous pouvons citer d'autres travaux d'apprentissage profond appliqués aux images satellite comme les auteurs de [8] qui proposent un réseau de neurones à convolutions pour la détection de nuages et de leurs ombres portées. Ce modèle a été capable de traiter en même temps les données en provenance de World-View 2 et de Sentinel-2. Dans [9] un modèle hybride conçu de deux classifieurs (un réseaux d neurones multi-couches et un réseau de neurones convolutif) a été proposé afin de traiter la problématique de classification multi-classes. Ce modèle a surpassé les deux classifieurs individuels (MLP et CNN) ainsi que le GLCM-MLP qui inclut les caractéristiques de texture GLCM. Ces approches d'apprentissage profond rapporte à la littérature une amélioration nette en performance, par contre elle nécessitent un travail d'étiquetage minutieux en amont de la phase d'apprentissage.

Pour contourner cela, notre idée est d'utiliser l'outil ACOLITE pour automatiser cette étape tout en assurant une qualité d'étiquetage très proche du travail de l'expert. En effet, nous avons pu configurer ACOLITE de façon à extraire à extraire des paramètres pertinents de la colonne d'eau. La combinaison de ces paramètres avec des seuils appropriés nous a permis d'étiqueter de manière efficace les images Sentinel-2 en classes (exploitable /non-exploitable) comme montré dans la section 3.

De plus, nous avons mis en œuvre des méthodes d'explicabilité sur le modèle d'apprentissage profond construit afin de qualifier les caractéristiques image utilisées par ce dernier dans son processus de prise de décision. Cela permet de s'extraire de la vision "boite noire" souvent utilisée à propos des modèles neuronaux. Dans ce sens, nous pouvons citer plusieurs recherches récentes comme [10] où les auteurs proposent une méthode d'explicabilité qui démontre son efficacité sur la mise en évidence des régions de l'image contributrices à la prise de décision du modèle. Les auteurs de [11] proposent une revue pertinente résumant les méthodes d'explicabilité récentes. Dans notre étude, nous utiliserons l'outil "Xplique" développé par l'IRT Saint Exupéry dans le cadre du projet Deel.ai [3]. Cet outil se base sur plusieurs méthodes locales adaptées aux réseaux de neurones. Ces méthodes sont associées à différentes métriques permettant d'évaluer les résultats de l'explication.

3 Méthodologie

Notre idée dans cette étude est d'utiliser l'outil ACOLITE pour labéliser les images Sentinel-2 et de créer un dataset d'apprentissage en image exploitables et non-exploitables. Par la suite, nous allons comparer plusieurs architectures et configurations de réseaux au travers de plusieurs apprentissages afin de sélectionner le meilleur modèle en qualité de précision et en temps d'inférence. Une fois le meilleur modèle sélectionné, des tests de validation et des analyses d'explicabilité sont mis en oeuvre avant de le mettre en production pour l'inférence et la classification des nouvelles images Sentinel-2 disponibles.

La méthodologie proposée dans cette étude, comme montrée à la figure 1, peut être répartie en trois étapes principales :

L'extraction des images Sentinel-2 originales et de leurs patchs, la labélisation via ACOLITE de ces patchs et la

3. DEpendable & Explainable Learning : https ://www.deel.ai

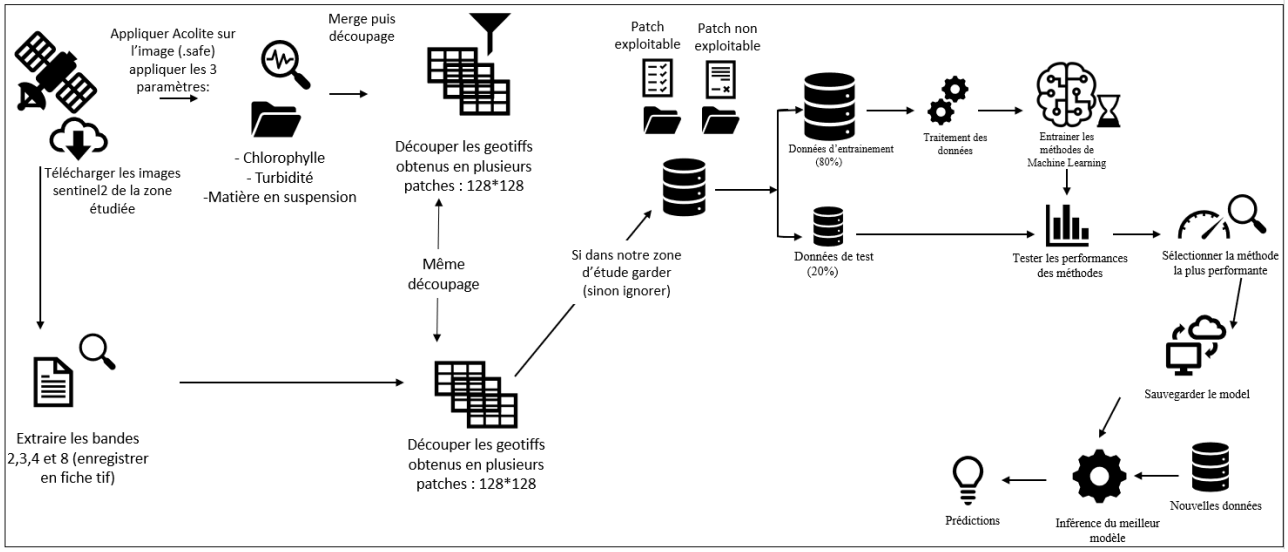


FIGURE 1 – Schéma du workflow global

phase d'affectation des labels.
Une fois l'apprentissage réalisé, nous avons par la suite approfondi l'analyse des résultats avec une étude d'explicabilité, notamment sur les mauvaises classifications. Elle nous a permis de mettre en avant le fait que les décisions du réseau sont majoritairement prises à partir des pixels du domaine continental et non du domaine océanique. De ce fait, nous avons décidé de masquer toutes les images avec un masque de terre pour contraindre l'apprentissage sur le domaine océanique. L'étude d'explicabilité sur ce nouvel apprentissage a bien permis de valider l'efficacité de ce masquage quant à la prise de décision.

3.1 Phase d'extraction de l'image originale et de ses patchs

Dans cette phase, l'image originale est extraite à partir des produits Sentinel-2, en exploitant les bandes 2,3,4, et 8 (bleu, vert, rouge et la proche-infrarouge). Ces spectres ont été sélectionnés afin d'exploiter la meilleure résolution possible offerte par Sentinel-2 sans avoir de traitement de sur ou sous échantillonnage à réaliser (10m). Une fois que l'image (raster) contenant ces 4 bandes est extraite, nous la découpons en plusieurs tuiles (patchs) de taille égale (128*128) afin de ne conserver que les tuiles faisant partie de notre zone d'étude : la bande littorale du département de l'Hérault sur 250km de largeur, voir la figure2.

3.2 Phase de passage d'ACOLITE et génération de ses patchs

Comme expliqué précédemment, nous utilisons ACOLITE pour labéliser les images Sentinel-2 en deux classes : exploitables et non-exploitables. Pour ce faire, nous nous intéressons aux produits de sortie mesures physiques ACOLITE qui contribuent directement ou indirectement à évaluer la qualité de l'image satellite pour le traitement d'inversion bathymétrique. Dans ce sens, nous avons identifié trois

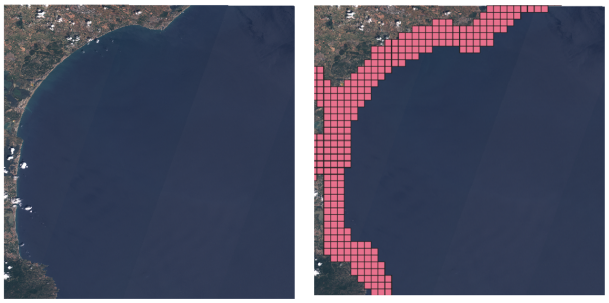


FIGURE 2 – Tuiles étudiées à partir d'une image S2 (la bande littoral héraultaise)

paramètres principaux qui sont : **la turbidité, la chlorophylle et la matière en suspension** ; paramètres calculables et produits au format geotiff pour chaque image Sentinel-2 de notre base de données d'apprentissage.

La turbidité : désigne la teneur d'un fluide en matières qui le troublent. Dans les cours d'eau, elle est généralement causée par des matières en suspension et des particules colloïdales qui absorbent, diffusent ou réfléchissent la lumière. Quand un fleuve turbide se jette en mer, il crée généralement un bouchon vaseux, parfois bien visible depuis un satellite, ceci veut dire qu'une eau trop turbide rend l'image inexploitable pour le calcul de la bathymétrie. A noter que dans ACOLITE, la turbidité est calculé en suivant l'algorithme de Nechad et al[12]. Nous avons fixé par expérimentations le seuil de turbidité acceptable inférieure à 20 mg/L par conséquent, les patchs qui dépassent cette valeur sont classés comme non exploitables.

La chlorophylle : En plus de la turbidité, il est important d'étudier dans les zones non turbides l'indicateur chlorophylle. En effet, cet indicateur reflète la biomasse phytoplanctonique présente dans la colonne d'eau. La chloro-

phylle est un pigment photosynthétique très répandu, il en existe différents types et le plus représenté dans les végétaux marins est la chlorophylle *a* [13]. Elle est de plus représentative de la matière organique végétale, vivante ou fraîchement morte. Sachant que la présence de cet indice impacte négativement l'étude de la bathymétrie, nous avons fixé le seuil de 1.6 µg chla/L par expérimentations pour différencier les images exploitables de celles non exploitables.

La matière en suspension (MES) : autre facteur qui a un impact négatif sur le calcul de la bathymétrie. En effet, MES désigne les matières solides insolubles visibles à l'œil nu présentes en suspension et peut atteindre jusqu'à 3 kg/m2 mais est surtout confinée près de la côte, pour des profondeurs inférieures à 20 mètres on constate que la zone où la concentration en MES est la plus forte correspond à la zone de déferlement des vagues [14]. Par expérimentations nous avons fixé les le seuil de 10 mg/L au générateur ACOLITE pour déterminer les tuiles acceptables de celles qui ne le sont pas.

3.3 Phase d'étiquetage

La phase d'expérimentations de la labélisation nous a permis d'affiner la construction du critère final de seuillage permettant de discriminer au mieux les 2 classes pour chaque patch. Nous nous sommes notamment appuyés sur la comparaison avec une labélisation manuelle réalisée par un expert en bathymétrie cotière. Le critère optimal retenu prend en considération en plus de la moyenne des concentrations des constituants de la colonne d'eau (matière en suspension (TSM), la turbidité (TUR) et la Chlorophylle (CHLA)), le nombre de pixels satisfaisant et ne satisfaisant pas ces conditions (pixel ok - pixel nok) afin de labéliser au patch. Il peut être décrit comme suit :

Algorithm 1 Méthode d'étiquetage des patchs

Pour chaque "i" dans patchs **faire :**
Si 0 < TSM < 10 et 0 < TUR < 20 et 0 < CHLA < 1.6 et Pixel_ok > Pixel_nok **alors** :
    $patchs[i] \leftarrow acceptable$
Sinon :
    $patchs[i] \leftarrow non\_acceptable$

Une fois que les patchs des images ACOLITE sont labélisés, nous étiquetons chaque patch correspondant de l'image originale (issue des bandes 2, 3, 4 et 8) avec le même label. L'enjeu est de réaliser l'apprentissage sur les images originales et par la suite (une fois le modèle entraîné) d'appliquer des prédictions directement sur des images originales sans avoir à passer par l'étape ACOLITE.

3.4 Phase d'apprentissage

TABLE 1 – Recherche de la meilleure architecture / hyperparamètres

Architecture	**Hyperparamètres**	**f1-score**
Conv2D_1(64) max_pooling2d_1 Conv2D_2(64) max_pooling2d_2 Flatten () Dense_1 (64) Dense_2 (2)	epochs=80 batch_size=16 optimizer=Adam learning_rate=0.00001	76%
Conv2D_1(64) max_pooling2d_1 Conv2D_2(64) max_pooling2d_2 Conv2D_3(64) max_pooling2d_3 Flatten () Dense_1 (64) Dense_2 (64) Dense_3 (2)	epochs=80 batch_size=16 optimizer=Adam learning_rate=0.00001	**95%**
Conv2D_1(64) max_pooling2d_1 Conv2D_2(64) max_pooling2d_2 Conv2D_3(128) max_pooling2d_3 Conv2D_4(128) max_pooling2d_4 Flatten () Dense_1 (128) Dense_2 (64) Dense_3 (2)	epochs=80 batch_size=16 optimizer=Adam learning_rate=0.00005	92%

Dans cette phase, nous avons implémenté et testé plusieurs modèles personalisés conçus depuis le début tout en variant les hyperparamètres. L'idée ici est de commencer par l'architecture la plus simple possible et de rajouter des couches et/ou des neurones par couches à chaque expérimentation jusqu'à l'obtention des résultats satisfaisants. L'objectif de cette démarche est de trouver le modèle avec les meilleures performances tout en gardant l'architecture du modèle la moins complexe possible afin d'éviter un besoin excessif en matière de ressources de calcul.
De plus, pour chaque architecture, nous cherchons les hyperparamètres optimums, ceci est facilement implémentable avec Keras_tuner [4]. Les résultats de cette expérimentation rapportés dans le tableau 1 nous ont permis de sélectionner l'architecture la plus simple et la plus efficace en termes de f1-score. Elle est constituée de 3 couches de convolutions suivies de 3 couches denses.
La Figure 3 résume l'architecture du modèle retenu.
Dans cette étude, nous avons rassemblé un dataset contenant 41 images Sentinel-2 datant de 2017 à 2021 de la zone côtière héraultaise, ces 41 images sont découpées en patchs de 128*128 pixels. L'idée ici est de traiter des tuiles de petites taille pour épouser les contours des nuages et autres panaches turbides et ainsi exploiter une plus grande partie des images. Ce découpage nous a donné un ensemble d'apprentissage de 17690 patchs. Nous nous sommes assurés

4. https ://www.tensorflow.org/tutorials/keras/keras_tuner

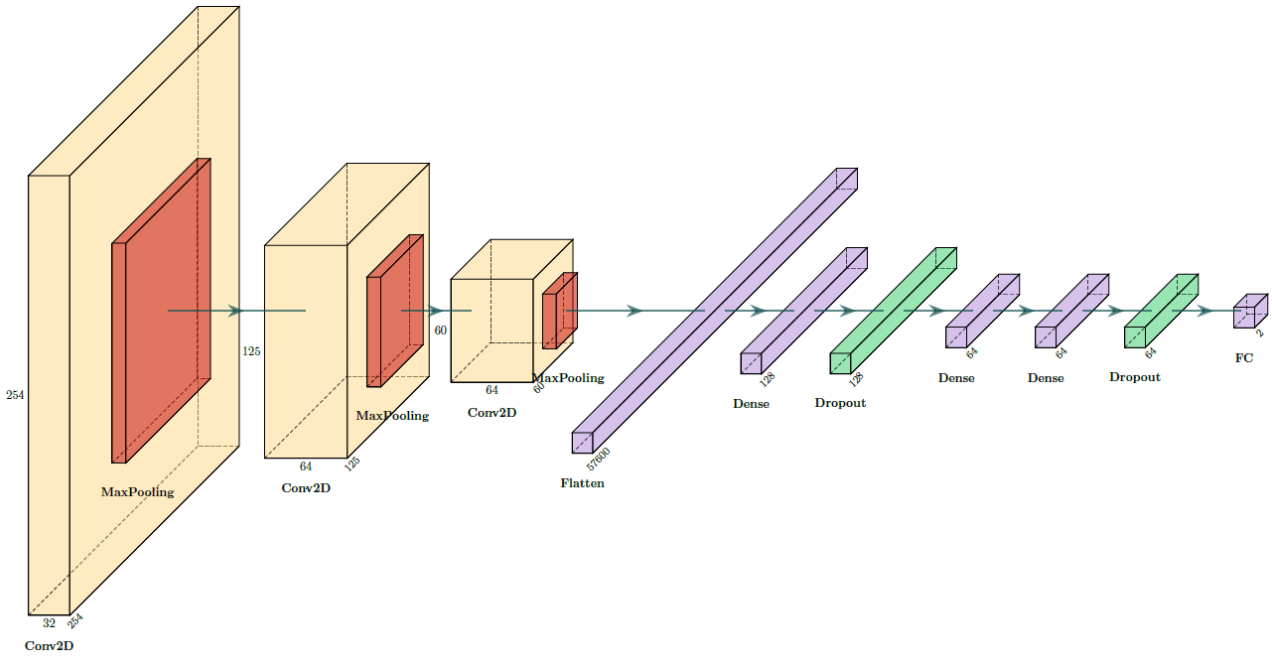


FIGURE 3 – Schéma de l'architecture du modèle CNN retenu

de constituer un dataset équilibré (autant d'images exploitables que non exploitables). Après labélisation, nous avons obtenu 9646 patchs labélisés en classe 0 (exploitable) et 8044 labélisés en classe 1 (non exploitable). La répartition entre l'ensemble d'entraînement et de test est de 85%, 15% respectivement comme montré sur la figure 4 :

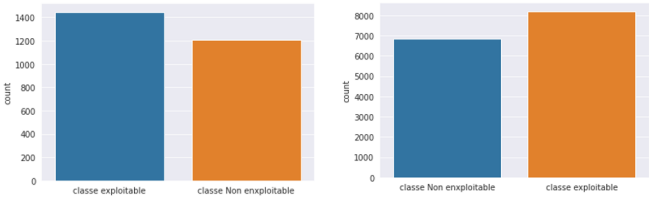


FIGURE 4 – Distribution de patchs d'entraînement / patchs de test

4 Résultats et discussion

Comme décrit précédemment, nous avons fixé les hyperparamètres du modèle par expérimentations par les méthodes de recherche automatique des hyperparamètres, les paramètres retenus sont :

- Nombre d'époch = 80,
- L'optimiseur = Adam,
- l'indice d'apprentissage = 0.00001,
- La taille du batch = 16.

Avec ces paramètres, l'apprentissage du modèle se réalise de façon optimale sans sur-apprentissage apparent (voir la figure 5).

En testant le modèle de classification obtenu sur notre ensemble de test, nous obtenons la matrice de confusion schématisée à la figure 6 ainsi que les évaluations des performances comme mentionnées à la figure 7.

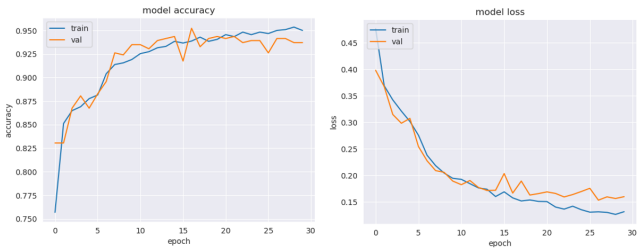


FIGURE 5 – Évolution de la précision et de la loss lors de l'entraînement

Pour rappel les formules d'évaluation sont définies mathématiquement comme suit :

$$Accuracy(Taux\_de\_justesse) = \frac{VP + VN}{VP + VN + FP + FN}$$

$$Prcision = \frac{VP}{VP + FP} \quad , \quad Rappel = \frac{VP}{VP + FN}$$

$$F1{-}score = \frac{2 * Prcision * Rappel}{Prcision + Rappel} = \frac{2 * VP}{2 * VP + FP + FN}$$

Sachant que :
VP : représente les vrais positifs,
VN : les vrais négatifs,
FP : les faux positifs,
FN : les faux négatifs.

Ces résultats montrent bien que le modèle obtenu est très efficace , nous remarquons bien à partir de la matrice de

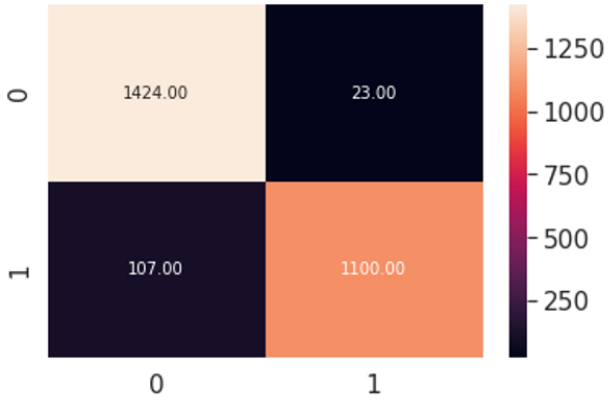


FIGURE 6 – Matrice de confusion

	precision	recall	f1-score	support
0	0.93	0.98	0.96	1447
1	0.98	0.91	0.94	1207
accuracy			0.95	2654
macro avg	0.95	0.95	0.95	2654
weighted avg	0.95	0.95	0.95	2654

FIGURE 7 – Évaluation et performance du modèle

confusion, que sur les 2654 patchs de test, le modèle ne s'est trompé que sur 130 patchs : 107 patchs prédits comme exploitables alors qu'ils ne le sont pas et sur 23 patchs prédits en non-exploitables alors qu'ils le sont.
En évaluant le modèle, ce dernier obtient une précision f1-score moyenne de 95% de précision, score très satisfaisant par rapport à notre objectif de mise en production régulière de la chaîne bathymétrique.

4.1 Explicabilité du modèle retenu

Les résultats obtenus paraissent satisfaisants, néanmoins nous avons approfondi l'analyse avec des méthodes d'explicabilité (avec la librairie Xplique[5] issue du projet DEEL réalisé par les équipes de l'IRT Saint Exupéry) afin de qualifier notamment d'où les erreurs de prédiction peuvent provenir. L'explicabilité permet de connaître les pixels qui influencent le plus la prise de décision du modèle obtenu. Voici pour exemple, 4 tuiles pour l'étude de l'explicabilité, ces 4 tuiles ont été classées correctement par notre modèle de la manière suivante : en Non-exploitables (patch1792_1536, patch1536_1536) et en Exploitables (patch768_2816, patch768_2560) et les résultats de l'explicabilité avec les méthodes Rise, Lime, Kernelshape et Occlusion sont schématisés à la figure 8.
D'après ces résultats, nous pouvons constater que les méthodes donnent des cartes de chaleur assez similaires, surtout lorsqu'il s'agit de la méthode 'Lime' et 'KernelShape' ; elles se basent toutes les deux sur des super-pixels pour schématiser les zones contributrices à la prise de décision lors de l'inférence du modèle. Nous remarquons aussi que la plupart des décisions sont prises à partir des pixels représentant des zones habitées (zone terre) au lieu de se concentrer sur les pixels de la zone côtière comme c'était le cas lors de la phase de labélisation avec les calculs des paramètres ACOLITE. La conclusion est donc que notre apprentissage est biaisé par l'environnement alors que nous souhaitons seulement nous concentrer sur des paramètres physiques maîtrisés de la colonne d'eau. Par conséquent, nous avons décidé de refaire l'apprentissage du modèle en appliquant un masque de terre (land mask) pour contraindre l'apprentissage à se faire exclusivement sur les pixels ne contenant que la colonne d'eau (représentant que des pixels mer).

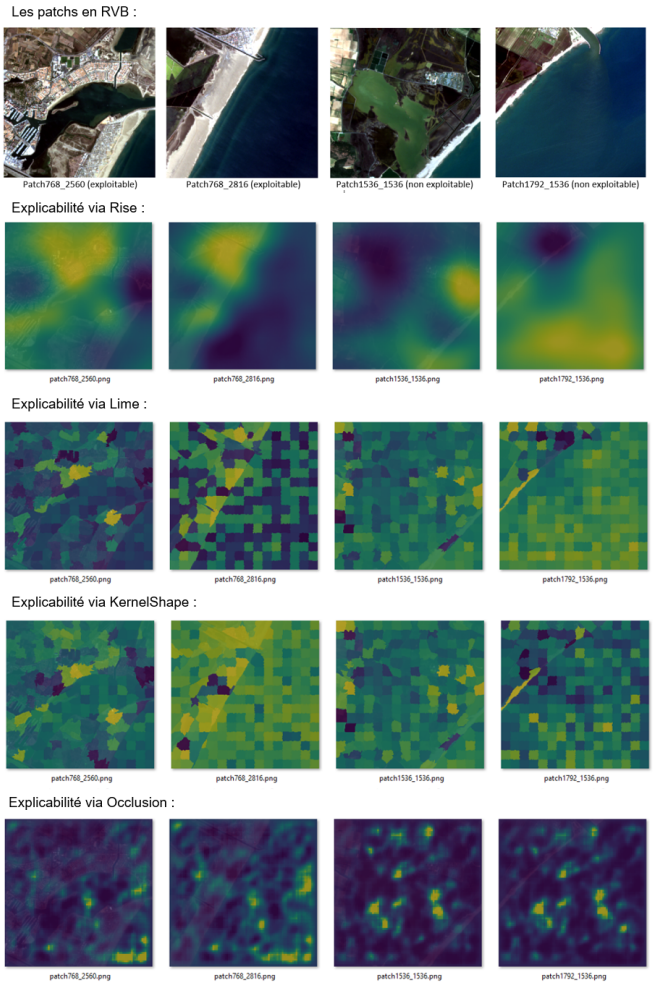


FIGURE 8 – Explicabilité du modèle

4.2 Expérimentations avec des images masquées (par le masque terre)

Notre idée en appliquant le masque terre (land mask) sur notre base d'apprentissage issue d'images Sentinel-2 est de pousser le modèle à ignorer les pixels représentants la

5. https ://github.com/deel-ai/xplique

terre et de raisonner uniquement sur les pixels représentant la mer. Pour ce faire, nous avons extrait une partie du masque terre depuis OpenStreetMap [6] qui correspond à notre zone d'intérêt. Nous montrons en figure 9 un exemple d'une image de notre base d'apprentissage en appliquant ce masque.

Après l'application du masque, nous suivons toutes les étapes précédentes comme pour les images non masquées : découpage en patchs 128*128 pixels, puis labélisation avec ACOLITE et découpage entre ensemble d'apprentissage et ensemble de test pour arriver à l'étape d'implémentation et d'expérimentation du modèle CNN. L'architecture du meilleur modèle retenu pour les images masquées en termes de précision f1-score est identique à celui retenu pour les images non masquées. La matrice de confusion et les résultats d'évaluation de ce modèle sont rapportés respectivement sur la figure 10 et la figure 11.

Nous remarquons que les résultats du modèle entraîné sur les images masquées sont légèrement supérieurs à ceux du modèle entraîné sur les images non masquées. D'après la matrice de confusion, nous pouvons noter que sur l'ensemble de 2707 images de test, il ne se trompe que sur 130 patchs : 77 patchs non exploitables prédits comme exploitables et 53 exploitables prédits comme non exploitables. Ceci donne un f-score moyen de 95% qui est très satisfaisant. Ces résultats illustrent que non seulement la prédiction est plus explicable, plus maîtrisée mais également que le masquage permet de réduire le nombre de prédictions en erreur.

6. https ://osmdata.openstreetmap.de/data/land-polygons.html

FIGURE 9 – Procédure de masquage d'image Sentinel-2 par le masque terre

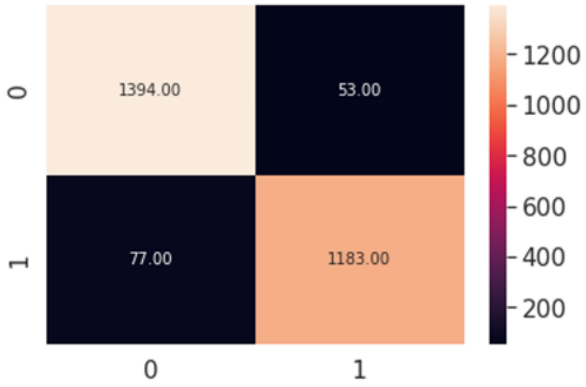


FIGURE 10 – Matrice de confusion du modèle sur les images masquées

	precision	recall	f1-score	support
0	0.95	0.96	0.96	1447
1	0.96	0.94	0.95	1260
accuracy			0.95	2707
macro avg	0.95	0.95	0.95	2707
weighted avg	0.95	0.95	0.95	2707

FIGURE 11 – Performances du modèle sur les images masquées

De la même manière que pour les images non masquées, en appliquant la méthode d'explicabilité "Rise" sur un ensemble de tuiles, nous obtenons les résultats affichés à la figure 12.

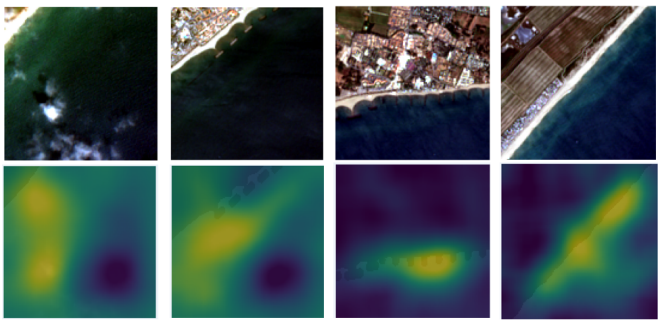


FIGURE 12 – Éxplicabilité du modèle via la méthode 'Rise' sur les images masquées

Malgré l'apport du masquage qui permet de s'absoudre des pixels continentaux dans la prise de décision comme illustré dans la figure 12 au travers de la méthode d'explicabilité Rise, il reste des cas en erreur et notamment des faux positifs. Étant donné que ces faux positifs engendrent des cal-

culs inutiles, il nous parait intéressant de pousser l'analyse de l'explicabilité en la corrélant aux images de paramètres physiques calculés afin de tenter de réduire leur occurence. Il existe également un intérêt sur l'optimisation des faux négatifs : l'optimisation de la couverture spatiale mais celle-ci apparaît comme moins dimensionnante car moins de cas d'erreur et un délai de revisite de 5 jours avec les satellites Sentinel-2.

5 Conclusion et perspectives

Dans cette étude, nous avons pu développer un modèle CNN capable d'identifier de manière très précise si une image satellite est exploitable ou pas par le processus d'inversion bathymétrique. En effet, nous avons pu réaliser un étiquetage très efficace à l'aide de l'outil ACOLITE en utilisant des paramètres physiques de la colonne d'eau tels que : la turbidité, la chlorophylle et les matières en suspension. Grâce à cet étiquetage, nous avons pu proposer un modèle CNN à 6 couches que nous avons entraîné sur des images masquées dans le domaine continental pour le pousser à raisonner uniquement sur les pixels de l'image représentant la mer et à ignorer les pixels terre. Avec cette approche, nous avons obtenu un modèle de classification d'image satellitaire efficace et explicable, car comme montré avec l'étude d'explicabilité, il génère des cartes de caractéristiques à partir des pixels mer et a obtenu un f1 score moyen de 95%.

Comme perspectives, nous prévoyons de poursuivre l'analyse d'explicabilité en étudiant les corrélations entre celle-ci et les paramètres physiques produit par ACOLITE. Nous prévoyons également d'exploiter lors de l'apprentissage du modèle les masques de nuages et de leurs ombres produits par ACOLITE. En effet, les travaux d'explicabilité ont montré des erreurs de classification liées à ceux-ci (voir la figure 13). Ces erreurs sont en particulier liées à la génération de valeurs aberrantes des paramètres physiques sur les nuages et leurs ombres. Les performances du modèle devraient être ainsi améliorées en masquant ces derniers. Par ailleurs, nous comptons évaluer les performances de ce modèle sur des images provenant d'autres satellites tels que les satellites Landsat de la NASA. Ces derniers possèdent des bandes spectrales comparables à celles des Sentinel-2 mais proposent des résolutions spatiales moins fines permettant d'évaluer les capacités du modèle à généraliser selon ce dernier critère. Enfin l'intégration dans la chaîne automatique d'inversion bathymétrique du modèle développé sera réalisée, elle permettra la sélection automatique d'images Sentinel-2 dans un contexte opérationnel.

Remerciements

Cette étude a été réalisée dans le cadre du projet de recherche SB : Monitoring environnemental à l'échelle locale & aide à la décision – Bassin de Thau. Mené par l'IRT Saint Exupery, il est co-financé par les sociétés TELESPAZIO, ATOS, INATYSCO, INSIDE, SUEZ EAU FRANCE (Rivages Pro Tech) et réalisé en collaboration avec le Syndicat Mixte du Bassin de Thau et Sète Agglopôle.

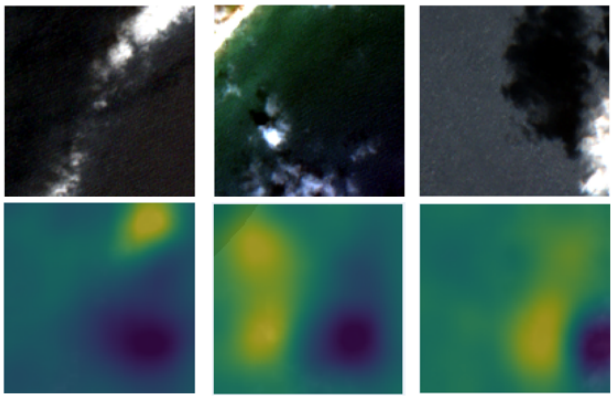


FIGURE 13 – Explicabilité du modèle via la méthode 'Rise' sur les images nuageuses

Références

[1] E. Salameh, F. Frappart, R. Almar, P. Baptista, G. Heygster, B. Lubac, D. Raucoules, L. P. Almeida, E. W. J. Bergsma, S. Capo, M. De Michele, D. Idier, Z. Li, V. Marieu, A. Poupardin, P. A. Silva, I. Turki, and B. Laignel, "Monitoring Beach Topography and Nearshore Bathymetry Using Spaceborne Remote Sensing : A Review," *Remote Sensing*, vol. 11, p. 2212, Jan. 2019. Number : 19 Publisher : Multidisciplinary Digital Publishing Institute.

[2] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications : A meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, June 2019.

[3] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, "Review of Image Classification Algorithms Based on Convolutional Neural Networks," *Remote Sensing*, vol. 13, p. 4712, Jan. 2021. Number : 22 Publisher : Multidisciplinary Digital Publishing Institute.

[4] L. Zhang, L. Zhang, and B. Du, "Deep Learning for Remote Sensing Data : A Technical Tutorial on the State of the Art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, pp. 22–40, June 2016. Conference Name : IEEE Geoscience and Remote Sensing Magazine.

[5] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep Learning in Remote Sensing : A Comprehensive Review and List of Resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, pp. 8–36, Dec. 2017. Conference Name : IEEE Geoscience and Remote Sensing Magazine.

[6] B. Wilson, N. C. Kurian, A. Singh, and A. Sethi, "Satellite-derived bathymetry using deep convolutional neural network," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2280–2283, 2020.

[7] B. Ai, Z. Wen, Z. Wang, R. Wang, D. Su, C. Li, and F. Yang, "Convolutional neural network to retrieve water depth in marine shallow water area from remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 2888–2898, 2020.

[8] M. Segal-Rozenhaimer, A. Li, K. Das, and V. Chirayath, "Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (cnn)," *Remote Sensing of Environment*, vol. 237, p. 111446, 2020.

[9] C. Zhang, X. Pan, H. Li, A. Gardiner, I. Sargent, J. Hare, and P. M. Atkinson, "A hybrid mlp-cnn classifier for very fine resolution remotely sensed image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 133–144, 2018. Geospatial Computer Vision.

[10] M. Brahimi, S. Mahmoudi, K. Boukhalfa, and A. Moussaoui, "Deep interpretable architecture for plant diseases classification," in *2019 Signal Processing : Algorithms, Architectures, Arrangements, and Applications (SPA)*, pp. 111–116, 2019.

[11] Q.-s. Zhang and S.-C. Zhu, "Visual interpretability for deep learning : a survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 27–39, 2018.

[12] B. Nechad, K. G. Ruddick, and Y. Park, "Calibration and validation of a generic multisensor algorithm for mapping of total suspended matter in turbid waters," *Remote Sensing of Environment*, vol. 114, no. 4, pp. 854–866, 2010.

[13] A. Minghelli-Roman and C. Dupouy, "Influence of Water Column Chlorophyll Concentration on Bathymetric Estimations in the Lagoon of New Caledonia, Using Several MERIS Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, pp. 739–745, Apr. 2013. Conference Name : IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

[14] Y. Leredde, H. Michaud, E. Berthebaud, C. Lauer-Leredde, P. Marsaleix, C. Estournel, B. Guerinel, S. Thorin, T. Schvartz, and C. Richard, "Modélisation numérique de l'hydrodynamique sédimentaire dans la baie d'Aigues-Mortes (Languedoc-Roussillon, France). Application à la géomorphodynamique et à la remise en suspension des sédiments rechargés sur les plages," in *XIIIèmes JNGCGC Dunkerque*, pp. 447–458, Editions Paralia, 2014.

Industrialisation d'algorithmes de *deep learning* pour l'extraction des caractéristiques des médicaments

S. Bento Pereira[1], R. Benassi[2], Y. Isaac[1], P. Sendorek[2], S. El Alami[2], R. Sagean[2], S. Lequeux[2], N. Cauvet[1]

[1] Vidal SA, 21 rue Camille Desmoulins, 92789 Issy-les-Moulineaux
[2] Publicis Sapient France, 94 avenue Gambetta, 75020 Paris

suzanne.bento-pereira@vidal.fr
romain.benassi@publicissapient.com

Résumé

Les systèmes d'aide à la décision pour le bon usage du médicament nécessitent de disposer d'informations structurées et normalisées pour décrire les caractéristiques des médicaments (indications, contre-indications, effets indésirables etc.). Par le passé ces données étaient saisies manuellement par les pharmaciens chez VIDAL, mais ce processus a été semi-automatisé ces deux dernières années pour faciliter et rendre plus rapide cette tâche. Nous avons implémenté en production un outil d'indexation semi-automatisé dont nous présenterons ici les performances.

Mots-clés

Traitement du langage naturel, Apprentissage automatique, Santé, Médicaments, Résumé des caractéristiques du produit, Terminologie comme sujet.

Abstract

Decision support modules for the proper use of drugs require structured and standardized information to describe the characteristics of drugs (indications, contraindications, adverse effects, etc.). In the past, this data was entered manually by pharmacists at VIDAL, but this process has been semi-automatized over the past two years to make this task easier and faster. We have implemented in production a semi-automatic indexing tool, its performances will be presented here.

Keywords

Natural language processing, Machine learning, Health, Drugs, Summary of product characteristics, Terminology as topic.

1 Introduction

Les modules d'aide à la décision utilisés dans les LAP (Logiciels d'Aide à la Prescription) permettent de sécuriser la prescription des médecins. Ils sont capables de détecter des anomalies qui peuvent mettre en danger les patients comme des contre-indications, des interactions médicamenteuses, des précautions d'emploi etc. Pour pouvoir fonctionner, et être conformes à la réglementation [1], ils doivent disposer au sein de leur base de connaissance des données structurées nécessaires sur tous les produits médicamenteux disponibles sur le marché comme la liste des concepts de contre-indications, d'indications, ou celle des effets indésirables pour chaque médicament.

1.1 Constitution de la base de connaissance sur les médicaments

Pour constituer cette base de données, les pharmaciens VIDAL doivent s'appuyer sur les textes officiels qui contiennent ces informations : les RCP (Résumé des Caractéristiques des Produits). De ces RCP sont extraites les données administratives sur les médicaments telles que le nom du produit, la date de commercialisation etc. et les données thérapeutiques :

- indications
- contre-indications
- précautions d'emploi
- effets indésirables
- etc.

Il existe dans la base de connaissance plus d'une 50^{ne} de type de données différents à renseigner pour un médicament. Et plus de $15\,000$ médicaments sur le marché pour lesquels ces données doivent être régulièrement mises à jour.

1.2 Indexation manuelle des caractéristiques des médicaments à partir des textes officiels

Cette analyse se fait de manière quotidienne, par la lecture des RCP reçus. Chaque RCP (voir Figure 1 pour un exemple) comprend plusieurs rubriques distinctes (ne sont citées que celles qui nous intéressent ici, il en existe une 30^{ne} en tout) :

- la rubrique *Indications thérapeutiques* : narre les maladies pour lesquelles le médicament peut être utilisé

RÉSUMÉ DES CARACTÉRISTIQUES DU PRODUIT

ANSM - Mis à jour le : 11/04/2011

1. DENOMINATION DU MEDICAMENT

DOLIPRANE 1000 mg, comprimé

2. COMPOSITION QUALITATIVE ET QUANTITATIVE

Paracétamol ..... 1000,00 mg

Pour un comprimé.

Pour la liste complète des excipients, voir rubrique 6.1.

3. FORME PHARMACEUTIQUE

Comprimé.

4. DONNEES CLINIQUES

4.1. Indications thérapeutiques

Traitement symptomatique des douleurs d'intensité légère à modérée et/ou des états fébriles.
Traitement symptomatique des douleurs de l'arthrose.

4.2. Posologie et mode d'administration

Mode d'administration
Voie orale.
Les comprimés sont à avaler tels quels avec une boisson (par exemple eau, lait, jus de fruit).
Posologie
Attention: cette présentation contient 1000 mg de paracétamol par unité: ne pas prendre 2 unités à la fois.
Cette présentation est réservée à l'adulte et à l'enfant à partir de 50 kg (environ 15 ans).
La posologie unitaire usuelle est de un comprimé à 1000 mg par prise, à renouveler au bout de 6 à 8 heures. En cas de besoin, la prise peut être répétée au bout de 4 heures minimum.
Il n'est généralement pas nécessaire de dépasser 3 g de paracétamol par jour, soit 3 comprimés par jour.
Cependant, en cas de douleurs plus intenses, la posologie maximale peut être augmentée jusqu'à 4 g (4 comprimés) par jour. Toujours respecter un intervalle de 4 heures entre deux prises.
Fréquence d'administration:
Les prises systématiques permettent d'éviter les oscillations de douleur ou de fièvre:

- chez l'adulte, elles doivent être espacées de 4 heures minimum.

Insuffisance rénale:
En cas d'insuffisance rénale sévère (clairance de la créatinine inférieure à 10 ml/min), l'intervalle entre deux prises sera au minimum de 8 heures.
Ne pas dépasser 3 g de paracétamol par jour, soit 3 comprimés.

4.3. Contre-indications

- Hypersensibilité au paracétamol ou aux autres constituants.
- Insuffisance hépatocellulaire.

4.4. Mises en garde spéciales et précautions d'emploi

Figure 1: Le RCP du DOLIPRANE 1 000 mg cp (`http://agence-prd.ansm.sante.fr`).

- la rubrique *Contre-indications* : décrit les situations dans lesquelles la prise du médicament est dangereuse
- la rubrique *Effets indésirables* : explicite les effets non souhaités, secondaires au traitement par le médicament et aboutissant à un résultat néfaste (gêne, allergie, complications graves, y compris le décès)

Après lecture et analyse, vient la saisie des données une à une suivant les terminologies existantes et gérées par ailleurs chez VIDAL (voir 3.1.1). Les pharmaciens vont choisir parmi les concepts disponibles de chaque terminologie ceux qui correspondent le mieux à la notion qu'ils souhaitent indexer. Cette saisie se fait via des formulaires dans des applications internes. En moyenne, les pharmaciens vont renseigner 44 termes pour les effets indésirables, 8 pour les contre-indications et 4 termes pour les indications d'un médicament ce qui est un travail assez fastidieux.

1.3 Faciliter l'indexation via l'IA

L'*apprentissage automatique*, ou *Machine Learning* (ML), nous permet de semi-automatiser cette tâche pour la rendre plus rapide et plus simple pour les pharmaciens. Nous avons implémenté en production un outil d'indexation semi-automatisé qui suggère les concepts potentiellement pertinents et positionnés sur les phrases du RCP. Son fonctionnement intègre aujourd'hui l'indexation des indications, contre-indications et effets indésirables. Sur la base de ces propositions, le pharmacien peut décider de valider, supprimer ou modifier les concepts proposés afin de renseigner ces données dans la base de connaissance. Nous présenterons dans cet article le fonctionnement de notre outil basé sur deux approches ML et une approche à base de règles. Nous présenterons également ses performances. Et nous terminerons par une discussion et conclusion.

2 État de l'art

L'indexation de concepts dans les documents consiste à y détecter la présence de concepts d'un référentiel termino-ontologique. Cela permet de rendre l'information qui était jusque-là inexploitable sous forme de texte brut, exploitable par des applications informatiques. Par exemple, dans le domaine médical, les concepts MeSH sont employés pour l'indexation et la recherche d'articles scientifiques dans la base MEDLINE[1], et ceux de la Classification Internationale des Maladies (CIM10) sont employés pour caractériser les séjours hospitaliers à des fins médico-économiques[2].

L'indexation automatique de concepts terminologiques dans des documents de santé a été abordée dans de nombreux travaux de recherches avec l'utilisation de plusieurs techniques. Il existe des approches à base de TAL (Traitement Automatique des Langues) avec l'utilisation de *regex*, dictionnaires et terminologies existantes [2]. Ce sont des approches qui sont assez efficaces lorsqu'elles s'appliquent à des terminologies riches en termes et synonymes et quand les termes sont proches de ceux que nous trouverons dans le texte à analyser. Ces approches ont été par exemple déjà utilisées pour l'extraction de concepts français dans les RCP [3]. Elles ont aussi le mérite de ne pas nécessiter de bases d'apprentissage importantes qui sont souvent inexistantes ou non disponibles.

Des méthodes à base de ML peuvent aussi être utilisées avec des classifieurs type *conditional random fields* (CRFs), *support vector machines* (SVM), *Convolutional Neural Network* (CNN) ou *Transformers* (BERT). Les résultats sont très bons lorsque ces méthodes sont appliquées sur des corpus parfaitement annotés et des terminologies assez restreintes [4, 5, 6, 7, 8, 9]. En particulier, Rubrichi et al. [4] qui a comparé deux approches à base de classifieurs CRFs et SVM pour l'indexation des interactions médicamenteuses en italien à partir des RCP.

Enfin il y a des approches hybrides mêlant *apprentissage automatique* et TAL. Par exemple, Zweigenbaum et al. [10] pour l'indexation de concepts CIM10 en français dans les certificats de décès utilise une méthode à base de dictionnaires puis des classifieurs SVM.

La plupart des études sont toutefois appliquées à la langue anglaise, le français l'est beaucoup moins. Et nous avons trouvé très peu d'articles sur l'indexation de RCP.

3 Méthodes

3.1 La base de documents et les données disponibles

Les données ont été récupérées à partir du système d'information VIDAL. Elles comprennent une base de documents RCP, l'indexation manuelle correspondante et les 3 terminologies d'indexation.

[1] `pubmed.ncbi.nlm.nih.gov`
[2] `epmsi.atih.sante.fr/welcomeEpmsi.do`

3.1.1 Les trois terminologies

Indications. Cette terminologie contient la liste des concepts décrivant les indications. Elle contient 4 047 concepts. Chacun d'entre eux a un identifiant unique, un libellé préféré et 0 à n synonymes.

Contre-indications. Elle liste 3 806 concepts décrivant les contre-indications. Chacun d'entre eux a également un identifiant unique, un libellé préféré et 0 à n synonymes.

Effets indésirables. Contient 4 714 concepts pour les effets indésirables avec pour chacun un identifiant unique, un libellé préféré et 0 à n synonymes.

3.1.2 La base de documents indexés manuellement

Nous disposons au début du projet d'une base de 9 353 RCP indexés, issus de 19 années d'historique d'indexation manuelle par les pharmaciens VIDAL.
L'indexation manuelle est réalisée au niveau du document, elle lie l'identifiant du document avec les identifiants des concepts indexés pour les indications, contre-indications et effets indésirables (voir un exemple dans le Tableau 1). Il n'existe pas de lien entre la rubrique ou la phrase exacte du document et les identifiants des concepts indexés.

Type de données\Produit	PD111 - DOLIPRANE 1000 mg cp
Indications	IND45 Fièvre; IND89 Douleur d'intensité légère à modérée
Contre-indications	CI02 Hépatopathie décompensée; CI46 Hypersensibilité au paracétamol; CI56 Insuffisance hépatique sévère
Effets indésirables	EI12 Céphalée; EI45 Anémie; EI89 Diarrhée; EI87 Confusion mentale; EI43 Malaise; EI15 Vertige

Table 1: Extrait des indications, contre-indications et effets indésirables renseignés avec leurs identifiants pour le médicament DOLIPRANE 1 000 mg cp

La base d'apprentissage contient l'extraction de 8 353 rubriques *Indications thérapeutiques*, 8 353 rubriques *Contre-indications* et 8 353 rubriques *Effets indésirables* indexées. Elle comprend aussi les indexations des concepts contre-indication, indication et effet indésirables correspondants. Elle sera utilisée pour entraîner nos algorithmes d'IA.

3.2 Une combinaison de trois approches

L'objectif est de trouver, parmi une liste pré-définie, l'ensemble des concepts positionnés sur une phrase. Chaque concept est représenté par un label préféré, c'est-à-dire un groupe de mot le caractérisant mais aussi par un ensemble de synonymes, entendus ici comme au sens de groupes de mots potentiellement différents du label préféré mais recouvrant la même réalité.
Nous avons fait le choix de combiner trois types d'approches différentes afin d'optimiser les performances. Deux de ces approches reposent sur une mécanique d'*apprentissage automatique* et sont ensuite combinés dans le cadre d'un fonctionnement dit *hybride*. La dernière approche, quant à elle, se situe en aval des deux premières. Elle s'inscrit dans une logique de rattrapage, à partir de règles métier, de cas bien identifiés par les équipes VIDAL.

3.2.1 Approche par similarité

Distance de Ratcliff-Obershelp. Afin de caractériser la présence d'un concept au sein d'un texte, nous utilisons la mesure de similarité de Ratcliff-Obershelp [11]. Pour deux chaînes de caractères S_1 et S_2, cette mesure se calcule selon la formule suivante

$$D = \frac{2K}{|S_1| + |S_2|} \tag{1}$$

où $|.|$ représente l'opérateur donnant le nombre de caractères d'une chaîne, et K le nombre de *caractères correspondants* entres les deux chaînes. Ce concept de nombre de *caractères correspondants* se définit récursivement par la somme de la taille de la plus grande sous-chaîne en commun et le nombre de *caractères correspondants* des deux côtés de ladite plus grande sous-chaîne en commun.

Dans le cadre de notre application. Afin d'identifier la présence de l'un des concepts d'intérêt au sein du texte donné en entrée, nous allons parcourir l'ensemble des libellés possibles. Pour chaque synonyme nous calculons la mesure de similarité entre chacun des m mots le constituant avec chacun des n mots du texte. Cela revient à construire un tableau de taille $[m, n]$ où la valeur à la $i^{\text{ème}}$ ligne et $j^{\text{ème}}$ colonne correspond à la similarité entre le $i^{\text{ème}}$ mot du synonyme, et le $j^{\text{ème}}$ mot du texte.
Par exemple, si nous recherchons un concept *AVC* représenté par le synonyme *accident vasculaire cérébral* dans un texte qui contiendrait consécutivement chacun des trois mots le constituant, et des mots très différents de part et d'autre, nous aurions un résultat proche du suivant

$$\begin{aligned}
&\ldots 0, 0, 1, 0, 0, 0, 0, 0 \ldots && \text{similarité avec } accident\\
&\ldots 0, 0, 0, 1, 0, 0, 0, 0 \ldots && \text{similarité avec } vasculaire\\
&\ldots 0, 0, 0, 0, 1, 0, 0, 0 \ldots && \text{similarité avec } cérébral.
\end{aligned}$$

Prendre le maximum par colonne nous permet d'aboutir au vecteur suivant

$$\ldots 0, 0, 1, 1, 1, 0, 0, 0 \ldots,$$

nous calculons alors une moyenne glissante sur une fenêtre de taille m, ce qui revient à considérer un calcul de convolution avec un filtre $[1/m, 1/m, \ldots, 1/m]$. Dans le cas de notre exemple, m est égal à 3, nous obtenons alors le filtre $[1/3, 1/3, 1/3]$ et le résultat suivant:

$$\ldots 0, 0, \mathbf{0.3}, \mathbf{0.6}, \mathbf{1}, \mathbf{0.6}, \mathbf{0.3}, 0 \ldots.$$

Le maximum de ce vecteur correspond à ce que nous appelons ici le *maximum de similarité* (et l'indice de ce maximum dans le vecteur peut permettre de récupérer sa position dans le texte).

Il suffit que la valeur de ce *maximum de similarité* soit supérieur à un seuil fixé pour que le concept soit considéré comme présent dans le texte. Nous avons fait le choix d'utiliser une valeur de seuil spécifique par synonyme et d'en faire l'apprentissage automatiquement.

Apprentissage des seuils par modèle. Pour un concept donné, et un synonyme s, l'objectif ici est de déterminer les seuils λ_s, qui vont maximiser la f-mesure associée à la détection du concept parmi un ensemble de textes d'entraînement. Pour chaque synonyme nous considérons un ensemble d'entraînement spécifique contenant tous les textes pour lesquels le concept n'est pas présent, mais seulement un sous-ensemble des textes pour lequel il est présent. Ce sous-ensemble de textes contenant le concept est choisi parmi les textes les plus proches du synonyme, c'est-à-dire dont la valeur de similarité est inférieure à λ_s. Il suffit alors de tester comme valeur de seuil l'ensemble des valeurs de similarité atteintes pour le synonyme donné et vérifier celle qui maximise la f-mesure.

3.2.2 Approche par réseaux convolutifs

Contexte. En complément de l'approche par similarité, et à l'instar de ce que ferait le pharmacien humain, nous avons développé une méthode d'Intelligence Artificielle (IA) dont l'objectif est de s'intéresser directement au sens du texte par opposition à une simple recherche de mots, ou groupe de mots. Nous nous appuyons sur les récentes avancées du *deep learning* dans le domaine du NLP (*Natural Language Processing*) [12, 13, 14, 15, 16, 17] afin d'obtenir un niveau de représentation particulièrement fin du langage.

***Embeddings* et CNN.** Ces méthodes permettent de produire des représentations numériques des mots via une approche nommée *plongement lexical*, ou *word embedding*, dont la spécificité est que deux mots de sens proche y ont une représentation numérique proche elle aussi (voir par exemple [12, 13]). L'état de l'art actuel de ces approches, se fonde sur des architectures de type *Transformers*, introduites dans [15] et dont le modèle BERT [16] est disponible en libre accès. Les RCP traités étant en français c'est logiquement CamemBERT [17], la déclinaison francophone de BERT, qui a été utilisée pour la génération des *embeddings*. Une fois les textes des RCP plongés dans un espace numérique, l'enjeu est toujours d'être capable d'y reconnaître, ou non, la présence des concepts. Plusieurs approches sont possibles (voir notamment [18]) parmi lesquelles nous avons fait le choix d'utiliser une architecture de type CNN par analogie entre la recherche d'un motif dans un texte et celle d'un objet dans une image, cas d'usage classique des CNN ([19, 20, 21]).

Mise en œuvre. La présence d'un concept au sein d'un texte est modélisée comme étant un événement indépendant de la présence respective des autres concepts. En particulier, la présence de l'un n'exclut pas la présence d'un autre. Nous sommes donc en présence d'une classification type *multi-labels* plutôt que *multi-classes*, cette dernière dénomination supposant un caractère exclusif des catégories les unes par rapport aux autres. Cette caractéristique ainsi que la présence d'un grand nombre de concepts à extraire, plusieurs milliers, nous incitent à construire une modélisation spécifique pour *chacun* des concepts.

Le réseau de neurones que nous considérons pour *un concept donné* peut s'écrire sous la forme

$$\hat{p} = \max_t F(c[t-k:t+k]) \tag{2}$$

où $\hat{p}$ est une probabilité caractérisant la présence du concept au sein du texte, F est une fonction correspondant à une succession de filtres convolutifs, c est la séquence d'*embeddings* de texte issue de CamemBERT, t l'indice sur lequel est centrée l'évaluation et k la taille du filtre de convolution. Nous appelons cette architecture *common denominator* puisqu'elle est conçue afin de reconnaître le dénominateur commun (en termes de sémantique) à chaque concept.

Le réseau contient n couches convolutives successives, chacune ayant un paramètre de dilatation égal à 2^i pour i son numéro dans l'ordre de succession. Le nombre n est choisi de manière à être pertinent par rapport à la taille des libellés. Le réseau (voir Figure 2) se compose ainsi de n couches de convolution avec pour chacune une fonction d'activation tangente hyperbolique, sauf pour la dernière qui se voit affectée une fonction sigmoïde afin d'obtenir des valeurs entre 0 et 1, caractérisant une probabilité de présence du concept. Le réseau se termine par une couche de *max pooling* finale qui permet de ne récupérer qu'une valeur de probabilité, la maximale caractérisant à elle seule la présence ou l'absence du concept sur l'ensemble du texte d'intérêt. Afin de construire le jeu d'apprentissage pour chacun des modèles, nous associons à chaque texte une valeur cible (ou *target*) valant 0 ou 1 en fonction respectivement de l'absence ou présence du concept (voir Figure 3). À noter qu'ici, contrairement à l'approche par similarité décrite en 3.2.1, nous n'utilisons que les textes bruts sans aucune information de synonymie. Charge au réseau d'apprendre à reconnaître le dénominateur commun sémantique d'un concept à partir des seuls textes et des valeurs cibles.

Le texte est ensuite transformé sous forme de vecteurs numériques (ou *embeddings*) grâce au modèle CamemBERT selon une procédure en deux temps. La première étape consiste à séparer le texte en *tokens*, des éléments plus petits mais porteurs d'unité de sens pour le modèle (ce ne sont pas nécessairement des mots, ce peut-être seulement des groupes de lettres successives). La seconde revient à associer à chaque *token* un vecteur numérique de taille M (ici, M vaut 768). Le modèle CamemBERT est pré-entraîné pour effectuer automatiquement ces deux tâches, c'est donc ainsi que nous l'utilisons pour transformer nos données d'entrée. Prenons l'exemple d'un texte réduit à une phrase

texte: *Ce médicament ne doit jamais être utilisé dans les états de rétention hydrosodée*,

la séparation en *tokens* nous donne ici vingt éléments sous la forme d'une série d'identifiants,

$$\mathbf{tokens} : \big(\ 5 \quad 44 \quad 31922 \quad 45 \quad 279 \quad 283 \quad \dots \quad 6\ \big).$$

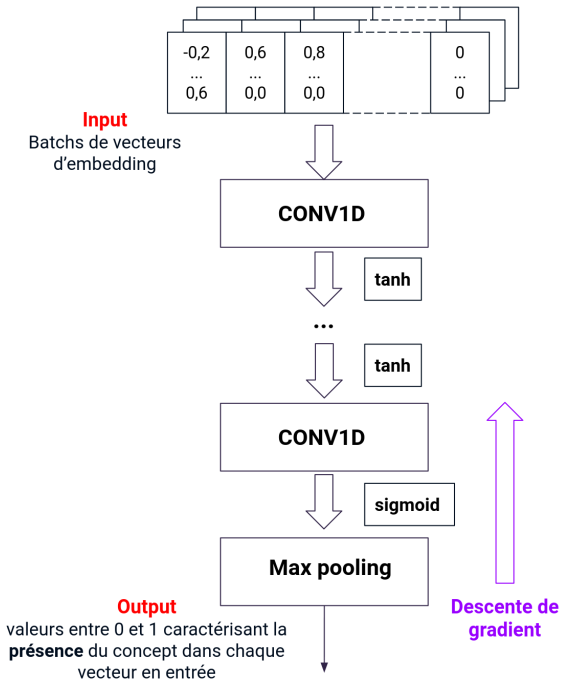


Figure 2: Architecture du réseau *common denonimator* conçu dans le cadre de ce projet.

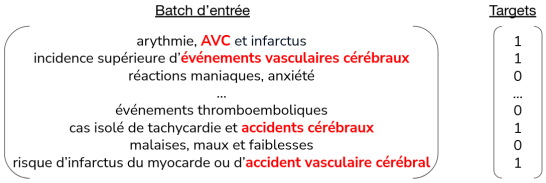


Figure 3: Illustration de la construction d'un jeu d'apprentissage pour un concept donné (ici *accident vasculaire cérébral*).

Chacun de ces *tokens* est ensuite plongé dans un espace vectoriel numérique de dimension 768 ce qui permet d'obtenir en sortie, sur cet exemple, une matrice de taille 768×20,

$$\textbf{embeddings} : \begin{pmatrix} -0.2 & 0.6 & \dots & -0.4 \\ 0.7 & -0.1 & \dots & -0.8 \\ \vdots & \vdots & \vdots & \vdots \\ 0.6 & 0.0 & \dots & -0.4 \end{pmatrix}.$$

Pour entraîner le réseau, nous cherchons à optimiser une fonction de coût qui correspond, à nouveau, à la f-mesure. Concrètement cela revient à optimiser la fonction suivante

$$-\log\left[f_m(\hat{\text{p}}, \text{targets})\right] \tag{3}$$

où f_m correspond à une version dite *soft* de la f-mesure, variante continue de la version traditionnelle, permettant d'aider la convergence lors de l'apprentissage.

3.2.3 Rattrapage à base de règles métiers

Les deux premières approches sont complétées par une troisième, qui va selon plusieurs règles métiers ajuster les propositions de concepts en en ajoutant ou en supprimant. Ces règles associent une portion de texte provenant de RCP à un ou plusieurs concepts provenant des trois terminologies avec une indication *à ajouter* ou *à supprimer*. Ces règles ont été créées à base de règles métiers (après discussions avec les pharmaciens) et à base de motifs en erreurs trouvés après plusieurs tests sur les approches IA. Il existe plus d'un millier de règles.
Un exemple de règle : *Infections et infestations* qui est un titre dans les tableaux de la rubrique *Effets indésirables* ne doivent pas être indexés avec le concept EI79 *infection*.

3.2.4 Combinaison des trois approches

Les trois approches considérées, pour rappel,

1. similarité
2. *common denominator*
3. rattrapage à base de règles métiers,

se combinent et se complètent. Les deux premières relèvent d'un apprentissage automatique et, comme vu précédemment, pour chaque concept deux modèles sont appris. L'un pour l'approche par similarité, l'autre via entraînement d'un réseau de neurones convolutif selon le principe de recherche du dénominateur commun. Dans les deux cas un calcul de la f-mesure est effectué, sur un jeu de données de test, afin d'évaluer les performances respectives des deux modèles. Les f-mesures sont sauvegardées en base, modèle par modèle (voir Figure 4).
Une fois en production, l'annotation des RCP se déroule en deux étapes. Dans un premier temps, un système *hybride* entre l'approche de similarité et l'approche *common denominator* se met en place autour de la table des f-mesures enregistrées durant l'apprentissage. Pour chaque concept d'intérêt c'est, parmi les deux disponibles, le modèle ayant eu la f-mesure la plus favorable sur le jeu de

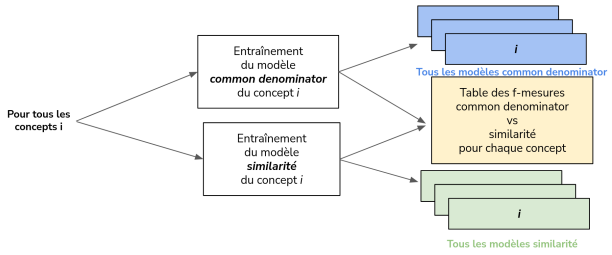


Figure 4: *Entraînement* - pour chaque concept, les modèles *similarité* et *common denominator* sont entraînés. Les valeurs respectives des f-mesures des deux modèles sont gardées en base de données afin de pouvoir décider ultérieurement lors de l'annotation lequel des deux modèles choisir concept par concept.

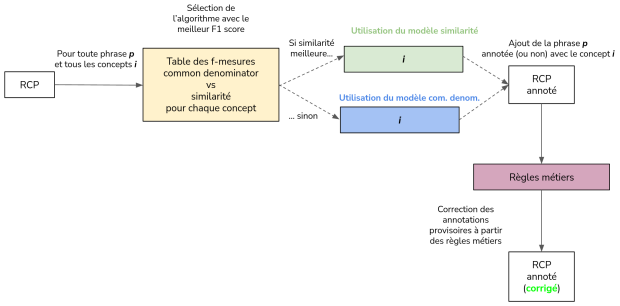


Figure 5: *Annotation* - pour chaque concept c'est, parmi les deux modèles statistiques entraînés, celui ayant eu la f-mesure la plus favorable qui est choisi pour annoter le RCP puis une correction à base de règles métiers déterministes est effectuée.

test qui est choisi. À l'issue de cette première itération, une version annotée *intermédiaire* du document est obtenue ; celle-ci sert alors de base à une seconde étape, le rattapage par règles métiers, qui permet d'amender le document et d'en faire une version *corrigée* (voir Figure 5), à valider par l'humain.

3.3 Intégration dans l'outil d'indexation semi-automatique

Le service de suggestion de concepts permettant l'indexation semi-automatique du RCP est disponible via une API qui prend en entrée le texte HTML de la rubrique et le nom de la rubrique concernée. En sortie, elle délivre le contenu de la rubrique découpée en phrases avec les indexations trouvées automatiquement selon la terminologie adéquate.
Cette API a été intégrée au sein de l'outil semi-automatique d'indexation qui présente le texte du RCP ainsi que les propositions de concepts faites par l'IA à valider par les pharmaciens. Pour le moment l'application ne propose que les concepts de contre-indication et d'effets indésirables (les indications seront intégrées dans les prochains mois).
Les propositions apparaissent sur les phrases encadrées de rouge et après passage de la souris. Et chaque proposition

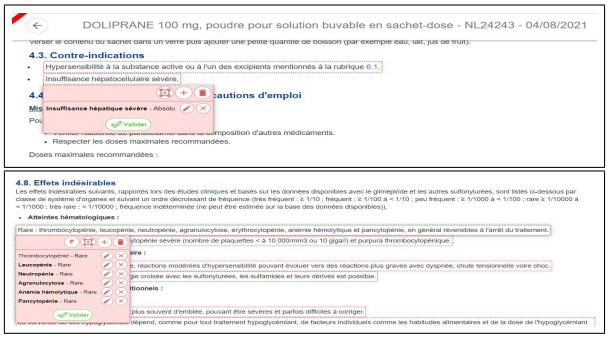


Figure 6: Copies d'écran de l'outil d'indexation semi-automatique (pour la rubrique *Contre-indications* en haut et la rubrique *Effets indésirables* en bas).

dispose d'une possibilité de suppression, modification et de validation (voir Figure 6).
L'indexeur pharmacien va analyser l'intégralité des deux rubriques contre-indication et effets indésirables et les propositions associées afin d'annoter l'ensemble des concepts nécessaires pour ce document.

3.4 La Boucle de *Feedback*

À chaque RCP analysé et son indexation complète validée, un document contenant le RCP indexé est enregistré dans les bases documentaires VIDAL.
Ces documents indexés sont récupérés tous les mois pour alimenter la base d'apprentissage et bénéficier de nouvelles entrées plus récentes. Les référentiels sont aussi mis à jour au même moment, et une nouvelle version de l'API est déployée.
Ceci nous permet de pouvoir apprendre en continu sur les nouvelles indexations.

3.5 Analyse des performances

Nous avons mesuré les performances de l'API d'indexation automatique.
La base d'évaluation contient 1 000 rubriques *Indications thérapeutiques*, 1 000 rubriques *Contre-indications* et 1 000 rubriques *Effets indésirables* indexées soit environ 10% du total de départ. Nous disposons comme *gold standard* des indexations manuelles correspondantes réalisées par les pharmaciens avec les concepts d'indication, de contre-indication et d'effets indésirables. Elle est utilisée pour évaluer la qualité de l'apprentissage.
L'API a été utilisée sur les textes des rubriques et nous avons comparé les résultats d'indexation automatique obtenus avec le *gold standard*. Les différentes mesures calculées sont :

- La précision : proportion de concepts pertinents parmi l'ensemble des concepts suggérés automatiquement par l'outil (niveau document)

$$\frac{|C_{manuel} \cap C_{auto}|}{|C_{auto}|} \quad (4)$$

avec C_{manuel} l'ensemble des concepts pertinents, C_{auto} celui des concepts obtenus via l'outil et $|.|$ l'opérateur donnant le cardinal d'un ensemble.

- Le rappel : proportion de concepts pertinents suggérés automatiquement parmi l'ensemble des concepts manuellement indexés

$$\frac{|C_{manuel} \cap C_{auto}|}{|C_{manuel}|} \tag{5}$$

- La f-mesure : moyenne harmonique de la précision à plat et du rappel

$$2\,\frac{\text{precision} \,.\, \text{rappel}}{\text{precision} + \text{rappel}} \tag{6}$$

Les mesures seront également détaillées par terminologie.

4 Résultats des performances de l'indexation automatique

	Nombre	Précision	Rappel	f-mesure
Indications	1 000	0.89	0.89	0.87
Contre-indications	1 000	0.92	0.87	0.88
Effets indésirables	1 000	0.81	0.86	0.83
TOTAL	1 000	0.87	0.87	0.86

Table 2: Résultats sur la base d'évaluation

Les évaluations sur les 1 000 rubriques indexées pour chaque rubrique *Indications*, *Contre-indications* et *Effets indésirables* a montré que l'IA obtient une performance moyenne totale de 86% de f-mesure (voir Tableau 2). La précision moyenne mesurée étant de 87% et le rappel moyen de 87%[3]. Le détail par rubrique, montre que les meilleurs résultats sont obtenus pour les *Contre-indications* avec 88% de f-mesure moyenne soit 1% de plus que pour les *Indications* et 5% de plus pour les *Effets indésirables*.

5 Discussion

Les résultats obtenus sont satisfaisants, l'IA permet d'extraire beaucoup de concepts pertinents et de faire gagner du temps aux utilisateurs dans la recherche de ces concepts.
L'avis des utilisateurs après utilisation en production et au quotidien du nouvel outil d'indexation semi-automatique est bon également.
Il reste encore toutefois une marge de progression pour cet outil. Afin de comprendre ces résultats, les erreurs ont été analysées afin de lister les causes et dégager des pistes d'amélioration. Ces analyses ont été menées

[3]Les valeurs de f-mesures du Tableau 2 ne correspondent pas directement à la moyenne harmonique des valeurs de précisions et de rappels indiquées, mais à la moyenne des f-mesures obtenues sur l'ensemble des rubriques. Numériquement, ceci implique que la f-mesure n'est pas nécessairement comprise entre les valeurs correspondantes de précision et de rappel moyens.

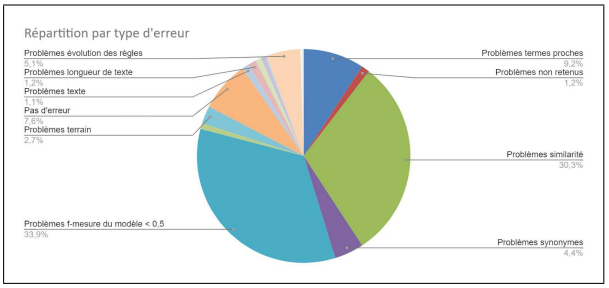


Figure 7: Répartition des erreurs sur les EI.

avec deux pharmaciens indexeurs experts sur l'indexation automatique et sur deux échantillons de rubriques :

- Un échantillon de 15 rubriques *Effets Indésirables*. Elles ont été choisies selon les critères suivants : une f-mesure globale d'environ 60%, une rubrique assez longue avec de nombreuses phrases et de nombreux concepts indexés. C'est ainsi 555 concepts en erreur (soit des concepts manquants soit des concepts erronés) qui ont été analysés.
- Et un échantillon de 17 rubriques *Contre-indications* choisies selon les mêmes critères. Ici c'est 171 concepts en erreur qui ont été analysés.

Concernant l'indexation des effets indésirables, plusieurs causes ont été identifiées (voir Figure 7) :

- problème de performance de certains modèles : leur performance propre ne dépasse pas 50% de f-mesure. Leurs sous-performances peuvent s'expliquer par différents facteurs liés à la base d'apprentissage. D'abord, il existe pour certains concepts peu d'exemples avec seulement une ou deux indexations, c'est le cas de concepts rares ou de concepts récemment ajoutés (voire pas d'indexation du tout pour les nouvelles notions). Ensuite certains termes sont ambigus, mêmes libellés ou synonymes alors que le sens est différent, ils ont tendance à sortir en même temps ou l'un à la place de l'autre (exemple : *mg* qui correspond à une unité de mesure et non au concept *magnésium*). Enfin il existe des difficultés pour les termes proches (*asthme/asthme sévère*). Les phrases et contextes étant souvent proches, il est difficile pour les modèles associés d'être performant
- problème de longueur de texte : la prédiction d'indexation se fait au niveau des phrases. Cette longueur de texte est parfois insuffisante pour retrouver la notion au complet ou son niveau de précision suffisant
- pas d'erreur : l'indexation manuelle est réalisée par différents indexeurs qui peuvent avoir des avis différents sur le choix du terme le plus pertinent
- les règles d'indexation peuvent aussi évoluer dans le temps : une indexation qui était considérée comme

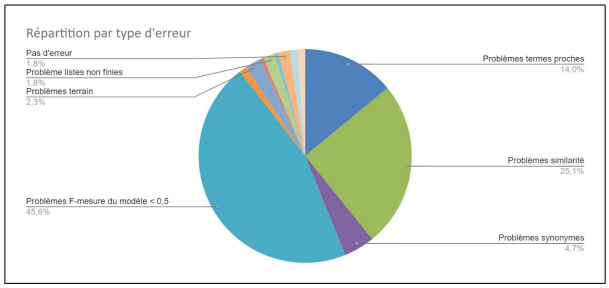


Figure 8: Répartition des erreurs sur les CI.

juste auparavant peut ne plus l'être après changement des règles éditoriales d'indexation par les pharmaciens

- problème de terrain : certains concepts ne sont pas à indexer alors qu'ils sont bien présents dans le texte. C'est le cas des terrains qui décrivent les facteurs favorisant la survenue d'un effet indésirable mais pas l'effet indésirable en lui même (exemple : *risque d'éruption cutanée chez les personnes porteuses du VIH*, seul le *risque d'éruption cutanée* est indexé par les indexeurs, pas le terrain *VIH*)
- problème de texte : la qualité du document, son orthographe, son format peuvent avoir une incidence sur l'indexation automatique
- problème de similarité : la méthode par similarité va avoir tendance à rapprocher des libellés proches (à une ou deux lettres près) ce qui va entraîner l'indexation d'un mauvais terme
- problème de synonymes : les terminologies utilisées ne sont pas exhaustives en matière de synonymes. Il peut manquer certains libellés qui vont empêcher l'indexation d'un concept particulier.
- listes non finies d'éléments : certains ne sont pas explicités dans le texte avec l'utilisation de notions telles que *autres*, *etc.* ou des listes entres parenthèses non finies
- il est également possible, en cas de nécessité, pour l'indexeur d'indexer une propriété clinique absente du RCP mais indiquée par d'autres sources d'information ou de ne pas retenir des termes présents dans le RCP (règles d'indexation particulières).

Concernant l'indexation des contre-indications, les mêmes types de causes peuvent être remontés avec des proportions un peu différentes (voir Figure 8) : La boucle de *Feedback* et l'ajout continu de règles nous permettent d'améliorer petit à petit les propositions automatiques, il sera donc intéressant à terme de faire une analyse de l'évolution des performances.
Pour la suite, plusieurs actions vont être menées. Côté gestion de terminologies, il est prévu d'enrichir les terminologies en synonymes. Côté indexeurs pharmaciens, il est prévu de réaligner les façons d'indexer pour ne plus avoir des soucis de différences d'indexation inter-indexeurs. Côté data science, un algorithme spécifique va être travaillé pour les problèmes de terrain.
Dans le futur, il est envisagé d'intégrer d'autres rubriques, toujours réalisées à la main actuellement par les pharmaciens (la rubrique *Composition qualitative et quantitative* avec l'indexation des substances, la rubrique des *Précautions d'emploi* avec l'indexation des précautions d'emploi).

6 Conclusion

L'objectif du projet était de développer des outils d'aide à l'indexation permettant de suggérer automatiquement des concepts à indexer aux pharmaciens pour l'indexation des propriétés thérapeutiques des médicaments dans la base de connaissances VIDAL. Trois méthodes ont été mises en œuvre : deux approches utilisant des algorithmes d'apprentissage automatique et une utilisant directement des règles métiers.
L'étude démontre qu'une automatisation d'une partie de l'indexation est possible à hauteur de 86% pour aider les indexeurs pharmaciens dans l'indexation quotidienne des RCP.

Remerciements

Remerciements aux équipes VIDAL : les pharmaciens, les développeurs et les *data scientists* qui ont participé aux développements de l'outil.

Références

[1] HAS[4]. Certification des logiciels des professionnels de santé. *Mis à jour le 03 mai 2021*.

[2] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. *O'Reilly Media, Inc.*.

[3] Bento Pereira, S. (2008). *Indexation multi-terminologique de concepts en Santé* (Doctoral dissertation, Rouen).

[4] Rubrichi, S., & Quaglini, S. (2012). Summary of Product Characteristics content extraction for a safe drugs usage. *Journal of Biomedical Informatics*, 45(2), 231-239.

[5] Hasan, M., Kotov, A., Carcone, A. I., Dong, M., Naar, S., & Hartlieb, K. B. (2016). A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. *Journal of biomedical informatics*, 62, 21-31.

[6] Blanco, A., Perez-de-Viñaspre, O., Pérez, A., & Casillas, A. (2020). Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity. *Computer methods and programs in biomedicine, 188*, 105264.

[4] `https://www.has-sante.fr`

[7] Remmer, S., Lamproudis, A., & Dalianis, H. (2021). Multi-label Diagnosis Classification of Swedish Discharge Summaries–ICD-10 Code Assignment Using KB-BERT. In *RANLP 2021: Recent Advances in Natural Language Processing, 1-3 Sept 2021, Varna, Bulgaria* (pp. 1158-1166). Association for Computational Linguistics.

[8] Amin, S., Neumann, G., Dunfield, K., Vechkaeva, A., Chapman, K. A., & Wixted, M. K. (2019, September). MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. In CLEF *(Working Notes)* (pp. 1-15).

[9] Biseda, B., Desai, G., Lin, H., & Philip, A. (2020). Prediction of ICD Codes with Clinical BERT Embeddings and Text Augmentation with Label Balancing using MIMIC-III. *arXiv preprint arXiv:2008.10492*.

[10] Zweigenbaum, P., & Lavergne, T. (2017). Détection de concepts et granularité de l'annotation (Concept detection and annotation granularity). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2-Articles courts* (pp. 226-233).

[11] Ratcliff, J.W., & Metzener, D. E. (1988). Pattern Matching: The Gestalt Approach. *Dr. Dobb's Journal*, 13(7), 46.

[12] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

[14] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[16] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[17] Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., & Sagot, B. (2020) CamemBERT: a Tasty French Language Model, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul, 2020, pages 7203–7219

[18] Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923*.

[19] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

[20] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

[21] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Session 2 : Apprentissage et Textes

Apprentissage multimodal pour le diagnostic de fautes sur données séquentielles non alignées et arbitrairement longues

Victor Pellegrain[1,2], Myriam Tami[2], Michel Batteux[1], Céline Hudelot[2]

[1] IRT SystemX
[2] Université Paris Saclay, CentraleSupélec, MICS

victor.pellegrain@irt-systemx.fr

Résumé

La complexité toujours grandissante des systèmes industriels amène de nouveaux verrous scientifiques pour les tâches liées à la maintenance prévisionnelle. Dans cet article nous présentons une revue des méthodes utilisées pour réaliser un diagnostic de fautes, et pointons leurs limites pour gérer des données multi-sources et hétérogènes, propres à l'industrie 4.0. Nous formalisons théoriquement ce nouveau cadre et proposons StreaMulT, une architecture permettant de gérer des séquences multimodales au fil de l'eau, non alignées et arbitrairement longues.

Mots-clés

Apprentissage multimodal, Diagnostic, Dépendances à long terme, Données non alignées, Séquences arbitrairement longues

Abstract

The industry 4.0 era brings more and more complexity to industrial systems, resulting in new challenges for predictive maintenance strategies. This work presents a review of the methods developed to perform fault diagnosis, along with their limitations to handle heterogeneous multi-sources data. We theoretically formalize this new applicative setting and propose StreaMulT, a Streaming Multimodal Transformer able to manage heterogeneous, unaligned and arbitrary long input sequences in a streaming fashion.

Keywords

Multimodal learning, Fault diagnosis, Long-term dependencies, Unaligned data, Arbitrary long sequences

1 Introduction

Un système industriel peut rencontrer des *défaillances*, se définissant par l'incapacité de ce système à réaliser au moins une de ses fonctions requises [17]. Ces défaillances peuvent mener à l'occurrence d'événements indésirables et redoutés, avec des conséquences plus ou moins importantes selon la criticité du système. Les occurrences de défaillances font souvent suite à la présence de *fautes*, à savoir une condition anormale du système caractérisée par la déviation d'une de ses caractéristiques par rapport à une valeur de référence acceptable. Afin d'éviter de telles occurrences, il est souvent nécessaire de considérer les techniques de diagnostic de fautes (DF) : leur détection puis leur isolation et identification, consistant à classifier le type de faute. Ainsi, les termes "isolation et identification" et "classification" sont interchangeables comme indiqué par [32]. Ces étapes sont essentielles dans une politique de maintenance préventive. De nombreuses approches de DF ont été utilisées, historiquement catégorisées entre les méthodes dites "basées modèle" et les méthodes dites "basées données".

Jusqu'à encore récemment, l'immense majorité des données acquises sur les systèmes était composée de séries temporelles, décrivant des grandeurs physiques locales comme la température, la pression, la vibration, etc. Aujourd'hui, l'ère de l'industrie 4.0 place l'interconnectivité et l'automatisation intelligente au centre du schéma de production industrielle. Cela se traduit essentiellement par l'intégration d'une multitude de capteurs connectés dans les machines ou systèmes industriels, dans le but de créer des systèmes de contrôle global appelés SCADA (Supervisory Control And Data Acquisition). Ces nombreux capteurs permettent ainsi l'acquisition d'une grande quantité de données issues d'une multitude de sources. En conséquence, ils fournissent plus d'information pour guider les modèles d'apprentissage et ainsi améliorer leurs performances. Cependant, ces flux de données multi-sources sont de nature hétérogène : séries temporelles issues de mesures de capteurs, textes issus de rapports d'intervention, images issues de prises de vue d'éléments du système. Ces différentes sources de données peuvent également présenter une hétérogénéité dans leur fréquence d'acquisition. En effet, si les grandeurs physiques sont mesurées régulièrement à une période de l'ordre de la seconde, des images sont acquises de façon plus éparse dans le temps, lorsque des rapports textuels d'intervention sont enregistrés encore plus rarement et à une fréquence sporadique. Cette double hétérogénéité entre les données multi-sources constitue un véritable verrou scientifique pour les modèles d'apprentissage usuels, qui sont généralement conçus pour exploiter la structure d'un type de données particulier, commune à tous les exemples d'entraînement. Ce verrou explique l'absence d'approches exploitant simultanément différents types de données dans la littérature de la surveillance et du diagnostic, pourtant es-

sentielles pour exploiter ces nouveaux jeux de données dans leur entièreté.
La prise en compte de données de natures hétérogènes est justement le problème auquel s'attaque la communauté de l'apprentissage multimodal [6, 13], et ce à différentes fins comme par exemple la fusion de modalités. En revanche, les méthodes actuelles de l'état de l'art se concentrent essentiellement sur des données statiques (une image et sa description textuelle par exemple), ou sur des données temporelles de taille fixe et généralement courte (clips vidéos de quelques secondes par exemple) et ne sont donc pas applicables en l'état sur des flux de données de capteurs industriels arbitrairement longs. Or, ce type de données est classique pour le cadre applicatif du DF. Notre contribution est donc triple, et dresse le plan de cet article :

— Dans la section 2 nous dressons un état de l'art des approches utilisées pour le DF, ainsi que pour l'apprentissage multimodal. Nous soulignons les forces et les limites de ces approches dans notre cadre applicatif.
— Nous formalisons ensuite dans la section 3 un nouveau cadre théorique modélisant la tâche de détection et classification simultanées de fautes à partir de données multimodales et arbitrairement longues. Ce cadre est essentiel pour exploiter de tels jeux de données propres à l'industrie 4.0, mais pas exclusivement.
— Enfin, dans la section 4 nous présentons StreaMulT, un modèle d'apprentissage profond basé sur une architecture Transformer [50], et capable de gérer des données multimodales, de fréquences d'acquisition hétérogènes, non alignées et arbitrairement longues. Nous validons également cette architecture expérimentalement.

2 Travaux antérieurs

2.1 Diagnostic

Un des premiers travaux à avoir listé et ordonné les différentes méthodes de DF est la série de trois articles de Venkatasubramanian et al. [51]. Cette revue, qui constitue le point de départ de notre étude, classifie les approches de DF selon la connaissance a priori que le concepteur a sur les différentes fautes pouvant survenir, ainsi que sur leur expression à travers les données acquises du système (symptômes de faute). Les stratégies utilisant cette connaissance a priori en représentant le système par un modèle physique sont dites "basées modèle", et différenciées entre qualitatives et quantitatives selon les représentations mathématiques utilisées ; les approches exploitant l'historique des données sont logiquement dites "basées données". Si les méthodes basées modèle fonctionnent bien lorsque le concepteur possède une bonne compréhension a priori des lois physiques régissant le système, elles sont difficilement exploitables dans le cas contraire. Ainsi, à un certain stade de complexité du système considéré, les interactions entre les composants sont difficilement modélisables. Dans ce cas, les méthodes basées données sont une alternative adéquate : le modèle utilisé apprend ces relations à partir de l'historique des données. Nous nous concentrerons sur les approches basées données dans cette étude, et plus précisément les approches de machine learning (ML).

Machine Learning. De nombreuses revues (présentées ci-après) ont répertorié les approches de ML appliquées au DF. Notre but n'est donc pas ici de donner une liste exhaustive des différents modèles utilisés, mais plutôt de dresser un état des lieux des différentes positions de ces articles, des challenges auxquels ils répondent, et leurs limites par rapport aux nouveaux défis de l'industrie 4.0.
Certaines revues de la littérature adoptent une position propre à un domaine applicatif industriel. C'est notamment le cas des articles [32, 72, 38], qui font part des méthodes de DF appliquées respectivement aux systèmes de traitements chimiques, aux roulements, ou aux systèmes d'air conditionné. Ces études motivent ainsi principalement leur démarche par les conséquences liées à l'apparition de fautes dans leurs domaines respectifs, comme la surconsommation d'électricité et des coûts économiques importants [38]. Les approches évoquées sont également présentées comme adaptées aux jeux de données propres à ces domaines : Zhang et al. [72] considèrent essentiellement des données de vibration et de courant de moteur, comme sur le jeu de données Paderborn [*] ; tandis que Rogers et al. [38] présentent des modèles utilisant essentiellement des données de température et d'humidité, et incitent la communauté à travailler sur des systèmes de thermostat. A l'opposé, d'autres travaux adoptent une posture plus méthodologique dans leur présentation de l'état de l'art des méthodes de DF [33]. Les plus récents [3, 37, 23] motivent leur démarche par l'apparition de nouveaux challenges pratiques liés à l'arrivée de l'industrie 4.0, comme notamment la capacité à gérer des quantités massives de données multi-sources en temps rapide.
Ces études présentent les approches de ML comme plus adaptées lorsque les profils de fautes sont complexes. Ainsi, Zhang et al. [72] font part de la limite des approches basées modèles à détecter précocement des fautes en raison de symptômes non traçables par ce type de modèles, ou à correctement démêler la présence de plusieurs fautes simultanément. Si certains travaux cités ne traitent que de la détection de fautes [28, 57], la majorité considère également la partie isolation et identification, bien que les auteurs de [3] utilisent parfois le terme de diagnostic pour évoquer la détection seule. En revanche, comme souligné par Reis et al. [37], en pratique deux méthodologies différentes coexistent : si la communauté de la maîtrise statistique des procédés (MSP) traite les tâches de détection et d'isolation et identification de fautes de manière séquentielle, la communauté du ML les traite parfois simultanément, sous la forme d'une classification en $C + 1$ classes, décomposées en une classe de fonctionnement normal, et C classes de fautes distinctes.
Comme présenté dans [23], les modèles de ML pour le DF

∗. Available online : https ://mb.uni-Paderborn.de/kat/forschung/datacenter/bearing-datacenter

sont généralement composés d'un module d'extraction de caractéristiques, fournissant à la suite du modèle des éléments pertinents depuis les données brutes, et d'un module de diagnostic. Certains modules d'extraction de caractéristiques se concentrent sur le domaine temporel pour capturer et caractériser l'information présente dans les séries temporelles fournies par les capteurs du système, par exemple via l'utilisation de réseaux de neurones [71]. Il est également d'usage d'avoir recours à des outils de traitement du signal, pour exploiter les caractéristiques du domaine fréquentiel des séries acquises : [26, 48] utilisent respectivement des transformées de Fourier et de Laplace. Enfin, d'autres approches travaillent dans le domaine temps-fréquence, par exemple via l'utilisation de transformée en ondelettes [73]. Ce choix de module d'extraction de caractéristiques est fortement influencé par l'application et le type de données d'entrée, et donc par la connaissance a priori du concepteur.

Le module de diagnostic est ensuite composé au choix :

- soit d'un premier sous-module de détection permettant de réaliser la surveillance, suivi par un second sous-module de classification effectuant ensuite la tâche d'isolation et identification de fautes.
- soit d'un unique modèle de classification assurant simultanément la détection et l'isolation et identification de fautes.

Dans le cas où le jeu de données est étiqueté, le module unique de classification recevant ces caractéristiques est libre d'utiliser le modèle de ML de son choix : SVM [21], forêt aléatoire [62], réseau de neurones peu profond [18], réseau de neurones récurrents [61], etc. Cette approche de détection et classification simultanées est cependant parfois critiquée [37] car pouvant mener à des problèmes pratiques :

- Les occurrences des fautes pouvant conduire à des défaillances et des événements redoutés sont souvent assez rares dans les jeux de données réelles. Cela amène un problème de déséquilibre entre les classes, amplifié si on considère une classification multi-classes.
- Pour ce type de tâche, une erreur de classification du modèle aura le même poids durant la phase d'apprentissage, quelle que soit cette mauvaise classification. Cependant selon la criticité du système, la performance de la détection d'une faute peut être bien plus importante que sa classification.

Pour pallier ces difficultés, une première tâche de surveillance peut être effectuée via des méthodes de détection d'anomalies [10]. De façon semblable au schéma utilisé dans les travaux issus de la communauté de la MSP, ces méthodes semi-supervisées modélisent le comportement normal du système dans la phase d'apprentissage, et classifient comme faute les points effectuant une déviation significative de ce modèle lors de la phase d'inférence. Ces modèles sont plus robustes au déséquilibre présent dans les jeux de données, et pourront être suivis par un modèle de classification pour réaliser l'isolation et identification de la faute.

Enfin, si les conditions de fonctionnement normal du système ne sont pas connues, il est également possible de concevoir le modèle de diagnostic en utilisant des approches de clustering. C'est ce que font Diaz-Rozo et al. dans [9], comparant les performances des algorithmes de mélanges de lois gaussiennes, de clustering hiérarchique agglomératif, et K-means.

Deep Learning. De la même façon que les méthodes basées modèle, les approches de ML classiques se retrouvent aujourd'hui assez limitées face aux données plus complexes de l'industrie 4.0. Ainsi, comme décrit par [72, 23, 35], les méthodes d'extraction de caractéristiques basées sur une connaissance des données acquises peuvent ne plus suffire à effectuer un diagnostic correct. Pour répondre à ces challenges, des modèles de Deep Learning (DL) sont ainsi utilisés, intégrant une phase d'apprentissage de représentation des données dans les premières couches, afin d'extraire automatiquement les caractéristiques les plus saillantes pour une tâche subsidiaire [7, 22], ici le diagnostic. Ainsi, de nombreux travaux ont montré la supériorité des modèles de DL pour le DF, utilisant aussi bien comme algorithme d'apprentissage de représentation des modèles discriminatifs (réseaux convolutifs [59, 56, 34], réseaux récurrents profonds [1, 12], Transformers [58], etc.) que des modèles génératifs (modèles probabilistes graphiques [65, 24], Autoencoders [20, 46, 39], GANs [60, 25]).

Diagnostic à partir de données multimodales. La complexité des données acquises s'intensifie encore de nos jours, avec des capteurs mettant à disposition des données multimodales. Si certains travaux s'attaquent au DF à partir d'images thermiques [8, 19, 47], de rayons X [36], de photographies [55, 54], ou de rapports textuels de maintenance [52, 42], l'application à des données multimodales (de natures hétérogènes) en est à son balbutiement. La majorité des travaux relatifs à la tâche de DF et mentionnant des données "multimodales" fait en fait référence à des modes de fonctionnement différents de l'appareil (comme un climatiseur fonctionnant en mode économique) [43]. Pour Zhou et al. [74], le terme "multimodal" fait référence aux différentes dérivées de leurs séries numériques. A notre connaissance, seuls deux travaux considèrent des données multimodales (au sens "hétérogènes") dans une optique de maintenance. Mian et al. [29] fusionnent des données numériques de signaux vibratoires avec des images thermiques du système afin d'améliorer les performances de classification. Ils utilisent une approche de ML classique, réalisant l'extraction de caractéristiques grâce à une transformée de Hilbert, et utilisant la concaténation comme technique de fusion. Malheureusement, leur jeu de données n'est pas mis à disposition de la communauté, empêchant de se comparer à leur approche. Yang et al. [63] appliquent un modèle multimodal à une tâche connexe de la notre : le pronostic de défaillances. Leur objectif est ainsi de prévoir le temps restant avant l'occurrence d'une défaillance du système. En ce sens, la tâche finale est une régression, mais leur cadre d'étude peut se transposer à celui que nous considérons. Leur approche traite trois modalités (données numériques de capteurs, images et textes) sous la forme de trois

branches distinctes, apprenant une représentation propre à chaque modalité (à l'aide de couches convolutives pour les images et le texte, et d'une couche linéaire pour la modalité numérique). Ils adoptent une approche de fusion tardive, par concaténation de chaque sortie de branche, avant d'appliquer une dernière couche de régression. Cet article est le travail s'appuyant sur un jeu de données public le plus proche de notre problème considéré. Cependant, ce jeu de données comporte quelques points négatifs pour notre cadre. Premièrement, les images considérées ne sont en réalité que des graphiques correspondant aux courbes acquises dans la modalité numérique. Elles ne représentent donc pas réellement des images issues d'une prise de vue du système, qui ont une structure bien différente à l'échelle locale, et n'apportent de surcroît pas d'information supplémentaire sur l'état du système. Deuxièmement, le jeu de données est simulé. Cela implique un manque de richesse et de diversité pour la modalité textuelle. On retrouve beaucoup de fois les mêmes phrases au mot près dans les exemples et on perd ainsi une partie de la nature non-structurée du texte brut.

Les mécanismes de fusion de ces deux seules contributions existantes sur cette application sont relativement simples (concaténation). Nous passons en revue dans la section suivante les enjeux et avancées de l'apprentissage multimodal, afin de tirer parti des meilleures architectures existantes pour notre problème.

Par ailleurs, devant l'absence de jeu public de données multimodales et réelles dans les communautés liées aux systèmes industriels, nous nous sommes tournés vers des jeux de données issus d'autres domaines (voir section 5). Par conséquent, nous invitons la communauté industrielle à mettre à disposition un jeu de données représentatif de ce problème afin d'encourager à la réalisation de futurs travaux sur cette tâche à forts enjeux.

2.2 Apprentissage multimodal

L'accès à différentes sources d'observation d'un même phénomène nous donne de l'information complémentaire et/ou supplémentaire (parmi d'autres types de relation entre modalités [11]). Ce gain d'information est en général bénéfique pour les performances du modèle utilisé pour une tâche considérée [16]. Cependant, cette hétérogénéité de nature entre les différentes sources de données se traduit par des espaces de définition et des propriétés hétérogènes : données structurées et continues (séries temporelles de grandeurs physiques) ou non structurées, discrètes et parcimonieuses (one-hot encodings de texte libre) par exemple. Ainsi, un même concept aura des représentations vectorielles également très différentes dans chaque espace propre à une modalité, ce qui implique une difficulté à mesurer une similarité entre des points de modalités différentes. Cette difficulté est définie sous le nom de fossé d'hétérogénéité [13]. De ce verrou scientifique découlent plusieurs enjeux comme la transduction ou l'alignement entre modalités, tous décrits dans différentes revues [13, 6]. Le challenge nous intéressant ici est celui de la fusion entre modalités, souvent lié à l'apprentissage de représentation jointe multimodale. Cette tâche a pour but de projeter des représentations unimodales dans un sous-espace sémantique joint, afin de combler ce fossé d'hétérogénéité.

Pour ce faire, les premières approches ont adopté des stratégies de fusion précoce en concaténant [30] ou multipliant [69] les caractéristiques de chacune des modalités ; ou de fusion tardive, combinant les décisions de modèles unimodaux par système de vote [53].

A l'opposé de ces méthodes agnostiques à un type de modèle, certaines architectures de DL modélisent les interactions inter-modalités et intra-modalité afin d'apprendre les représentations jointes les plus pertinentes. Des approches génératives utilisent par exemple des variantes multimodales de machines de Boltzmann [44], ou des réseaux de croyance profonds [45], apprenant une distribution jointe sur les deux modalités d'entrée. Ces architectures entraînables de façon non supervisée peuvent ainsi générer des modalités à partir d'une autre, et sont donc plus robustes aux modalités manquantes. Cependant, le coût élevé d'approximation d'inférence des algorithmes est souvent rédhibitoire à leurs usages.

L'autre grande famille de modèles génératifs utilisés est celle des autoencoders. Ils visent à apprendre une représentation condensée, qui capture les éléments essentiels à la reconstruction de l'objet initial. Des adaptations multimodales ont été développées [31], dans lesquelles la couche intermédiaire commune prend en entrée les deux modalités, et essaye de les reconstruire à partir de ce vecteur intermédiaire commun. Cependant, ces architectures ne se basant que sur la reconstruction des données d'entrée, les représentations apprises sont agnostiques à une tâche précise et donc génériques. Cela peut impliquer une baisse de performances si l'apprentissage n'est pas guidé par des contraintes supplémentaires [41].

Par ailleurs, des modèles discriminatifs utilisant des mécanismes d'attention [5] ont également été utilisés à la fois dans un but d'amélioration de performances mais aussi de gain d'interprétabilité. En effet, à une échelle intra-modalité, le mécanisme d'attention est utilisé pour sélectionner les composantes les plus pertinentes de chaque modalité, dépendant du contexte donné par les autres modalités, comme sur une tâche de Visual Question Answering [64]. A une échelle inter-modalités, ce type de mécanismes permet de pondérer la contribution de chacune des modalités dans la prise de décision finale [27].

Plus récemment, ce mécanisme d'attention cross-modal a été étendu aux architectures de Transformers, avec pour but l'apprentissage de représentations contextuelles [70, 49, 2]. Ces approches ont également l'avantage de pouvoir gérer des modalités non alignées, comme le mettent en avant Tsai et al. [49]. Effectivement, les données que nous considérons dans le cadre du DF ne sont pas alignées temporellement. Par exemple, une faute apparaissant à un temps spécifique t pourra être corrélée à la fois à des points récents de données de capteurs (de l'ordre des secondes précédentes), ainsi qu'à des points de données bien plus antérieurs pour la modalité textuelle (rapport de maintenance de la semaine précédente). Les approches multimodales classiques utilisées pour des données séquentielles (basées sur des ré-

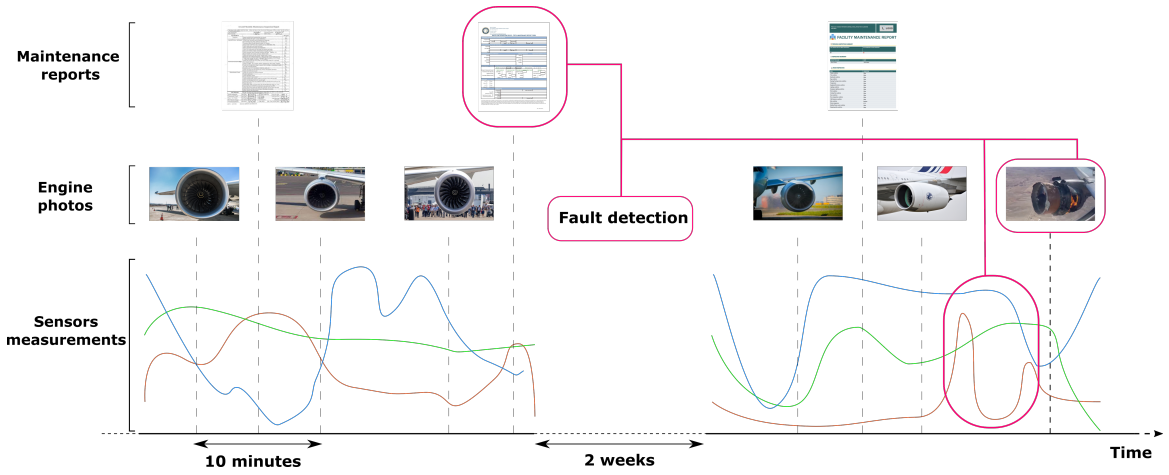


FIGURE 1 – Exemple d'un système industriel produisant des données hétérogènes, non alignées, arbitrairement longues, dans le cadre d'une tâche de diagnostic de faute.

seaux récurrents par exemple), ne gèrent pas ce problème de non-alignement [67, 68]. L'approche utilisée par Tsai et al. [49] s'attaque à ce challenge grâce à son module de représentation cross-modale, et plus précisément au produit matriciel requêtes-clés, modélisant toutes les corrélations entre deux séquences de modalités différentes. L'architecture StreaMulT que nous proposons en section 4 s'inspire de ce modèle. Elle présente l'avantage de s'adapter au cadre de données multimodales arbitrairement longues, que nous introduisons dans la prochaine section.

3 Cadre théorique

Dans cette section, nous définissons le problème auquel s'attaque notre méthode. Nous nous plaçons dans le contexte applicatif du DF et considérons des flux de données hétérogènes à la fois par leur nature (séries numériques, texte brut, images, son, etc.) et par leurs fréquences d'acquisition. Nous supposons que ces différents flux sont a priori non alignés et que l'historique des données peut être arbitrairement long. Enfin, nous considérons le cas où un système industriel peut ne jamais s'arrêter de fonctionner et nécessite donc que les séquences d'entrée soient traitées au fil de l'eau (en "streaming"). Cet exemple est illustré dans la Fig. 1.

Pour des besoins de clarté, et sans perte de généralité, nous considérons trois modalités notées α, β, γ. Soient, trois séries temporelles $(X_\alpha, X_\beta, X_\gamma)$ de différentes modalités. Chaque série temporelle est indicée par le temps, possède ses propres temps d'acquisition et son propre espace de définition. Ainsi, pour la modalité α,

$$X_\alpha := (X_\alpha(t))_{t\in\mathcal{T}_\alpha} \text{ et } \forall t \in \mathcal{T}_\alpha,\ \ X_\alpha(t) \in \mathbb{R}^{d_\alpha}$$

où $\mathcal{T}_\alpha$ et d_α sont respectivement les ensembles dénombrables contenant les temps d'acquisition de la modalité α et sa dimension de caractéristiques associée.

Notre objectif est de réaliser une tâche de prédiction au cours du temps. Soit $\mathcal{X}$ l'ensemble d'entrée défini par : $\mathcal{X} := \left\{ [X(s)]_{s\leq t}, t \in \mathbb{R} \right\}$, où $[X(s)]_{s\leq t} = \bigcup_{j\in\{\alpha,\beta,\gamma\}} \{X_j(s), s \leq t\}$ sont les données de toutes modalités acquises avant le pas de temps t. Formellement, étant donné un espace de labels $\mathcal{Y}$ commun à toutes les modalités, notre but est de trouver la fonction de prédiction optimale $h^* \colon \mathcal{X} \mapsto \mathcal{Y}$ minimisant une fonction de perte L sur un espace d'hypothèse $\mathcal{H}$:

$$h^* = \arg\min_{h\in\mathcal{H}} L(h)$$

avec $\quad L(h) := \frac{1}{|\mathcal{T}_y|} \sum_{t\in\mathcal{T}_y} l\left(h\left([X(s)]_{s\leq t}\right), y_t\right)$

où l est une fonction de score mesurant l'erreur entre la prédiction de h au temps t et la vérité terrain y_t, et où $\mathcal{T}_y$ représente l'ensemble dénombrable des temps d'acquisition des labels, dont la définition dépend de la tâche. Par exemple, pour une tâche de DF, $\mathcal{T}_y := \mathcal{T}_\alpha \cup \mathcal{T}_\beta \cup \mathcal{T}_\gamma$ car l'objectif est de détecter et classifier une faute à toute nouvelle acquisition de données.

A notre connaissance, ce cadre de données multimodales, possédant des temps d'acquisition différents, potentiellement non alignées et arbitrairement longues n'a jamais été introduit et traité auparavant.

4 Modèle proposé

Nous proposons StreaMulT (Streaming Multimodal Transformer), un modèle prenant avantage des architectures de Multimodal Transformer [49] et Emformer [40]. La longueur arbitraire des séquences d'entrée est contrôlée par un mécanisme de traitement par blocs (voir Fig. 2), et la multimodalité est gérée par des modules de Transformers cross-modaux fonctionnant en streaming (voir Fig. 3).

Transformer cross-modal. Le module d'attention cross-modale, défini dans [49], traite le fossé d'hétérogénéité des données d'entrée [14] en exprimant une modalité cible α avec les caractéristiques brutes d'une modalité source β. Formellement, en considérant nos séquences d'entrée X_α et X_β, l'attention cross-modale exprimant X_α à partir de

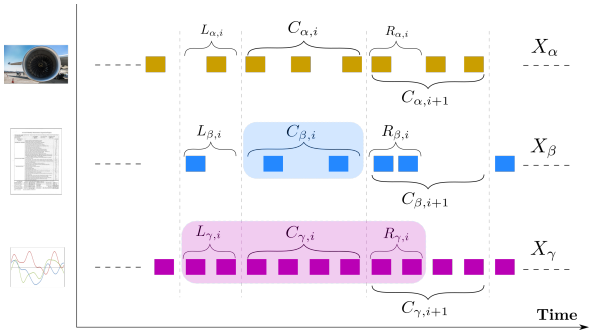


FIGURE 2 – Mécanisme de traitement par blocs pour l'apprentissage multimodal. Pour la modalité α : $X_\alpha, C_{\alpha,i}, L_{\alpha,i}$ et $R_{\alpha,i}$ correspondent respectivement à la séquence d'entrée entière, au i-ème segment central initial et aux contextes gauche et droit associé à ce segment central, afin de former le i-ème segment contextuel. La zone bleue représente un segment central pour la modalité β et la zone rose représente un segment contextuel pour la modalité γ.

X_β , notée $X_{\beta\to\alpha}$ se calcule comme suit :

$$X_{\beta\to\alpha} := Attn(Q_\alpha, K_\beta, V_\beta) = \text{softmax}\left(\frac{Q_\alpha K_\beta^T}{\sqrt{d_k}}\right) V_\beta$$

$$= \text{softmax}\left(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^T X_\beta^T}{\sqrt{d_k}}\right) X_\beta W_{V_\beta}$$

avec Q_α la matrice de requêtes de la modalité α, K_β, V_β les matrices de clés et valeurs de la modalité β et $W_{Q_\alpha}, W_{K_\beta}, W_{V_\beta}$ des poids appris. Cette "scaled dot-product attention", inspirée par le mécanisme de self-attention du Transformer original [50], modélise les dépendances à long terme à travers son produit matriciel et gère ainsi des données non alignées de la même façon [49].

Mécanisme de traitement par blocs. Cependant, la longueur arbitraire des séquences d'entrée de notre cadre implique deux verrous majeurs. Premièrement, l'entraînement du modèle est insoluble en raison de la complexité quadratique de l'architecture Transformer, et deuxièmement l'inférence ne peut s'effectuer au fil de l'eau, l'architecture Transformer ayant besoin de la séquence d'entrée complète pour effectuer le produit matriciel. Pour résoudre ces problèmes, nous adoptons un mécanisme de traitement par blocs, découpant les séquences d'entrée en plus petits segments disjoints $(C_i)_{i\geq 0}$ (voir Fig. 2). Nous calculons ensuite l'attention sur ces segments et réduisons ainsi la complexité du modèle durant le calcul de l'attention crossmodale. Pour éviter les effets de bords, nous ajoutons à ces segments disjoints des blocs de contextes gauche et droit, concaténés aux segments initiaux afin de former des segments contextuels $X_i = [L_i : C_i : R_i]$. Enfin, pour véhiculer l'information entre les segments, nous utilisons une banque de mémoire, à la manière d'Emformer [40].

Architecture globale. Notre architecture globale *end-to-end* combine ainsi les avantages des deux architectures précédentes [49, 40] et est illustrée dans la Fig. 4.

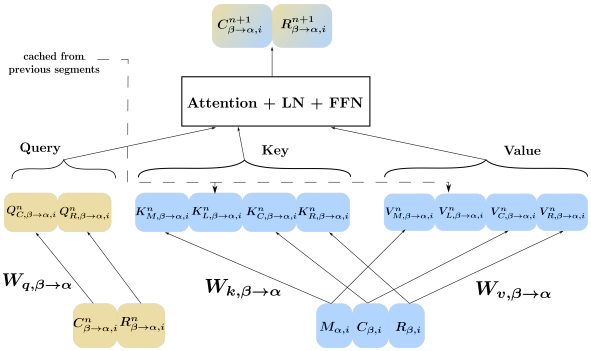


FIGURE 3 – Module de Streaming Crossmodal Transformer

Nous décrivons ici le traitement de la modalité α.
X_α passe d'abord par une couche convolutive 1D afin de modéliser une première structure locale temporelle, et de projeter l'ensemble des modalités dans un espace commun de dimension d. Les bornes des segments sont ensuite délimitées, et en suivant l'approche de traitement par blocs, tous les segments contextuels $X_{\alpha,i}$ sont traités en parallèle. Ils passent premièrement à travers un module Emformer unimodal afin d'initialiser la banque de mémoire propre à cette modalité. Ensuite, chaque paire de modalités source/cible (β / α) est traitée par son propre module Streaming Crossmodal Transformer (SCT), dont le fonctionnement est illustré dans la Fig. 3. Plus spécifiquement, chaque segment de la modalité cible $X_{\alpha,i}$ est exprimé en utilisant le segment temporel correspondant de la modalité source $X_{\beta,i}$ ainsi que la banque de mémoire de cette même modalité source $M_{\beta,i}$, contenant de l'information compressée des segments précédents.
Ainsi, pour chaque couche n du module SCT $\beta \to \alpha$:

$$\left[\hat{C}^n_{\alpha,i}, \hat{R}^n_{\alpha,i}\right] = \text{LN}([C^n_{\alpha,i}, R^n_{\alpha,i}])$$
$$\left[\hat{C}^n_{\beta,i}, \hat{R}^n_{\beta,i}\right] = \text{LN}([C^n_{\beta,i}, R^n_{\beta,i}])$$
$$K^n_{\beta,i} = [K^n_{M,\beta\to\alpha,i}, K^n_{L,\beta\to\alpha,i}, K^n_{C,\beta\to\alpha,i}, K^n_{R,\beta\to\alpha,i}]$$
$$V^n_{\beta,i} = [V^n_{M,\beta\to\alpha,i}, V^n_{L,\beta\to\alpha,i}, V^n_{C,\beta\to\alpha,i}, V^n_{R,\beta\to\alpha,i}]$$
$$Z^n_{C,\beta\to\alpha,i} = Attn(Q^n_{C,\beta\to\alpha,i}, K^n_{\beta,i}, V^n_{\beta,i}) + C^n_{\beta\to\alpha,i}$$
$$Z^n_{R,\beta\to\alpha,i} = Attn(Q^n_{R,\beta\to\alpha,i}, K^n_{\beta,i}, V^n_{\beta,i}) + R^n_{\beta\to\alpha,i}$$
$$\left[\hat{C}^{n+1}_{\alpha,i}, \hat{R}^{n+1}_{\alpha,i}\right] = \text{FFN}(\text{LN}([Z^n_{C,\beta\to\alpha,i}, Z^n_{R,\beta\to\alpha,i}]))$$
$$[C^{n+1}_{\alpha,i}, R^{n+1}_{\alpha,i}] = \text{LN}(\left[\hat{C}^{n+1}_{\alpha,i}, \hat{R}^{n+1}_{\alpha,i}\right] + [Z^n_{C,\beta\to\alpha,i}, Z^n_{R,\beta\to\alpha,i}])$$

où,

$$\left[K^n_{M,\beta\to\alpha,i}, K^n_{C,\beta\to\alpha,i}, K^n_{R,\beta\to\alpha,i}\right] = W_{k,\beta\to\alpha}\left[M_{\beta,i}, \hat{C}^n_{\beta,i}, \hat{R}^n_{\beta,i}\right]$$
$$\left[V^n_{M,\beta\to\alpha,i}, V^n_{C,\beta\to\alpha,i}, V^n_{R,\beta\to\alpha,i}\right] = W_{v,\beta\to\alpha}\left[M_{\beta,i}, \hat{C}^n_{\beta,i}, \hat{R}^n_{\beta,i}\right]$$
$$\left[Q^n_{C,\beta\to\alpha,i}, Q^n_{R,\beta\to\alpha,i}\right] = W_{q,\beta\to\alpha}\left[C^n_{\beta\to\alpha,i}, R^n_{\beta\to\alpha,i}\right]$$

et $\left(K^n_{L,\beta\to\alpha,i}, V^n_{L,\beta\to\alpha,i}\right)$ sont les copies des clés et valeurs (mises en cache) correspondant aux segments précédents, dont la taille est fixée par la taille du contexte gauche. LN, FFN, Attn correspondent respectivement à des

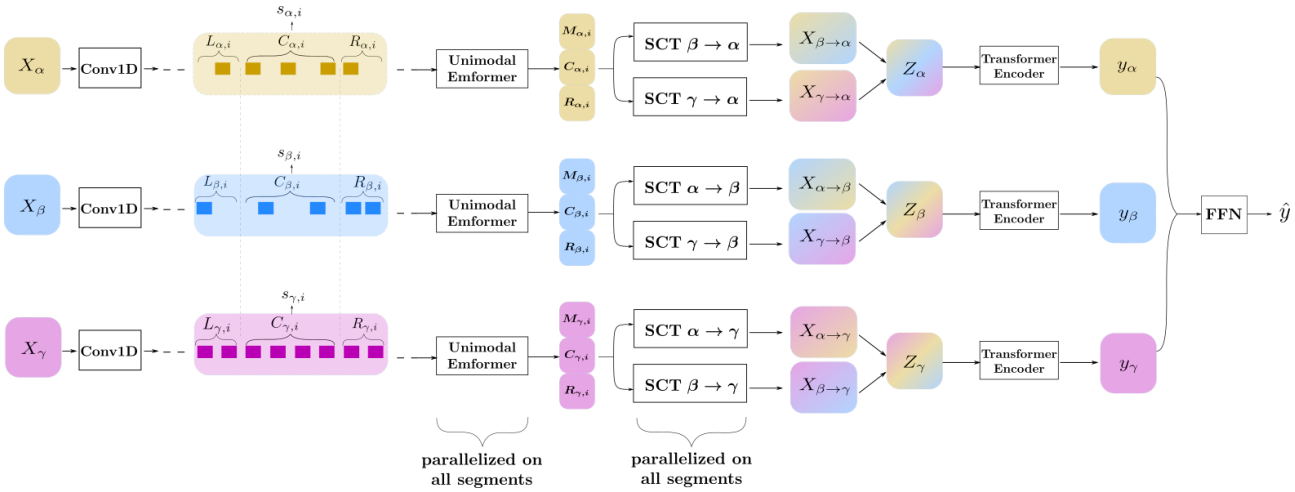


FIGURE 4 – Architecture globale du Streaming Multimodal Transformer. SCT signifie Streaming Crossmodal Transformer. Les couleurs différentes représentent les natures hétérogènes des différentes modalités, et les dégradés expriment les caractéristiques cross-modales.

couches "Layer Normalization", "Feed-Forward" et "scaled dot-product Attention".

Après la dernière couche N, les représentations des contextes droits $(R^N_{\beta\to\alpha,i})_{i>0}$ sont défaussées. Les segments $(C^N_{\beta\to\alpha,i})_{i>0}$ sont concaténés pour former la représentation cross-modale finale $X_{\beta\to\alpha}$. Les représentations cross-modales correspondant à la même modalité cible α sont alors concaténées selon la dimension des caractéristiques dans un vecteur $Z_\alpha := \begin{pmatrix} X_{\beta\to\alpha} \\ X_{\gamma\to\alpha} \end{pmatrix}$, qui est à son tour traité par un encoder de Transformer classique afin d'exploiter la nature séquentielle des données. La sortie y_α de cette couche est finalement concaténée avec celles des autres modalités, et une couche linéaire résulte en la prédiction $\hat{y}_t$.

5 Expériences et résultats

Comme abordé en section 2, à notre connaissance il n'existe pas de jeu de données public représentatif de notre cadre. Nous décidons donc de conduire nos expériences sur le jeu de données CMU-MOSEI [4], afin d'évaluer empiriquement notre architecture StreaMulT et la comparer avec des approches existantes sur un cadre proche de celui recherché, à savoir des données séquentielles réelles, multimodales et non alignées. Le jeu de données CMU-MOSEI est composé de 23,454 clips vidéos témoignant de l'opinion de plus de 1000 orateurs sur plus de 250 sujets divers. De ces clips sont extraits des caractéristiques audio, vidéo et textuelles, utilisées pour créer une version brute non-alignée du jeu de données ainsi qu'une version réalignée entre les trois modalités. La longueur des phrases alignées est fixée à 50 tokens, utilisant un éventuel padding.

La tâche associée à ce jeu de données est une analyse de sentiments sur ces clips vidéos, étiquetés par des annotateurs humains avec un score de sentiment allant de -3 (sentiment très négatif) à 3 (sentiment très positif). Comme dans les travaux précédents [49], nous évaluons les performances de notre modèle suivant 5 métriques : l'accuracy d'une classification sur 7 classes, l'accuracy binaire (sentiment positif ou négatif), le score F1, l'erreur moyenne absolue et la corrélation entre les prédictions du modèle et les étiquettes.

Pour souligner la valeur ajoutée de StreaMulT, nous conduisons nos expériences dans deux cadres différents. (1) Nous considérons d'abord les clips vidéo comme nos séquences d'entrée complètes, et observons les performances de StreaMulT lorsque nous divisons ces clips en segments plus courts. Pour pouvoir définir les mêmes bornes de segments entre les modalités, nous réalisons ces expériences sur la version alignée de CMU-MOSEI. Nous décidons de diviser chaque séquence en 5 segments de 10 pas de temps. (2) Nous concaténons ensuite tous les clips vidéos d'un même orateur et considérons cette suite de clips comme notre séquence d'entrée longue, afin de construire artificiellement des séries arbitrairement longues. Dans cette configuration, nous choisissons comme segments les clips initiaux et pouvons donc utiliser la version non-alignée.

StreaMulT n'a pas pour objectif de battre les architectures les plus performantes sur la tâche d'analyse de sentiments multimodal [66, 15], sa valeur ajoutée étant sa capacité à gérer des séquences multimodales non alignées **arbitrairement longues**. Ainsi nous ne reportons ici que les métriques concernant le Transformer multimodal données dans [49], pour une comparaison équitable avec une architecture similaire. Nous avons également utilisé le code officiel de cette approche mis à disposition [†], en gardant les valeurs d'hyperparamètres données dans [49]. Nous ne sommes cependant pas parvenus à reproduire exactement les résultats communiqués dans l'article, et nous présentons donc ceux que nous avons obtenus. Toutes les métriques sont moyennées sur 5 trajectoires d'entraînement.

†. https ://github.com/yaohungt/Multimodal-Transformer

Métrique	Acc_7^h	Acc_2^h	$F1^h$	MAE^l	$Corr^h$
MulT [49]	**51.8**	**82.5**	**82.3**	**0.580**	**0.703**
MulT‡(1)	49.32	81.05	81.42*	0.615	0.666
StreaMulT‡ (1)	50, 08*	81.08*	81.01	0.608*	0.671*
MulT‡(2)	-	-	-	-	-
StreaMulT‡ (2)	49.25	80.55	80.84	0.621	0.665

TABLE 1 – Résultats sur CMU-MOSEI. Les meilleurs résultats sont en gras. ‡ : notre implémentation ou reproduction depuis le code officiel, avec les valeurs d'hyperparamètres fournies. * : meilleur score parmi la catégorie ‡. (1) et (2) font références aux deux environnements d'expérimentation définis plus haut.

Le tableau 1 montre que notre architecture reproduit les résultats du Transformer Multimodal dans l'environnement (1) (fait même un peu mieux sur 4 des 5 métriques), ce qui démontre la capacité de la banque de mémoire à véhiculer l'information pertinente à la classification entre les différents segments de taille 10, tandis que MulT a accès à la séquence de taille 50 dans son intégralité. Pour l'environnement (2), les performances baissent légèrement. Cependant, ce cadre d'évaluation artificiel permet de mettre en évidence la plus-value de notre architecture qui est sa capacité à traiter au fil de l'eau des données arbitrairement longues. A l'inverse, on observe que MulT rencontre une erreur de mémoire et n'est donc pas capable de gérer ces longues séquences en raison de sa complexité.

Pour valider qualitativement notre modèle, nous affichons la carte de chaleur des différents poids d'attention du modèle dans la Fig. 5. Cette carte représente les différents poids d'attention du module SCT associé aux modalités images (requêtes) / texte (clés), pour une séquence d'entrée de taille 50.

Cette carte de chaleur nous rappelle premièrement que les séquences de langage sont non alignées entre les modalités : à l'inverse d'une diagonale monotone, nous observons différentes activations sur des lignes verticales, correspondant à certains embeddings textuels corrélés à plusieurs images. Si certains non-alignements restent dans le champ d'un même segment, comme représenté dans le quatrième segment par le rectangle vert, l'accès à la banque de mémoire permet au modèle d'accéder à des données à plus longue portée, comme illustré dans le troisième segment par les deux rectangles jaunes. Celui de droite indique des dépendances non alignées au sein du troisième segment, tandis que celui de gauche met en lumière l'activation de certaines images de ce segment par des caractéristiques textuelles venant du passé, sauvegardées dans la banque de mémoire.

Ces comportements différents témoignent de la capacité de l'architecture StreaMulT à adapter sa stratégie selon le contexte, accédant à des données non-alignées du passé via la banque de mémoire lorsque nécessaire.

6 Conclusion

Notre état des lieux des méthodes utilisées pour le diagnostic de fautes dresse leurs limites pour répondre aux défis posés par l'industrie 4.0 : gérer des séquences hétérogènes, non alignées et arbitrairement longues. Notre architecture StreaMulT combine l'attention cross-modale et le mécanisme de traitement par blocs parallélisé afin de traiter ces séquences multimodales en streaming. Les expériences conduites sur le jeu de données CMU-MOSEI ont montré des résultats prometteurs : une conservation des performances couplée à une capacité à gérer des séquences arbitrairement longues durant la phase d'entraînement, et à traiter ces flux de données en streaming à l'inférence.

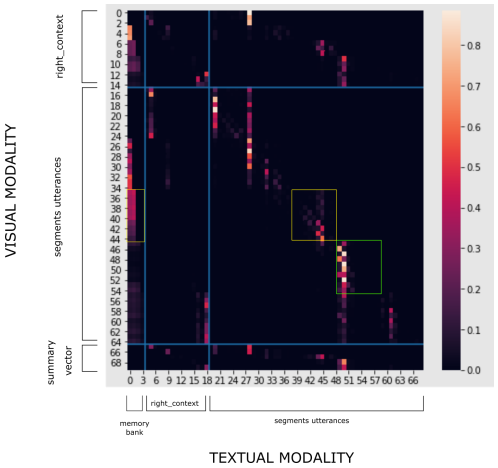


FIGURE 5 – Carte de chaleur des poids d'attention de StreaMulT pour le module cross-modal images (requêtes) / texte (clés). La séquence de taille 50 est découpée en segments de taille 10, avec des contextes gauche et droit de tailles respectives 10 et 3. Les lignes bleues délimitent les interactions entre les différents composants des séquences (contextes, segments, mémoire).

Remerciements

Victor Pellegrain est financé par l'IRT SystemX en collaboration avec CentraleSupélec. Ce travail a été réalisé grâce aux ressources du centre de calcul Mésocentre de CentraleSupélec et de l'ENS Paris-Saclay, soutenu par le CNRS et la Région Ile-de-France.

Références

[1] Abed, W. : A robust bearing fault detection and diagnosis technique for brushless dc motors under non-stationary operating conditions. JCAES (2015)

[2] Akbari, H., et al. : VATT : transformers for multimodal self-supervised learning from raw video, audio and text. CoRR abs/2104.11178 (2021)

[3] Angelopoulos, A., et al. : Tackling faults in the industry 4.0 era—a survey of machine-learning solutions and key aspects. Sensors 20(1), 109 (2019)

[4] Bagher Zadeh, A., et al. : Multimodal language analysis in the wild : CMU-MOSEI dataset and interpretable dynamic fusion graph. In : ACL 2018. pp. 2236–2246

[5] Bahdanau, D., et al. : Neural machine translation by jointly learning to align and translate. In : ICLR 2015

[6] Baltrusaitis, T., et al. : Multimodal Machine Learning : A Survey and Taxonomy. IEEE TPAMI 41(2), 423–443 (2019)

[7] Bengio, Y., et al. : Representation learning : A review and new perspectives. IEEE TPAMI 35, 1798–1828 (2013)

[8] Choudhary, A., et al. : Bearing fault diagnosis of induction motor using thermal imaging. pp. 950–955 (2018)

[9] Diaz Rozo, J., et al. : Machine learning-based cps for clustering high throughput machining cycle conditions. Procedia Manufacturing 10, 997–1008 (2017)

[10] Goldstein, M., Uchida, S. : A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PloS one 11, e0152173 (2016)

[11] Grifoni, P. : Multimodal human computer interaction and pervasive services (2009)

[12] Guo, L., et al. : A recurrent neural network based health indicator for remaining useful life prediction of bearings. Neurocomputing 240 (2017)

[13] Guo, W., et al. : Deep Multimodal Representation Learning : A Survey. IEEE Access 7, 63373–63394 (2019)

[14] Guo, W., et al. : Deep multimodal representation learning : A survey. IEEE Access 7, 63373–63394 (2019)

[15] Han, W., et al. : Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In : EMNLP 2021. pp. 9180–9192. ACL (2021)

[16] Huang, Y., et al. : What makes multimodal learning better than single (provably). ArXiv abs/2106.04538 (2021)

[17] Isermann, R. : Fault-Diagnosis Systems From Fault Detection to Fault Tolerance, vol. 28 (2006)

[18] Jafar, R., et al. : Application of artificial neural networks (ann) to model the failure of urban water mains. Mathematical and Computer Modelling 51, 1170–1180 (2010)

[19] Janssens, O., et al. : Thermal image based fault diagnosis for rotating machinery. Infrared Physics Technology 73, 78–87 (2015)

[20] Jia, F., et al. : Deep neural networks : A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mechanical Systems and Signal Processing 72-73 (2015)

[21] Konar, P., et al. : Bearing fault detection of induction motor using wavelet and neural networks. pp. 798–809 (2009)

[22] LeCun, Y., et al. : Deep learning. Nature 521, 436–44 (2015)

[23] Li, Z. : Deep learning driven approaches for predictive maintenance : A framework of intelligent fault diagnosis and prognosis in the industry 4.0 era (2018)

[24] Liang, T., et al. : Bearing fault diagnosis based on improved ensemble learning and deep belief network. Journal of Physics 1074, 012154 (2018)

[25] Liu, H., et al. : Unsupervised fault diagnosis of rolling bearings using a deep neural network based on generative adversarial networks. Neurocomputing 315 (2018)

[26] Liu, Y., et al. : Application to induction motor faults diagnosis of the amplitude recovery method combined with fft. Mechanical Systems and Signal Processing 24, 2961–2971 (2010)

[27] Long, X., et al. : Multimodal keyless attention fusion for video classification. AAAI 2018 pp. 7202–7209

[28] Luo, B., et al. : Early fault detection of machine tools based on deep learning and dynamic identification. IEEE TIE 66(1), 509–518 (2019)

[29] Mian, T., et al. : A sensor fusion based approach for bearing fault diagnosis of rotating machine. Journal of Risk and Reliability 0(0), 1748006X211044843 (0)

[30] Morency, L.P., et al. : Towards multimodal sentiment analysis : Harvesting opinions from the web. In : ICMI 2011. p. 169–176. ACM

[31] Ngiam, J., et al. : Multimodal Deep Learning. ICML 3(3), 194–203 (2011)

[32] Nor, N., et al. : A review of data-driven fault detection and diagnosis methods : Applications in chemical process systems. Reviews in Chemical Engineering 36 (2019)

[33] Palade, V., et al. : Computational Intelligence in Fault Diagnosis (2006)

[34] Pan, J., et al. : Liftingnet : A novel deep learning network with layerwise feature learning from noisy mechanical data for fault classification. IEEE TIE PP, 1–1 (2017)

[35] Peng, Y., et al. : Current status of machine prognostics in condition-based maintenance : A review. International Journal of Advanced Manufacturing Technology 50, 297–313 (2010)

[36] Reid, A., et al. : Fault location and diagnosis in a medium voltage epr power cable. IEEE TDEI 20, 10 – 18 (2013)

[37] Reis, M.S., Gins, G. : Industrial process monitoring in the big data/industry 4.0 era : from detection, to diagnosis, to prognosis. Processes 5(3) (2017)

[38] Rogers, A., et al. : A review of fault detection and diagnosis methods for residential air conditioning systems. Building and Environment 161, 106236 (2019)

[39] Shao, H., et al. : A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders. MSSP 102, 278–297 (2018)

[40] Shi, Y., et al. : Emformer : Efficient memory transformer based acoustic model for low latency streaming speech recognition (2020)

[41] Silberer, C., et al. : Learning grounded meaning representations with autoencoders. ACL 2014 1, 721–732

[42] Sipos, R., et al. : Log-based predictive maintenance. In : ACM SIGKDD 2014. p. 1867–1876

[43] Sipple, J. : Interpretable, multidimensional, multimodal anomaly detection with negative sampling for detection of device failure. In : ICML 2020. vol. 119, pp. 9016–9025. PMLR (2020)

[44] Srivastava, N. : Multimodal Learning with Deep Boltzmann Machines 15, 2949–2980 (2014)

[45] Srivastava, N., et al. : Learning representations for multimodal data with deep belief nets. ICML Workshop (2012)

[46] Sun, J., et al. : Intelligent bearing fault diagnosis method combining compressed data acquisition and deep learning. IEEE TIM PP, 1–11 (2017)

[47] Taheri-Garavand, A., et al. : An intelligent approach for cooling radiator fault diagnosis based on infrared thermal image processing technique. Applied Thermal Engineering 87, 434–443 (2015)

[48] Taneja, G., et al. : Reliability modelling and analysis of a single machine subsystem of a cable plant (2017)

[49] Tsai, Y.H.H., et al. : Multimodal transformer for unaligned multimodal language sequences. ACL 2019 pp. 6558–6569

[50] Vaswani, A., et al. : Attention is all you need. In : NIPS 2017. vol. 30

[51] Venkatasubramanian, V., et al. : A review of process fault detection and diagnosis. part i : Quantitative model-based methods 27(3), 293–311. part ii : Qualitative models and search strategies 27(3), 313–32. part iii : Process history based methods 27(3), 327–346. Computers Chemical Engineering (2003)

[52] Wang, F., et al. : Bilevel feature extraction-based text mining for fault diagnosis of railway systems. IEEE TITS 18(1), 49–58 (2016)

[53] Wang, H., et al. : Select-additive learning : Improving cross-individual generalization in multimodal sentiment analysis (2016)

[54] Wang, J., et al. : Machine vision intelligence for product defect inspection based on deep learning and hough transform. Journal of Manufacturing Systems 51, 52–60 (2019)

[55] Wang, S., et al. : Panoramic crack detection for steel beam based on structured random forests. IEEE Access 6, 16432–16444 (2018)

[56] Wen, L., et al. : A new convolutional neural network based data-driven fault diagnosis method. IEEE TIE PP, 1–1 (2017)

[57] Wen, L., et al. : A new snapshot ensemble convolutional neural network for fault diagnosis. IEEE Access 7, 32037–32047 (2019)

[58] Wu, B., et al. : Simultaneous-fault diagnosis considering time series with a deep learning transformer architecture for air handling units. Energy and Buildings 257, 111608 (2021)

[59] Xia, M., et al. : Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks. IEEE/ASME Transactions on Mechatronics PP, 1–1 (2017)

[60] Xie, Y., Zhang, T. : Imbalanced learning for fault diagnosis problem of rotating machinery based on generative adversarial networks. pp. 6017–6022 (2018)

[61] Yam, R., et al. : Intelligent predictive decision support system for condition-based maintenance. IJAMT 17, 383–391 (2001)

[62] Yang, B.S., et al. : Random forests classifier for machine fault diagnosis. JMST 22, 1716–1725 (2008)

[63] Yang, Z., et al. : A multi-branch deep neural network model for failure prognostics based on multimodal data. Journal of Manufacturing Systems 59, 42–50 (2021)

[64] Yang, Z., et al. : Stacked attention networks for image question answering. IEEE CVPR 2016(1), 21–29

[65] Yu, K., et al. : A bearing fault and severity diagnostic technique using adaptive deep belief networks and dempster–shafer theory. Structural Health Monitoring (2019)

[66] Yu, W., et al. : Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. arXiv (2021)

[67] Zadeh, A., et al. : Memory fusion network for multi-view sequential learning. AAAI 2018 pp. 5634–5641

[68] Zadeh, A., et al. : Multimodal language analysis in the wild : Cmu-mosei dataset and interpretable dynamic fusion graph. ACL 2018 1, 2236–2246 (2018)

[69] Zadeh, A., et al. : Tensor Fusion Network for Multimodal Sentiment Analysis pp. 1103–1114 (2018)

[70] Zadeh, A., et al. : Factorized Multimodal Transformer for Multimodal Sequential Learning pp. 1–13 (2019)

[71] Zarei, J., et al. : Vibration analysis for bearing fault detection and classification using an intelligent filter. Mechatronics 24 (2014)

[72] Zhang, S., et al. : Deep learning algorithms for bearing fault diagnostics—a comprehensive review. IEEE Access 8, 29857–29881 (2020)

[73] Zhang, Z., et al. : Fault diagnosis and prognosis using wavelet packet decomposition, fourier transform and artificial neural network. Journal of Intelligent Manufacturing 24 (2013)

[74] Zhou, F., et al. : A multimodal feature fusion-based deep learning method for online fault diagnosis of rotating machinery. Sensors 18, 3521 (2018)

Apprentissage automatique avec peu d'exemples pour l'extraction du contenu des documents non structurés

M. KANDI, L. NICOLAIEFF, Y. ZEGAOUI, C. BORTOLASO

Berger-Levrault, 64 Rue Jean Rostand, 31670 Labège, France

{mohamed.kandi, lina.nicolaieff,younes.zegaoui,christophe.bortolaso}@berger-levrault.com

Résumé

De nos jours, les entreprises déploient des mécanismes complexes pour automatiser la collecte, le stockage et le traitement de données. Néanmoins, certaines données sont sous un format non structuré : factures, bons de commande, ordonnances, etc. Il est possible de construire un modèle entraîné sur des exemples annotés pour extraire le contenu utile de ces documents. Cependant, dans de nombreux cas, il y a beaucoup de variations dans les types de documents. L'annotation d'exemples est une tâche fastidieuse et répétitive effectuée régulièrement lorsque de nouveaux types de documents arrivent. Pour minimiser ce travail de supervision, nous présentons dans ce papier une méthode permettant de sélectionner un sous-ensemble pertinent de documents non structurés à annoter. Pour évaluer la méthode, nous avons entraîné un modèle de type Faster R-CNN avec cinq jeux de données différents. Nous avons comparé les performances avec différents types de documents et taille de jeux de données. Nous montrons qu'un choix pertinent et automatisé d'exemples de documents peut éviter un effort considérable d'annotation.

Mots-clés

Détection et extraction de contenu, apprentissage avec peu d'exemples, Faster R-CNN, Triplet-loss.

Abstract

Nowadays, companies deploy complex mechanisms to automate data collection, storage, and processing. Some of this data is in an unstructured format : invoices, medical prescriptions... It is possible to build a model trained on annotated examples to extract useful information from these documents. However, in many cases, there is a lot of variation in document templates. Annotation is a tedious and repetitive task done regularly when new document templates arrive. We present in this paper a method to select a small and relevant subset of unstructured documents to annotate. To evaluate the method, we trained a model with five datasets. We compared the performance with different choices of document templates and dataset size. We show that a relevant and automated choice of document examples can avoid a huge annotation effort.

Keywords

Content localization and extraction, Few-shot learning, Faster R-CNN, Triplet-loss.

1 Introduction

Les données constituent un élément clé de la prise de décision. De nos jours, de nombreuses entreprises déploient des processus métier complexes pour automatiser la collecte, le stockage et le traitement d'une énorme quantité de données. Malheureusement, certaines de ces données ont un format non structuré : factures, emails, devis, bons de commande, tickets, documents scannés, cartes d'identité, ordonnances, etc. Cette représentation non structurée est difficile à exploiter par la machine, ce qui complique l'automatisation des processus métier et reporte un effort considérable sur les agents administratifs qui doivent ressaisir les informations dans les logiciels de gestion.

Le traitement intelligent de documents (Intelligent Document Processing, IDP) est un ensemble de moyens, méthodes et technologies permettent de capturer, extraire et traiter des données à partir de nombreux formats de documents [1]. Avec le traitement intelligent de documents, il est possible de transformer des données non-exploitables en données structurées facilement manipulables par un processus métier automatisé. Un framework IDP utilise des techniques d'analyse d'image, de traitement du langage naturel et d'apprentissage automatique profond pour remplir cette tâche. Ces derniers ont connu un grand succès ces dernières années, grâce à une grande quantité de données générées, à la disponibilité de capacités de calcul à la demande à des coûts raisonnables, et aux méthodes et architectures neuronales fournies par la communauté scientifique.

Les applications du traitement intelligent des documents sont nombreuses. Dans ce travail, nous prenons, comme cas d'usage, le contrôle de conformité des factures générées par les logiciels de gestion financière. En France, les factures générées au niveau des municipalités sont envoyées à une autorité centrale, qui les imprime et les envoie aux entités concernées. La réglementation impose de nombreuses règles concernant la forme et le contenu des factures. Par exemple, l'adresse de l'expéditeur et celle du destinataire doivent figurer dans des cases bien définies, certaines zones doivent être vides, le logo ne doit pas chevaucher sur les

marges, etc. Un document non-conforme est rejeté. Dans le cas contraire, cela entraîne des conséquences négatives pour de nombreuses personnes, comme le fait de ne pas recevoir un document à cause d'une adresse illisible, tronquée ou mal positionnée.

La réglementation évolue régulièrement, les éditeurs de logiciels de gestion financière doivent donc s'adapter rapidement aux changements. Pour cela, il est crucial de contrôler les factures et de détecter automatiquement les éléments pertinents (tels que l'adresse de l'expéditeur, l'adresse du destinataire et le logo) et d'extraire le contenu. Dans les travaux existants, il existe deux approches pour y parvenir [2] : le pattern-matching et l'apprentissage automatique. La première approche nécessite la mise en place de règles de correspondance qui sont difficiles à maintenir. De plus, cette approche ne se généralise pas. L'objectif d'un framework IDP est d'adopter une solution qui pourrait être généralisée à d'autres types de documents pour de futurs cas d'usage. C'est pourquoi nous avons décidé de baser notre travail sur les approches d'apprentissage automatique.

L'un des principaux inconvénients des méthodes d'apprentissage automatique tient dans le besoin de documents annotés pour entraîner le modèle. L'annotation est une tâche fastidieuse et coûteuse. Nous appliquons des techniques d'apprentissage avec peu d'exemples (Few-Shot Learning ou *FSL*) pour réduire l'effort d'annotation. Pour sélectionner un sous-ensemble suffisant de documents pertinents, nous proposons une méthode basée sur le calcul d'embeddings avec un modèle Triplet-Loss [1] [3, 4, 5], puis un clustering avec une méthode de k-means. Lorsque les documents similaires se retrouvent dans le même cluster, nous considérons qu'il n'est pas pertinent de les annoter tous. Nous construisons notre jeu de données avec seulement quelques exemples de chaque cluster. Nous annotons ce jeu de données puis nous entraînons un modèle suivant une architecture Faster R-CNN [6] pour détecter les éléments pertinents dans les documents.

Ce papier est organisé comme suit. D'abord, nous donnons un aperçu des travaux existants et nous positionnons notre travail dans la section 2. Puis, nous présentons la description du système proposé dans la section 3. Ensuite, nous détaillons les expériences réalisées pour évaluer notre contribution dans la section 4. Enfin, nous concluons le papier et donnons nos perspectives de recherche dans la section 5.

2 Travaux antérieurs

Il existe deux approches pour extraire des données de documents non structurés [2] : le pattern-matching (sous-section 2.1) et l'apprentissage automatique (sous-section 2.2). L'approche de l'apprentissage automatique s'appuie sur l'apprentissage avec peu d'exemples pour construire des modèles efficaces avec peu d'effort d'annotation (sous-section 2.3). Pour positionner notre travail, nous présentons, dans ce qui suit, le principe et les limites des méthodes existantes.

1. https ://www.tensorflow.org/addons/tutorials/losses_triplet

2.1 Méthodes orientées pattern-matching

Ces méthodes consistent à identifier des motifs dans les documents et à les utiliser pour extraire des informations [7, 8]. Plusieurs types de documents sont prédéfinis, et le but est de vérifier dans quelle mesure les documents sources correspondent aux modèles cibles. Cependant, la création et la maintenance des types demandent du temps et de l'expertise. De plus, ces méthodes ne fonctionnent pas bien dans le cas de petites différences entre les documents sources et les modèles, qui sont difficiles à interpréter pour la méthode. Par exemple, si nous avons plusieurs formats de facture avec des éléments éventuellement mal placés, il pourrait être difficile de déterminer si un désalignement avec un modèle est dû au fait que le document n'appartient pas à ce modèle ou aux éléments mal placés. Une solution rapide serait d'ajouter des règles qui peuvent être évaluées à tout moment pour vérifier si chaque élément est bien placé. Néanmoins, l'établissement de toutes les règles possibles serait fastidieux et déraisonnable pour une approche de type pattern-matching.

2.2 Méthodes basées sur l'apprentissage automatique

Une autre approche consiste à utiliser l'apprentissage automatique. Il s'agit d'entraîner un modèle avec un ensemble d'exemples annotés. Certains travaux considèrent la tâche comme une classification de mots. Pour chaque mot du document, nous décidons de l'extraire ou non. Si nous devons détecter plusieurs éléments, la tâche devient une classification multi-classes. Ces travaux ont opté, pour résoudre le problème, soit des modèles classiques d'apprentissage automatique comme les SVMs [9], soit avec des réseaux de neurones [10].

Nous pouvons également considérer les documents comme des images. Dans ce cas, il est possible de tirer parti des architectures *CNN* bien établies pour la détection d'objets : YOLO [11], Single Shot MultiBox Detector [12], Fast R-CNN [13], Faster R-CNN [6], Feature Pyramid Networks [14], etc.

L'un des principaux inconvénients des méthodes d'apprentissage automatique c'est la nécessité de disposer de nombreux exemples annotés pour entraîner le modèle. L'apprentissage avec peu d'exemples est un domaine de recherche qui vise à résoudre ce problème.

2.3 Apprentissage avec peu d'exemples : Few-shot learning

L'annotation manuelle est une tâche coûteuse, et il est difficile de créer un grand ensemble de données annotées. Nous devons sélectionner le plus petit sous-ensemble possible de documents tout en étant pertinents et variés en termes d'exemples.

Les travaux existants sur l'apprentissage avec peu d'exemples se répartissent en trois grandes catégories [15]. La première catégorie se concentre sur les données. Il s'agit de commencer par l'ensemble des données disponibles. Ensuite, on utilise les connaissances antérieures pour appliquer des transformations qui génèrent un ensemble de don-

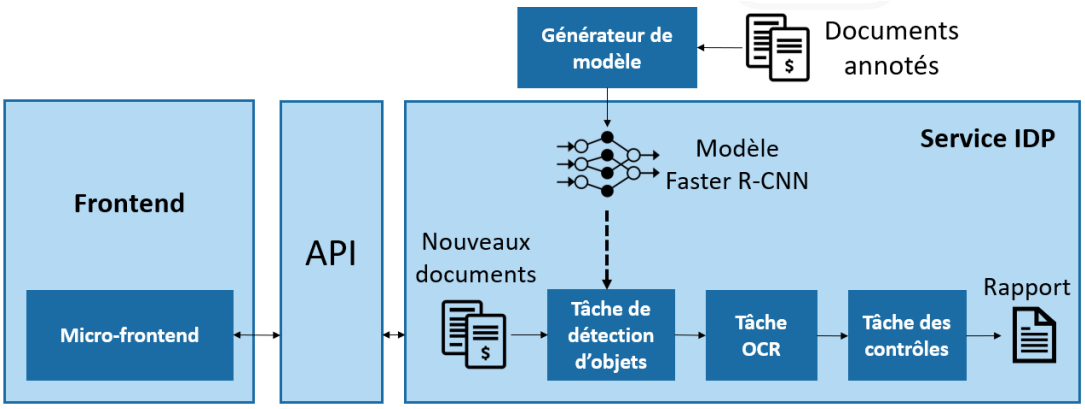


FIGURE 1 – Une vue globale du framework de traitement intelligent des documents

nées plus important. La deuxième catégorie se concentre sur le modèle. Il s'agit d'utiliser les connaissances préalables pour réduire l'espace des hypothèses et limiter la complexité du problème à résoudre. On peut résoudre ce dernier avec un jeu de données qui n'est pas volumineux. La troisième catégorie se concentre sur l'algorithme. Il s'agit d'utiliser les connaissances préalables pour modifier la stratégie de recherche des meilleurs paramètres du modèle : en donnant une bonne initialisation ou/et en guidant les étapes de la recherche.

La première catégorie de travaux applique des transformations pour augmenter les données. Nous pouvons faire l'augmentation avec une liste de règles faites à la main. Des exemples typiques sont l'augmentation d'images avec des traductions [16], des recadrages [17], des rognages [18], des rotations [18] etc. Cependant, les règles sont souvent spécifiques à l'ensemble de données disponible et sont difficilement applicables à d'autres ensembles de données. Par conséquent, l'augmentation des données ne nous permet pas de résoudre complètement le problème d'apprentissage avec peu exemples.

La deuxième catégorie de travaux tente de réduire l'espace des hypothèses et de limiter la complexité du problème. Nous pouvons classer les méthodes appartenant à cette catégorie en deux techniques [15] : (1) l'apprentissage multi-tâches et (2) l'apprentissage par embeddings. Lorsque nous avons plusieurs tâches liées, l'apprentissage multi-tâches permet d'apprendre plusieurs tâches simultanément en exploitant à la fois les informations génériques communes et les informations spécifiques à la tâche. Cette technique suppose que certaines tâches n'ont que quelques exemples alors que d'autres en ont beaucoup. Les paramètres à apprendre pour chaque tâche dépendent des autres tâches. Nous pouvons le faire avec le partage [19, 20] ou la fixation des paramètres [21, 22]. L'inconvénient de l'apprentissage multi-tâches est qu'il suppose l'existence de plusieurs tâches similaires, certaines avec de nombreux exemples. Ce qui n'est pas toujours le cas. Souvent, nous nous trouvons dans un cas où nous considérons une seule tâche ou un ensemble de tâches avec peu d'exemples. Il faut également noter que l'apprentissage simultané de toutes les tâches est nécessaire. Lorsqu'une nouvelle tâche arrive, l'ensemble du modèle multi-tâches doit être réentraîné, ce qui est coûteux et lent.

Avec l'apprentissage basé sur les embeddings, nous représentons chaque document dans un espace de plus petite dimension. Dans cet espace, les documents similaires sont proches, tandis que les documents différents sont éloignés. La fonction de transformation est entraînée sur des connaissances antérieures [23], ou avec des connaissances spécifiques de la tâche à accomplir [24, 25]. Il peut également être entraîné sur une combinaison des deux [26, 27, 28, 29]. Nous ne pouvons utiliser les méthodes d'embeddings que si nous disposons d'un grand ensemble de données, contenant suffisamment d'exemples de différentes classes génériques de connaissances préalables, ou d'un modèle pré-entraîné. Malheureusement, ce n'est pas toujours le cas. En outre, l'efficacité de ces méthodes est incertaine si les données spécifiques à la tâche ne sont pas liées aux données génériques. Enfin, la manière de combiner les informations génériques et spécifiques dépend de la nature des données, et il n'existe pas de stratégie bien établie.

La troisième catégorie de travaux utilise les connaissances préalables pour influencer l'algorithme d'exploration des paramètres du modèle. Il existe trois techniques [15] : (1) choisir des paramètres initiaux endossés à l'aide de données provenant d'autres tâches, puis affiner avec les données de la tâche cible [30, 31, 32, 33], (2) choisir des paramètres initiaux approuvés par un méta-algorithme à partir d'un ensemble de tâches qui ont la même distribution que la tâche cible [34], (3) trouver des méta-algorithmes pour guider intelligemment la direction de recherche ou l'étape d'itération en fonction des connaissances préalables [35]. La première technique est utile pour accélérer l'apprentissage du modèle, mais elle sacrifie la précision. Or, la précision est un objectif clé de nos cas d'utilisation. Les deuxième et troisième techniques présentent plusieurs problèmes ouverts et non résolus dans l'état de l'art : en particulier, comment gérer différentes granularités, comme la classification de documents au sens large, par opposition à la classification de modèles de factures, ou différentes sources de données, comme les images par opposition aux textes.

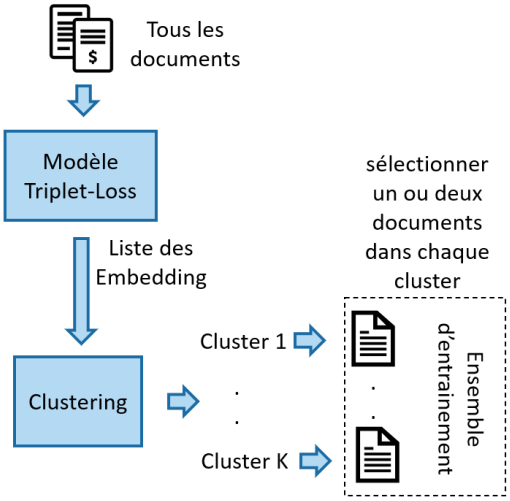


FIGURE 2 – Sélection de documents avec la Triplet-loss et le clustering

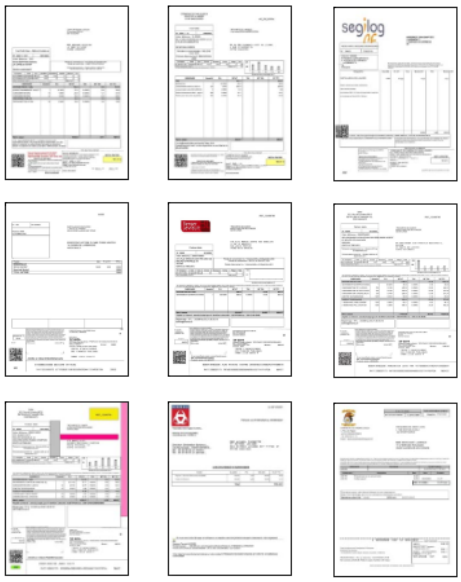


FIGURE 3 – Types de factures

3 Description du système et méthodologie

Dans nos travaux, nous considérons les documents en entrée comme des image et nous adoptons un réseau *CNN* détecteur d'objets standard. Pour entraîner (fine-tuning) ce modèle, nous avons besoin d'un ensemble de documents annotés. Nous montrons dans la figure 1, l'architecture de notre framework IDP. Celui-ci se présente sous la forme d'une API. Les clients peuvent soumettre de nouveaux documents sous forme de flux. Ces documents sont soumis à un module qui détecte et localise les éléments pertinents. Dans notre cas, nous avons limité l'étude à la détection et localisation de : (1) l'adresse expéditeur, (2) l'adresse destinataire, (3) le logo et (4) la datamatrix (QRCode contenant l'ensemble des informations de la facture). Une fois l'objet détecté, nous utilisons une méthode OCR pour extraire le texte qu'il contient. L'avantage est que le framework peut être généralisé pour de nombreux modèles de documents avec des jeux de données annotés. Enfin, le flux passe par une liste de contrôles qui permettant de vérifier le respect de la réglementation. Un rapport est généré qui contient la liste des éléments, leur emplacement dans le document et leur valeur. Le rapport contient également les résultats des contrôles. Pour les contrôles non respectés, on peut voir un commentaire qui aide à la correction.
La figure 4 montre un exemple de détection et extraction de contenu à partir d'une facture. Dans le résultat affiché, nous pouvons voir les éléments détectés, leur score de confiance, ainsi qu'un tableau des contrôles de conformité.

3.1 Détection d'éléments pertinents

Le module de détection d'éléments est basé sur un modèle *CNN* de détection d'objets et a pour objectif de produire plusieurs régions, ou imagettes, à partir de l'image du document en entrée. Ces régions doivent être centrées le plus possible sur les éléments d'intérêt du document. Les régions sont ensuite transmises en entrée à une méthode OCR pour extraire le contenu textuel. Les modèles *CNN* détecteur d'objets existants se répartissent en deux catégories : les méthodes en deux étapes et les détecteurs en un coup (single shot detector). Alors que les premiers sont censés produire des résultats plus robustes, les seconds peuvent traiter les images plus rapidement et réaliser une détection d'objets en temps réel. Dans notre cas, le temps réel n'est pas une contrainte car nous souhaitons généralement traiter les documents par lots. Nous avons alors décidé d'utiliser l'architecture Faster R-CNN [13] car c'est le détecteur à deux étapes le plus largement utilisé.

Le modèle Faster R-CNN [13] est composé de 2 modules :

— le premier, appelé le réseau de proposition de régions, traite l'image d'entrée et produit plusieurs régions d'intérêt avec différentes tailles et proportions, appelées propositions ;
— le second module prend en entrée ces propositions et vise à les ajuster au mieux autour de l'objet qu'elles contiennent, à l'aide d'une fonction de régression, ainsi qu'à leur attribuer une classe, à l'aide d'une fonction de classification classique.

En raison de la présence du premier module, nous nous attendons à ce que le modèle ait une meilleure précision par rapport aux détecteurs en un coup tout en obtenant des valeurs de rappel similaires. De plus, pour obtenir de meilleures performances, nous avons utilisé un modèle pré-entraîné sur le célèbre corpus COCO (Common Objects in Context) [36]. Il nous est possible d'obtenir un grand nombre de documents de factures, cependant le coût d'annotations serait trop important. Il est alors nécessaire de disposer d'un moyen pour sélectionner un petit sous-ensemble de documents aussi varié que possible afin d'optimiser au mieux le travail d'annotations.

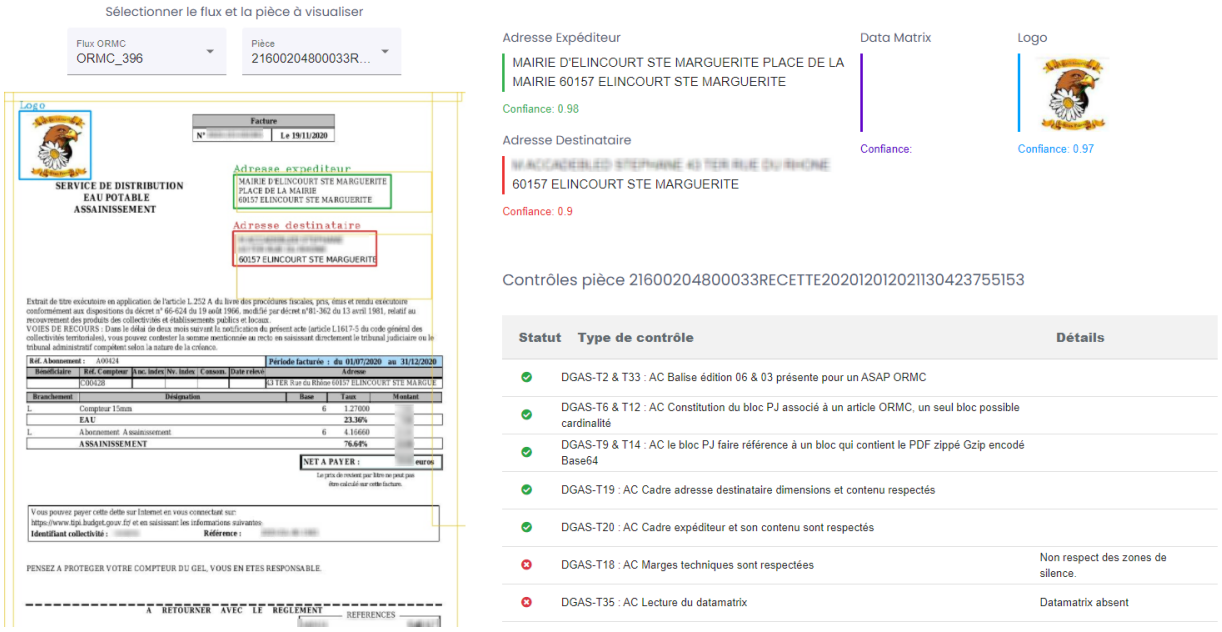


FIGURE 4 – Exemple d'extraction d'extraction de contenu d'une facture

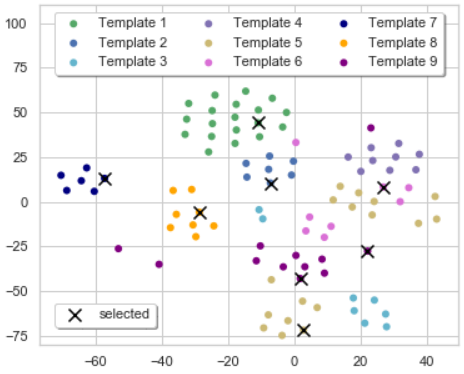
3.2 Sélection des meilleurs candidats pour l'entraînement

Bien qu'il puisse être coûteux en temps d'annoter chaque document à notre disposition avec des boîtes englobantes autour des éléments d'intérêt, il est intéressant de regrouper les documents en fonction des similarités de leurs structure graphique (template). Étant donné que les documents de chaque groupe sont très similaires, notre objectif est de sélectionner les exemples qui capturent le mieux les caractéristiques sous-jacentes du modèle afin que le réseau puisse en tirer parti lors de l'entraînement sans avoir à traiter de nombreux documents quasi-identiques. La figure 2 illustre le processus de sélection des documents candidats. Tout d'abord, nous utilisons un modèle Triplet-loss [37] pour projeter les documents dans un espace vectoriel. Le but du modèle est que les vecteurs embeddings associés aux documents d'un même template soient proches les uns des autres dans l'espace latent tout en étant éloignés des documents des autres templates. Le réseau *CNN* est entraîné comme tel : pour chaque image en entrée, appelée ancre, deux autres images sont passées en entrée au modèle. L'une est l'exemple positif, appartenant au même template que l'ancre, l'autre étant l'exemple négatif, appartenant à un template différent. La fonction de coût calculée minimise la distance dans l'espace latent entre l'ancre et l'exemple positif tout en maximisant la distance entre l'ancre et l'exemple négatif. Une fois le réseau entraîné, nous calculons un vecteur embedding pour chaque document de notre ensemble de données et utilisons l'algorithme du k-means pour regrouper ceux-ci en clusters. Nous pouvons alors vérifier à quel point les clusters obtenus recoupent la répartition des documents en templates. Il est ainsi utile de montrer à quel point nous pouvons utiliser le même espace latent afin d'y projeter des documents appartenant à template jamais vu sans avoir à entraîner de nouveau le modèle Triplet-loss. Enfin, nous sélectionnons pour chaque cluster des documents candidats. Les documents sélectionnés sont alors annotés manuellement et constituent notre ensemble d'apprentissage pour le modèle de détection d'éléments Faster R-CNN. Pour un cluster donné, nous pouvons choisir le document le plus proche du centroïde ou la paire de documents qui ont une distance maximale entre eux.

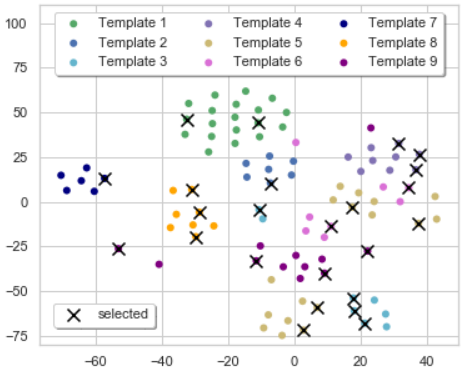
4 Experimentations

Pour évaluer notre approche, nous avons dans un premier temps pré-entraîné un modèle Triplet-loss avec le jeu de données RVL-CDIP[2] [38]. Puis dans un second temps, nous avons affiné le modèle avec un ensemble de données spécifique à notre cas d'utilisation (27 factures issues de 9 types de structures différentes). La figure 3 montre un exemple de chaque type de facture. Puis, nous avons calculé la représentation d'un ensemble de test contenant 87 nouvelles factures, avec le modèle Triplet-loss déjà entraîné. Enfin, nous avons exécuté un algorithme de clustering K-means sur les documents de l'ensemble de test. Nous avons fait varier la stratégie de sélection de documents (le document le plus proche du centroïde par rapport à la paire de documents présentant la distance maximale entre eux) et le nombre de documents sélectionnés (8, 16 et 24 documents). Pour visualiser le résultat, nous avons projeté les représentations sur un espace bi-dimensionnel avec T-SNE [39]. La figure 6 montre les résultats de cette méthode. Chaque point correspond à une facture et les couleurs permettent de distinguer les types de facture. On constate que les documents appartenant à un même type sont, dans la plupart des cas, proches dans l'espace des représentations. Chaque docu-

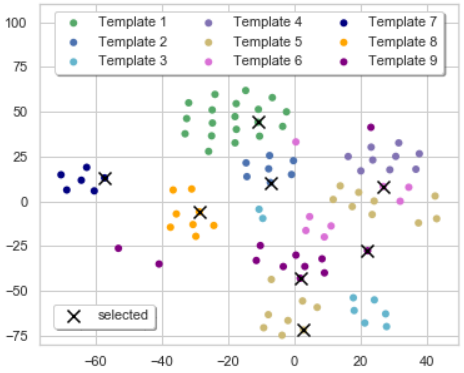
2. https ://www.cs.cmu.edu/ aharley/rvl-cdip/



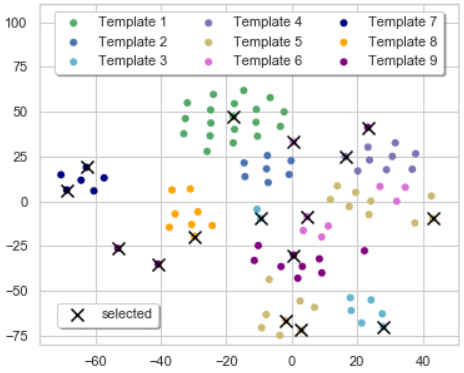
(a) Document le plus proche du centroïde - distance euclidienne - 8 documents sélectionnés



(b) Document le plus proche du centroïde - distance euclidienne - 24 documents sélectionnés

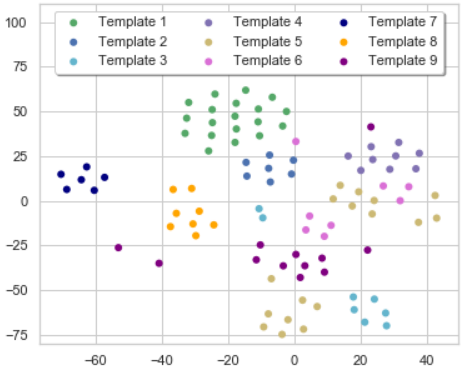


(c) Document le plus proche du centroïde - similarité cosinus - 8 documents sélectionnés

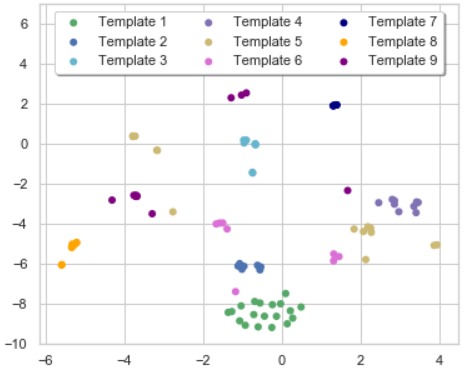


(d) Paire de documents avec distance maximale - distance euclidienne - 16 documents sélectionnés

FIGURE 5 – Projection des vecteurs de documents par T-SNE selon la fonction de distance adoptée (cosinus ou euclidienne), le nombre de documents sélectionnés et le choix des documents (plus proche du centroïde ou paire la plus éloignée).



(a) 1 epoch



(b) 5 epochs

FIGURE 6 – Projection des vecteurs embeddings à l'aide du T-SNE : après 1 epoch d'entraînement et après 5 epochs.

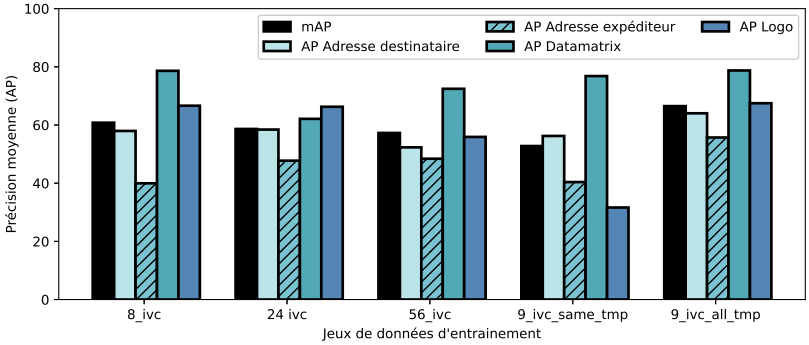


FIGURE 7 – Résultat de la précision moyenne (AP) du modèle de détection d'éléments Faster R-CNN

Entraînement	**mAP**	AP Destinataire	AP Expéditeur	AP Datamatrix	AP Logo
8_ivc	**60.789**	57.954	39.953	78.620	66.627
24_ivc	**58.634**	58.428	47.734	62.111	66.264
56_ivc	**57.267**	52.315	48.406	72.444	55.901
$9\_ivc\_same\_tmp$	**52.766**	56.230	40.383	76.818	31.634
$9\_ivc\_all\_tmp$	**66.486**	64.025	55.696	78.742	67.479

TABLE 1 – Résultats détaillés de la précision moyenne (AP)

Entraînement	Adresse destinataire	Adresse expéditeur	Datamatrix	Logo
8_ivc	8	8	6	4
24_ivc	24	24	19	12
56_ivc	56	56	36	38
$9\_ivc\_same\_tmp$	9	9	5	6
$9\_ivc\_all\_tmp$	9	9	4	9

TABLE 2 – Nombre de factures contenant chaque objet dans les jeux d'entrainement

ment d'un type se rapproche des autres dans l'espace après cinq époques d'entrainement en comparaison à une seule époque.

Sur les figures 5, les factures sélectionnées pour l'annotation sont identifiées par une croix. Nous remarquons que la similarité cosinus sélectionne les mêmes documents que la distance euclidienne. Les points sélectionnés sont bien répartis dans l'espace des représentations, ce qui garantit la variété des documents.

D'un autre côté, nous avons évalué la sensibilité du modèle de détection d'objets Faster R-CNN au jeu de données d'entraînement. Pour ce faire, nous avons calculé la performance avec la métrique de précision moyenne (AP) sur plusieurs sous-ensembles en faisant varier le nombre de documents et la diversité des types de structure des documents.

Pour évaluer l'importance de la diversité, nous avons construit cinq jeux d'entraînement :

- Les jeux d'entrainement 8_ivc, 24_ivc, 56_ivc contiennent 8 types de factures. Ce qui les différencie, c'est leur taille. Le premier contient 8 factures, le second 24 documents, et le troisième 56 documents.
- Nous disposons alors d'un ensemble d'apprentissage $9\_ivc\_all\_tmp$ contenant une facture pour chaque type de documents (neuf types).
- Enfin, un ensemble d'apprentissage $9\_ivc\_same\_tmp$ contenant neuf factures du même type.

Notre jeu de test contient des exemples des neuf types de facture. Nous avons calculé la métrique *mAP* sur quatre objets (adresse du destinataire, adresse de l'expéditeur, DataMatrix et logo). Les résultats sont présentés dans la Figure 7 et le Tableau 1. Nous remarquons des scores *mAP* similaires lorsqu'on utilise un sous-ensemble de 8 factures seulement (60,789%) par rapport à un sous-ensemble de 24 factures (58,634%), et qu'il est même légèrement inférieur lorsqu'on utilise 56 factures (57,267%). Nous pensons que cela est dû au fait que la plupart des exemples du même format dans l'ensemble de données sont fortement similaires les uns aux autres, n'apportant donc pas plus d'informations au modèle. Nous notons que les factures ne contiennent pas le même nombre d'objets, ce déséquilibre peut aussi expliquer la diminution du score *mAP* lorsque l'on augmente le nombre de factures (Tableau 2).

Cette hypothèse est accentuée par l'expérience sur la diversité des sous-ensembles : le score *mAP* est le plus élevé (66,486%) lors de l'utilisation d'un sous-ensemble comprenant seulement 9 factures sélectionnées selon notre critère

de diversité. Ces résultats nous amènent à penser que dans le cas spécifique des factures où les données ont tendance à être homogènes, il est préférable de trouver quelques exemples avec des structures graphiques variées plutôt que d'ajouter simplement des exemples aléatoires au jeu de données d'entraînement.

5 Conclusion

Dans cet article, nous avons montré que le modèle basé sur la Triplet-loss combiné au clustering peut être utilisé pour sélectionner un sous-ensemble de documents pertinents pour annoter et former un modèle de location d'objets. Dans des travaux futurs, nous mènerons des expériences sur un plus grand nombre de types de documents. Nous prévoyons également d'utiliser les vecteurs calculés pour surveiller les performances du modèle. Lorsqu'un nouveau document arrive, nous calculons la similarité de sa représentation avec les documents de l'ensemble d'apprentissage courant. Si la similarité est faible, alors nous déclenchons une alerte pour permettre à un validateur de vérifier que l'élément est correctement détecté. Si ce n'est pas le cas, il annote le nouveau document et l'ajoute au jeu de données de la prochaine mise à jour du modèle.
Nous prévoyons également d'étendre notre travail en concevant de nouvelles expériences qui pourraient nous aider à obtenir de meilleurs résultats sur la partie module de détection d'éléments de notre architecture : (1) unifier le modèle Triplet-loss avec le modèle de détecteur *CNN* en leur faisant partager certaines de leurs caractéristiques, (2) comparer le modèle Triplet-loss + k-means avec une approche unifiée de clustering à intégration profonde (*DEC*) [40], (3) aller plus loin dans la direction *FSL* en tirant parti des méthodes existantes telles que les réseaux de correspondance [41] pour aider notre modèle à obtenir le plus d'informations de notre ensemble de données, à la fois annotées et brutes, pendant l'entraînement.

Références

[1] "Grobid," https ://github.com/kermitt2/grobid, 2008–2021.

[2] R. B. Palm, F. Laws, and O. Winther, "Attend, copy, parse end-to-end information extraction from documents," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 329–336.

[3] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the iEEE conference on computer vision and pattern recognition*, 2016, pp. 1335–1344.

[4] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification." *Journal of machine learning research*, vol. 10, no. 2, 2009.

[5] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet : A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn : Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[7] E. Riloff, "Automatically constructing a dictionary for information extraction tasks," in *Proceedings of the eleventh national conference on Artificial intelligence*, 1993, pp. 811–816.

[8] I. Muslea *et al.*, "Extraction patterns for information extraction tasks : A survey," in *The AAAI-99 workshop on machine learning for information extraction*, vol. 2, no. 2. Orlando Florida, 1999.

[9] Y. Li, K. Bontcheva, and H. Cunningham, "Svm based learning system for information extraction," in *International Workshop on Deterministic and Statistical Methods in Machine Learning*. Springer, 2004, pp. 319–339.

[10] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of NAACL-HLT*, 2016, pp. 260–270.

[11] J. Redmon and A. Farhadi, "Yolo9000 : better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd : Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[13] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[14] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[15] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples : A survey on few-shot learning," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[16] S. Benaim and L. Wolf, "One-shot unsupervised cross domain translation," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 2108–2118.

[17] H. Qi, M. Brown, and D. G. Lowe, "Low-shot learning with imprinted weights," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5822–5830.

[18] Y. Zhang, H. Tang, and K. Jia, "Fine-grained visual categorization using meta-learning optimization with

sample selection of auxiliary data," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 233–248.

[19] S. Motiian, Q. Jones, S. M. Iranmanesh, and G. Doretto, "Few-shot adversarial domain adaptation," *arXiv preprint arXiv :1711.02536*, 2017.

[20] Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun, "Few-shot charge prediction with discriminative legal attributes," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 487–498.

[21] W. Yan, J. Yap, and G. Mori, "Multi-task transfer methods to improve one-shot learning for multimedia event detection." in *BMVC*, 2015, pp. 37–1.

[22] Z. Luo, Y. Zou, J. Hoffman, and L. Fei-Fei, "Label efficient learning of transferable representations across domains and tasks," *arXiv preprint arXiv :1712.00123*, 2017.

[23] E. Triantafillou, R. Zemel, and R. Urtasun, "Few-shot learning through an information retrieval lens," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 2252–2262.

[24] G. Koch *et al.*, "Siamese neural networks for one-shot image recognition," 2015.

[25] L. Yan, Y. Zheng, and J. Cao, "Few-shot learning for short text classification," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29 799–29 810, 2018.

[26] L. Bertinetto, J. Henriques, P. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers." International Conference on Learning Representations, 2019.

[27] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Advances in neural information processing systems*, 2016, pp. 523–531.

[28] B. N. Oreshkin, P. Rodriguez, and A. Lacoste, "Tadam : task dependent adaptive metric for improved few-shot learning," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 719–729.

[29] F. Zhao, J. Zhao, S. Yan, and J. Feng, "Dynamic conditional networks for few-shot learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 19–35.

[30] S. Ö. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *NeurIPS*, 2018.

[31] R. Keshari, M. Vatsa, R. Singh, and A. Noore, "Learning structure and strength of cnn filters for small sample size training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9349–9358.

[32] D. Yoo, H. Fan, V. N. Boddeti, and K. M. Kitani, "Efficient k-shot learning with regularized deep networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[33] J. Kozerawski and M. Turk, "Clear : Cumulative learning for one-shot one-class image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3446–3455.

[34] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[35] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas, "Learning to learn by gradient descent by gradient descent," in *Advances in neural information processing systems*, 2016, pp. 3981–3989.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco : Common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740–755.

[37] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition*, A. Feragen, M. Pelillo, and M. Loog, Eds. Cham : Springer International Publishing, 2015, pp. 84–92.

[38] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2015.

[39] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[40] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. JMLR.org, 2016, p. 478–487.

[41] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016.

Compréhension narrative semi-automatique pour le *debriefing* de session de simulation

A. Bitoun[1], A.-G. Bosser[2], Martín Diéguez[3], F. Legras[4]

[1] MASA Group, S.A.
[2] ENIB, Lab-STICC CNRS UMR 6285
[3] Université d'Angers, LERIA
[4] Chercheur indépendant

ariane.bitoun@masagroup.net

Résumé

Dans cette communication, nous décrivons les résultats du projet STRATEGIC autour de la mise-en-histoire semi-automatique des traces de l'activité lors d'une simulation utilisée dans le cadre de la formation. De nouveaux outils de représentation permettront de générer des vignettes correspondant aux moments décisifs identifiés par l'opérateur dans le graphe narratif et de proposer ainsi une synthèse du déroulement global du scénario, illustrée par des "photographies" instantanées de la situation tactique. Ceci facilitera le débriefing*, et favorisera l'apprentissage coopératif.*

Mots-clés

Sensemaking*, Storification*

Abstract

In this paper, we describe the results of the STRATEGIC project about the semi-automatic storification of activity traces of simulation sessions used for training. New representation tools will make it possible to generate thumbnails corresponding to the decisive moments identified by the operator in a narrative graph and thus to propose a synthesis of the overall progress of the scenario illustrated by instantaneous "photographs" of the tactical situation. This will facilitate debriefing, and promote cooperative learning.

Keywords

Sensemaking, Storifications

1 Introduction

L'entraînement par la simulation est une forme de pédagogie très efficace. Elle peut utiliser des logiciels (la famille des jeux sérieux dédiés à l'apprentissage), associés à une phase de *debriefing* qui permet aux participants de comprendre ce qui s'est passé pendant l'entraînement [15]. Une telle solution, souvent avantageuse financièrement, fournit un contexte sécurisé dans lequel les participants peuvent apprendre en constatant les effets de leurs décisions et de leurs actions.

L'entraînement militaire, en particulier, repose souvent sur des simulations qui peuvent être instrumentées pour de l'entraînement collectif ou dédiées à l'entraînement des postes de commandement. Ainsi l'entraînement peut s'effectuer en utilisant des simulations sur des champs de batailles réels, virtuels, constructifs, ou mixtes. Dans un environnement réel, les unités utilisent des équipements opérationnels pour combattre des ennemis composés d'individus ou de cibles. Dans les environnements virtuels, les unités utilisent des simulateurs pour représenter l'équipement et les armes. Les effets des armes, le terrain et les forces ennemies sont générés par ordinateur. Dans des environnements constructifs, les résultats du champ de bataille sont déterminés par une simulation informatique afin de fournir la situation de combat nécessaire à la formation du commandement et de l'état-major. Quelle que soit la nature de la simulation, celle-ci doit fournir aux entraînés un retour les informant de la manière dont leurs actions ont contribué au succès ou à l'échec de la mission. En général, une session de formation commence par une phase de *préparation*, lors de laquelle l'environnement opérationnel réaliste est créé, suivie d'une phase d'*exercice* lors de laquelle les participants prennent part à la simulation et, enfin, une phase de *debriefing* lors de laquelle les participants échangent (aidés et guidés par un modérateur) afin de comprendre ce qui s'est passé pendant l'entraînement et pourquoi et comment améliorer ou maintenir leurs performances dans des situations similaires à l'avenir. Le moment où cette discussion a lieu et sa durée sont très importants [4] : trop de détails conduisent à un manque de concentration des participants, et si on la retarde trop, les participants peuvent oublier les raisons qui les ont poussés à adopter une ligne de conduite spécifique.

Cependant, l'énorme quantité de données générées durant l'entraînement complique cette tâche [1] : il est très difficile de mettre en forme les données à temps pour la préparation du *debriefing*.

Pour proposer une solution à ce problème, nous avons développé un prototype et des cas d'usage d'un outil d'aide à la création narrative, qui permet à un humain de four-

1. Une simulation courte peut générer 60000 rapports à traiter

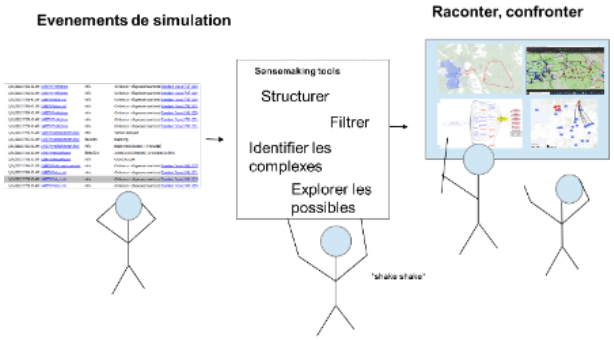


FIGURE 1 – Métaphore d'utilisation d'un outil de *sensemaking* narratif pour le *debriefing* de session de simulation.

nir des explications, sous la forme d'un récit illustré, aux autres participants de la simulation. L'idée est de fournir une analyse semi-automatisée du déroulement de la simulation sous la forme d'une reconstruction narrative des causes potentielles des événements qui se sont produits. Suivant une longue tradition de la représentation de la causalité sous forme de graphes [28], nous utilisons des diagrammes pour représenter les événements, leurs causes et conséquences. Ceux-ci peuvent ensuite être intégrés à l'interface de rejeu de la simulation au travers d'un système de *vignettes* qui fournit également aux participants d'autres informations de nature instantanée sur la simulation (figure 1).

2 Utiliser les récits dans le cadre de l'apprentissage par la simulation

Les humains ont de tout temps utilisé les récits pour faire sens du monde et expliquer le déroulement d'événements passés. Certaines approches de l'enseignement par le numérique ont donc tout naturellement utilisé la narration [13]. Dans de telles approches, raconter une histoire implique de formuler des relations causales entre des événements sélectionnés [33, 2]. Dans le domaine de l'éducation et des jeux sérieux, la *storification* [3] est le terme utilisé pour décrire la création d'une structure causale en établissant des liens entre les événements du récit. Un des défis dans ce domaine est de pouvoir automatiser, ou semi-automatiser cette activité afin de pouvoir s'adapter à différents profils d'utilisateurs. D'autre part, des recherches en psychologie au sujet de la compréhension des histoires ont aussi montré l'importance de la perception des relations causales entre les événements du récit [31].

Alors que la modélisation de la causalité occupe une place centrale en Intelligence Artificielle [28], le point de vue de l'Intelligence Narrative est plus proche de la notion de *commonsense reasoning* : celui ou celle qui raconte doit sélectionner les évènements dignes d'être rapportés, exprimer les liens de cause et conséquence entre eux, et décider du niveau de granularité dans la définition des évènements afin que le récit prenne son sens. L'idée est de fournir une explication sous une forme qui est compréhensible par les humains [29]. Cet objectif rejoint certains enjeux qu'on retrouve dans le domaine de l'IA explicable [25]. Ajoutons cependant que dans un récit, il y a également un point de vue (ce que Genette appelle la voix narrative par exemple [17]). Cette conception des histoires racontées correspond bien à notre contexte : à partir du matériel de base on veut semi-automatiser la création d'un récit en prenant en compte la subjectivité potentielle de celui ou celle qui raconte.

Notre objectif n'était donc pas de produire un système complètement automatisé : notre système doit permettre la confrontation de plusieurs points de vue lors du *debriefing* ou d'activités d'apprentissage coopératif. Notre système doit pouvoir permettre à chaque utilisateur de construire et expliquer leur propre récit, subjectif, qui dépendra des informations auxquelles ils ont eu accès (selon le rôle des participants de la simulation cela peut varier grandement), et des raisons de leurs décisions. Le tuteur pourra avoir accès à toutes les informations et son récit sera également construit différemment.

3 Description de l'architecture proposée

Dans ce projet, nous combinons une approche de construction narrative qui a fait ses preuves dans le domaine de la génération narrative à partir de spécifications formelles, et des représentations instantanées liées à l'état de la simulation pour les enrichir sémantiquement. Le tout est construit sur la description des simulations définies pour SWORD, la simulation de champ de bataille de la société MASA [2], ainsi que sur les rapports produits par chaque session de simulation. La figure 2 décrit l'architecture de la solution que nous proposons et que nous justifions plus avant dans la suite de cette section.

3.1 Une approche de la construction et de l'analyse narrative fondée sur la logique linéaire

La formalisation des récits est un problème qui a souvent été traité en Intelligence Artificielle du point de vue de la représentation des connaissances *Knowledge representation* et du raisonnement au sujet de l'action et du changement, en partant de la modélisation des actions narratives et en décrivant son impact sur l'environnement. Dans [8] on utilise la Logique Linéaire [18] pour leur modélisation, ce qui a donné lieu à des approches formelles autour de l'analyse et la vérification de propriétés des histoires en Logique Linéaire Intuitioniste [9, 11]. Un événement sera modélisé sous forme d'action narrative de base à l'aide de l'implication linéaire $\multimap$, qui permet de décrire comment la consommation de ressources peut en créer de nouvelles. Par exemple, on peut modéliser ainsi la réaction qui produit une molécule d'eau : $H_2 \otimes O \multimap H_2O$

Cette approche permet de modéliser de manière déclarative chaque événement, en décrivant son impact sur l'environnement en termes de consommation ou de production

2. https ://masasim.com/en/notre-metier/defense/

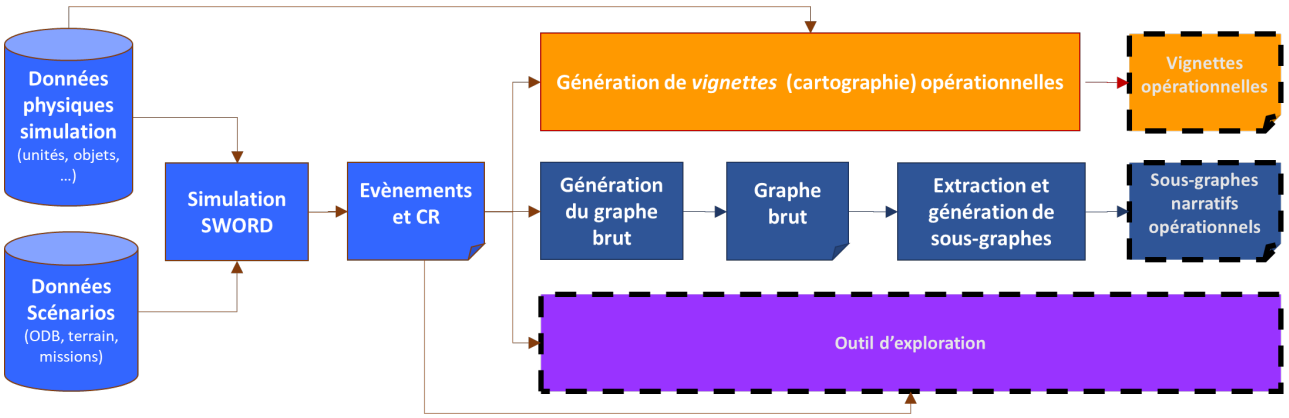


FIGURE 2 – La base de donnée de SWORD décrivant la session de simulation permet un encodage automatique des événements de la simulation, rendant possible une analyse narrative enrichie avec des vignettes fournissant différentes vues stratégiques pour un moment donné

des ressources. Cela a été appliqué également pour analyser d'un point de vue causal des séquences d'événements produites à partir d'une spécification sous forme d'actions en logique linéaire, comme au sein du système TeLLer[3] : en suivant, au fil du déroulement, quelle action va effectivement consommer une ressource produite par une autre action, on peut établir une relation de cause possible entre les deux événements ainsi modélisés. Cette approche a également été utilisée dans le cadre de systèmes de génération narrative fondés sur la programmation en logique linéaire pour étudier les récits ainsi créés sous la forme de leur représentation diagrammatique [22, 23].
Une des contribution dans ce projet a été d'appliquer ces mécanismes, dans un contexte où nous pouvions intégrer deux formes importantes de la compréhension humaine, toujours sous l'angle du récit : l'exploration contrefactuelle et la granularité des événements.

— *le raisonnement contrefactuel* est la simulation mentale de scénarios alternatifs de type "et si ...", ou la modification d'un ou plusieurs événements qui se sont produits mènent à la déduction d'une situation alternative. Cette forme de raisonnement joue un rôle central dans les jugements de causalité associés aux événements [24]. En IA, le raisonnement contrefactuel a été formalisé par Lewis [21], qui a fourni une sémantique claire fondée sur des *spheres* et a donné lieu à de nombreux résultats en argumentation [30], causalité [27] et raisonnement hypothétique [19]. La question a été récemment revisitée par [7] qui a fourni une nouvelle formalisation en *Answer Set Programming* [10]. Explorer des scénarios contrefactuels implique de pouvoir analyser des variants produits par la simulation en mode*rejeu* qui est disponible avec l'outil SWORD, et peut être modifié à cet effet.
— *Granularité :* plusieurs travaux se sont intéressés au lien entre le nombre de relations causales au sein d'un ensemble d'événements et l'importance perçue de cet événement [24, 32, 31]. Ces travaux nous permettent de formuler des heuristiques fondées sur la centralité de certains événements par exemple, et de les regrouper en un seul événement composé, de niveau sémantique plus élevé. D'autres heuristiques, spécifiques au domaine, ont aussi été explorées.

3. https ://github.com/jff/TeLLer

3.2 SWORD, une simulation constructive pour l'entraînement militaire

Le logiciel de formation que nous utilisons repose sur une simulation constructive, qui est utilisée pour entraîner les personnels de commandement de brigade et de division à l'aide de scénarios de conflits de grande envergure tels que des opérations de déstabilisation, des attaques terroristes, ou des catastrophes naturelles. Il simule des situations variées dans des environnements réalistes, et permet aux entraînés de mener des milliers d'unités autonomes sur un terrain virtuel. Des agents peuvent recevoir des ordres d'opérations à mener donnés par les entraînés, et les exécutent en adaptant leur comportement au fur et à mesure de l'évolution de la situation.
Les modèles qui capturent ces comportements sont composés d'algorithmes qui permettent aux agents de percevoir, se déplacer, de communiquer et de tirer, et de la description des capacités de l'équipement correspondant, stockée dans une base de donnée. La base de donnée pour une session de simulation contient trois types d'informations :

— **Données des éléments physiques :** La constitution des unités est décrite. La simulation étant constructive, la plupart des caractéristiques des équipements ou unités sont décrites par leurs effets ou leurs capacités.
— **Données d'initialisation du scenario :** ceci inclut les informations suivantes : terrain, ordre de bataille, météo, données fournies par la simulation telles que les événements, connaissances obtenues par les agents.
— **Données générées par la simulation, décrivant l'évolution de la situation :** elles incluent tous les événements, les connaissances à propos de l'environnement, et tous les rapports liés aux missions.

Toutes ces informations sont présentées aux participants sous forme de messages échangés par les agents durant la session de simulation. Ci-dessous un exemple de ce qui est présenté aux participants :

```
[07:29:47] - Report - ENG.Counter mobility
             platoon: Disembarkment started
.....
[07:30:17] - Report - INF.Mortar troop: Unit
             detected at ...
.....
[07:30:17] - Report - INF.Rifle platoon:
             Unit detected at ...
```

3.3 Processus de traitement proposé

Après avoir traduit les données de SWORD, les éléments de la simulation et les messages échangés en descriptions d'actions formelles, on procède à une première analyse du flot des ressources manipulées au travers de ces actions, sous une forme similaire à ce qui a été proposé précédemment pour la génération narrative [22]. Puis, nous montrons comment le graphe obtenu peut être traité pour soutenir le discours d'un participant qui raconte son expérience. L'outil SWORD est équipé d'une fonction de *rejeu* qui peut être améliorée et intégrer ces analyses graphiques, qui peuvent alors être elles mêmes complétées : les outils traditionnels des postes de commande militaires sont un bon point de départ pour enrichir l'interface d'un tel outil (calques cartographiques affichant des symboles spécifiques), mais afin de vraiment faire sens, certains points de vue devront être utilisés pour chacun des noeuds, afin de refléter la situation, et les relations entre les noeuds. Nous proposons d'utiliser des *Vignettes* pour représenter les noeuds les plus saillants du graphe narratif[4]. La suite de cet article détaille ces différents traitements

4 Construction du graphe narratif

La production de graphe narratif comporte deux étapes : dans un premier temps, un graphe causal brut est construit à partir des rapports de la simulation pour créer un graphe dont les noeuds sont des événements liés par des relations causales, puis des sous-graphes peuvent être extraits par l'utilisateur, et automatiquement retravaillés afin d'être plus lisibles.

4.1 Graphe causal brut

Nous avons produit une description formelle des traces des événements SWORD en actions atomiques, qui correspondent aux noeuds du graphe causal brut. Ces actions expriment ce qui, dans la simulation, a été modifié lors de la production de l'événement. Ce composant fournit à la fois le moyen de traiter les événements et les traduire en actions automatiquement à partir des bases de données décrivant la session de simulation et à la fois la production du graphe causal *brut* décrivant les causes contributives aux événements. L'algorithme effectue un traitement séquentiel de l'état de la session de simulation au cours de son déroulement, ce qui nous permet d'envisager pour l'implémentation finale au sein de SWORD une construction au fur et à mesure du déroulement de l'exercice (actuellement le prototype est un composant séparé). En effet, l'état de la session de simulation est maintenu en permanence, et permet de connaître pour chacun de ses *instants* le statut opérationnel complet de chaque unité, ainsi que les connaissances qu'ont les unités les unes par rapport aux autres. L'implémentation actuelle est faite en Go et le processus de génération du graphe à partir des traces prend environ 500ms sur un ordinateur portable de milieu de gamme pour nos scénarios les plus complexes. Le graphe produit est décrit dans les formats dot et json, ce qui facilite leur traitement avec des outils standard de traitement et de visualisation.
Les différents types d'événements et de rapports produits par la session de simulation (position et état des unités, échanges de tirs, événement de détection, de déplacement des unités, partage de connaissances entre unités,...) ont donc tous été traduits en formules de logique linéaire utilisant l'implication linéaire comme connecteur central. Les diagrammes obtenus à partir de l'analyse du flot des ressources dans la succession d'événements permet ainsi d'obtenir le réseau des causes contributives entre ces événements. La figure 3 montre un exemple de graphe calculé à l'aide d'un exemple très simple : une unité se déplace, et rencontre une zone minée.

4.2 Exploiter le graphe brut : identification de complexes d'événements, requêtes

Les résultats d'analyse narrative sur les exercices réalistes donnent des graphes de très grande taille, même sur des exemples simples. Le scenario nommé Ménil Annelle est le plus gros scenario d'exercice :

Scénario :	Egypt	Sweden	Ménil Annelle
noeuds	1902	5760	11891
liens	4021	12620	22668

Les graphes résultants ne sont donc ni représentables, ni a fortiori manipulables ou interprétables par un utilisateur, même expert. Nous avons travaillé sur des factorisation automatiques de granularité des noeuds présentés ainsi que fait quelques aménagements d'affichage simples.
Nous avons testé des heuristiques de traitement, en commençant par des simplifications (unification des déplacements atomiques par exemple), nous avons travaillé également sur les contraintes de placement des noeuds du graphe pour en faciliter la lisibilité d'un point de vue temporel, et ainsi faciliter la perception de trajectoires événementielles. Nous nous sommes reposés sur des algorithmes et heuristiques classiques de représentation des graphes, et avons identifié cet aspect comme un enjeu futur nécessitant un travail spécifique, intégré aux cas d'utilisations futurs et à des enjeux ergonomiques. En jaune sur la figure 4, on peut voir un noeud représentant un échange de tirs, dont la conséquence est la destruction d'unités. Ce sont des missions ayant entraîné des déplacements, qui ont provoqué des détections mutuelles d'unités, qui ont mené à cet échange.

4. La selection des noeuds du graphe importants à explorer se fera à l'aide de routines automatiques ou au choix de l'utilisateur

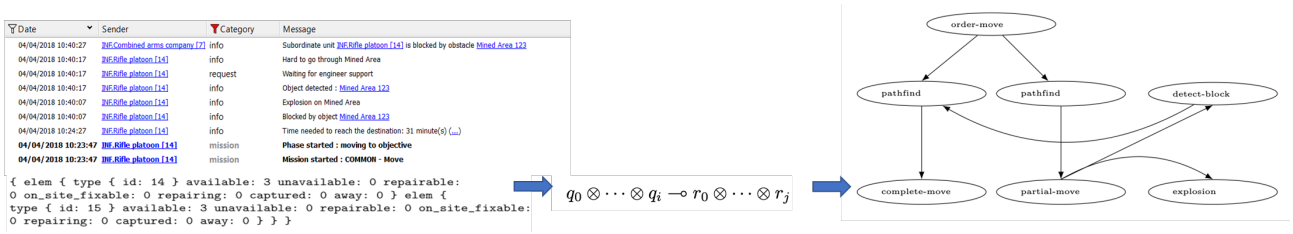


FIGURE 3 – A partir des rapports de la simulation, encodage des événements en actions narratives, en décrivant leurs effets en logique linéaire, puis production du graphe exprimant les relations causales entre ces dernières

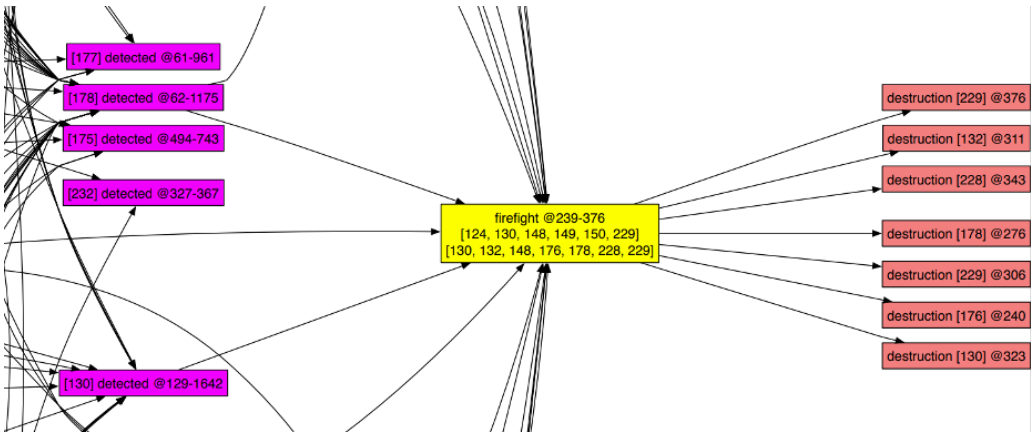


FIGURE 4 – Un extrait de graphe post-traitement automatique sur la granularité des événements présentés

Certaines heuristiques que nous considérions prometteuses, comme l'utilisation des bascules de rapport de force, se sont finalement révélées inefficaces et redondantes pour repérer et factoriser les événements marquants par rapport à une analyse de la centralité causale de certains nœuds. En revanche, des mécanismes simples de rassemblements de complexes d'actions à partir de leur proximité spatio-temporelle, ainsi que mettant en valeur la saillance nécessaire de certains événements d'un point de vue métier (destructions, pertes humaines), se sont révélés très efficaces. Du point de vue des performances, la production de ce graphe narratif peut prendre quelques minutes sur le plus gros scénario mais une fois généré, les requêtes se font dans des temps interactifs. De plus, nous n'avons pas essayé d'optimiser.

Une fois ces traitements appliqués, le graphe *narratif* global obtenu est d'une taille plus à même d'être compréhensible, comme l'indique le tableau ci-dessous, mais pas encore autoporteur (pour vous en convaincre, vous pouvez également consulter de loin le graphe narratif du scénario Egypt, le plus simple de nos scénarios réalistes, sur la figure 5 :

Scénario :	Egypt	Sweden	Ménil Annelle
	Graphe brut		
noeuds	1902	5760	11891
liens	4021	12620	22668
	Graphe narratif		
noeuds	326	973	2072
liens	503	1429	2798

Des filtres sont ensuite appliqués, et offrent des vues partielles du graphe narratif. Par exemple, on peut demander l'histoire d'une unité spécifique, ou inventorier les événements qui forment les causes d'un événement donné. Le graphe en résultant détaille les missions des unités concernées, leurs déplacements, les ennemis détectés, les échanges de tir et les dommages. Les vues peuvent alors être enrichies de vignettes décrivant les contextes opérationnels.

5 Vignettes pour l'enrichissement du graphe narratif

Notre étude nous a menés à identifier quatre grandes familles de vignettes :

— les synthèses automatiques de la situation tactique courante et de la manœuvre en cours ;

— des analyses métier permettant à un officier d'une cellule d'un poste de commandement d'accéder en un instant aux principaux indicateurs ;

— les vues plus abstraites fondées directement sur des critères décisionnels doctrinaux ;

— des vues permettant des calculs prévisionnels, de délais par exemple, prenant en compte la situation tactique.

Nous donnons dans cette partie quelques exemples de vignettes. Les vignettes ont été validées par des experts.

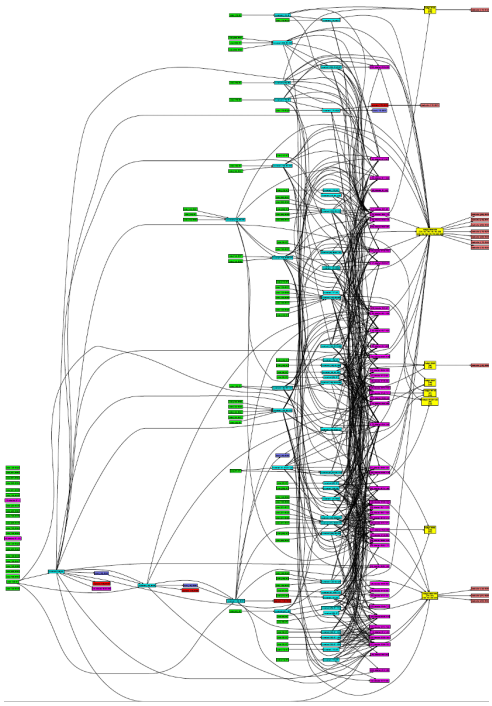


FIGURE 5 – Le graphe narratif complet pour le scénario réaliste le plus simple, Egypt (pas encore autoporteur !)

5.1 Synthèse : cartographie de l'avancement de la manoeuvre

Après avoir étudié la façon dont on travaille au sein d'un poste de commandement, en particulier lors de la préparation de *briefing*, il est apparu nécessaire de générer une cartographie de l'avancement des troupes contenant les principales lignes de bataille (FLOT (*Forward Lines of Own Troops*), la LOA (*Lines Of Attack*), LC (*Lines of Contact*), ...). La simulation permettant de calculer la situation tactique attendue, il est possible, en la comparant à la situation courante, de proposer un calcul de l'avancement des missions en cours et de proposer des indicateurs de succès des missions. Un exemple d'une telle vignette est présenté figure 6.

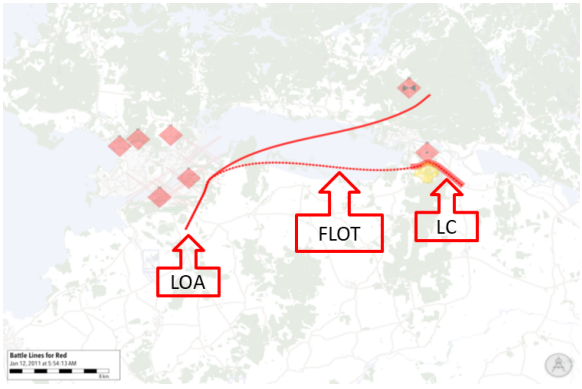


FIGURE 6 – Vignette de Synthèse : cartographie de l'avancement de la manoeuvre

5.2 Synthèse : cartographie contextualisée de l'emprise terrain des troupes

Grâce à un travail de contextualisation des données capacitaires brutes des unités réalisé par la simulation, il est possible de calculer des valeurs réalistes des capacités des unités (portée de tir, de portée de capteur, ...) en les combinant à des données plus abstraites. Par exemple, la portée de détection effective est fonction non seulement de sa portée théorique mais également de la vitesse de déplacement courante d'une unité, sa mission, de sa posture ou de la dangerosité du lieu.

5.3 Synthèse : cartographie des effets courants des troupes sur le terrain

Doctrinalement, chaque mission est caractérisée par un effet principal attendu et un ou deux effets secondaires. En enrichissant et combinant les vignettes précédentes, il est possible de proposer une cartographie des principaux effets appliqués sur le terrain : acquisition d'information, avance offensive, aide à la mobilité, contre-mobilité, contrôle de zone, etc. Ceci permet d'obtenir une vue simplifiée de la situation et ainsi d'apprécier la manœuvre en cours. Un exemple d'une telle vignette est présenté figure 7.

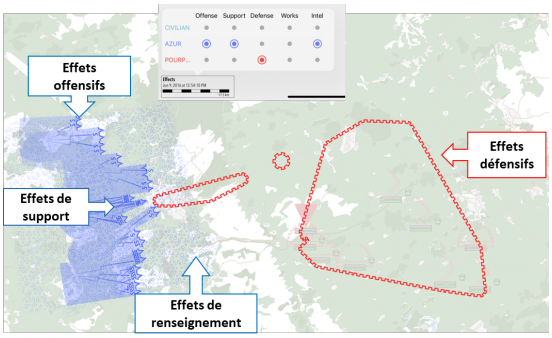


FIGURE 7 – Vignette de Synthèse : cartographie des effets courants des troupes sur le terrain

5.4 Analyse métier : cartographie des rapports de forces locaux

Il est possible également de proposer un autre contexte fondé sur une cartographie de la dangerosité locale de l'ennemi. Une estimation de cette dangerosité est calculée par la simulation pour chaque unité. Doctrinalement, le comportement à adopter étant, selon les missions, fonction de l'estimation du rapport de force et de l'environnement (zone urbaine, zone ouverte), il est donc possible d'extrapoler les risques concernant les missions en cours et d'alerter les officiers en charge. Un exemple d'une telle vignette est présenté figure 8.

5.5 Outils de projection : soutien à l'élaboration d'hypothèses d'évolution de la situation

En reprenant l'exemple précédent, une nouvelle détection ennemie risque de faire basculer le rapport de force local,

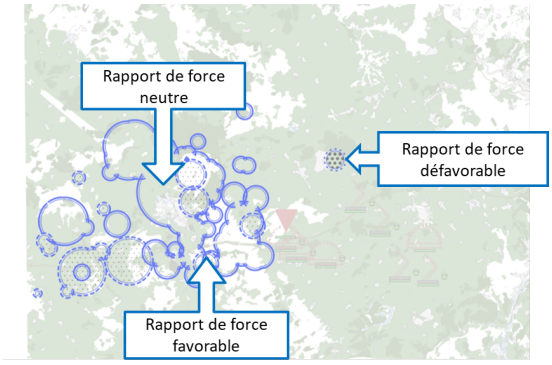


FIGURE 8 – Vignette d'analyse métier : cartographie des rapports de forces locaux

il est peut-être nécessaire de renforcer le dispositif rapidement. En combinant les capacités de déplacement et de tir, le terrain et ses évolutions, les obstacles connus (zones minées, etc.), les connaissances ennemies (et dans l'ambiance de vitesse à adopter localement), la météo courante et prévisionnelle, etc., la simulation peut calculer des délais de déplacement réalistes (et non théoriques). La vue interactive propose une sélection triée des unités les plus à même de soutenir une unité bleue en difficulté ou d'appliquer des tirs sur un ennemi désigné. Le but n'est pas ici d'avoir une estimation précise des délais mais de comparer les délais potentiels entre les unités. Un exemple d'une telle vignette est présenté figure 9.

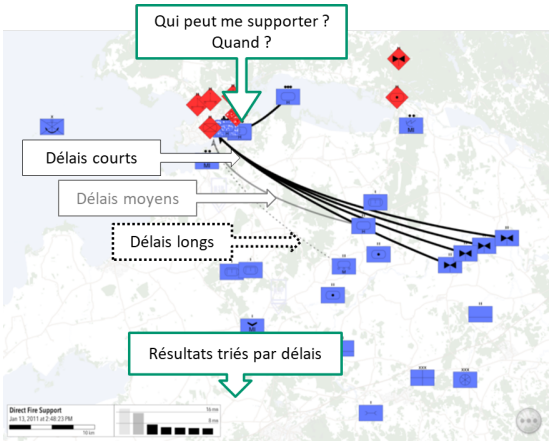


FIGURE 9 – Vignette de projection : soutien à l'élaboration d'hypothèses d'évolution de la situation

6 Cas d'utilisation proposé

Les heuristiques de type bascules de rapport de force (rebondissements, conflits et adversité) ainsi que d'autres que nous avons identifiées lors de travaux parallèles au champ d'investigation principal du projet n'ont pas toutes donné des résultats très intéressants du point de vue de la factorisation des événements dans le graphe, mais elles peuvent toujours se révéler très intéressantes pour une identification des moments saillants d'une simulation, proposant ainsi à l'utilisateur des points d'entrées méritant une investigation plus poussée pour l'analyse de la simulation.
En guise d'illustration de l'intégration future de nos résultats dans la simulation SWORD, nous décrivons ici un cas d'utilisation.

6.1 Scenario de la simulation

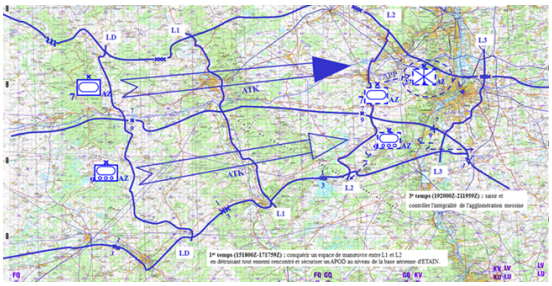


FIGURE 10 – Manœuvre entraîné : Conquérir la zone comprise entre L1 et L2 en neutralisant tout ennemi rencontré. Livrer une ligne de débouché sur la L2 pour le 17 juin

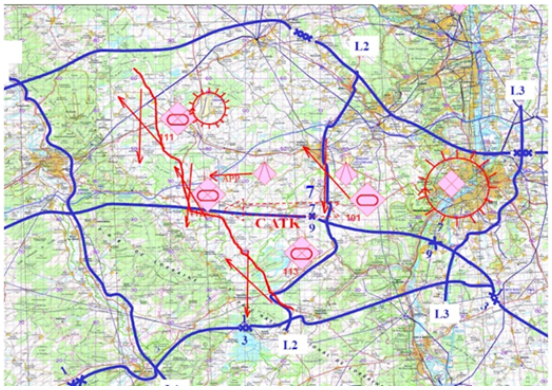


FIGURE 11 – Manœuvre Ennemie : Deux bataillons installés sur la ligne de défense. Réseaux d'obstacles minés entre la Meuse et la ligne de défense rouge. L'objectif de l'ennemi installé en défensive est d'interdire les accès Ouest de Metz tout en essayant de défendre la base aérienne d'Étain.

Les figures 10 et 11 décrivent un scénario de simulation pour respectivement : la manoeuvre objectif des entraînés (en bleu) et la manoeuvre ennemie (en rouge). L'opération bleue est un succès mais il y a de nombreuses pertes. Il apparaît que l'essentiel des pertes est constitué d'un bataillon de reconnaissance. Un rejeu de SWORD peut alors proposer ce bataillon comme élément saillant à étudier, pour essayer d'identifier les causes de ces pertes et déterminer si elles auraient pu être évitées.

6.2 Comprendre les causes des pertes

Pour comprendre l'origine des pertes, nous générons un sous-graphe narratif centré sur l'une des patrouilles d'éclairage détruite (figure 12. Cette extrait permet de comprendre facilement son histoire et de la raconter : La patrouille

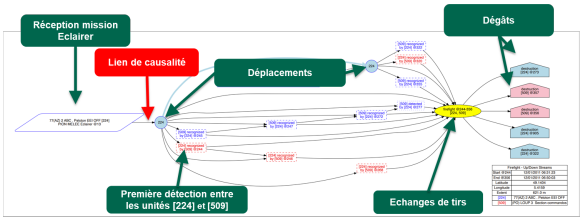


FIGURE 12 – La portion de graphe narratif relatif à l'unité 224

d'éclairage AT4CS [224] reçoit la mission « éclairer » à 11h35 (tic 224), elle se déplace et rencontre l'unité ennemie [509] deux minutes plus tard (tic 245). Les deux unités se détectent mutuellement, échangent des tirs et se créent mutuellement des dégâts entre 11h43 (tic 273) et 12h21 (tic 500). Pour une meilleure compréhension, il semble alors utile de regarder une synthèse de la situation tactique à cet instant. Les vignettes présentées figure 13 proposent une visualisation du contexte tactique lorsque l'unité [224] a été prise à partie : la vignette du rapport indique que l'unité était isolée et les deux autres vignettes montrent qu'elle était en première ligne, éloignée de plus de 8km des plus proches forces de combat.

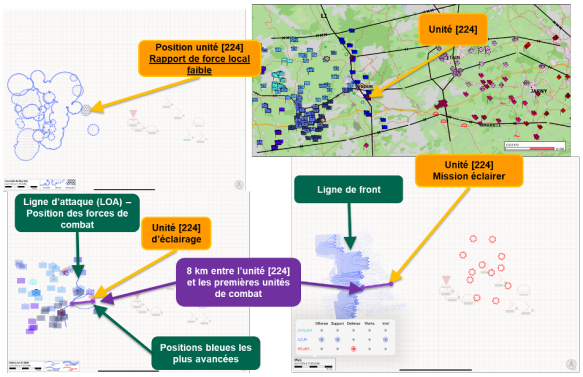


FIGURE 13 – Contexte tactique de l'unité 224 durant l'échange de tir fatal : capacités de soutien plusieurs-contre-un vers l'unité bleue centrale contre les unités rouges (à gauche). Capacités de soutien un-contre-plusieurs de l'unité en bas à gauche vers d'autres unités amies (à droite). Sur ces vignettes, la capacité d'assistance rapide et efficace est codée à l'aide de couleurs : de bonne (verte) à médiocre (rouge). Nous utilisons la *symbologie militaire interarmées* de l'OTAN [1]

7 Travaux proches

Niehaus et al. [26] expliquent que l'utilisation de l'intelligence narrative pour le *sensemaking* est une piste de recherche prometteuse, avec des applications potentielles dans le domaine militaire, de la santé, ou le milieu des affaires. Ils proposent de suivre une approche fondée sur l'apprentissage machine qui prend en compte la structure narrative ainsi que les liens causaux entre les différents événements d'un récit.

Nos travaux diffèrent de [26] dans l'approche : nous voulons que les participants de la simulation soient capables de produire eux-mêmes (avec l'assistance de l'outil) ce qui s'est passé, de leur propre point de vue. Bien que nous ayons envisagé d'incorporer dans le futur localement un peu d'apprentissage machine supervisé pour faciliter les tâches répétitives, nous ne recherchons pas à réaliser un système de storification complètement automatique basé sur ces méthodes.

Le système *Bardic* [6] propose des *narrativisations* pour décrire l'activité dans des domaines complexes afin que l'information devienne accessible à des non-experts. Ce système traduit un fichier de log donné en une théorie logique du premier ordre exprimée avec Impulse [14] [5] et à partir de ces représentations, le système produit un graphe causal à partir des préconditions des actions et de leurs effets. Il y a plusieurs similarités entre ces travaux et les nôtres : les deux systèmes sont fondés sur des logiques, et les deux outils sont orientés vers la production de graphes causaux à partir de données brutes. La manière d'extraire ce graphe brut est un peu différente. Bardic repose sur une approche semblable à STRIPS [16] qui fonctionne sur des modification d'états. Nous travaillons à partir d'une représentation des ressources et de leur consommation, ce qui nous fournit une représentation plus détaillée et générique (les états aussi peuvent être encodés sous forme de ressources), plus à même de représenter les liens de causalité au sein du réseau, moins ad-hoc.

8 Perspectives et travaux futurs

8.1 Reception du projet

Suite à de multiples démonstrations et présentations du projet auprès des opérationnels et de la DGA, des profils variés se sont montrés très intéressés par les résultats. Les équipes en charge du programme de formation SOULT ont commandé des « smart diagrams » pour la supervision d'exercices mobiles et l'AAR. La section EMAT, en charge de l'analyse et de la recherche opérationnelle, a également commandé des schémas intelligents liés à l'amélioration du renseignement. Cette section a pour mission d'analyser les données numériques, tant organiques qu'opérationnelles, pour présenter des vues objectives et consolidées.

8.2 Intégration dans SWORD

L'algorithme de construction du graphe brut procède séquentiellement sur l'état de la simulation au fur et à mesure du temps, ce qui laisse envisager une construction et une sauvegarde du graphe au fur et à mesure du déroulement de la simulation (au travers d'une intégration avec SWORD au lieu d'un composant dissocié). Cet état maintient à chaque instant de la simulation l'ensemble du statut opérationnel des unités de la simulation ainsi que les connaissances des unités les unes par rapport aux autres. Outre l'impression de gain de performance qu'une intégration plus synergique

5. Impulse fournit un cadre temporel (intervalles de Allen [5]) et épistémique ([34])

à SWORD pourrait procurer à l'utilisateur final, cela permet également de prévoir des cas d'utilisation différents (consultation des graphes narratifs en cours de simulation) et une exploration contrefactuelle en rejeu de session de simulation, qui permettrait à l'utilisateur d'aller au delà d'une analyse fondée sur des causes contributives en construisant différents graphes à partir de moments donnés de la simulation. Il pourrait ainsi retrouver les causes "réelles" (actual causes) au sens de [20], qui décrit bien (et caractérise formellement) comment causes et responsabilités telles que perçues par les humains font intervenir non seulement le raisonnement sur ce qui s'est produit, mais aussi sur ce qui aurait pu être si certains événements ou choix n'avaient pas eu lieu.

8.3 Perspectives de recherche

Plusieurs pistes de travaux futurs ont été identifiées à l'issue du projet. Des travaux concernant l'introduction de la notion de ressources dans le paradigme sous-tendant ASP (sémantique *stable models*) ont été envisagés un temps. Cette proposition permettrait à terme de combiner les avantages d'une représentation des ressources qui a fait ses preuves en narration computationnelle [23] avec un langage de programmation logique de niveau industriel. À terme cela faciliterai l'intégration de travaux similaires aux nôtres dans d'autres applications que SWORD. Nous avons également identifié plusieurs pistes pour l'amélioration du graphe narratif : par exemple, des techniques de classification ont été envisagées dans l'objectif de déterminer la probabilité d'apparition d'une séquence d'actions, de détecter les séquences peu probables, voire de prédire la suite d'une séquence, avec comme objectif à terme d'identifier des anomalies. Enfin, nous désirons combiner des techniques d'énumération de motifs dans les graphes (tels que [12]) avec des motifs dynamiques issus de la narratologie structurale pour proposer des factorisations d'événements d'un niveau sémantique élevé.

9 Conclusion

Le projet STRATEGIC a permis d'explorer l'utilisation d'outils de storification fondés sur la logique linéaire pour le *debriefing* de sessions de formation par la simulation. Ces outils assistent la construction de supports pour les personnes entraînées et leurs tuteurs, mêlant des représentation causales sous forme de diagrammes à des vues instantanées décrivant le contexte à des moments choisis (sous forme de vignettes). Nous remarquons que ces techniques peuvent être adaptées à d'autres types de simulation que celle que nous considérons, qui relève du domaine militaire. Dans une simulation SWORD, les graphes narratifs obtenus permettent notamment :

— d'expliquer l'histoire d'une unité ;

— de comprendre le rôle de toutes les unités impliquées dans une phase donnée, par exemple des échanges de tir, et évaluer leur importance dans ce contexte ;

— de visualiser le rôle des unités dans la manoeuvre, et évaluer leur importance (une mission peut avoir des conséquences multiples, ou n'avoir aucun effet) ;

Le développement du prototype et ses performances en termes de passage à l'échelle sont représentatifs de l'intégration finale : le projet prend en compte toutes les missions, mises à part quelques-unes liées à la logistiques. En ce qui concerne le graphe narratif, nous avons identifié les pistes pour permettre aux utilisateurs d'améliorer son exploitation par une intégration dans SWORD : points d'entrée des requêtes grâce à la proposition de moments-clés, modifications du mode rejeu pour l'exploration de scénario contrefactuels à partir de moments-clés).

De plus, afin d'appréhender le contexte tactique global pour un événement du graphe, nous avons proposé des vignettes permettant d'avoir une vision multifacette de la situation tactique à un moment donné :

— un calcul des capacités des unités au sol en fonction du contexte tactique (mission en cours et vitesse de l'unité, météo, expérience de l'unité, ...)

— un calcul des principaux effets appliqués sur le terrain en fonction des missions assignées aux unités ;

— un calcul du rapport de force (ou des forces locales des unités) à partir de leur connaissance de l'ennemi ;

— un calcul des principales lignes tactiques (FLOT, LC, LOA).

Les premiers résultats sont prometteurs et certains travaux ont déjà été commandés par l'armée dans un cadre qui dépasse largement le cadre du *debriefing* post-session de simulation. L'armée de terre française a commandé de nombreux diagrammes intelligents dans un cadre de formation, mais aussi pour l'évaluation du renseignement. À l'avenir, nous pensons que ces travaux pourraient également être utilisés au sein de systèmes d'aide à la décision et d'alerte. L'approche de *debriefing* narratif peut également être transférée à la simulation dans d'autres domaines.

Remerciements

Le projet STRATEGIC (2017-2021) a été financé par la Direction Générale de l'Armement au travers du programme ASTRID Maturation.

Références

[1] Nato joint military symbology (MIL-STD-2525D), 2014. Department of Defense Interface Standard.

[2] Mitchel Y Abolafia. Narrative construction as sensemaking : How a central bank thinks. *Organization Studies*, 31(3) :349–367, 2010.

[3] Sanne Akkerman, Wilfried Admiraal, and Jantina Huizenga. Storification in History education : A mobile game in and about medieval Amsterdam. *Computers & Education*, 52(2) :449–459, 2009.

[4] G. Allen and R. Smith. After action review in military training simulations. In *Proceedings of Winter Simulation Conference*, pages 845–849, 1994.

[5] J. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, (11) :832–843, 1983.

[6] Camille Barot, Michael Branon, Rogelio E Cardona-Rivera, Markus Eger, Michelle Glatz, Nancy Green, James Mattice, Colin M Potts, Justus Robertson, Makiko Shukonobe, et al. Bardic : Generating multimedia narrative reports for game logs. In *10th International Workshop on Intelligent Narrative Technologies*, 2017.

[7] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. Event-based and scenario-based causality for computational ethics. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pages 147–155, 2018.

[8] Anne-Gwenn Bosser, Marc Cavazza, and Ronan Champagnat. Linear Logic for non-linear storytelling. In *ECAI 2010*, volume 215 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2010.

[9] Anne-Gwenn Bosser, Pierre Courtieu, Julien Forest, and Marc Cavazza. Structural analysis of narratives with the Coq proof assistant. In *Proceedings of Interactive Theorem Proving - Second International Conference (ITP-2011)*, 2011.

[10] Gerhard Brewka, Thomas Eiter, and Mirosław Truszczyński. Answer set programming at a glance. *Commun. ACM*, 54(12) :92–103, 2011.

[11] Thomas Cabioch, Ronan Champagnat, Anne-Gwenn Bosser, Jean-Noël Chiganne, and Martin Dieguez. Timing Interactive Narratives. In *2019 IEEE Conference on Games (CoG)*, pages 1–8, August 2019. ISSN : 2325-4289.

[12] Raphaël Charbey, Cécile Bothorel, and L. Brisson. Énumération de motifs dans un graphe d'évolution de communautés (Pattern mining in community evolution graph). In *MARAMI*, 2020.

[13] Giuliana Dettori and Ana Paiva. *Narrative Learning in Technology-Enhanced Environments*, pages 55–69. Springer Netherlands, Dordrecht, 2009.

[14] M. Eger, C. Barot, and R. M. Young. Impulse : a formal characterization of story. In *CMN'15*, 2015.

[15] Ruth M Fanning and David M Gaba. The role of debriefing in simulation-based learning. *Simulation in healthcare*, 2(2) :115–125, 2007.

[16] Richard E. Fikes and Nils J. Nilsson. STRIPS : A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2(3) :189 – 208, 1971.

[17] Gerard Genette. *Narrative Discourse : An Essay in Methods*. 1979.

[18] Jean-Yves Girard. Linear logic. *Theoretical Computer Science*, 50 :1–102, 1987.

[19] J. Y. Halpern. Hypothetical knowledge and counterfactual reasoning. *International Journal of Game Theory*, 28(3) :315–330, 1999.

[20] Joseph Y. Halpern. *Actual Causality*. The MIT Press, 2016.

[21] D. Lewis. *Counterfactuals*. Blackwell, 1973.

[22] Chris Martens, Anne-Gwenn Bosser, Joao F Ferreira, and Marc Cavazza. Linear logic programming for narrative generation. In *International Conference on Logic Programming and Nonmonotonic Reasoning*, pages 427–432. Springer, 2013.

[23] Chris Martens, Joao F Ferreira, Anne-Gwenn Bosser, and Marc Cavazza. Generative story worlds as linear logic programs. In *Seventh Intelligent Narrative Technologies Workshop*, 2014.

[24] Lawrence J Mazlack. Granular causality speculations. In *Fuzzy Information, 2004. Processing NAFIPS'04. IEEE Annual Meeting of the*, volume 2, pages 690–695. IEEE, 2004.

[25] Tim Miller. Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence*, 267 :1–38, February 2019.

[26] James Niehaus, R. Michael Young, Scott Neal Reilly, Peter Weyhrauch, and James Tittle. Towards intelligent narrative-based interfaces for information discovery. In *10th International Workshop on Intelligent Narrative Technologies*, 2017.

[27] C. L. Ortiz. Explanatory update theory : Applications of counterfactual reasoning to causation. *Artificial Intelligence*, 108(1) :125 – 178, 1999.

[28] Judea Pearl. *Causality : Models, Reasoning and Inference*. Cambridge university press, 2009.

[29] M. O. Riedl. Computational narrative intelligence : a human-centered goal for artificial intelligence. In *Proceedings of the CHI 2016 Workshop on Human Centered Machine Learning*, 2016.

[30] C. Sakama. Counterfactual reasoning in argumentation frameworks. In *Computational Models of Argument - Proceedings of COMMA'14*, pages 385–396, 2014.

[31] Tom Trabasso and Linda L Sperry. Causal relatedness and importance of story events. *Journal of Memory and Language*, 24(5) :595 – 611, 1985.

[32] Tom Trabasso and Paul van den Broek. Causal thinking and the representation of narrative events. *Journal of Memory and Language*, 24(5) :612 – 630, 1985.

[33] Susan W. van den Braak, Herre van Oostendorp, Henry Prakken, and Gerard A. W. Vreeswijk. Representing narrative and testimonial knowledge in sensemaking software for crime analysis. In *Proceedings of the 2008 Conference on Legal Knowledge and Information Systems*, pages 160–169. IOS Press, 2008.

[34] H. van Ditmarsch, J.Y. Halpern, W. van der Hoek, and B.P. Kooi. *Handbook of Epistemic Logic*. College Publications, 2015.

Session 3 : Textes

Apprentissage automatique pour la surveillance de marques

J. Tytgat[1, 2], G. Wisniewski[1], A. Bétrancourt[2]

[1] Université Paris Cité, LLF, CNRS 75 013 Paris, France
[2] IPSIDE, 31 100 Toulouse, France

julie.tytgat@etu.u-paris.fr ; guillaume.wisniewski@u-paris.fr ; a.betrancourt@ipside.com

Résumé

Les marques sont un des piliers de la propriété intellectuelle, dont la protection et la surveillance sont des véritables enjeux industriels. L'augmentation massive des dépôts de marques motive l'utilisation de techniques d'apprentissage automatique pour assister dans cette surveillance, traditionnellement assurée par des processus manuels particulièrement coûteux. Cet article présente plus formellement la tâche induite – en insistant sur les problématiques et contraintes industrielles –, puis introduit et discute les méthodes déployées et premiers résultats obtenus.

Mots-clés

Propriété intellectuelle, marques, legaltech, mesure de similarité

Abstract

Trademarks are a cornerstone of intellectual property. Their protection and monitoring are major industrial challenges. The massive increase of trademark filings advocates for the use of NLP techniques in this protection, usually performed via manual and laborious processes. This article presents more formally the resulting task – insisting on the industrial constraints –, then introduces and discusses the methods used and obtained preliminary results.

Keywords

Intellectual property, legaltech, similarity metrics

1 Introduction

L'objectif de ce travail est de présenter une nouvelle application de l'apprentissage statistique : la surveillance automatique de marques, une tâche essentielle pour la protection de la propriété intellectuelle d'une entreprise, chronophage et coûteuse.

Une marque, au sens de la propriété intellectuelle, est un « signe » qui permet de différencier les produits et services d'une entreprise de ceux de ses concurrents. Une marque protégée est un actif qui peut être exploité, vendu, transféré. C'est un actif stratégique qui synthétise l'image d'un produit ou d'un service auprès du public et participe à la valeur d'une entreprise.

Protéger une marque permet à une entreprise d'agir contre des contrefacteurs, de créer de la valeur et d'accroître la crédibilité de celle-ci face à ses partenaires. À l'inverse, un conflit entre marques peut avoir diverses conséquences que ce soit sur l'image de la marque, ou une perte de revenus. C'est pourquoi le travail de surveillance autour de cet actif est essentiel : il faut pouvoir détecter parmi les nouveaux dépôts ceux qui risquent de remettre en question l'avantage conféré par une marque.

Dans le monde, il y a eu 13,4 millions de demandes de dépôts de marques en 2020. Cela représente 1,9 million de plus qu'en 2019 (soit 16,5% d'augmentation) et ce dans un contexte de pandémie [9]. En France, au sein de l'Institut National de la Propriété Industrielle (INPI), on compte plus de 100 000 dépôts sur la même année, soit 7,2% de plus que l'année précédente [1].

Pour apporter des éléments de réponse à cette problématique industrielle de la surveillance, nous présentons dans cet article une tâche de classification binaire visant à détecter la similarité entre deux marques dites *verbales* (cf. 3.1), dans un cadre de surveillance. Nous commençons par présenter dans la partie 2 des travaux similaires récents. Puis un récapitulatif des critères constituant la notion de similarité est proposé à la section 3. Nous décrivons également dans cette section, une première approche visant à déterminer automatiquement cette similarité en nous inspirant de travaux antérieurs. La partie 4 explique comment ces représentations sont combinées au sein de modèles d'apprentissage. Nous présentons également en section 5 nos premiers résultats avec une analyse critique des représentations et des modèles étudiés, et identifions les problématiques scientifiques soulevées par le passage à l'échelle de ce type de méthodes dans un contexte industriel. Enfin, nous concluons dans la section 6.

2 Travaux antérieurs

L'utilisation de technologies d'intelligence artificielle dans le domaine de la propriété intellectuelle est un champ de recherche très actif [2]. Dans le cadre de la similarité entre marques en particulier, on peut identifier différentes approches se focalisant sur les multiples caractéristiques concourant à la définition d'une marque.

Concernant l'aspect sémantique, on peut citer des approches basées sur les représentations vectorielles [8] avec *word2vec* ou [1] sur ontologie. Sur le versant phonétique,

1. https://www.inpi.fr/fr/nationales/chiffres-cles-de-la-propriete-industrielle

[3] utilise une approche neuronale, tandis que [4] se base sur un *soundex pondéré*.
On retrouve également des approches pluri-factorielles, par exemple [6] avec de la *fuzzy logic*, ou [8], qui combine l'analyse de la partie verbale de marques figuratives à une analyse visuelle du logo en général.
À la différence de ces précédentes contributions, nous changeons le paradigme avec lequel la notion de similarité est considérée. Nous présentons nos méthodes dans un cadre industriel de surveillance des marques, avec la volonté de passer à l'échelle, et non dans la reproduction de jugements obtenus en cours de justice, plus courants dans la littérature. Si des méthodes d'apprentissage statistique et/ou neuronale sont par ailleurs déjà utilisées pour calculer les *features* ou similarités sur un aspect de la marque, nous sommes les premiers à notre connaissance à utiliser ces méthodes pour les combiner.

3 Marques et représentation

3.1 Notion de marque

Nous allons commencer par expliquer ce qui définit une marque protégée et le risque de confusion, en nous concentrant sur les marques dites *verbales*, qui sont l'objet du travail que nous décrivons.
Des marques de différente nature peuvent être déposées : du logo à l'hologramme en passant par une marque sonore, une couleur... Ne peut être déposé ce qui ne peut faire l'objet d'une représentation graphique (comme une odeur), quelque chose désignant directement le produit ou service, ou des termes pouvant tromper le consommateur.
Une marque verbale est une combinaison de caractères typographiques standards. Cela peut être un nom, un ou plusieurs mots (Ex. : *Yoplait, Guy Degrenne*), un slogan (Ex. : *Parce que vous le valez bien (L'Oréal)*), des chiffres et/ou des lettres (Ex. : *307 (Peugeot), 24 Faubourg (Hermès)*). [2]

Une marque est enregistrée comme représentant des biens et/ou services spécifiques, dont la liste est l'un des principaux critères la définissant. La protection juridique à laquelle cette marque prétend s'étend donc précisément sur ces produits et services. Pour des questions de clarté et d'efficacité, cette liste est associée à un système de classification, dit de Nice, institué par l'*Arrangement de Nice concernant la classification internationale des produits et des services aux fins de l'enregistrement des marques* en 1957. Le système, mis à jour tous les 5 ans, recoupe actuellement 45 catégories, pour 34 classes de biens et 11 classes de services.
Tous ces éléments permettent de protéger un bien/service du risque de confusion correspondant à une situation dans laquelle une personne tierce utilise une marque qui ne serait pas suffisamment distincte d'une autre. Ce risque est évalué par un juge. Celui-ci prend en compte la similitude des signes, l'identité des produits et la notoriété de la marque. L'impression d'ensemble produite par ses critères auprès d'un consommateur moyennement attentif est le niveau à partir duquel on va juger ce risque de confusion. Le domaine d'activité, symbolisé par les classes de Nice, est donc l'un des facteurs clefs de ce risque de confusion, en plus de la similitude des signes.

Réussir à déterminer la similarité entre deux marques uniquement à partir du nom de celles-ci est donc une approximation de la tâche effectuée par le juge quand il estime le risque de confusion. Lorsqu'il évalue la similarité de deux marques, notre système ne prend pas en compte les critères extrinsèques que peuvent être la notoriété de la marque, la nature du public exposé à la marque, etc. C'est pourquoi notre travail se situe toujours en amont et en complément de celui, plus complexe, fourni par des experts du domaine.

Pour une marque verbale, la similitude entre les signes va s'apprécier au regard des différents critères les caractérisant. Plus précisément, il va s'agir d'une combinaison à divers degrés des aspects phonétique, graphique et sémantique.

Il faut également prendre en compte le caractère distinctif plus ou moins élevé de la marque antérieure en fonction des produits pour la désignation desquels elle a été enregistrée. À titre d'exemple, pour une marque fictive *Michel menuiserie*, enregistrée pour les classes de produits 19 et 20 [3], il convient de se concentrer sur la première partie du signe étant usuellement celle qui porte le caractère distinctif de la marque. Certains mots, par ex. des adjectifs mélioratifs comme *super*, *top*, ne constituent des éléments distinctifs et ce, quels que soient les produits et services. Le caractère distinctif doit donc s'apprécier sur la globalité du signe.

3.2 Représentation d'une marque

Nous allons maintenant présenter rapidement les différentes caractéristiques pouvant, intuitivement être utiles pour déterminer automatiquement la similarité entre deux marques. La caractéristique la plus naturelle sont les classes de Nice qui correspondent à une sur-approximation des produits et services de la marque et peuvent facilement être symbolisées sous forme de représentations vectorielles, auxquelles nous avons adjoint une fonction de similarité (cosinus, distance de Tanimoto [7]). Nous avons toutefois fait le choix de retirer cette caractéristique de notre modèle car nos premières expériences ont montré que celle-ci entraînait un sur-apprentissage. Cette approche présente également le désavantage de ne pas prendre en compte l'hétérogénéité des écarts conceptuels entre paires de classes.

Concernant le signe des marques, on se repose sur des méthodes de comparaison éprouvées : des distances d'édition (Levenshtein, Levensthein pondéré, Jaro-Winkler) et des similarités sur les chaînes de caractères en commun. Concernant ces dernières, on distinguera cependant deux approches : les mesures identifiant la plus longue sous-chaîne commune et celles comptabilisant les sous-mots communs.

2. exemples tirés de l'INPI : https://www.inpi.fr/fr/comprendre-la-propriete-intellectuelle/la-marque/les-differents-types-de-marque

3. Classe 19 : Matériaux de construction non métalliques, Classe 20 : Meubles, glaces (miroirs), cadres,

On choisit de ne pas développer de traitements spécifiques pour les marques de différentes tailles malgré leurs différences de fond. Une première motivation derrière ce choix est, comme pour les classes de Nice, l'observation d'une tendance à la sur-considération de cette caractéristique lors des premiers tests d'apprentissage. La seconde est que la longueur est de fait implicitement prise en compte par certaines méthodes de comparaisons du signe évoquée précédemment, comme les distances d'édition.

Pour améliorer ces mesures de similarités, il est nécessaire, même si l'on ne se trouve pas dans le cadre de marques complexes, mêlant figuratif et verbal, de prendre en compte l'aspect visuel de la marque. Typiquement, on peut citer l'utilisation d'un caractère visuellement proche pour en remplacer un autre, pour un effet de style ; sans oublier qu'une ponctuation ou un nombre peut être typographié en toutes lettres. Pour gérer ces cas, nous générons lors d'une phase de pré-traitement des normalisations possibles, telles que :

- Un " !" peut être réécrit en "i"
- Un "3" peut être réécrit en "e"
- Un "4" peut être réécrit en "A", "for" ou "four"
- Un "2" peut être réécrit en "to", "two"
- Un "@" peut être réécrit en "a" ou "at"
- Un "&" peut être réécrit en "et", "and" ou "N"
- Un "+" peut être réécrit en "t" ou "plus"

Ces règles sont appliquées indépendemment de façon à générer le maximum d'écritures possibles. On sélectionne la variante maximisant le résultat de similarité lors de nos autres traitements.

De même, pour la phonétique, nous nous sommes basés sur des algorithmes déjà étudiés dans ce contexte, soundex et metaphone. On remarque néanmoins que dans le cas de marques déposées en France, on peut concevoir une prononciation française autant qu'anglaise.

4 La tâche de surveillance de marque

Pour reproduire une situation de surveillance, nous avons considéré deux sources de données présentées ci-dessous. Nous parlerons ensuite des modèles d'apprentissage que nous avons testés.

En premier lieu, nous avons accès à une base de marques issues d'un contexte industriel de surveillance. Celle-ci comporte $68\,518$ paires de marques identifiées comme étant similaires. Issus d'un système expert, il s'agit des exemples positifs servant de base à notre apprentissage.

La seconde source de données est la base de marques déposées fournies par l'INPI. L'institut publie hebdomadairement une mise-à-jour de cette base sous une licence libre [4]. Afin de générer des exemples négatifs nous avons, pour chaque occurrence d'une marque surveillée dans notre liste de positifs, sélectionné aléatoirement 10 marques au sein de cette base. Notre corpus contient donc au total $753\,698$ paires dont environ 9% sont similaires.

4. https ://www.inpi.fr/fr/open-data-marques-francaises

Cette méthodologie cherche à simuler la notion de surveillance en contexte industriel, c.-à-d. détecter d'éventuels conflits entre les marques surveillées et une liste de nouvelles marques déposées.

Notre objectif est double : nous souhaitons, dans un premier temps, être capable d'inférer à partir de cet échantillon un système capable de détecter si une paire de marque non encore observée est potentiellement en conflit ; nous souhaitons également, à plus long terme, d'être capable de hiérarchiser la qualité (c.-à-d. la similarité) de ces exemples.

Ce corpus a été généré de manière à reproduire certaines des caractéristiques de la surveillance que nous avons observées : elle permet notamment de simuler le déséquilibre entre positifs (paires de marques similaires) et négatifs (paires de marques non similaires) ; elle a également l'avantage de reproduire l'hétérogénéité du nombres d'exemples qu'une marque en surveillance va produire : une marque très connue, ou avec un signe simple (court et/ou peu distinctif) va générer beaucoup plus de signalements qu'ils soient mérités ou erronés.

L'une d'une caractéristique de cette tâche est que le rapport de classe est déséquilibré : il y a énormément de marques déposées, pour très peu que l'on puisse considérer comme similaires. Cependant, dans ce contexte industriel, le moindre faux négatif peut s'avérer extrêmement critique.

Le ratio de marques similaires ne va pas forcément être homogène : une marque verbale de seulement 3 lettres par exemple va avoir beaucoup de marques similaires sur son signe ; il convient alors d'effectuer la distinction sur les produits et services.

Nous avons d'abord considéré un modèle de régression logistique, qui en plus d'être simple à mettre en place pour une première approche, a le mérite de permettre de contrôler le compromis entre rappel et précision relativement facilement (en faisant varier le seuil de décision), facteur clef dans notre tâche. Nous avons également testé un classifieur par forêt aléatoire qui, si il peut avoir tendance a sur-apprendre, présente l'avantage de fournir un certain degré d'explicabilité dans ses décisions, permettant d'avoir un premier retour critique sur les caractéristiques utilisées.

Nous avons utilisé, dans toutes nos expériences, l'implémentation de ces deux classifieurs de la bibliothèque `sklearn` et estimé les paramètres et les hyper-paramètres de manière standard.

5 Évaluation et discussion

Cette section présente nos premiers résultats et propose une discussion sur les différents points d'intérêt qu'ils soulèvent, à partir desquels sont émises des hypothèses pouvant les expliquer.

Pour évaluer les différentes méthodes évoquées, nous nous sommes servis du score F_1 et de la courbe ROC, deux métriques adaptées pour l'évaluation de classifieurs dans le contexte de problèmes déséquilibrés. Nous avons également considéré l'aspect qualitatif des résultats, toujours dans le cadre d'une coopération avec les experts.

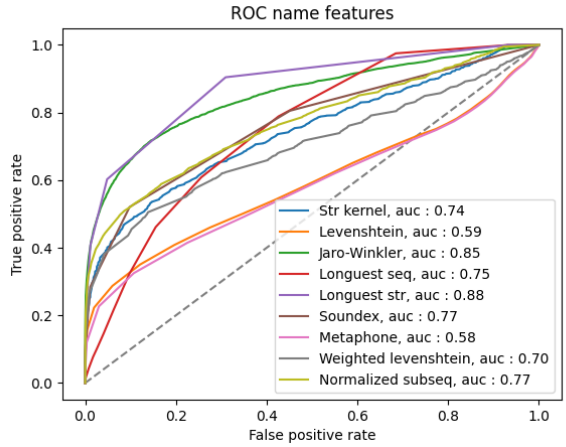


FIGURE 1 – Résultats des différentes *features* en isolation

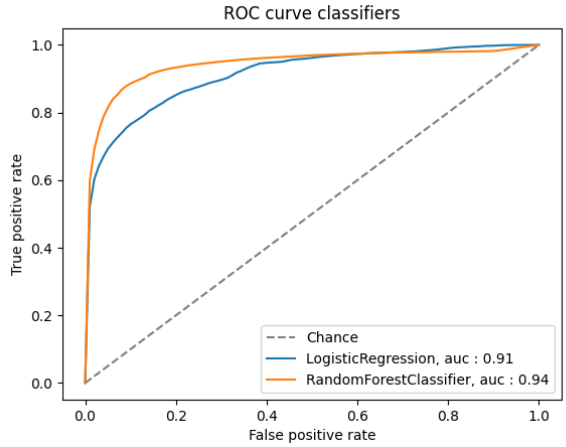


FIGURE 2 – Comparaison des modèles d'apprentissage

5.1 Features entre elles

La Figure 1 introduit, sous forme de courbes ROC, les résultats obtenus par les différentes métriques en isolation. Nous présentons et commentons ici les leçons qui peuvent en être tirées.

Distance d'édition Parmi les distances d'édition, celle obtenant les meilleurs résultats est Jaro-Winkler, loin devant la distance de Levenshtein. A noter que cette dernière s'améliore une fois pondérée [5] : ce qui est cohérent avec les résultats de Jaro-Winkler, qui met l'emphase sur le fait d'avoir préfixes similaires. Ces résultats correspondent à ceux observés dans la littérature [4, 5].

Similarité de sous-chaîne Si on compare la présence du plus long sous-mot avec la plus longue sous-séquence, on peut penser que cette dernière est plus pertinente. Toutefois, dans une optique de minimisation des faux négatifs, la mesure du plus long sous-mot devient aussi valable. Le fait de normaliser la plus longue sous séquence commune par la taille de la plus longue des deux chaînes permet de minimiser le nombre de faux positifs, mais s'avère à l'inverse moins performante quand il s'agit de minimiser le nombre de faux négatifs. Les résultats d'une implémentation naïve d'une fonction de *string kernel* suivent d'ailleurs ceux de la version normalisée de la plus longue sous-séquence commune, en légèrement moins bons.

Phonétique On peut observer que soundex obtient de meilleurs résultats que metaphone. La courbe de résultats de metaphone est très proche de celle obtenue pour la distance de Levenshtein. Cela peut s'expliquer par le fonctionnement de metaphone qui va traduire les séquences observées avec des modifications plus mineures et fines que soundex. L'application de Levenshtein pour les comparer reproduit, au final, le comportement de Levenshtein sans soundex. D'autre part, soundex a initialement été conçu pour des questions de recensement, augmentant l'importance donnée au début du signe, ce qui tend à améliorer les résultats comme observé sur les différentes distances d'édition.

En tentant d'expliciter les différentes *features* pour comprendre le fonctionnement de nos classifieurs, on ne peut que constater que certaines d'entre elles, pourtant classiquement utilisées dans la littérature, n'obtiennent pas les résultats escomptés (Levenshtein). D'autres, bien que très simples, obtiennent de bons résultats (plus longue séquence commune). Bien qu'à interpréter précautionneusement, ces retours sont cruciaux pour l'explicabilité du système, et donc son acceptation par les experts amenés à l'utiliser et les retours critiques qu'ils pourraient formuler.

A partir de ces différentes composantes, on peut transformer une paire de marques en un vecteur de caractéristiques dont chaque entrée représente une similarité ou une distance précédemment mentionnée. Leur analyse nous permettent de déterminer une première approche sur la façon de constituer les vecteurs de caractéristiques. Nos systèmes d'apprentissage reçoivent comme représentation d'une paire de marque un vecteur de taille 4 (édition, sous-chaîne1, sous-chaîne2, phonétique) de manière à varier les types d'information tout en restant explicable.

5.2 Modèle d'apprentissage

La régression logistique a obtenu un score F_1 de 0.60 et la forêt aléatoire de 0.72. Les résultats sont détaillés à la Figure 2.

Le déséquilibre entre les classes est considéré à l'apprentissage par une recherche paramètrique sur la pondération des classes, ainsi que par le choix des métriques d'évaluation.

Nous avons observé, à l'aune des résultats de ces deux modèles, un progrès par rapport à une *baseline* consistant à ne prendre qu'une seule des caractéristiques (la meilleure) sur le nom. Autrement dit, on peut supposer que nos modèles sont bien capables d'apprendre des informations fournies. On peut également déduire que les différentes caractéristiques apportent des informations diverses, complémentaires, dans la définition de la similarité de deux marques.

5. On augmente le poids des premiers mots, et le poids des mots se trouvant dans des positions proches.

Concernant les modèles testés, on ne peut que noter que la forêt aléatoire obtient de meilleurs résultats que la régression logistique, que ce soit en terme de F1 score, ou analysant la courbe ROC, ce qui n'est pas surprenant étant données les capacités de ces deux classifieurs.

5.3 Conclusion sur les résultats

Ces premiers résultats montrent qu'il y a bien un intérêt à utiliser des méthodes d'apprentissage pour déterminer la similarité entre marques. Cependant, on peut également conclure qu'il est essentiel de revoir la façon d'évaluer ces méthodes, avec une plus grande interaction avec les experts. Le fait d'obtenir de bons résultats selon les métriques classiques ne garantit pas dans les faits une applicabilité du système développé, ni sa pérennité dans le temps.
Il semble primordial de prendre en compte ces problématiques avant d'envisager le développement de nouveaux systèmes d'apprentissage dans le traitement de cette tâche. Les nouvelles méthodes proposées se heurteraient à la même difficulté d'évaluation et d'appréciation des capacités effectives.
Cette nécessité de repenser les métriques d'évaluation, et d'améliorer la façon de modéliser les différents aspects des marques, met en évidence des problématiques complexes à l'interface entre statistiques et considérations industrielles.

6 Conclusion

Dans ce travail, nous avons présenté une nouvelle approche pour déterminer la similarité entre marques verbales, basée sur des méthodes d'apprentissage automatique. Les résultats obtenus sont encourageants, mais ouvrent la voie à de multiples questions et améliorations, tant sur les plans scientifiques qu'industriels.

Features Un travail sur les *features* a été amorcé mais peut encore être approfondi, sur tous les critères. Malgré le fait que l'approche naïve pour intégrer les classes de Nice n'a pas porté ses fruits, il n'est pas exclu qu'une méthode plus approfondie contribue à améliorer les performances du modèle. Un travail sur l'aspect distinctif de la marque peut également être considéré : celui-ci est souvent au début, améliorant les résultats des mesures accordant un poids supérieur au début du signe (soundex, weighted Levensthein). On peut toutefois affiner sa prise en compte, avec une mesure statistique sur le poids calculé par *tf idf* par exemple.

Système d'apprentissage Un meilleur score d'optimisation du système que la F1 pour éviter les faux négatifs, problème crucial dans de nombreux champs d'application de l'IA. Le système d'active learning est pour l'instant à l'état de prototype et nécessite davantage de travail pour être déployé et soumis aux experts. Leurs retours seront déterminants pour définir dans quelle direction les efforts doivent se concentrer.

Evaluation L'évaluation doit s'affiner pour mieux saisir les retours qualitatifs des experts. Réussir à définir à partir de combien d'exemples le système est suffisamment fiable et surtout, capable de généraliser à de nouvelles marques présentant potentiellement d'autres caractéristiques. Par ailleurs, le déploiement du système sur le long terme mérite une évaluation continue pour éviter des écueils comme le *data drift*.

Références

[1] F. M. Anuar, R. Setchi, and Y. Lai. Semantic retrieval of trademarks based on conceptual similarity. *IEEE Transactions on Systems, Man, and Cybernetics : Systems*, 46(2) :220–233, 2016.

[2] Leonidas Aristodemou and Frank Tietze. The state-of-the-art on intellectual property analytics (ipa) : A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (ip) data. *World Patent Information*, 55 :37–51, 2018. Advanced Analytics of Intellectual Property Information for TechMining.

[3] Kyung Pyo Ko, Kwang Hee Lee, Mi So Jang, and Gun Hong Park. 2-gram-based phonetic feature generation for convolutional neural network in assessment of trademark similarity. *CoRR*, abs/1802.03581, 2018.

[4] Fatahiyah Mohd Anuar, Rossitza Setchi, and Yu-Kun Lai. Trademark retrieval based on phonetic similarity. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1642–1647, 2014.

[5] Taoxin Peng, Lin Li, and Jessie Kennedy. A comparison of techniques for name matching. *GSTF Journal on Computing (JoC)*, 2(1), 2014.

[6] Rossitza Setchi and Fatahiyah Mohd Anuar. Multi-faceted assessment of trademark similarity. *Expert Syst. Appl.*, 65 :16–27, 2016.

[7] Taffee T Tanimoto. Ibm internal report. *Nov*, 17 :1957, 1957.

[8] C. Trappey, A. Trappey, and Sam C. Lin. Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies. *Adv. Eng. Informatics*, 45 :101120, 2020.

[9] Wipo. *World intellectual property indicators 2021*. World Intellectual Property Organization, November 2021.

Recommandation d'objets d'apprentissage basée sur des objectifs d'apprentissage en utilisant les modèles de plongement de phrases

M. Tounsi Dhouib[1], C. Faron[1], O. Rodriguez Rocha[2]

[1] Université Côte d'Azur, Inria, CNRS, I3S, Sophia Antipolis, France
[2] Teach on Mars, France

{dhouib, faron}@i3s.unice.fr, oscar.rodriguez@teachonmars.com

Résumé

Avec la transformation numérique, l'adaptation et le développement des compétences sont devenus des facteurs majeurs pour améliorer les performances des collaborateurs et des entreprises. Comprendre les besoins des collaborateurs et les aider à atteindre leurs objectifs de développement de carrière est un véritable défi aujourd'hui. Dans ce travail, nous partageons notre expérience pour mettre en place un système de recommandation automatique permettant aux apprenants de trouver des objets d'apprentissage pertinents en fonction de leurs objectifs d'apprentissage. Cette tâche de mise en correspondance se base principalement sur la détermination de la similarité sémantique entre les objectifs et le contenu textuel des objets d'apprentissage. Nous avons évalué de manière comparative trois modèles pré-entraînés de plongement de phrases de l'état de l'art pour la tâche de la recommandation d'objet d'apprentissage. Les résultats des expérimentations montrent que l'utilisation de ces modèles de plongement de phrases dans le processus de recommandation est plus performante que le modèle BM25 d'Elasticsearch classiquement utilisé dans l'industrie.

Mots-clés

Apprentissage intelligent, Recommandation de cours de formation pour les collaborateurs, Systèmes de recommandation.

Abstract

With the digital transformation, adaptation and development of skills have become key factors in improving employee and company performance. Understanding the needs of employees and helping them achieve their career development goals is a real challenge today. In this work, we share our experience to implement an automatic recommendation system that allows learners to find relevant learning objects according to their learning goals. This matching task is mainly based on determining the semantic similarity between goals and the text of learning objects. We comparatively evaluated three state-of-the-art pre-trained sentence embeddings models for the learning object recommendation task. Experimental results show that using sentence embeddings models in the recommendation process outperforms the Elasticsearch BM25 model generally used in industry.

Keywords

Intelligent learning, Recommender systems, Recommendation of training courses for collaborators.

1 Introduction

Aujourd'hui, nous sommes confrontés à un contexte de mutation des modes de travail. Les entreprises doivent répondre aux attentes et aux besoins des collaborateurs en matière de formation et de développement des compétences, en les aidant à choisir la formation qui convient à leur parcours, à leurs compétences actuelles, aux besoins du projet mais aussi à leurs objectifs d'apprentissage.

Dans le cadre d'un projet de recherche collaboratif entre l'équipe de recherche WIMMICS [1] et la société *Teach on Mars* [2], nous souhaitons mettre en place un système de recommandation d'objets d'apprentissage qui correspondent aux besoins et aux objectifs des collaborateurs d'une entreprise. *Teach on Mars* est spécialisée dans l'e-learning et l'apprentissage mobile et développe une plate-forme d'apprentissage spécialisée dans la formation continue au sein des entreprises. L'objectif de *Teach on Mars* est de relier les collaborateurs des entreprises à la formation et aux communautés qui sont essentielles pour améliorer leur travail et leur performance. Notre système de recommandation repose sur deux composants : (i) d'un côté, le profil de l'apprenant qui est construit à partir de l'ensemble des compétences déjà acquises et de l'ensemble des compétences qu'il souhaite apprendre, (ii) de l'autre côté, une base de données qui contient 760 objets d'apprentissage manuellement identifiés par un expert de *Teach on Mars*, principalement écrits en français ou en anglais.

Dans cet article, nous proposons une approche basée sur le calcul de la similarité sémantique entre les objectifs d'apprentissage fournis par l'apprenant et l'ensemble des contenus textuels des objets d'apprentissage de la base. Nous rapportons le résultat des expériences que nous avons menées pour répondre au cas d'utilisation de *Teach on Mars* : nous comparons les performances de trois modèles pré-entraînés

1. https ://team.inria.fr/wimmics/
2. https ://www.teachonmars.com/fr/

de plongement de phrases de l'état de l'art pour la tâche de la recommandation des objets d'apprentissage par rapport au modèle BM25 [3] qui est l'algorithme de similarité utilisé par Elasticsearch pour représenter la pertinence d'un document par rapport à la requête.

Nos principales questions de recherche sont : (i) Quel est le meilleur modèle que nous pouvons utiliser dans notre cas, pour construire des représentations vectorielles des objets et objectifs d'apprentissage afin d'améliorer la qualité des recommandations ? (ii) Quels sont les meilleurs paramètres et méta-données à utiliser pour améliorer la qualité de notre système de recommandation ? (iii) Les modèles de plongement peuvent-ils vraiment améliorer la qualité de notre système de recommandations ?

Le reste de cet article est organisé comme suit. La section 2 donne un aperçu de l'état de l'art sur les systèmes de recommandation dans le domaine de l'éducation. La section 3 détaille notre approche de recommandation pour répondre au cas d'usage de *Teach on Mars*. La section 4 rapporte et discute les résultats de nos expériences menées sur les données de *Teach on Mars*. La section 5 conclut et donne quelques orientations pour des travaux futurs.

2 État de l'art

Les systèmes de recommandation (RS) sont des systèmes de filtrage d'information ayant comme objectif d'aider les utilisateurs à trouver des contenus, des produits ou des services en se basant sur les préférences des autres utilisateurs [13, 3]. En se basant sur plusieurs études sur les RS [7, 6, 8, 18], nous pouvons distinguer quatre techniques de recommandation : (i) recommandation basée sur le contenu qui se base sur l'analyse d'un ensemble des attributs des articles précédemment aimés par les utilisateurs afin de recommander ceux dont le contenu est le plus similaire [21]. Plusieurs travaux dans le domaine de l'e-learning ont utilisé cette technique [11, 15]. (ii) recommandation basée sur le filtrage collaboratif qui se base principalement sur l'analyse de l'utilisateur pour recommander des éléments appréciés par des utilisateurs similaires [10]. Parmi les travaux dans le domaine de l'e-learning, nous pouvons citer [22, 19]. (iii) recommandation basée sur la connaissance qui utilise des ontologies pour suggérer des articles à l'utilisateur en fonction du contexte de l'utilisateur, du contexte de l'article et de leurs relations modélisées par l'ontologie. Parmi les travaux dans le domaine de l'e-learning, nous pouvons citer [1, 17]. (iv) recommandation hybride qui combine deux ou plusieurs techniques citées ci-dessus pour améliorer la recommandation. Parmi les travaux dans le domaine de l'e-learning, nous pouvons citer [2, 9].

Dans ce travail, nous adoptons un approche de recommandation basée sur le contenu et plus précisément sur la mesure de similarité textuelle sémantique (STS) entre le profil de l'utilisateur qui est représenté par un ensemble d'objectifs d'apprentissage et le contenu textuel des objets d'apprentissage. Une méthode de correspondance stricte des mots ou des modèles TF-IDF entre les descriptions textuelles donnerait de mauvais résultats, car elle ne prendrait pas en compte les relations syntaxiques et sémantiques des mots telles que les synonymes ou les polysémies. Pour pallier cela, les techniques de plongement de mots (word embedding) ont été utilisées avec beaucoup de succès dans les tâches de STS. Le plongement de mots est une représentation distribuée des mots qui exploite la sémantique des mots en les faisant correspondre à des vecteurs de nombres réels. L'inconvénient de cette méthode est l'incapacité de ces modèles à prendre en compte le contexte des mots et à approfondir les relations entre les mots de la phrase [14, 12]. L'état de l'art sur les modèles de plongement de mots a récemment évolué vers ce que l'on appelle le plongement de mots contextuel. Les modèles d'apprentissage profond basés sur des architectures de transformeurs, tels que USE (Universal Sentence Encoder) [4], BERT (Bidirectional Encoder Representations from Transformers) [5] et SBERT (Sentence-BERT) [16], ont montré les meilleures performances sur les benchmarks de la tâche STS.

Dans ce travail, nous avons choisi d'évaluer le bénéfice de l'utilisation de ces modèles de plongement contextuel de l'état de l'art par rapport au modèle BM25 et cela pour plusieurs raisons : (i) ils ont démontré leur efficacité pour la tâche de STS, (ii) ils fournissent des modèles multilingues, et permettent ainsi la définition d'une approche générique de recommandation qui peut être étendue à plusieurs langues, (iii) ils n'ont pas besoin de passer par une étape d'extraction d'entités nommées car ils permettent de représenter sous forme de vecteurs des phrases entières avec leurs informations sémantiques.

3 Approche proposée

Dans notre scénario de recommandation, lorsqu'un nouvel objectif d'apprentissage est exprimé par l'apprenant, un ensemble d'objets d'apprentissage les plus pertinents pour cet objectif devrait être automatiquement suggéré. Dans notre approche, nous supposons qu'un objet d'apprentissage est pertinent pour un objectif d'apprentissage si ces deux derniers sont sémantiquement similaires.

La figure 1 présente l'approche que nous proposons. Elle comprend deux étapes principales : (i) la représentation vectorielle des objets et objectifs d'apprentissage, et (ii) le calcul de la similarité sémantique entre objets et objectifs d'apprentissage.

3.1 Représentation vectorielle des objectifs et objets d'apprentissage

Cette première étape permet de représenter chaque objet ou objectif d'apprentissage par un vecteur qui capture la sémantique de ses phrases de telle manière que l'objectif d'apprentissage et tous les objets d'apprentissage pertinents soient proches dans l'espace vectoriel.

3.1.1 Modèles de plongement de phrases

Nous avons commencé par étudier les différents modèles contextuels et multilingues pour générer ces représentations vectorielles :

3. https ://www.elastic.co/fr/blog/practical-bm25-part-1-how-shards-affect-relevance-scoring-in-elasticsearch

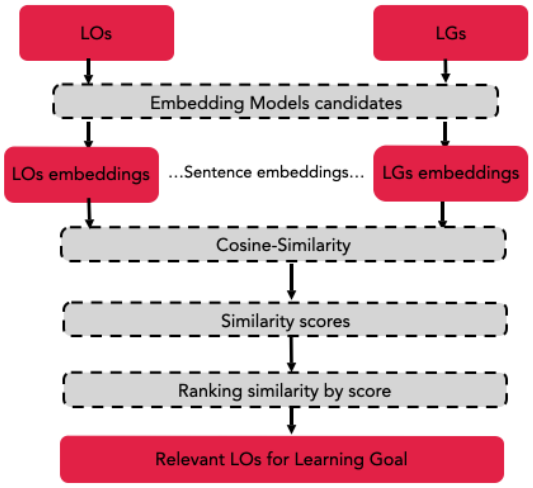


FIGURE 1 – Processus de recommandation des objets d'apprentissage.

Multilingual Universal Sentence Encoder (MUSE). MUSM [4][20] convertit un texte de longueur variable en un vecteur de 512 dimensions. Il est destiné à être utilisé pour des tâches de classification de textes, clustering, recherche de similarités textuelles sémantiques, etc.

Sentence-BERT. Le modèle SBERT représente une modification de l'architecture du modèle BERT pré-entraîné qui utilise des structures de réseau siamois et triplet pour dériver des incorporations de phrases sémantiquement significatives qui peuvent être comparées en utilisant la cosinus-similarité. En effet, le principal inconvénient des modèles BERT est que la recherche de la paire la plus similaire est coûteuse en terme de temps de calcul [16].

SBERT fournit deux modes principaux pour la recherche sémantique : (i) La recherche sémantique symétrique où la requête et les textes du corpus ont la même longueur. Parmi les modèles de cette catégorie, on peut citer *Paraphrase* [5], *Distiluse* [6]. (ii) La recherche sémantique asymétrique pour des requêtes courtes (c'est-à-dire une question ou un mot-clé) mais où les entrées dans le corpus sont plus longues. Parmi les modèles de cette catégorie, nous pouvons citer *MsMarco*.

Pour le traitement des objets d'apprentissage, si nous considérons uniquement leur titre, nous sommes dans le cas d'une recherche symétrique. Mais dans le cas où nous utilisons le contenu textuel de l'objet d'apprentissage, nous sommes plutôt dans le cas de la recherche asymétrique. L'inconvénient de la recherche asymétrique avec SBERT est qu'il n'existe pas de modèle multilingue pour générer les représentations vectorielles, ce qui ne répond pas à nos besoins par rapport aux données que nous avons, qui sont principalement en français et en anglais. Pour cette raison, nous avons décidé d'utiliser les modèles de recherche symétriques, non seulement pour encoder le titre de l'objet d'apprentissage mais aussi pour encoder son contenu textuel. Nous avons identifié deux modèles multilingues à évaluer : (i) Paraphrase fait correspondre les phrases et les paragraphes à un espace vectoriel dense de 768 dimensions et peut être utilisé pour des tâches telles que le clustering ou la recherche sémantique. (ii) Distiluse fait correspondre les phrases et les paragraphes à un espace vectoriel dense de 512 dimensions et peut être utilisé pour des tâches telles que le clustering ou la recherche sémantique.

3.1.2 Méthode de calcul des représentations vectorielles des objets et objectifs d'apprentissage

Représentation vectorielle simple. La représentation vectorielle d'un objet d'apprentissage (ou d'un objectif d'apprentissage) x est un vecteur $V(x)$ qui représente le résultat direct obtenu à partir du modèle utilisé.

Représentation vectorielle moyenne des objectifs d'apprentissage en se basant sur les niveaux d'un thésaurus. Nous considérons ici un thésaurus qui permet d'organiser l'ensemble des objectifs d'apprentissage selon différents niveaux hiérarchiques et dont les feuilles correspondent aux mots clés que nous utilisons pour annoter les contenus textuels. En utilisant un tel thésaurus, nous pouvons enrichir la représentation vectorielle simple d'un texte basée sur les seuls mots clés, avec les concepts des niveaux supérieurs dans le thésaurus.

La représentation vectorielle $V(lg)$ d'un objectif d'apprentissage lg est la moyenne des représentations vectorielles en considérant N niveaux du thésaurus :

$$V(lg) = \frac{1}{N} \sum_{i=1}^{N} V(c_i), \tag{1}$$

où $V(c_i)$ est la représentation vectorielle simple du concept c_i apparaissant dans lg et N est le nombre de niveaux considérés pour enrichir les représentations.

3.2 Similarité sémantique

Afin de pouvoir mesurer la similarité sémantique entre les représentations vectorielles V d'un objectif d'apprentissage LG et l'objet d'apprentissage LO, nous avons utilisé la métrique basée sur le cosinus.

$$sim(LO, LG) = \frac{V(LO) \cdot V(LG)}{||V(LO)|| \cdot ||V(LG||}, \tag{2}$$

Un objet d'apprentissage lo est recommandé pour un objectif d'apprentissage lg si la valeur de $sim(lo, lg)$ est supérieure à un seuil donné.

4 Expérimentations

4.1 Données et protocole d'évaluation

Le jeu de données est composé de deux éléments principaux : les objectifs d'apprentissage (LGs) et les objets d'apprentissage (LOs).

4. https ://tfhub.dev/google/universal-sentence-encoder-multilingual/3
5. https ://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2
6. https ://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

4.1.1 Objectifs d'apprentissage

Nous avons utilisé un référentiel interne à *Teach on Mars* qui a été défini par les experts de contenus de l'entreprise. Ce référentiel contient 166 concepts répartis en trois niveaux : (i) 5 catégories, (ii) 25 thématiques et (iii) 133 mots clés. Le tableau 1 montre le nombre de thématiques et de LOs par catégorie.

TABLE 1 – Référentiel interne des LGs

Catégories	Mots clés par catégorie	Thématiques par catégorie	Los par catégorie
Développement Personnel	33	6	194
Management and Leadership	36	6	181
Responsabilité Sociale d'entreprise	15	4	141
Business Performance	15	3	78
Innovation	39	6	190

4.1.2 Objets d'apprentissage

Un expert de *Teach on Mars* a identifié manuellement sur le Web (crawling) 1350 LOs. Il s'agit d'articles, de vidéos, ou de broadcasts. Chaque LO est associé manuellement par l'expert à une thématique unique du référentiel de *Teach on Mars*. Nous ne considérons dans cette étude que les LOs de type article car le texte est généralement plus long qu'une description de vidéo ou de broadcast. Nous avons obtenu 760 LOs dont 400 en langue française et 360 en anglais. La première étape du traitement consiste à récupérer le titre et le contenu textuel de chaque article.
Voici un exemple d'objet d'apprentissage :

```
Title: Learn to make decisions that last
take this quick test
Learning object text: [...] Everyday we
are faced with a multitude of decisions
that alter our lives in small or signi-
-ficant ways. How you weigh up the pros
and cons of each decision and decide
which direction to take isn/'t always
easy [...]
```

4.1.3 Protocole d'évaluation

Afin de déterminer le meilleur modèle et la meilleure représentation vectorielle des LOs et LGs, nous avons défini 20 expérimentations dont les paramètres sont décrits dans la Table 2. Comme *baseline* nous avons utilisé le modèle BM25 d'Elasticsearch.
Afin de mesurer la correspondance entre les recommandations produites automatiquement et les recommandations produites manuellement par les experts, nous avons utilisé les métriques de précision, rappel, F1 et "précision à N" en considérant les N LOs les mieux classés :

$$P@N = \frac{\text{EP parmi le top } N \text{ des ER}}{N}. \quad (3)$$

où EP représente les éléments pertinents et ER représente les éléments recommandés.
Les expériences ont été menées en utilisant la méthodologie de validation croisée 5 fois. La figure 2 présente la performance de notre système pour chaque paramètre testé en terme de précision, rappel, F1 score et precision@N.

TABLE 2 – Cadre expérimental

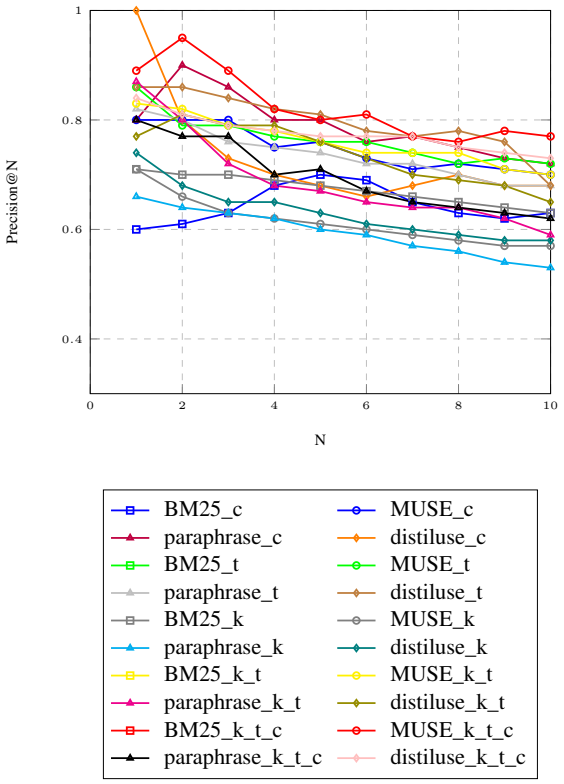
Expérimentations	Modèles	Paramètres
BM25_c	BM25	catégorie
BM25_t	BM25	thématique
BM25_k	BM25	mot clé
BM25_k_t	BM25	mot clé/ thématique
BM25_k_t	BM25	mot clé/ thématique catégorie
MUSE_c	MUSE	catégorie
MUSE_t	MUSE	thématique
MUSE_k	MUSE	mot clé
MUSE_k_t	MUSE	mot clé/ thématique
MUSE_k_t_c	MUSE	mot clé/ thématique/ catégorie
paraphrase_c	Paraphrase	catégorie
paraphrase_t	Paraphrase	thématique
paraphrase_k	Paraphrase	mot clé
paraphrase_k_t	Paraphrase	mot clé/ thématique
paraphrase_k_t_c	Paraphrase	mot clé/ thématique/ catégorie
distiluse_c	Distiluse	catégorie
distiluse_t	Distiluse	thématique
distiluse_k	Distiluse	mot clé
distiluse_k_t	Distiluse	mot clé/ thématique
distiluse_k_t_c	Distiluse	mot clé/ thématique/ catégorie

4.2 Résultat et discussion

Nous obtenons les meilleures performances du système en utilisant le niveau "Thématique" du référentiel interne de *Teach on Mars*, avec la meilleure valeur de précision en utilisant le modèle MUSE et la meilleure valeur de rappel en utilisant le modèle "Paraphrase". La meilleure valeur de F1 est obtenue on utilisant le modèle MUSE avec le niveau le plus précis du référentiel interne c.à.d "mot clé". En se basant sur la performance de notre système en terme de

Expérimentations	Seuil	P	R	F1
BM25_c	1.016	0.193	0.040	0.067
MUSE_c	0.130	0.350	0.370	0.360
paraphrase_c	0.240	0.310	0.280	0.290
distiluse_c	0.080	0.340	0.380	0.310
BM25_t	2.130	0.211	0.114	0.140
MUSE_t	0.190	**0.610**	0.450	0.520
paraphrase_t	0.270	0.400	**0.610**	0.480
distiluse_t	0.190	0.500	0.490	0.500
BM25_k	2.54	0.215	0.191	0.202
MUSE_k	0.220	0.520	0.590	**0.550**
paraphrase_k	0.360	0.390	0.61	0.480
distiluse_k	0.260	0.450	0.580	0.500
BM25_k_t	2.54	0.205	0.165	0.182
MUSE_k_t	0.130	0.370	0.570	0.470
paraphrase_k_t	0.270	0.330	0.590	0.430
distiluse_k_t	0.150	0.380	0.570	0.460
BM25_k_t_c	3.86	0.211	0.150	0.175
MUSE_k_t_c	0.080	0.360	0.560	0.440
paraphrase_k_t_c	0.300	0.380	0.510	0.440
distiluse_k_t_c	0150	0.430	0.510	0.470

(a) Precision, Rappel et F1-scores.



(b) Précision@N

FIGURE 2 – Performances du système en fonction des différents paramètres expérimentés .

P@N, nous avons obtenu le meilleur résultat en rajoutant du contexte au troisième niveau du référentiel c.à.d "mot clé", et ce en utilisant la moyenne des représentations vectorielles des trois niveaux du référentiel, c.à.d "mot clé", "thématique" et "catégorie" avec le modèle MUSE (MUSE_t en vert avec un rond). Nous remarquons aussi que le modèle Distiluse a une meilleure valeur de cette mesure pour la P@1 (distiluse_c en orange) en utilisant le niveau le plus générique du référentiel (catégorie) et P@5 et P@8 en utilisant les thématiques (distiluse_t en marron).

Sans surprise, les modèles de plongement contextuel sont de loin plus performants que les modèles classiques qui se basent sur BM25 pour identifier les éléments pertinents. Le modèle MUSE permet une légère amélioration de la recommandation par rapport aux autres modèles de SBERT. De plus l'enrichissement contextuel des objectifs d'apprentissage en utilisant les relations hiérarchiques du référentiel interne de *Teach on Mars* ou d'un thésaurus d'une manière générale donne clairement de meilleurs résultats que l'utilisation de simples mots clés. L'introduction de la connaissance du domaine dans le processus de recommandation est bénéfique et permet d'améliorer les performances du système même en utilisant des modèles de plongement contextuel bien réputés.

5 Conclusion

Dans cet article, nous avons proposé un système de recommandation des objets d'apprentissage en fonction des objectifs de l'apprenant en nous basant sur le calcul de leur distance de similarité. Nous avons étudié en particulier la représentation vectorielle des descriptions, et évalué les performances du système en utilisant trois modèles différents de plongement de phrases, et en étudiant la meilleure façon pour générer ces représentations vectorielles en nous basant sur le référentiel interne de *Teach on Mars*. Nos expériences montrent que la précision de la recommandation des objets d'apprentissage en utilisant le modèle MUSE et avec l'enrichissement contextuel des objectifs d'apprentissage avec des relations hiérarchiques est de loin la meilleure configuration par rapport aux modèles classiques de recherche de correspondance comme BM25.

Les perspectives de ce travail sont tout d'abord de déterminer la meilleure représentation vectorielle des objets d'apprentissage. Les expériences montrent que la moyenne des représentations vectorielles des objectifs d'apprentissage permet d'obtenir la meilleure précision P@N. A court terme nous allons évaluer ce type de représentation pour les objets d'apprentissage. Deux autres perspectives intéressantes sont d'étudier comment présenter ces recommandations à l'apprenant et de définir des parcours d'apprentissage personnalisés, en prenant en compte les niveaux de difficulté des objets d'apprentissage et le niveau et l'histo-

rique d'apprentissage de l'utilisateur.

Références

[1] Eiman Aeiad and Farid Meziane. An adaptable and personalised e-learning system applied to computer science programmes design. *Education and Information Technologies*, 24(2) :1485–1509, 2019.

[2] Soulef Benhamdi, Abdesselam Babouri, and Raja Chiky. Personalized recommender system for e-learning environment. *Education and Information Technologies*, 22(4) :1455–1477, 2017.

[3] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-based systems*, 46 :109–132, 2013.

[4] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder, 2018.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

[6] Manqing Dong, Feng Yuan, Lina Yao, Xianzhi Wang, Xiwei Xu, and Liming Zhu. Trust in recommender systems : A deep learning perspective. *arXiv preprint arXiv :2004.03774*, 2020.

[7] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. A survey on knowledge graph-based recommender systems. *arXiv preprint arXiv :2003.00911*, 2020.

[8] Deepani B Guruge, Rajan Kadel, and Sharly J Halder. The state of the art in methodologies of course recommender systems—a review of recent research. *Data*, 6(2) :18, 2021.

[9] Mushtaq Hussain, Wenhao Zhu, Wu Zhang, Syed Muhammad Raza Abidi, and Sadaqat Ali. Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review*, 52(1) :381–407, 2019.

[10] Shristi Shakya Khanal, PWC Prasad, Abeer Alsadoon, and Angelika Maag. A systematic review : machine learning based recommendation systems for e-learning. *Education and Information Technologies*, 25(4) :2635–2664, 2020.

[11] Sucheta V Kolekar, Radhika M Pai, and Manohara Pai MM. Rule based adaptive user interface for adaptive e-learning system. *Education and Information Technologies*, 24(1) :613–641, 2019.

[12] Goutam Majumder, Partha Pakray, Alexander Gelbukh, and David Pinto. Semantic textual similarity methods, tools, and applications : A survey. *Computación y Sistemas*, 20(4) :647–665, 2016.

[13] Deuk Hee Park, Hyea Kyeong Kim, Il Young Choi, and Jae Kyeong Kim. A literature review and classification of recommender systems research. *Expert systems with applications*, 39(11) :10059–10072, 2012.

[14] Elvys Linhares Pontes, Stéphane Huet, Andréa Carneiro Linhares, and Juan-Manuel Torres-Moreno. Predicting the semantic textual similarity with siamese cnn and lstm. *arXiv preprint arXiv :1810.10641*, 2018.

[15] Mohammad Mustaneer Rahman and Nor Aniza Abdullah. A personalized group-based recommendation approach for web search in e-learning. *IEEE Access*, 6 :34166–34178, 2018.

[16] Nils Reimers and Iryna Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv :1908.10084*, 2019.

[17] John K Tarus, Zhendong Niu, and Ghulam Mustafa. Knowledge-based recommendation : a review of ontology-based recommender systems for e-learning. *Artificial intelligence review*, 50(1) :21–48, 2018.

[18] María Cora Urdaneta-Ponte, Amaia Mendez-Zorrilla, and Ibon Oleagordia-Ruiz. Recommendation systems for education : Systematic review. *Electronics*, 10(14) :1611, 2021.

[19] Dianhui Wang and Ming Li. Stochastic configuration networks : Fundamentals and algorithms. *IEEE transactions on cybernetics*, 47(10) :3466–3479, 2017.

[20] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv :1907.04307*, 2019.

[21] Philip S Yu. Data mining and personalization technologies. In *Proceedings. 6th International Conference on Advanced Systems for Advanced Applications*, pages 6–13. IEEE, 1999.

[22] Yuwen Zhou, Changqin Huang, Qintai Hu, Jia Zhu, and Yong Tang. Personalized learning full-path recommendation model based on lstm neural networks. *Information Sciences*, 444 :135–152, 2018.

VAGO: un outil en ligne de mesure du vague et de la subjectivité

Benjamin Icard[1], Ghislain Atemezing[2], Paul Égré[1]

[1] Institut Jean-Nicod, CNRS, ENS, EHESS, PSL University, France
[2] MONDECA, France

Résumé

VAGO est un outil en ligne de mesure du vague et de la subjectivité dans le discours, fondé sur une base de données lexicale annotée ainsi que sur des règles expertes. VAGO est développé dans le cadre d'une coopération entre l'INSTITUT JEAN-NICOD (UMR 8129 du CNRS) et la société MONDECA. VAGO repose sur une quadruple typologie des expressions vagues, distinguant la généralité, l'approximation, le vague unidimensionnel et le vague multidimensionnel. Nous utilisons la typologie pour étiqueter les expressions comme marqueurs de subjectivité ou d'objectivité. Dans cette démonstration, nous présentons (i) les motivations de la typologie de VAGO, (ii) la chaîne technologique mise en place dans la réalisation de VAGO, et (iii) l'utilisation de VAGO pour l'aide à la détection d'informations fausses ou peu fiables. Vidéo de démonstration : https://youtu.be/L6cc05SlA5E

Mots-clés

Vague, précision, subjectivité, objectivité, faits vs opinions, qualité informationnelle, fausses informations, TAL.

Abstract

VAGO is an online tool relying on an annotated lexical database and expert rules to provide a measure of vagueness and subjectivity in textual documents. The development of VAGO is the result of the cooperation between the INSTITUT JEAN-NICOD (UMR 8129 of CNRS) and the MONDECA company. VAGO is based on a four-fold typology of vague expressions, distinguishing generality, approximation, one-dimensional vagueness, and multidimensional vagueness. In this demonstration, (i) we introduce the user to the motivations behind the VAGO typology, (ii) we make explicit the technological chain used for the implementation of VAGO, and (iii) we show how VAGO can help in the detection of false or unreliable information. Online demo : https://youtu.be/L6cc05SlA5E

Keywords

Vagueness, precision, subjectivity, objectivity, facts vs opinions, informational quality, fake news, NLP.

1 Introduction

Comment établir si une information est factuelle ou si elle rapporte une simple opinion ? Cette question est déterminante pour évaluer et améliorer la qualité de l'information partagée sur le web et dans les autres médias. La notion d'information factuelle comporte deux composantes : la *véridicité* d'une part, *l'objectivité* de l'autre. Un énoncé peut décrire une situation de façon objective, tout en rapportant des informations fausses (ex : "Emmanuel Macron est né le 12 juin 1960"). Inversement, un énoncé peut rapporter des informations en partie véridiques mais de façon biaisée ou subjective (ex : "la charge virale a baissé de façon spectaculaire"). Si garantir la véridicité d'une information suppose une enquête empirique, garantir l'objectivité en revanche obéit à des normes discursives qui sont du ressort d'une analyse du langage (cf. [8] sur le "conditionnement linguistique" des faits par rapport aux opinions).

Afin de rendre explicites les indices linguistiques permettant d'identifier les aspects objectifs ou subjectifs du discours, cet article présente l'outil VAGO, conçu pour fournir une mesure de la qualité informationnelle des documents textuels, selon deux axes. Le premier axe concerne la mesure du vague par rapport à la précision dans le discours. Le second concerne la mesure de l'objectivité par rapport à la subjectivité dans le discours. Bien que les deux dimensions soient logiquement indépendantes, l'observation principale qui sous-tend VAGO est qu'une sous-classe d'expressions vagues, composée en particulier d'adjectifs vagues multidimensionnels, constitue un marqueur fiable de la subjectivité.

Nous présentons ici les principes de VAGO, son architecture, puis ses applications à la détection de textes soupçonnés de véhiculer des informations fausses ou peu fiables. Cet article est structuré comme suit : la section 2 décrit la typologie utilisée pour l'annotation des termes, ainsi que des règles de quantification du degré de vague et de subjectivité. La section 3 décrit l'implémentation de l'outil suivie de la description d'un scénario d'utilisation en section 4. La section 5 présente ensuite le potentiel d'utilisation de VAGO pour le traitement des fausses informations.

2 Typologie du Vague

Un terme vague est une expression dont le sens est indéterminé du fait de sa relation plurivoque à un ensemble d'interprétations possibles [15, 4]. L'interprétation d'un mot vague est compatible avec un éventail ouvert de significations possibles [13, 14].

2.1 Vague pragmatique vs. sémantique

Dans la lignée de [11, 9], nous distinguons deux variétés principales de vague, à savoir l'imprécision *pragmatique* et l'indétermination *sémantique*. Lorsque le vague est pragmatique, les expressions ont des conditions de vérité définies, mais elles peuvent être utilisées avec relâchement en fonction du contexte. Alors que dans les cas d'indétermination sémantique, les expressions ont des conditions de vérité intrinsèquement incertaines.

En adaptant une typologie proposée dans [5], nous distinguons deux autres types d'expressions : les expressions d'*approximation* (catégorie V_A), et les expressions de *généralité* (catégorie V_G). Les expressions d'approximation comprennent des mots tels que "*environ*" ou "*presque*" qui modifient des expressions précises (lieux, chiffres) et rendent plus large leur signification. Les expressions de généralité comprennent des mots comme "*certains*" et "*ou*", qui ont des conditions de vérité précises mais sous-spécifiques (comparez "*certains navires sont partis*" et "*trois navires sont partis* ").

Du côté de l'indétermination sémantique, on distingue les expressions de vague de degré (catégorie V_D) et les expressions de vague combinatoire (catégorie V_C) (terminologie reprise de [1]). Les premières comprennent principalement des adjectifs unidimensionnels (par exemple, *"vieux"*, *"grand"*), et les secondes des adjectifs multidimensionnels (par exemple, *"beau"*, *"intelligent"*, *"extraordinaire"*, etc.). Dans les deux cas, ces expressions sont fondamentalement sensibles au contexte et donnent lieu à des désaccords ou des divergences d'opinion entre des locuteurs compétents qui peuvent les interpréter de manière très différente [10, 12, 16].

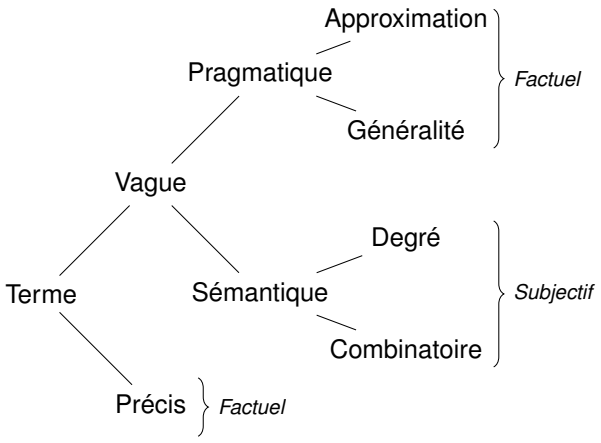


FIGURE 1 – Typologie du vague.

Partant de cette typologie, VAGO fournit un inventaire des expressions vagues et subjectives au sein des énoncés en leur affectant l'une de ces quatre catégories. L'hypothèse principale est alors que les expressions de type V_D et V_C sont des marqueurs de subjectivité au sein des énoncés, en raison de leurs conditions de vérité incertaines ([8] les appellent subjectif simple vs subjectif complexe). En revanche, les expressions de type V_A et V_G ne sont pas traitées comme subjectives, même si elles introduisent une incertitude. En d'autres termes, les expressions sémantiquement vagues sont traitées comme des marqueurs de subjectivité, et les expressions pragmatiquement vagues comme des marqueurs factuels (d'objectivité), de la même façon que le sont les expressions précises (comparez "*J'ai lu un livre de 100 pages / d'environ 100 pages / extraordinaire*").

2.2 Mesure du vague et de la subjectivité

Dans sa version actuelle, la base de données support de VAGO répertorie exclusivement des termes vagues, ce qui signifie que les termes précis ne sont pas inventoriés en tant que tels. Sur la base du lexique, on quantifie le degré de vague et de subjectivité des énoncés du discours.

Les scores de vague et de subjectivité d'une phrase sont définis comme la proportion des marqueurs concernés au sein de la phrase, selon les Équations 1 et 2.

$$R_{vague}(\phi) = \frac{|V_A|_\phi + |V_G|_\phi + |V_D|_\phi + |V_C|_\phi}{N_\phi} \quad (1)$$

$$R_{subjective}(\phi) = \frac{|V_D|_\phi + |V_C|_\phi}{N_\phi} \quad (2)$$

Ici N_ϕ désigne le nombre total de mots dans l'énoncé ϕ et $|V_G|_\phi$, $|V_A|_\phi$, $|V_D|_\phi$ et $|V_C|_\phi$ le nombre de termes relatifs à chaque catégorie du lexique VAGO dans ϕ.

Une phrase qui reçoit une score de subjectivité de 0 est classée comme factuelle, tandis qu'une phrase dont le score de vague est 0 est classée comme précise. Pour des ensembles de phrases (textes), les scores de vague et de subjectivité indiquent respectivement la proportion de phrases ayant des scores de vague et de subjectivité non nuls (voir Figure 4).

2.3 Règles de modulation du vague

VAGO dispose en outre de règles visant à moduler la classification des termes selon le contexte, en particulier :

- Règles d'annulation du vague de degré concernant les syntagmes de mesure "*5 feet*", "*27 years*", etc., quand ils modifient des adjectifs de degré comme "*tall*", "*old*", etc. VAGO traite "*John is old*" comme vague mais "*John is 27 years old*" comme précis.
- Règles d'annulation du vague combinatoire lorsqu'un adjectif multidimensionnel est présent au sein d'un nom consacré ou d'un idiome. Par exemple, VAGO ne relèvera pas les adjectifs "*générale*" ou "*suprême*" dans les expressions "*Direction Générale de l'Armement*" ou "*Cour suprême*", ou encore "*cher*" dans la formule d'adresse "*Cher Pierre*".

3 Implémentation

Cette section présente l'architecture de l'outil VAGO et décrit les modules utilisés pour l'implémenter. Une version de VAGO testable en ligne est disponible au lien https://research.mondeca.com/demo/vago/.

3.1 Architecture

La Figure 2 présente une vue globale de l'architecture de VAGO. Le backend est construit autour du cadre de traitement de la langue GATE [3]. Il contient également une couche de gestion de contenu sémantique selon l'outil d'annotation CA-Manager (CA-M) [2]. L'outil est configuré pour la détection automatique de langue du corpus à traiter avec TextCat [1], pour le moment limitée au français et à l'anglais.

L'interface de VAGO [2], établie en langage JavaScript, fournit une interface graphique pour la représentation des scores à l'aide de deux baromètres. L'un des baromètres représente le degré de vague ou de précision d'un texte. L'autre baromètre indique à quel point le texte évalué exprime une opinion ou un énoncé factuel (la proportion de vocabulaire subjectif vs objectif). L'interface présente également une section détaillée de résultats où sont indiqués pour chaque phrase les marqueurs vagues relevés et la catégorie correspondante (V_X). Il est aussi possible d'utiliser VAGO de manière programmatique avec son API REST.

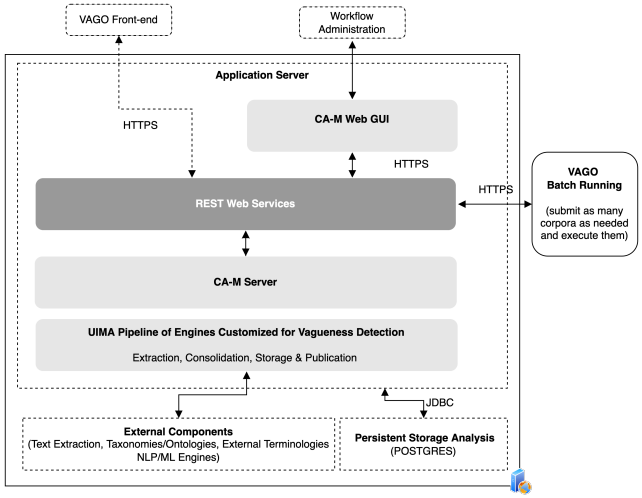


FIGURE 2 – Architecture de l'outil VAGO.

3.2 Gestion de la terminologie

Le lexique (1436 entrées dans chaque langue, version mars 2022) est représenté sous forme de thésaurus en RDF où chaque concept est capturé par le vocabulaire SKOS [3]. Un concept contient les attributs principaux qui sont le nom, les synonymes, et si possible des relation sémantiques avec d'autres concepts. Par exemple, pour le terme "*angry*", étiquetté V_C, on obtient la représentation machine ci-dessous en RDF/Turtle dans le Listing 1. Ces concepts SKOS sont ensuite convertis en gazetteers pour GATE.

```
1  PREFIX sch: <http://data.diekb.fr/id/scheme/>
2  PREFIX :<http://data.diekb.fr/id/vague/comb/>
3  sch:100081
4    a skos:ConceptScheme ;
5    rdfs:label "Concepts vague combinatoire"@fr.
6  :109690
7    a skos:Concept ;
8    skos:altLabel "angrily"@en ;
9    skos:inScheme sch:100081 ;
10   skos:prefLabel "angry"@en, "en colere"@fr.
```

Listing 1 – Exemple en Turtle d'une entrée dans la base de gestion des terminologies de VAGO.

4 Utilisation de VAGO

Supposons qu'on veuille évaluer un document tel que le texte [4] reproduit en Figure 3. Il suffit pour cela de se rendre sur l'interface VAGO accessible en ligne et de copier le texte dans la fenêtre de test (la langue est détectée automatiquement, mais peut aussi être choisie). L'interface propose également quelques exemples de textes pour avoir un aperçu de la sortie de VAGO.

> **"Good** news, Wuhan's corona virus can be cured by one bowl of freshly boiled garlic water. **Old** Chinese doctor has proven it's efficacy. **Many** patients has also proven this to be **effective**. Eight (8) cloves of chopped garlics add seven (7) cups of water and bring to boil., Eat and drink the boiled garlic water, overnight improvement and healing. **Glad** to share this."

FIGURE 3 – Exemple de texte soumis pour évaluation (les erreurs de syntaxe sont d'origine). Entrées VAGO en gras.

Le résultat est décliné en deux sections comme le montre la Figure 4. La première section présente deux baromètres qui indiquent la classification du document selon les axes vague/précis et subjectif/factuel, ainsi qu'un décompte du nombre de phrases vagues ou précises, subjectives ou factuelles, présentes au sein du corpus. La seconde section détaille les différentes phrases décrites comme vagues, ainsi que les marqueurs lexicaux et les catégories ayant présidé à cette classification.

5 Analyse des fausses informations

L'une des applications de VAGO concerne l'évaluation du caractère biaisé ou douteux d'un texte (*fake news*, opinions). Le texte de la Figure 3 constitue un tel exemple de message douteux qui repose sur des déclarations vagues à plusieurs niveaux. Si le vague n'implique pas le faux (ni le faux le vague), il peut donc constituer un indice de fausseté ou de manque de fiabilité, selon la manière dont les termes vagues sont utilisés et leur prévalence.

Afin d'évaluer cette hypothèse, dans [7], la version princeps de VAGO a été appliquée à un large corpus de documents en langue anglaise (environ 28000 documents), réparti en différents corpus académiques étiquetés comme *légitimes* ou comme *biaisés*. Le but de cette étude était de

1. https://www.let.rug.nl/vannoord/TextCat/index.html
2. https://research.mondeca.com/demo/vago/
3. https://www.w3.org/TR/skos-reference/
4. https://www.factcheck.org/2020/02/fake-coronavirus-cures-part-2-garlic-isnt-a-cure.

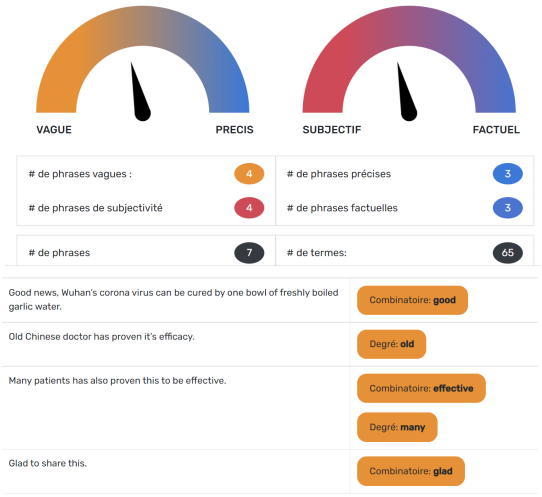


FIGURE 4 – Analyse en ligne par VAGO du texte présenté en Figure 3 (les marqueurs identifiés par VAGO sont aussi rapportés en gras ; en général, les baromètres ne donnent pas des mesures de vague et de subjectivité identiques).

comparer les mesures opérées par VAGO et la catégorisation des documents à l'aide d'un classifieur de type CNN utilisant des méthodes d'apprentissage profond [6]. Les résultats de cette analyse indiquent une corrélation positive entre la présence de marqueurs de vague subjectif et la catégorisation des articles comme biaisés. Ils manifestent en outre que parmi les adjectifs et les adverbes, les entrées déterminantes pour la classification de textes comme biaisés entrent dans la catégorie V_C. Ces résultats confirment que la détection des marqueurs de vague subjectif constitue un indice utile pour le traitement automatique des informations douteuses, même si, comme il convient de le souligner à nouveau, l'emploi de termes relevant du vague subjectif n'implique pas nécessairement que les informations rapportées sont fausses (le contexte doit être pris en compte).

6 Perspectives

L'outil VAGO repose sur une base de données annotée qui continue d'être enrichie et qui appelle encore des améliorations. Dans sa version actuelle, la limite principale de VAGO réside dans le caractère encore restreint du lexique et dans la prise en compte limitée du contexte s'agissant de la modulation du vague des termes. Les catégories existantes demandent à être affinées pour traiter de lexiques spécialisés ou de format brefs (type tweets). Il est également prévu de relever les marqueurs de précision, sachant que dans la version existante de VAGO, ce qui n'est pas vague est considéré comme précis par défaut. Nos recherches en cours visent à combler ces limites en combinant les avancées des réseaux de neurones avec les techniques de traitement du langage naturel actuellement utilisées dans VAGO.

Remerciements

Programmes DIEKB (DGA01D19018444) (Mondeca, CNRS, Airbus) et HYBRINFOX (ANR-21-ASIA-0003). Les auteurs remercient deux rapporteurs anonymes ainsi que le programme ANR-17-EURE-0017 (FrontCog).

Références

[1] W. P. Alston. *Philosophy of Language*. Prentice Hall, 1964.

[2] H. Cherfi, M. Coste, and F. Amardeilh. CA-manager : a middleware for mutual enrichment between information extraction systems and knowledge repositories. In *4th workshop SOS-DLWD*, pages 15–28, 2013.

[3] H. Cunningham. GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2) :223–254, 2002.

[4] P. Égré. *Qu'est-ce que le vague ?* Librairie philosophique J. Vrin, 2018.

[5] P. Égré and B. Icard. Lying and vagueness. In J. Meibauer, editor, *Oxford Handbook of Lying*. OUP, 2018.

[6] G. Gadek and P. Guélorget. An interpretable model to measure fakeness and emotion in news. *Procedia Computer Science*, 176 :78–87, 2020.

[7] P. Guélorget, B. Icard, G. Gadek, S. Gahbiche, S. Gatepaille, G. Atemezing, and P. Égré. Combining vagueness detection with deep learning to identify fake news. In *Proceedings of 24th International Conference on Information Fusion*, page 8, 2021.

[8] E. Kaiser and C. Wang. Packaging information as fact versus opinion : Consequences of the (information-) structural position of subjective adjectives. *Discourse Processes*, pages 1–25, 2021.

[9] C. Kennedy. Vagueness and grammar : The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, 30(1) :1–45, 2007.

[10] M. Kölbel. Faultless disagreement. In *Proceedings of the Aristotelian society*, volume 104, pages 53–73. Oxford University Press Oxford, UK, 2004.

[11] P. Lasersohn. Pragmatic halos. *Language*, 75(3) :522–551, 1999.

[12] L. McNally and I. Stojanovic. Aesthetic adjectives. In J. O. Young, editor, *The Semantics of Aesthetic Judgment*, pages 17–37. Oxford University Press, 2017.

[13] M. Pinkal. *Logic and Lexicon : a Study of the Indefinite*. Springer, 1995.

[14] D. Raffman. *Unruly words : A study of vague language*. Oxford University Press, 2013.

[15] B. Russell. Vagueness. *The Australasian Journal of Psychology and Philosophy*, 1(2) :84–92, 1923.

[16] S. Solt. Multidimensionality, subjectivity and scales : Experimental evidence. In *The Semantics of Gradability, Vagueness, and Scale Structure*, pages 59–91. Springer, 2018.

Session 4 : Explicabilité / confiance

Méthodologie d'anonymisation dès la conception d'un jeu de données en imagerie médicale

J. Clech[1,5], A. Gotlieb[2], F. Sève[3,5], F. Didout[4,5], P. Malléa[1,5]

[1] NEHS Digital, 1 rue Augustine Variot 92240 Malakoff, France
[2] Simula Research Laboratory, KA 23, 0164 Oslo, Norway
[3] Kalhyge, 4-6 rue Truillot 94200 Ivry sur Seine, France
[4] MNH, 331, avenue d'Antibes, 45200 Amilly, France
[5] groupe NEHS, 185, rue de Bercy 75012 Paris, France

jeremy.clech@groupe-nehs.com

Résumé

La recherche en santé s'appuie notamment sur des bases de données d'imagerie médicale. Les données personnelles qu'elles contiennent doivent être évacuées afin d'empêcher toute réidentification ultérieure des patients. Dans cet article, nous présentons notre méthodologie d'anonymisation de données d'imagerie médicale. Les leçons apprises dans cette expérience nous ont permis 1) de créer un premier outil qui peut anonymiser ce type d'imagerie et 2) de mettre à la disposition de la communauté IA cette base de données au travers de la plateforme européenne AI4Europe.

Mots-clés

Imagerie médicale, Apprentissage automatique, Qualité des données, Anonymisation, RGPD.

Abstract

Health research relies in particular on medical imaging databases. The personal data they contain must be removed in order to prevent any subsequent re-identification of patients. In this article, we present our methodology for anonymizing medical imaging data. The lessons learned in this experience allowed us 1) to create a first tool that can anonymize this type of imagery and 2) to make this database available to the AI community through the European platform AI4Europe.

Keywords

Medical imaging, Machine learning, Data quality, Anonymization, GDPR.

1 Introduction

Le Règlement Général sur la Protection des Données (RGPD) apporte à travers ses principes et ses obligations un cadre juridique pour l'exploitation de données à caractère personnel. L'objectif de ce papier est de proposer un retour d'expérience sur la mise en application d'un traitement d'anonymisation sur des données à caractère personnel liée à une base de données d'imagerie médicale : comment conjuguer les contraintes réglementaires, de structuration et mise en qualité des données collectées afin de pouvoir proposer en un temps court des volumes de données et les exploiter à des fins de recherche en IA.

L'imagerie médicale produit annuellement plusieurs dizaines de millions d'examens en France. Ainsi en 2019, 56,7 millions d'actes d'imageries médicales ont été réalisés par les radiologues libéraux[1] [1, p. 134]. Ces données, stockées en France dans des infrastructures informatiques sécurisées contre la violation de données (divulgation, perte et altération), sont un formidable vivier et sont sources d'innovations médicales afin de lutter contre la perte de chance d'un patient en réalisant trop tardivement des examens, d'améliorer la productivité des radiologues en garantissant la qualité et la constance du diagnostic ou encore de personnaliser les soins en fonction du patient, de ses souhaits et de son contexte.

Alors même que l'apprentissage automatique est arrivé à un haut niveau de maturité méthodologique et industrielle, l'accès à ces larges volumes de données est malaisé. En effet, tant en France qu'en Europe, leur accès requiert d'une part des déclarations et autorisations auprès des autorités compétentes (e.g. CNIL) et information auprès des patients et d'autres part que le recours à des sous-traitants (e.g. prestataires, intervenants) garantisse le bon respect des exigences du RGPD. Dès lors, accéder à ces bases de données requiert un investissement important sur les plans juridique et financier et sur un temps long. Les difficultés rencontrées par le gouvernement français pour la mise en place du *Health Data Hub*, avec la remise en cause de l'hébergeur de données de santé Microsoft en raison de risques de transferts de données vers les États-Unis [2], sont un exemple flagrant.

La pandémie du COVID-19 apporte un éclairage fort sur cette problématique : la Chine a pu proposer dès mars 2020 une solution de triage des patients à partir d'un jeu de données composé de 3 191 patients (dont 1 000 non atteints du COVID-19) [3] alors même que de nombreuses sociétés

[1] Ce nombre d'actes ne prend donc pas en compte ceux réalisés par les hôpitaux publics et ni ceux réalisés par les autres spécialités comme par exemple la cardiologie, la médecine nucléaire ou encore la gynécologie.

existent en France (startup et PME établies) en imagerie médicale, et que l'écosystème français maîtrise toute la chaîne de traitements de l'IA (laboratoires de recherche, centres de calculs…).

Nous souhaitons lever ce verrou à l'innovation en proposant un cadre méthodologique permettant un accès à des données de qualité aux acteurs de cet écosystème, dans des délais courts, de manière maîtrisée et respectueuse de la réglementation européenne. Ainsi, nos travaux visent à réaliser dès la conception du projet une collecte anonymisée de données. La finalité de cette démarche a permis de mettre à disposition une base de données d'imagerie médicale de forte volumétrie mais également une preuve de concept avec un outil d'anonymisation automatique de rapports radiologiques.

L'anonymisation est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et ce de manière irréversible. De fait, l'anonymisation permet de sortir du cadre du RGPD au prix d'une perte relative d'information puisque supprimant les données à caractère personnel. Comme rapidement décrit en section 2, de nombreuses méthodes existent afin d'altérer ces données en vue d'écarter la possibilité de réidentification ultérieure. Toutefois, ces dernières sont réalisées *a posteriori* de la collecte et peuvent introduire des biais.

Après présentation en section 3 des motivations et enjeux à la collecte d'une base de scanners thoraciques, nous décrivons en section 4 les éléments clefs du traitement portant sur des données anonymisées. Nous abordons en section iv la conception et les opérations mises en œuvre pour la collecte des données anonymisées. La section 6 est consacrée à la vérification de l'efficacité de l'anonymisation. La section 7 décrit les éléments mis à la disposition de la communauté IA tandis que la section 8 discute de la pertinence de l'approche proposée consistant à définir l'anonymisation des données lors de la définition du projet de recherche. Enfin, la section 9 conclue ce travail et propose quelques perspectives.

2 Rappel des travaux antérieurs

Le RGPD met en place des contraintes, telles que le recueil du consentement préalable des personnes, rendant parfois impossible l'exploitation des données et l'anonymisation est la seule méthode (lorsque ce recueil ne peut être envisagé) permettant de réaliser ces exploitations dans un cadre licite [4]. De nombreuses techniques, rappelées dans [5] ont été développées et existent aujourd'hui. Les approches classiques de suppression [6] et de généralisation [7] ou encore de recodage global [8] visent en premier objectif à supprimer le caractère personnel des données respectivement par suppression pure ou simple ou par réduction de l'espace des valeurs possibles afin de diminuer les valeurs singulières. Ces approches ont été enrichies afin de mieux intégrer la diversité des données dans ces données généralisées comme avec « *Anotomy* » [9] ou bien par des permutations d'une valeur au sein d'un groupe d'individus évalués comme similaire [10].

Ces techniques réalisées *a posteriori* de la collecte sont utiles en fonction des cas d'usage et des spécificités des données à anonymiser mais présentent plusieurs inconvénients comme i) la lenteur de réalisation, car les traitements imposent une intervention manuelle pour les prises de décision critique ; ii) l'impact sur les modèles d'apprentissage car une généralisation trop importante peut diminuer drastiquement la pertinence d'un modèle et que des permutations mal maîtrisées peuvent entraîner un biais d'apprentissage.

Pour contrer ces problématiques et partant du principe que les données ne peuvent pas être anonymisées tout en restant utiles, l'approche de la confidentialité différentielle[2] [11] propose de contourner l'obstacle en empêchant d'accéder directement à la donnée initiale. Ces solutions sont de plus en plus utilisées, notamment par les GAFAM[3], pour garantir la confidentialité. Cette approche est intéressante mais conduit à la production de biais potentiels liés à l'utilisation d'un générateur, qui rendent les approches d'apprentissage automatique inopérantes.

L'Apprentissage Fédéré[4] est utilisé sur les données de santé [12] car proposant également de ne pas accéder directement aux données en permettant de distribuer l'apprentissage sur les différents sites gérant les données et sous la responsabilité des DPD de chacun de ces sites. Cette approche offre un solide niveau de sécurité et de traçabilité des accès mais induit pour ce faire un prérequis technique non négligeable avec la mise en place d'une infrastructure sur chacun des sites concernés. Ceci génère 2 difficultés : une organisationnelle et une de ressources. En effet, il est nécessaire d'impliquer chacune des Direction des Systèmes Informatiques (DSI) des établissements participants afin qu'elles mettent à disposition les données dans une base distincte de celle de production, accorder de la puissance de calcul et autoriser et contrôler les flux.

Nous pensons que de définir dès la conception du traitement de collecte cet objectif clair d'anonymisation permet de minimiser l'utilisation de ces techniques ou du moins d'en avoir une meilleure maîtrise. C'est cette approche de la confidentialité par construction[5] pour les données d'imagerie médicale que nous présentons et défendons dans cet article

3 Genèse de FIDAC

Lors de la première vague de la pandémie de maladies liée à la propagation du coronavirus, il était difficile de déterminer rapidement si un patient était atteint du COVID-19 ou bien d'une autre maladie pulmonaire. Or, à ce stade limité de nos connaissances sur cette nouvelle maladie, il était crucial de pouvoir séparer les flux de patients (covid et non covid) au sein des établissements de santé pour réguler au mieux la propagation de l'épidémie. Toutefois, nous ne disposions ni de tests antigéniques, ni d'autotests permettant une réponse rapide. À cette période, seuls les tests PCR étaient disponibles mais ceux-ci étaient en nombre limité et rendent leur résultat sous 24h. Dans ce contexte, le scanner thoracique a beaucoup été employé car il permet de caractériser la pathologie du patient par la présence de signes radiologiques typiques, comme par exemple de « *crazy-paving pattern* » [13].

[2] *Differential Privacy*
[3] Google Apple Facebook Amazon Microsoft
[4] *Federated Learning*
[5] *Privacy-by-design*

Afin de favoriser l'effort collectif de lutte contre la crise liée à la pandémie, NEHS Digital, la Société Française de Radiologie (SFR) et le Collège des Enseignants en Radiologie de France (CERF) se sont mobilisés pour mettre en place une base de données anonymisées de référence nationale et européenne permettant l'amélioration des connaissances et le développement de solutions innovantes pour le diagnostic, le pronostic et le suivi des conséquences de la COVID-19. Cette base de données anonymisées a été baptisée FIDAC pour ***F**rench **I**maging **D**atabase **A**gainst **C**oronavirus*.

Nous avons mis en place un système de collecte de scanners thoraciques avec des données complémentaires de type clinique, virologique et radiologique sur une cohorte de patients présentant des signes cliniques d'infection au COVID-19. Dès le départ du projet, les partenaires avaient pour exigence que ces données soient anonymisées afin de favoriser les échanges et le partage des connaissances. Au final, la base anonymisée est constituée de l'imagerie de 5 843 patients adultes.

Ces scanners sont enrichis de métadonnées médicales et les données sont anonymisées par les établissements de santé volontaires. Cette base est destinée à être mettre mise à disposition de tiers à des fins statistiques ou de recherche.

Afin de rendre cette base exploitable pour la recherche en IA, nous avons proposé un mécanisme d'anonymisation qui assure l'impossibilité de réidentification d'une personne physique et ce, en considérant les principes de régulation du RGPD tels que définis dans l'article 26 [14]. Il est important de noter que le traitement amont de ces données (de leur collecte à leur anonymisation) doit lui répondre strictement aux exigences du RGPD. Dans le reste de ce document, nous nous focalisons sur le processus d'anonymisation et proposons un retour d'expérience sur son utilisation dans cette base de données d'imagerie médicale.

4 Conception du processus d'anonymisation des données

4.1 Les grands principes

Pour concevoir un processus d'anonymisation pertinent, nous avons suivi les recommandations de l'autorité nationale, c'est-à-dire la CNIL [15] [16] :

1. Supprimer les éléments d'identification directe ainsi que les valeurs rares qui pourraient permettre une réidentification aisée des personnes ;
2. Distinguer les informations importantes des informations secondaires ou inutiles (i.e., supprimables) ;
3. Définir la finesse acceptable pour chaque information conservée (e.g., conserver l'année de naissance des patients n'est pas possible, mais conserver la décennie dans laquelle ils sont nées est acceptable) ;
4. Définir les priorités (e.g. est-il plus important de conserver une grande finesse sur telle information ou de conserver telle autre information ?).

La suppression d'éléments d'identification est triviale lorsqu'ils sont explicites (e.g. nom et prénom) et ceux-ci sont soit écartés lors de la collecte, soit ils sont immédiatement supprimés. Évaluer le degré d'importance des données ainsi que définir leur finesse et priorité relative requièrent de réaliser des arbitrages. C'est pourquoi, il est indispensable de bien définir les objectifs d'études ainsi que le domaine métier concerné.

Ce processus d'anonymisation correspond ainsi à identifier et réduire les données à collecter mais également de baisser leur granularité informationnelle. La conséquence de l'anonymisation correspond donc une nécessaire perte de précision des données.

Cette seule phase de conception est loin d'être suffisante. En ce sens, le RGPD [17, pp. 11-12] définit trois critères qui permettent de s'assurer qu'un jeu de données est véritablement anonyme :

1. La non-individualisation : il ne doit pas être possible d'isoler un individu dans le jeu de données ;
2. La non-corrélation : il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu ;
3. La non-inférence : il ne doit pas être possible de déduire de façon quasi-certaine de nouvelles informations sur un individu.

Cette seconde phase de robustesse de l'anonymisation ne peut pleinement s'évaluer qu'au regard des données collectées. Elle est donc à réaliser *a minima* à la fin de la collecte, bien qu'il puisse être intéressant de la faire en cours de collecte pour disposer de premières tendances. Enfin, les techniques d'anonymisation et de réidentification étant amenées à évoluer régulièrement, il est indispensable d'effectuer une troisième phase de surveillance afin de préserver dans le temps le caractère anonyme des données produites.

4.2 Un processus d'anonymisation par construction

La FIGURE 1 ci-dessous illustre la méthodologie d'anonymisation par construction que nous avons mise en place pour le traitement des données personnelles issues des jeux de données en imagerie médicale. Cette méthodologie, comme nous allons le voir, s'appuie sur les principes énoncés ci-dessus.

En s'inspirant des bonnes pratiques utilisées en confidentialité par construction, nous distinguons cinq phases successives, c'est à dire :

i. la phase de conception du projet de diffusion des données d'imagerie médicale, essentiellement portées par les experts du domaine concerné et par les responsables du projet
ii. la phase de traitement qui porte sur la collecte des données en établissements de santé ;
iii. la phase de vérification de l'anonymisation des données, principalement réalisée par des experts du traitement des données ;

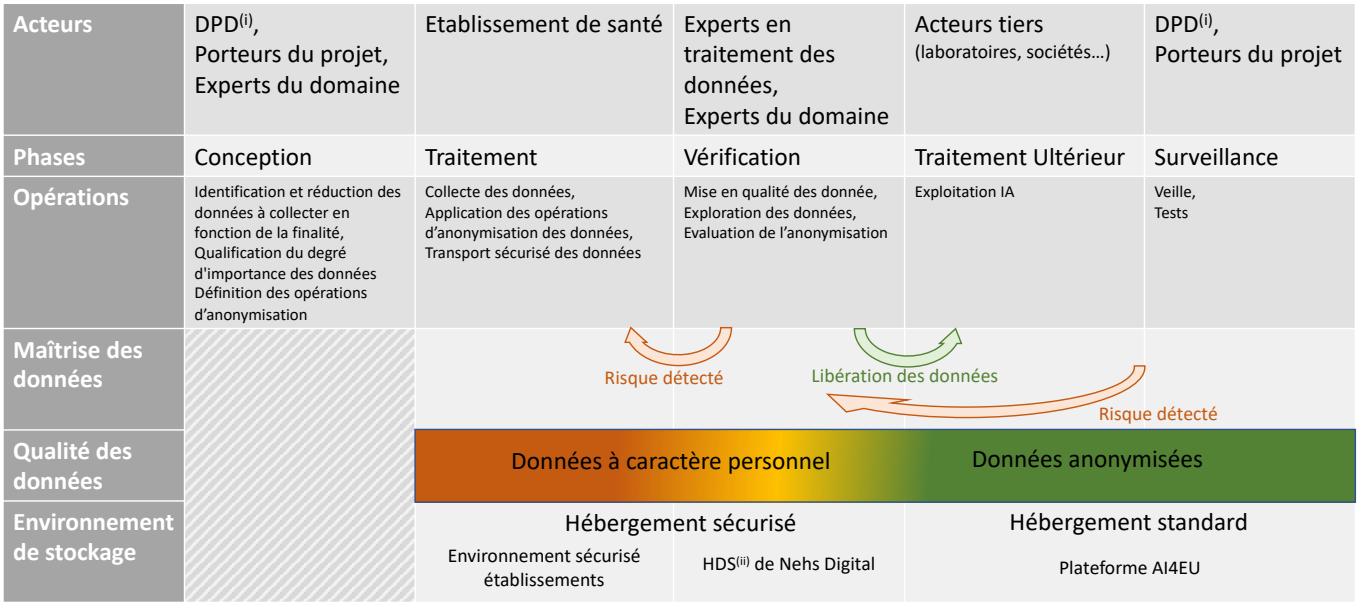


FIGURE 1 - METHODOLOGIE D'ANONYMISATION PAR CONSTRUCTION

iv. la phase de traitement ultérieur des données utilisant massivement l'apprentissage automatique et qui concerne les acteurs tiers tels que les laboratoires privés ou publiques, les sociétés privés qui développent des solutions innovantes, etc. ;

v. la phase de surveillance qui incombe principalement aux Délégués à la Protection des Données (DPD)[6] des entités concernées et qui interagissent de concert avec les responsables de projet.

Les opérations concernées par ces différentes phases sont indiquées dans la figure car elles représentent véritablement le cœur des activités du projet de diffusion des données d'imagerie médicale. Il est intéressant de noter que le basculement entre l'hébergement sécurisé et privé vers un hébergement des données accessible aux acteurs finaux et au public (par exemple, par l'entremise de plateformes d'IA) n'est réalisé qu'après la phase de vérification (libération des données). En effet, il n'est pas rare que cette phase identifie des risques majeurs qu'il faut traiter en appliquant des processus d'élimination ou bien d'offuscation des données.

Les risques identifiés dans les environnements de production sont quant à eux surveillés et conduisent à une amélioration de la phase de vérification.

5 Mise en œuvre du processus d'anonymisation pour la base FIDAC

5.1 Identification et réduction des données à collecter

Pour définir la pertinence des données à collecter, NEHS Digital s'est appuyé sur la Société Française de Radiologie (SFR) et le Collège d'Enseignants Radiologues de France (CERF) qui ont défini, outre la série de scanner et leur compte-rendu radiologique, un ensemble de 4 informations médicales supplémentaires à récolter : (i) l'indication, (ii) le délai entre le début des symptômes et la réalisation du scanner thoracique, (iii) le diagnostic radiologique et (iv) le résultat du test PCR. En outre, 2 informations complémentaires relatives à l'établissement ayant réalisé l'examen ont également été demandées : (v) le nom de l'établissement et (vi) son identifiant.

Un travail de conception a été mené afin de modéliser les opérations d'anonymisation des données, le « transport » de ces dernières ainsi que leur hébergement. La sélection des données à retenir a principalement porté sur les métadonnées présentes dans les images composant les séries de scanner thoracique.

Concernant FIDAC, les données collectées sont : (i) des données médicales définies par la SFR, (ii) le compte-rendu radiologique produit par le radiologue et (iii) l'imagerie médicale au standard DICOM [18].

La finalité est de mettre à disposition un jeu de données conséquent et adapté pour la recherche en IA sur l'identification automatique de la COVID. En s'appuyant sur les experts du domaine que sont la SFR et le CERF, les données médicales sont jugées prioritaires ainsi que l'imagerie alors que les données techniques sont jugées secondaires. Par ailleurs, la conservation du format DICOM a également été jugée prioritaire car elle permet d'exploiter toute la dynamique de l'image acquise. Pour résumer, les données suivantes ont été ciblées :

— Les 6 données définies par les radiologues (sans identification directe des patients) et mentionnées dans le premier paragraphe de cette sous-section ;

— 1 donnée textuelle sous la forme du Compte Rendu Radiologique (ne devant pas contenir d'entête ni de paragraphe d'identification du patient) ;

[6] *Data Protection Officer (DPO)*

— 1 série d'images au standard DICOM contenant 1 donnée image et plus d'une centaine de données associées (métadonnées ou tags DICOM).

La partie conséquente du travail d'analyse des données à conserver ou non, a donc principalement porté sur les métadonnées DICOM. Ce standard encapsule l'image acquise dans un fichier contenant en outre un ensemble de métadonnées d'identification (patient, médecin), de réalisation de l'examen (date et heure, dosimétrie), de dynamique et caractéristiques de l'image acquises (résolution, mot-machine alloués), de description du matériel (marque, modèle, composants) et d'un ensemble de descripteurs permettant l'interopérabilité avec les systèmes d'information hospitaliers et radiologiques. Ces métadonnées sont nombreuses, de l'ordre de centaines, et certaines sont obligatoires pour conserver la conformité à ce standard[7] [19].

Le standard DICOM prévoit ces opérations de déidentification et propose un cadre [20] qui définit les éléments obligatoires de manière la plus minimaliste possible. Suite à l'application de ce profil anonymisé, il en résulte un sous-ensemble de 93 métadonnées à analyser en dehors de l'image elle-même.

5.2 Qualification du degré d'importance des données

À partir de ce travail d'identification et de réduction des données à collecter, nous avons passé en revue chacune de ces données afin de qualifier leur niveau d'importance. Le résultat de cette revue est synthétisé dans la TABLE 1.

Les critères retenus sont les suivants :

— Information principale pour l'objectif : l'absence de la donnée dégrade fortement la pertinence du projet ou le remet en cause ;

— Information secondaire pour l'objectif : l'absence de la donnée diminue la finesse des analyses mais sans remettre en cause l'objectif principal ;

— Information inutile pour l'objectif : l'absence de la donnée n'impacte pas les objectifs principaux ou secondaires définis ;

— Information techniquement requise pour conformité au standard DICOM : la donnée est nécessaire pour exploiter l'image au standard DICOM.

	Données médicales	Compte-rendu	Image
Information principale	4	0	25
Information secondaire	2	1	23
Information inutile	0	0	20
Information techniquement requise	0	0	26

TABLE 1 – Qualification de l'importance des données

Les 20 informations évaluées comme inutiles ont été soit supprimées soit mises à la valeur par défaut en fonction des exigences du standard DICOM. Par exemple la description de l'examen a été supprimée alors que sa date et heure de réalisation ont été vidées.

5.3 Données techniquement requises

Les informations requises techniquement ont nécessité une analyse complémentaire pour déterminer si elles pouvaient contribuer à réidentifier le patient. Pour cela, nous les avons passées en revue pour évaluer s'il était nécessaire de retraiter leurs valeurs. Le résultat de cette évaluation est présenté dans la Table 2.

Par exemple, le type de modalité (ici un scanner) induit la valeur CT[8] pour la donnée *Image Type*. Cette valeur étant constante par construction du jeu de données, aucun retraitement n'est donc nécessaire. Ou encore, les données relatives au patient (son nom et son identifiant) sont automatiquement inscrites dans le fichier image. Étant contraint de disposer de nom unique de patients pour une bonne comptabilité DICOM, nous avons ré-encodé ces valeurs.

Le point d'attention a porté sur les données de type *Instance UID*. Le standard DICOM [21] a défini l'UID selon le schéma d'identification basé sur l'identification objet de l'OSI comme défini par le standard ISO 8824. Chaque identifiant est unique et enregistré selon l'ISO 9834-1 afin d'assurer son unicité. Chaque UID est ainsi composé de 2 parties : une racine provenant de l'organisation émettrice et un suffixe :

UID = <organisation émettrice>.<suffixe> (1)

Un type « *Instance UID* » est un UID mais utilisé pour chacune des instances d'un élément DICOM. Dès lors, par construction chacune des images d'une série d'un scanner possède un identifiant unique à travers le monde. Ce mécanisme permet d'assurer la traçabilité et de garantir que 2 images distinctes ne soient pas associées par erreur à un même patient. En raison de sa construction, il est possible de retrouver beaucoup d'informations comme le fabricant de la modalité mais également la période au cours de laquelle les images ont été réalisées.

Pour évacuer toute possibilité de déduire ces diverses informations, ces identifiants ont été régénérés par séquence aléatoire lors de leur export à partir des établissements contributeurs vers les serveurs NEHS. Un même préfixe d'UID a été utilisé afin de généraliser cette valeur.

6 Vérification de l'efficacité de l'anonymisation

La réception des données anonymisées s'effectue à ce stade dans un environnement certifié Hébergeur de Données de Santé (HDS) requis par la législation française pour toute entité hébergeant des données de santé. Cette certification s'appuie principalement sur les normes ISO 27001 [22] et ISO 27018 [23].

Au cours de la collecte, nous procédons à des opérations de contrôle qualité des données (e.g. détection de valeurs aberrantes, manquantes) et de l'anonymisation. D'une part,

[7] Ces métadonnées permettent par exemple de garantir la bonne interprétation des données avec les appareils compatibles (console de diagnostic ou de revue) et de réaliser des mesures (distances, volumes, angles) correctes.

[8] *Computed Tomography*

nous avons mis en place une procédure portant sur les métadonnées contenues dans les images DICOM et d'autre, nous procédons à une vérification portant sur les compte-rendus radiologiques.

6.1 Vérification des métadonnées

Comme dans tout contrôle de qualité des données, les statistiques descriptives sont utilisées afin d'appréhender et comprendre le jeu de données.

Typiquement, la distribution de l'âge par décennie a fait apparaître rapidement une rareté d'individus aux classes d'âges extrêmes (18-19 ans et plus de 100 ans) comme illustré dans la FIGURE 2 réalisé au cours de la collecte. Pour minimiser le risque de réidentification pour ces classes extrêmes, nous avons évalué la possibilité de les regrouper avec les classes suivantes ou précédentes auprès des radiologues. Ces derniers ont validé que cette baisse de finesse n'avait pas d'impact majeur sur l'interprétation des images.

Tag	Donnée	Action de retraitement
0002:0001	File Meta Information Version	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0002:0002	Media Storage SOP Class UID	aucune : valeur identique pour tout le jeu de données
0002:0003	Media Storage SOP Instance UID	valeur régénérée lors de l'anonymisation
0002:0010	Transfer Syntax UID	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0002:0012	Implementation Class UID	aucune : valeur générique pour ce type d'image
0002:0013	Implementation Version Name	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0008:0005	Specific Character Set	aucune : valeur courante sans être exclusive à une modalité
0008:0008	Image Type	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0008:0016	SOP Class UID	aucune : valeur identique pour tout le jeu de données
0008:0018	SOP Instance UID	valeur régénérée lors de l'anonymisation
0008:0050	Accession Number	valeur régénérée lors de l'anonymisation
0008:0060	Modality	aucune : valeur identique pour tout le jeu de données
0010:0010	Patient's Name	valeur régénérée lors de l'anonymisation
0010:0020	Patient ID	valeur régénérée lors de l'anonymisation
0010:0021	Issuer of Patient ID	valeur régénérée lors de l'anonymisation
0018:0015	Body Part Examined	aucune : valeur normalisée
0018:9345	CTDIvol	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0020:000D	Study Instance UID	valeur régénérée lors de l'anonymisation
0020:000E	Series Instance UID	valeur régénérée lors de l'anonymisation
0028:0002	Samples per Pixel	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0028:1050	Window Center	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0028:1051	Window Width	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0028:1052	Rescale Intercept	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0028:1053	Rescale Slope	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0028:1054	Rescale Type	aucune : valeur non spécifique ne permettant pas d'identifier la modalité
0028:1055	Window Center & Width Explanation	aucune : valeur non spécifique ne permettant pas d'identifier la modalité

TABLE 2 – RETRAITEMENT DES DONNEES REQUISES TECHNIQUEMENT

Nous avons également évalué le volume de de données transféré par établissement pouvant donner lieu à un risque de réidentification. Il existe une forte disparité dans les contributions et la base de données comporte plusieurs établissements ayant transférés moins de 50 patients. Cette information ayant été évaluée comme secondaire, nous avons décidé de la généraliser et de ne retenir que la région d'appartenance comme illustré en FIGURE 3. Toutefois, à la fin de la collecte, certaines régions comportaient encore un nombre faible de patients (moins d'une centaine), nous les avons également regroupées dans une catégorie *Autre*.

FIGURE 2 – Distribution de l'âge des patients par genre et décennie

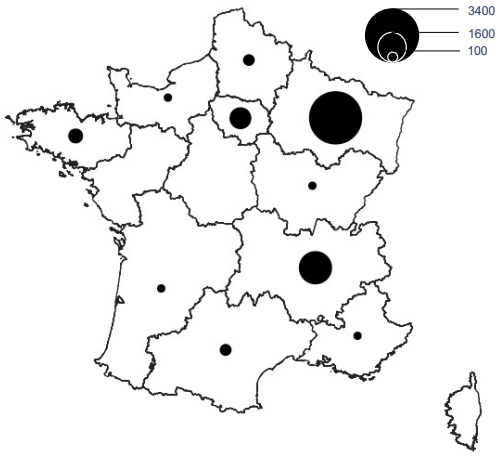


FIGURE 3 – Répartition des patients par région

AORTE THORACIQUE	0,14%
ABDOMINAL OU PEL	0,14%
THX COVID	0,12%
thorax ss iv cov	0,10%
THORAX SANS IV	0,10%
THORAX EP :	0,10%
THORAX COVID	0,10%
ABDOMINO PELVIEN	0,10%
TAP:	0,08%
HEAD	0,08%
CHEST_ABDOMEN	0,08%
THORAX:	0,07%
EP	0,07%

FIGURE 4 – Extrait de parties du corps recensée (tag 0018:0015) et leur proportion dans la base

D'autres analyses statistiques sont menées comme des Classifications Ascendantes Hiérarchiques, ou des mesures de corrélation afin de tenter d'identifier des croisements de données pouvant faire apparaître des moyens d'inférences pour réidentifier des patients. De ces travaux, nous avons identifié que pour les établissements ayant transmis peu de cas, il pouvait y avoir un risque (faible) de réidentification par une personne en contact avec les établissements concernés. Puisque nous avons supprimé l'information directe du nom de l'établissement, nous aurions pu croire avoir écarter ce risque. Néanmoins, nos travaux ont montré plusieurs corrélations fortes entre le nom de l'établissement et d'autres informations comme par exemple le modèle de scanners employés ou encore les libellés des parties du corps. En effet, comme l'illustre la FIGURE 4, certains noms sont très peu utilisés et présents uniquement dans un seul cas, malgré la standardisation DICOM.

6.2 Vérification des compte-rendus radiologues

Un compte-rendu radiologique (CRR) contient par nature beaucoup d'informations à caractère personnel de santé et donc sensibles. Les contributeurs avaient pour consigne d'extraire le corps du compte-rendu car cette partie ne contient que les informations médicales et sans données personnelles.

Afin de garantir l'absence de ces éléments, une vérification automatique à la soumission s'assure de l'absence de termes introduisant le nommage d'une personne (e.g. Monsieur, Mme, Dr, Confrère). En cas de détection de ces éléments, le contributeur est alors alerté et cette soumission refusée.

Afin de s'assurer qu'aucune autre donnée à caractère personnel ne se trouve présente, des contrôles aléatoires suivant une stratégie d'échantillonnage par établissement et médecin sont réalisés par des personnels soumis à un engagement de confidentialité. La procédure prévoit une vérification quotidienne de 5% des compte-rendus de la veille, avec un minimum de 5 CRR pour attester de l'anonymisation de ces derniers.

Après plusieurs semaines, 404 CRR ont été collectés dont 140 ont été vérifiés. Parmi ces 140, 33 cas (23,6%) ont été remontés comme contenant les informations suivantes :

- Modèle et marque du scanner utilisé ;
- Date du précédent examen ;
- Nom de l'établissement de l'examen ;
- Âge du patient.

Puisqu'à l'occasion du travail sur les métadonnées, nous avons déterminé un risque de réidentification sur ces données et qu'il n'est pas possible de retraiter l'ensemble des documents, nous avons décidé de supprimer définitivement ces derniers.

6.3. Synthèse des actions correctives

Tag	Donnée	Action corrective
0008,0070	Fabricant	Suppression
0008,1090	Modèle	Suppression
0010,1010	Âge du patient	Regroupement en tranche d'âges élargies sur les extrêmes
0018,0015	Partie du corps examinée	Regroupement en 4 catégories
0018,1020	Version logicielle	Suppression
NA	Établissement	Regroupement en grandes régions
NA	Rapport Radiologique	Suppression

TABLE 3 – Actions correctives réalisées pour baisser le risque de réidentification

Suite à l'évaluation des risques de réidentification, nous avons entrepris au cours de la collecte de supprimer certaines données ou bien de diminuer leur finesse en les regroupant (cf. TABLE 3). Dès lors, une opération de réencodage a été menée sur les données collectées, et les outils d'anonymisation mis à disposition des établissements contributeurs ont été mis à jour.

7 Mise à disposition du jeu de données pour la communauté IA

Le jeu de données anonymisé que nous avons créé a été mis à la disposition de la communauté de recherche en IA sous forme d'un accès indirect depuis la plateforme AI4Europe[9].

En effet, dans le cadre du projet H2020 AI4EU, un pilote industriel en santé a été conduit par Nehs avec ses partenaires académiques afin d'évaluer la capacité de la plateforme à mobiliser les acteurs européens de l'IA autour de problématiques fortes lié à la santé et démontrer le potentiel de la plateforme dans sa capacité à mettre en œuvre des solutions IA pour le monde industriel. Ainsi, ce jeu de données a fait l'objet d'une présentation et d'une mise à disposition.

La base de données FIDAC est ainsi accessible à travers la plateforme AI4Europe [24] et ouverte aux projets d'IA. Afin d'en assurer la surveillance, les acteurs souhaitant exploiter ces données doivent s'identifier et accepter une licence d'utilisation des données à titre non-exclusif qui leur donne le droit d'accéder, d'utiliser et d'exploiter la base de données et les données qu'elle contient conformément à leur destination.

Les retours de ces acteurs sur les données et leurs analysent contribuent à maintenir un niveau de veille sur ces dernières et nous pouvons alerter les licenciés de tout problèmes ultérieurs.

De cette mise à disposition, Thales et NEHS Digital ont exploité la base FIDAC dans le cadre du projet d'aide au diagnostic COVID-19 par IA appliquée sur des scanners tomographiques thoraciques (CT SCAN) financé par l'Agence de l'Innovation de Défense. L'IA permet d'apporter une aide à la décision pour le triage des cas covid / non covid en effectuant un 1er diagnostic probable.

Nous avons réalisé également un prototype de recherche pour anonymiser automatiquement les compte-rendus radiologiques qui systématise certains des principes mentionnés ci-dessus. En ce sens, nous avons développé le projet *medical Imaging Report Anonymiser* (mIRA). Ce projet au statut de PoC propose une API REST permettant de soumettre un compte-rendu et de recevoir sa version anonymisée. Ce dernier est quant à lui disponible à travers la plateforme AI4Europe [25].

8 Discussion

Le travail de conception sur les projets de données est une étape clef depuis l'avènement du RGPD. Dans le contexte d'un traitement en vue d'anonymiser les données, cette phase en devient majeure : l'effort de minimisation des données simplifie drastiquement ces opérations d'anonymisation ultérieure car cela diminue fortement les corrélations potentielles entre les différentes variables.

Pour mener à bien ce travail, il convient de définir clairement les objectifs de l'exploitation future de ce jeu de données à anonymiser. En outre, nous pensons que cette démarche permet de mieux exploiter les outils d'anonymisation existants en motivant les choix et paramétrages de ces algorithmes en fonction de la finalité d'exploitation plutôt que par une décision *a posteriori*.

Cette démarche peut sembler de prime abord rentrer en opposition avec les méthodes d'apprentissage automatique. En effet, diminuer de fait les données d'entrées pourrait laisser à penser que l'on va appauvrir les espaces des possibles et perdre en performance. Toutefois, une grande phase réalisée au cours des projets d'*apprentissage automatique* consiste à améliorer la qualité les données d'entraînement, de réaliser des opérations d'analyse d'impact des variables, de sélection de variables, d'échantillonnage ou encore de construction de variables synthétiques. Ce sont typiquement les opérations que nous avons menées lors de l'évaluation de l'efficacité de l'anonymisation.

Nous avons vu que les analyses de l'efficacité de l'anonymisation peuvent mettre en exergue des faiblesses permettant de réidentifier les individus et donc amener à réaliser des actions correctives sur le jeu de données. En conséquence, tant que les données ne sont pas libérées, il nous paraît primordial que ces données anonymisées mais non encore vérifiées disposent du même niveau de sécurité que si ces données n'étaient pas anonymisées. Dans notre cas, nous avons continué à utiliser l'environnement labellisé Hébergement de Données de Santé (HDS) jusqu'à la libération des données.

Une fois les données libérées, elles sont réputées anonymisées jusqu'à ce qu'elles ne le soient plus. Cela peut arriver par l'apparition de nouvelles techniques d'apprentissage, mais également par l'apparition ultérieure d'autres données qui, mises en lien avec le jeu initial, permet de faire émerger des schémas de réidentification. L'apparition de ces données complémentaires peut faire suite à des mises à disposition légales ou bien via des fuites suite à des violations de données. Ces pourquoi maîtriser la traçabilité des acteurs exploitant les jeux de données est nécessaire.

9 Conclusion et perspectives

Anonymiser des données sensibles n'est pas une opération triviale tant elle peut être lourde de conséquences. L'investissement en temps et en moyen peut paraître conséquent mais finalement les opérations dédiées s'intègrent bien aux étapes de conduite d'un projet basé sur la donnée. En effet, la phase de conception permet d'intégrer l'identification et la réduction des données à collecter alors que la phase de mise en qualité des données peut être utilisée pour s'assurer de l'efficacité de l'anonymisation.

L'étape de conception est importante et permet de faciliter les prises de décisions en cas de risque de réidentification. Toutefois, lorsque la finalité est générique comme dans notre cas (mettre à disposition des données pour la recherche en IA), il peut être difficile d'arbitrer quelles données conserver et avec quelle finesse.

Ce projet nous a permis de confirmer qu'il est primordial de

[9] https://www.ai4europe.eu/node/107

prévoir un espace sécurisé pendant la phase de collecte et tant que l'efficacité de l'anonymisation n'a pas été éprouvée sur le jeu de données à publier. Cela permet le cas échéant de mettre en place des actions de retraitement ou de suppression avant d'effectivement libérer les données.

La poursuite de nos travaux vise à appliquer cette méthodologie à d'autres jeux de données de santé afin d'en éprouver la capacité de généralisation.

10 Remerciements

Nous remercions nos partenaires que sont la SFR et le CERF pour leur volonté et dynamisme à contribuer à aider la recherche et la prise en soin des malades. Un grand merci également aux radiologues, DSI et personnels des établissements qui ont participé activement à la constitution de la base FIDAC.

11 Références

[1] J.-P. Laboureix, «Les comptes de la sécurité sociale 2020-2021,» 2021.

[2] CNIL, «La Plateforme des données de santé (Health Data Hub),» 09 02 2021. [En ligne]. Available: https://www.cnil.fr/fr/la-plateforme-des-donnees-de-sante-health-data-hub. [Accès le 03 03 2022].

[3] M. Wang, C. Xia, L. Huang, S. Xu, C. Qin, J. Liu, Y. Cao, P. Yu, T. Zhu, H. Zhu, C. Wu, R. Zhang, X. Chen, J. Wang, G. Du, C. Zhang, S. Wang, K. Chen, Z. Liu, L. Xia et W. Wang, «Deep learning-based triage and analysis of lesion burden for COVID-19: a retrospective study with external validation,» *Lancet Digit Health,* vol. 2(10), pp. e506-e515, Oct. 2020.

[4] L.-P. Sondeck, «Anonymisation des données, une nécessité à l'ère du RGPD,» *Sécurité des systèmes d'information,* 10 Nov. 2019.

[5] F. Ben Fredj, «Méthode et outil d'anonymisation des données sensibles,» Conservatoire national des arts et metiers - CNAM; Université de Sfax (Tunisie). Faculté des Sciences économiques et de gestion, 2017.

[6] L. H. Cox, «Suppression methodology and statistical disclosure analysis,» *Journal of the American Statistical Association,* vol. 75(370), p. 377–385, 1980.

[7] P. Samarati, «Protecting respondents identities in microdata release,» *IEEE transactions on Knowledge and Data Engineering,* vol. 13(6), pp. 1010-1027, 2001.

[8] J. Domingo-Ferrer et V. Torra, «A quantitative comparison of disclosure control methods for microdata,» *Confidentiality, disclosure and data access: theory and practical applications for statistical agencies,* pp. 111-134, 2001.

[9] X. Xiao et Y. Tao, «Anatomy: Simple and effective privacy preservation,» chez *Proceedings of the 32nd international conference on Very large data bases*, Seoul (Korea), 2006.

[10] T. Dalenius et S. P. Reiss, «Data-Swapping: A Technique for Disclosure Control,» *Journal of Statistical Planning and Inference,* vol. 6(1), pp. 73-85, 1982.

[11] C. Dwork, F. McSherry, K. Nissim et A. Smith, «Calibrating Noise to Sensitivity,» *Private Data Analysis Journal of Privacy and Confidentiality,* vol. 7(3), 2016.

[12] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis et W. Shi, «Federated learning of predictive models from federated Electronic Health Records,» *International Journal of Medical Informatics,* vol. 112, pp. 59-67, 2018.

[13] S. E. Rossi, J. J. Erasmus et M. Volp, «Crazy-Paving Pattern at Thin-Section CT of the Lungs: Radiologic-Pathologic Overview,» *RadioGraphics Vol. 23, No. 6,* 2003.

[14] Parlement européen et du conseil, «Règlement 2016/679 du parlement Européen et du Conseil,» *Journal officiel de l'Union européenne,* 27 04 2016.

[15] CNIL, «L'anonymisation des données, un traitement clé pour l'open data,» 17 10 2019. [En ligne]. Available: https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data. [Accès le 10 10 2021].

[16] CNIL, «Fiche n°1 : Identifier les données à caractère personnel,» chez *Guide RGPD du développeur, v1.0.1*, LaboCNIL, 2020, pp. 5-6.

[17] Article 29 Data Protection Working Party, «Opinion 05/2014 on Anonymisation Techniques,» European Commission, 2014.

[18] NEMA, «DICOM,» The Medical Imaging Technology Association, [En ligne]. Available: https://www.dicomstandard.org. [Accès le 11 10 2021].

[19] Innnolitics, LLC, «DICOM Standard Browser,» 2016. [En ligne]. Available: https://dicom.innolitics.com/ciods. [Accès le 11 10 2021].

[20] DICOM Standards Committee, «DICOM PS3.15 2022a - Security and System Management Profiles,» DICOM, [En ligne]. Available: http://dicom.nema.org/medical/dicom/current/output/html/part15.html. [Accès le 11 10 2021].

[21] DICOM Standards Committee, «Unique Identifiers (UIDs),» [En ligne]. Available: https://dicom.nema.org/dicom/2013/output/chtml/part05/chapter_9.html.

[22] ISO/IEC JTC 1/SC 27, ISO/IEC 27001:2013, 2013, p. 23.

[23] ISO/IEC JTC 1/SC 27, ISO/IEC 27018:2019, 2019, p. 23.

[24] NEHS Digital, «Covid-19 chest CT-scan Dataset,» 12 2021. [En ligne]. Available: https://www.ai4europe.eu/research/ai-catalog/covid-19-chest-ct-scan-dataset-0.

[25] J. Clech et G. Martial, «mIRA - medical Imaging Report Anonymiser,» 11 2021. [En ligne]. Available: https://www.ai4europe.eu/research/ai-catalog/mira-medical-imaging-report-anonymiser.

L'IA au travail : propositions pour outiller la confiance[1]

Y. Ferguson[1], C. Pecoste[2], avec la collaboration de A. Leblanc[3], P. Crespin[4]

[1] Icam site de Toulouse 1, LaborIA
[2] Icam site de Toulouse
[3] Renault Group, mis à disposition de l'Institut de Recherche Technologique SystemX dans le cadre du programme Confiance ai
[4] Spix Industry

yann.ferguson@icam.fr
charly.pecoste@icam.fr

Résumé

De nombreux travaux appréhendent les enjeux du déploiement de l'IA dans les organisations et les métiers. Au-delà des dynamiques destructions/créations d'emplois, elles dégagent des questions de transformation du travail et de leur acceptabilité. Elles pointent en ce sens sur la nécessité d'établir une confiance technique et sociale envers l'IA. Cet article décrit un outil de diagnostic social des applications de l'IA au travail en vue de faciliter l'instauration d'une confiance située. Cet outil a vocation à être spécifié, expérimenté et formalisé dans le programme Confiance AI.

Mots-clés

Travail, Outil de diagnostic, Confiance, Socialisation.

Abstract

Several studies have examined the challenges of AI deployment in organizations and professions. Beyond the dynamics of job destruction/creation, they highlight the issues of work transformation and their acceptability. In this sense, they point to the need to establish technical and social trust in AI. This article describes a tool for social diagnosis of AI applications at work in order to facilitate the establishment of trust linked to specific working context. This tool will be specified, experimented and formalized in the Confiance AI program.

Keywords

Work, Diagnostic Tool, Trust, Socialization

1 Introduction

L'essor actuel de l'intelligence artificielle (IA) autour notamment de l'approche connexionniste conduit à des projections souvent alarmantes en matière d'impact sur l'emploi et le travail. Les premières études dites « centrées sur l'emploi » [1] ont prédit des destructions massives d'emplois à court, moyen et long terme. Les plus récentes, davantage « centrées sur la tâche » [2] ont réduit les estimations de destruction d'emplois mais ont pointé sur des transformations importantes du travail. La France, en particulier, s'appuie sur l'enquête de l'OCDE [3], qui évalue à 32% la part des emplois profondément transformés par l'IA au cours de vingt prochaines années.

Au-delà des enjeux politiques et économiques qui imposent d'anticiper ces bouleversements afin d'assurer simultanément la compétitivité des entreprises et l'employabilité des travailleurs, la transformation profonde d'un tiers des emplois constitue un « momentum anthropologique » : une période d'interrogation sur ce que signifie socialement le travail. Pour l'anthropologue James Suzman, « *le travail que nous accomplissons définit aussi ce que nous sommes ; il détermine nos perspectives futures, il dicte avec qui nous passons la majeure partie de notre journée, et où, il nous procure un sentiment de dignité ; il influence beaucoup de nos valeurs et oriente nos loyautés politiques* » (p.8-9) [4]. Cette fonction sociale du travail rend toute transformation particulièrement sensible et, par ricochet, l'introduction d'une technologie porteuse de changements importants. L'IA ajoute à cette généralité deux spécificités :

- Elle brasse un imaginaire riche de fantasmes puisés notamment dans la fiction,
- Elle adresse des tâches cognitives jusque-là considérées comme « proprement humaines » et donc préservées de l'automatisation engagés depuis la révolution industrielle.

Ces spécificités de l'IA ne sont pas ignorées. Des études pour qualifier ses effets en vue de définir des systèmes d'IA (SIA) à la fois performants et « encapacitant » sont en cours. La France a notamment initié avec le Canada le Partenariat Mondial pour l'Intelligence Artificielle (PMIA) au sein duquel le Future of Work Working Group a pour mandat d'identifier des bonnes pratiques en matière de SIA responsable au travail [5]. Elle a également lancé le programme LaborIA, centre de ressources d'expérimentations sur l'IA dans le milieu professionnel afin de « *mieux appréhender l'intelligence artificielle et ses effets sur le travail, l'emploi, les compétences et le dialogue social, dans*

[1] Ce travail a bénéficié d'une aide de l'État au titre du programme France 2030 dans le cadre de l'Institut de Recherche Technologique SystemX.

l'objectif de faire évoluer les pratiques des entreprises et l'action publique ». Il s'agit d'établir des principes de complémentarité entre les SIA et les travailleurs autour de deux idées : « *elle devrait être saine (notamment au regard des enjeux éthiques) et capacitante pour l'être humain (c'est-à-dire qu'elle doit lui permettre de faire mieux, de se développer et ne pas le mettre dans une posture d'avilissement ou d'asservissement)* » [6]. Les industriels sont évidemment associés à ces initiatives. Ils sont notamment rassemblés dans un collectif, baptisé « Confiance AI » [7], qui réunit des acteurs académiques et industriels français majeurs des domaines de la défense, des transports, de l'industrie manufacturière et de l'énergie. Ses membres ont décidé de mutualiser leurs savoir-faire scientifiques et technologiques de pointe pour « *concevoir et industrialiser des systèmes à base d'IA de confiance* ». Pilier technologique du Grand Défi « *Sécuriser, certifier et fiabiliser les systèmes fondés sur l'intelligence artificielle* » lancé par l'Etat, Confiance.ai est le plus gros programme de recherche technologique du plan #AIforHumanity qui doit faire de la France un des pays leader de l'IA. Son originalité repose sur sa stratégie intégrative : il traite les défis scientifiques relatifs à l'IA de confiance et apporte des solutions tangibles, applicables dans le monde réel et industrialisables. La confiance est instituée techniquement (« *by design* ») et socialement, entre un SIA et le professionnel qui l'utilise.

Cet article présente un outil en cours d'élaboration, « MAIAT » (Mesure de l'Acceptabilité sociale de l'IA au travail), pour qualifier, évaluer et accompagner la construction sociale de la confiance entre un SIA et ses utilisateurs dans différents métiers et organisations. MAIAT se présente plus précisément comme un instrument de « dérisquage » de la construction de la confiance en attirant l'attention sur des points des vigilance que nous avons identifiés à partir du catalogue mondial de cas d'usage du PMIA que nous coordonnons. L'outil est développé dans le cadre du programme Confiance AI en appui de l'intégration d'un assistant vocal intelligent, une application speech-to-text destinée à faciliter la saisie de données à un poste de travail. MAIAT propose des catégories de critères basés sur des indicateurs à suivre tout au long du processus d'intégration que nous qualifions de « processus de socialisation », à savoir l'appropriation progressive d'une technique par un métier, une communauté professionnelle fondés sur le partage de valeurs, de normes, de pratiques, d'une identité. Nous partons du postulat qu'un SIA vient nécessairement bouleverser cette communauté et, plutôt que de stigmatiser une « résistance au changement », nous essayons d'identifier les ressorts sociaux d'un processus de rejet/acceptation. Selon nous, la compréhension de ces ressorts est la condition pour baliser une trajectoire d'adoption basée sur la confiance. Cet outil de diagnostic et de suivi de la socialisation d'un SIA se base sur deux familles de trois critères. La première rassemble des critères qui ont essentiellement trait au bien-être des travailleurs : la reconnaissance, l'engagement relationnel et la surveillance (2). La seconde comporte des critères qui renvoient davantage à l'engagement : l'autonomie, le savoir-faire, la responsabilité (3). Nous présentons enfin comment nous envisageons de le déployer (4).

2 Evaluer les effets des SIA sur le bien-être au travail

La première famille de points de vigilance touche le bien-être des travailleurs. Le bien-être met l'accent sur la perception personnelle et collective des situations et des contraintes de la sphère professionnelle. Il fait référence à un sentiment général de satisfaction et d'épanouissement dans et par le travail qui dépasse l'absence d'atteinte à la santé. Le sens de ces réalités a, pour chacun, des conséquences physiques, psychologiques, émotionnelles et psychosociales et se traduit par un certain niveau d'efficacité pour l'entreprise. Le bien-être est une notion à la fois vague, subjective et englobante qui se répercute évidemment sur l'engagement que nous traitons par ailleurs. Par souci d'efficience, nous l'avons résumé dans notre outil à trois critères : la reconnaissance, les relations sociales et la surveillance.

2.1 La reconnaissance

Au travail, la reconnaissance prend la forme de revendications de salaires, de statuts, mais surtout d'une demande plus générale et plus diffuse qui porte sur la personne elle-même, le « respect » et la dignité que chacun estime dus. La quête de considération et de prestige ainsi que le souci de paraître font partie des mobiles fondamentaux qui guident nos vies : une des motivations principales de l'existence humaine réside dans le désir d'être « reconnu par autrui ». Dans le travail, la reconnaissance s'allie souvent au mérite qui a pris une place considérable dans les organisations. Le mérite est une norme de justice centrale pour établir les inégalités justes. Il justifie en effet les inégalités de gratification symbolique ou matérielle, de positions, d'attributions et les rend acceptables aux yeux des membres du collectif. C'est pourquoi les organisations construisent des indicateurs et des systèmes d'évaluation sophistiqués qui vont organiser, objectiver, la reconnaissance du mérite.

Brun et Dugas [8] distinguent quatre dimensions de la reconnaissance au travail :

- La reconnaissance existentielle qui porte sur l'individu et non sur le seul salarié ;
- La reconnaissance de la pratique qui porte sur la manière d'exécuter le travail ;
- La reconnaissance des efforts consentis, qui portent sur l'engagement, l'intensité voire les risques encourus ;
- La reconnaissance des résultats effectifs, observables, mesurables et contrôlables.

L'IA au travail peut fragiliser directement ces quatre formes de reconnaissance :

- Reconnaissance existentielle : l'IA peut élever la substituabilité des travailleurs en déplaçant la valeur vers la machine ;
- Reconnaissance de la pratique : l'IA exécute les tâches ou renforce le caractère procédurier du travail ;
- Reconnaissance des efforts : l'IA facilite le travail ;
- Reconnaissance des résultats : l'IA rend illisible la contribution propre au travailleur.

Au final, ce critère reproblématise les enjeux de déplacement

de la valeur du travail en pointant sur la nécessité pour les organisations de produire des nouveaux indicateurs qui permettront de définir et de positionner la reconnaissance au travail.

Exemple de SIA portant une problématique de reconnaissance : *Une tâche de fixation de porte sur un avion repose sur un petit nombre de techniciens capables d'assurer cette pose après une dizaine d'itérations. Un SIA divise ce nombre par deux en calculant des points que le technicien doit ensuite suivre. Cette tâche n'implique plus d'expertise particulière, le SIA rend les techniciens substituables. Ceux-ci refusent de l'utiliser.*

Critère 1 : La fragilisation de la reconnaissance

1.1. Le SIA réduit-il la distinction entre les travailleurs (réduction de l'écart entre l'expert et le novice) ? (Oui =1 ; Non=0)[2]

1.2. Des tâches requérant auparavant de l'expertise sont-elles désormais partiellement ou totalement automatisées ? (O=1, N=0)

1.3. Le SIA supprime-t-il des tâches pénibles, ingrates, répétitives ou dangereuses ? (O=0 ; N=1)

1.4. L'introduction de la technologie rend-elle moins visible le résultat de l'activité du travailleur ? (O=1 ; N=0)

Mots-clés : Reconnaissance, singularité du travailleur, pratique, efforts, substituabilité, mérite, valorisation, inégalités justes.

2.2 Les relations sociales

Depuis les travaux classiques de l'Ecole des relations humaines (notamment Mayo et Maslow dans les années 30'), il est admis que la qualité des relations entre les collaborateurs et entre eux et leur hiérarchie est déterminante tant pour leur bien-être que pour leur performance. Dans les sociétés modernes, marquées par l'individualisme et la fragilité des engagements dans les collectifs, le travail continue d'incarner ce lieu où émergent des formes collectives dans lesquels les individus se construisent et s'engagent. Or, l'intermédiation technologique dans les relations sociales n'est évidemment pas sans incidence. Aussi, pour Bobillier Chaumon [9], « *tout dispositif qui viserait, d'une manière ou d'une autre, à remettre en cause les équilibres sociaux en place, les réseaux de travail constitués (formels et informels), les sentiments d'appartenance à une communauté, aurait de grandes difficultés à être accepté. Le rejet des technologies relèverait ici davantage d'une stratégie de défense ou de protection face au danger de désorganisation et de fragilisation que peut faire peser la technologie sur le collectif de travail* ».

Supposées établir une « société de communication », les Technologies de l'Information et de la Communication ont transformé la façon dont les travailleurs communiquent au travail, avec des résultats mitigés. Dominique Wolton [10] souligne notamment qu'elles alimentent une confusion entre communication fonctionnelle et technique, qui relève de l'efficacité informationnelle, et la communication normative et humaine, qui désigne un processus d'intercompréhension et de réciprocité. La gestion de messagerie est en outre une tâche chronophage une « *procrastination structurée* » [11] dont certaines organisations ont décidé de règlementer l'usage voire de l'interdire dans des activités nécessitant un haut niveau de coordination collective [12]. L'appauvrissement de la communication normative est en effet souvent le prix de l'efficacité de la communication fonctionnelle.

Cet appauvrissement peut doublement fragiliser le « travail collectif » et le « collectif de travail », pour reprendre la distinction opérée par Clot et Caroly [13] :

- Le travail collectif décrit comment des personnes exerçant différents métiers, se réunissent autour d'un objectif commun,
- Le collectif de travail renvoie au partage de règles de métiers et de critères sur la qualité du travail. Il désigne la possibilité pour un collectif de « *soigner le travail* » via une « *coopération conflictuelle* », des « *disputes sur la qualité du travail* » qui aboutissent à des référentiels communs [14].

De nombreux SIA automatisent ou interfèrent dans des tâches sociales, c'est-à-dire des tâches qui se basent essentiellement sur une communication humaine. Certains SIA lui substituent des relations machine-machine, humain-humain via des machines, ou humain-machine. Ces dernières peuvent être source de satisfaction et leur part peut croître au détriment des relations humaines. Les chatbots, par exemple, ont pour eux leur disponibilité, leur immédiateté, l'homogénéité de leur réponse, leur stabilité émotionnelle et n'implique pas de réciprocité. Pour le psychiatre Serge Tisseron [15], cette prévisibilité de la communication avec une machine risque de rendre moins acceptable l'irréductible écart entre les attentes et la réalité des relations humaines. Cela pourrait conduire les utilisateurs à privilégier la communication avec une machine, de la même façon que de nombreux travailleurs préfèrent la communication asynchrone de la messagerie à la confrontation directe. Victor Scardigli [16] a lui montré comment les premiers concepteurs de l'automatisation du pilotage des avions se méfiaient de l'usage de la voix dans les interactions des pilotes avec les contrôleurs aériens. Les dangers qu'ils associent aux erreurs de prononciation, de traduction, d'interprétation d'inattention leur font préférer un « automatisme social », un remplacement du lien social par un lien technique entre ordinateurs au sol et à bord, jugé plus sûr et efficace. Pourtant les contrôleurs expriment leur attachement aux informations paraverbales contenues dans la voix du pilote, révélatrices du climat du cockpit et l'équipage ressent le lien humain avec la terre comme un élément de stabilité émotionnelle en cas d'incident à bord.

France Stratégie pointe alors sur un risque de désengagement relationnel, en raison d'une « *déshumanisation des pratiques et un appauvrissement des interactions sociales, lesquels constituent très souvent la raison d'être de certains métiers* » [17] La communication pourrait en effet être appauvrie par les

[2] Le système de notation de MAIAT est détaillé dans la section 4.2.

stratégies de dialogue standardisées des machines qui déterminent les réponses et parfois mêmes les questions à partir d'architecture de choix.

Exemple de SIA portant une problématique de relations sociales : Le « voice picking » ou « commande vocale » qui guide les préparateurs de commande dans les entrepôts, a généralement une compréhension réduite à une cinquantaine de mots et son usage tend de surcroît à limiter la réponse humaine à deux : « répétez » et « OK ». En outre, sortir de ce lexique configuré, pour saluer un collègue par exemple, provoque un message d'erreur de type « chiffre contrôle faux ».

Critère 2 : Le désengagement relationnel

2.1 Le SIA introduit-il une communication entre des machines ? (O=1 ; N=0)

2.2 Le SIA crée-t-il une interaction humain-machine au détriment d'une communication entre personnes ? (O=1 ; N=0)

2.3 Le SIA intervient-il dans la communication entre plusieurs personnes ? (O=1 ; N=0)

2.4 Le SIA impose-t-il des lexiques et des syntaxes standardisés pour communiquer ? (O=1 ; N=0)

Mots-clés : Communication, coopération, travail collectif, collectif de travail, appauvrissement du langage, interactions sociales.

2.3 La surveillance et le contrôle

Dans l'imaginaire collectif, les machines sont associées à des systèmes de surveillance et de contrôle. C'est ce qu'on appelle l'effet « Big Brother », en référence au roman de George Orwell, 1984. Le droit pour l'employeur de contrôler et de surveiller l'activité de ses salariés est admis dans son principe, tout en faisant l'objet d'un encadrement autour de la distinction activité professionnelle/vie privée. Dans certaines situations, non seulement l'employeur peut, mais doit surveiller l'activité du salarié, lorsque la finalité est de protéger des installations comportant un risque élevé d'explosion ou de diffusion de matières dangereuses ou de détournement de celles-ci par des tiers non autorisés, et d'assurer la protection de personnes exposées à des risques particuliers en raison de ces activités. En dehors de ces situations spécifiques, ce droit est soumis à un contrôle de proportionnalité : il doit être justifiable par les intérêts légitimes de l'employeur. Ainsi, la commission nationale informatique et liberté (Cnil) a censuré un dispositif de surveillance par caméra au-dessus d'un poste de travail. Le dispositif fonctionnait la journée en mode visualisation et plaçait l'employé sous surveillance permanente et constante. Le gérant pouvait accéder en temps réel aux images depuis son téléphone et donc exercer cette surveillance à distance [18].

Les SIA modernes mobilisent des algorithmes apprenant à partir de données collectées. Ces données peuvent concerner directement les travailleurs et être exploitées en vue d'évaluer leur productivité. Plusieurs technologies sont en ce sens dédiées à « la mesure de soi » (quantified self) en collectant et en analysant des données individuelles. Ces méthodes sont largement déployées dans les entrepôts : le nombre de commandes par journée, par semaine, le nombre moyen d'articles par commande, le temps écoulé entre le moment de la commande et sa réception, entre l'arrivée du client et son départ, le nombre d'articles collectés par employé et par heure, le parcours de l'employé… Toutes ces données sont enregistrées, analysées, transformées en indicateurs de productivité. Head [19] décrit ainsi l'émergence d'un « *management numérique* » qui transforme les travailleurs encadrés en « *représentations électroniques* » d'êtres humains, ces « *nombres, mots codés, cônes, carrés, et autres triangles qui nous incarnent sur les écrans des managers* » qui assurent la micro-gestion du travail de chaque salarié ou de chaque équipe. Ces représentations électroniques sont de plus en plus appliquées aux cadres intermédiaires qui, privés de leur rôle traditionnel de supervision, se retrouvent soumis au même contrôle tatillon de leur temps et de leurs performances que celui que certains exerçaient auparavant sur leurs subordonnés.

Le SIA peut également solliciter une identification pour des raisons de sécurité, de paramétrage individuel ou de traçabilité des problèmes rencontrés par la machine. Quel que soit l'usage prévu de ces données, elles peuvent être exploitées pour surveiller ou contrôler le travailleur, du moins peut-il le ressentir ainsi. La Réglementation Générale des Données Personnelles limite l'utilisation des données à une finalité unique et précise et impose des principes de transparence sur la collecte, l'utilisation et la conservation des données personnelles. Mais peu de travailleurs (et de citoyens en général) maîtrisent ce cadre.

Exemple de SIA portant une problématique de surveillance : Une entreprise a expérimenté l'usage du robot social Pepper de la société SoftBank Robotics, qui déambulait notamment dans les salles de pause pour leur proposer de les divertir. Les salariés s'en méfiaient en raison de la caméra sur son front qui lui sert de capteur d'interaction, mais qu'ils imaginaient être une caméra de surveillance. Les pilotes de l'expérimentation ont alors collé un post-it qui disait « Je ne vous filme pas » pour lever les inquiétudes.

Critère 3 : La surveillance

3.1 Le SIA intègre-t-il une caméra/micro susceptible de filmer/écouter le travailleur ou d'être perçu comme tel ? (O=1 ; N=0)

3.2 Le SIA implique-t-il des identifiants permettant de collecter des données sur son utilisateur ? (O=1 ; N=0)

3.3 Les données collectées par le SIA sont-elles exploitées pour mesurer la productivité de son utilisateur ? (O=1 ; N=0)

3.4 La finalité de l'utilisation des données est-elle transparente ? (O=1 ; N=0)

Mots-clés : Surveillance, contrôle, traçabilité, mesure de productivité individuelle, privacy, RGPD, proportionnalité.

3 Evaluer l'effet des SIA sur l'engagement au travail

L'engagement professionnel se définit par une « *attitude qui traduit la force des liens unissant l'individu à son travail. L'engagement implique l'attachement affectif (s'identifier à l'organisation), l'attachement instrumental (coût*

d'opportunité), enfin, l'attachement moral (obligation envers l'organisation) » [20]. Vecteur majeur de la performance d'une organisation, il est dès lors un objectif des managers qui multiplient les méthodes pour stimuler l'implication de leurs employés. Mais l'engagement professionnel est souvent tributaire de ressorts internes, où il est généré par l'individu lui-même. Il est en effet aujourd'hui communément admis que les motivations autodéterminées ou intrinsèques sont bien plus efficaces que les motivations extrinsèques ou contrôlées. Elles correspondent au fait d'accomplir une ou plusieurs tâches au travail par intérêt, par plaisir ou encore par satisfaction. Plus la motivation serait autodéterminée, plus les performances au travail seraient fortes. Si la motivation contrôlée permet la performance, c'est essentiellement à court terme, tandis que la motivation autodéterminée la favoriserait à long terme et soutiendrait le bien-être au travail. Elle renvoie aux aspirations des individus à se réaliser, notamment dans le travail, que Clot attribue essentiellement à la liberté qu'accorde l'organisation pour assurer un travail de qualité [14]. Notre outil aborde l'engagement au prisme de trois critères : l'autonomie, le savoir-faire et la responsabilité.

3.1 L'autonomie

L'autonomie qualifie la possibilité pour le travailleur de devenir sujet, de s'éprouver comme l'auteur de ses œuvres, d'affirmer ses choix, d'agir de lui-même. Le principe d'autonomie oppose le travail authentique, expressif et personnel au travail mécanique, déshumanisé et « abstrait ». L'autonomie est devenue un principe de justice central au travail : chaque travailleur est conduit à juger de la justice de son travail en fonction de la liberté, de l'autonomie et de la réalisation de soi qu'il lui permet [21]. A l'opposé, le sentiment d'injustice résulte de la fatigue, de l'usure, de l'absence d'intérêt pour la tâche, du sentiment de mépris et d'impuissance sur sa propre activité. L'autonomie consacre la réalisation de soi comme juste et l'aliénation comme injuste. L'autonomie est aujourd'hui un principe d'organisation (les « équipes semi-autonomes » des pays nordiques, les « cercles de qualité » et les « groupes de projets » venus du Japon) et de management (avec notamment la généralisation de la direction par objectifs).

Tout projet d'automatisation d'une activité porte en germe une réduction de l'espace d'autonomie du travailleur (mais peut, à termes, en générer un nouveau). Pour une machine, l'autonomie est la capacité d'opérer indépendamment d'un opérateur humain dans un environnement dynamique complexe. Cette diminution des interventions réduit l'espace d'autonomie des travailleurs. Consécutivement, les SIA modifient de la répartition de l' « intelligence » du travail [22]. Pour un salarié, l'importance de l'activité de réflexion dans sa tâche délimite son autonomie dans l'organisation du travail, c'est-à-dire son pouvoir, sa valeur marchande, l'intérêt du travail, la maîtrise de son itinéraire professionnel et, partant, de son avenir personnel. Un intégrateur de SIA nous révélait d'ailleurs bannir la notion de « machines intelligentes » de son vocabulaire : « *ce sont les travailleurs qui sont intelligents, mais les machines* ». Les SIA portent toutefois un risque de « *paternalisme technologique* » aux multiples visages - alertes, recommandations, aides à la décision, rappels à l'ordre, blocages, interdictions qui placerait progressivement le travailleur dans une situation de dépendance. Les SIA développeraient plus généralement une « *logique rationnalisante* » [23] mettant le travailleur en situation d'obéissance dans des organisations « algocratiques » recourant à un « management algorithmique » : « *obéir aux ordres d'une intelligence artificielle, perdre le contrôle sur les processus, déléguer les décisions à la machine sont autant de modes de complémentarité, qui, au niveau individuel et collectif, seront susceptibles de créer de la souffrance au travail* » (p. 186) [24]. En deçà de l'obéissance à un algorithme, le SIA peut dégrader la flexibilité cognitive du travailleur qui, suivant le déroulement de la séquence opératoire, est moins en capacité de s'interrompre, de s'adapter à un environnement changeant, de répondre à une sollicitation extérieure. Il peut aussi se retrouver prisonnier d'un un tunnel attentionnel qui l'empêche d'envisager des solutions alternatives.

Exemple de SIA portant une problématique d'autonomie : Dans les entrepôts qui recourent au voice picking, les opérateurs suivent un itinéraire fixé par le SIA. Auparavant, les opérateurs expérimentés le définissaient de façon autonome pour élaborer une « belle palette », c'est-à-dire une palette équilibrée dont la construction facilite la mobilité.

Critère 4 : La perte d'autonomie

4.1 Le SIA détermine-t-il un déroulement de l'action du travailleur ? (O=1 ; N=0)

4.2 Le SIA émet-il des notifications à l'adresse du travailleur ? (O=1 ; N=0)

4.3 Le SIA réduit-il ou rend-il plus difficile la prise d'initiative pour le travailleur ? (O=1 ; N=0)

4.4 Le travailleur dispose-t-il de marge manœuvre convenue dans l'utilisation ou l'interprétation du SIA ? (O=0 ; N=1)

Mots-clés : Procédures, initiative, travail réel, liberté, notification, pratiques professionnelles, flexibilité cognitive.

3.2 Le savoir-faire

L'identité professionnelle se construit en rapport à un savoir-faire : une école, un diplôme puis l'expérience contribuent chez l'individu à se construire une identité pour soi et pour autrui. Ce savoir-faire sert de base à sa légitimité, elle-même constitutive de son statut et de sa position, parfois durement acquise. L'histoire des organisations industrielles est ainsi traversée par plusieurs transformations qui ont fragilisé cette identité en générant un sentiment de dépossession :

- L'organisation scientifique du travail qui dépossédait les opérateurs de la conception des modes opératoires et les spécialisait.
- Le toyotisme qui, à l'inverse, dépossédait les opérateurs de leur spécialité en promouvant la polyvalence.

Dans les deux cas, l'innovation de processus, en déplaçant la valeur du travail, a mis fin à un monde au sein duquel les opérateurs s'étaient constitués des règles, des routines efficaces, un statut, une identité. Cela renvoie à ce que l'ergonomie appelle l'« *acceptabilité située* » [9]. On y regarde, dans le contexte d'usage, ce que la technologie « *permet/autorise de*

faire » ou « *oblige à faire* », mais aussi ce qu'elle « *empêche de faire* » ou « *plus comme avant* » et ce, sur différentes dimensions de l'activité.

Les SIA sont potentiellement porteurs de quatre tendances corrélées entre elles :

- La valeur du travail pourrait se déplacer, être moins dans le faire, fonction de l'IA, que dans le contrôle, la vérification, l'approbation et la validation : une partie de l'activité s'oriente vers la machine « intelligente » elle-même, dont il s'agit de s'occuper [25]. Le travail est ainsi décentré, le travail direct et immédiat décline tandis que celui sur le système technique devient l'objet principal.
- Le SIA généralise des savoir-faire, c'est-à-dire baisse le niveau d'expertise nécessaire à l'exécution de tâches, les rendant plus accessibles et consécutivement moins distinctives. Les savoir-faire ne disparaissent pas mais sont moins valorisants professionnellement.
- Le SIA réduit l'espace des pratiques au profit des process [19], la capacité d'agir des travailleurs se limitant à suivre des instructions sans aucune autre finalité. Head distingue en ce sens le processus, « *une série d'opérations et la façon dont elles sont reliées les unes aux autres* » ; de la pratique, qui désigne « *l'accumulation de connaissances tacites et de compétences* » dont les employés ont besoin pour effectuer leur travail. Cette distinction évoque les concepts aristotéliciens de *praxis* et de *poïésis*. La *poïèsis* est la simple opération de faire, tandis que la *praxis* renvoie au but recherché derrière une action donnée [26]. Un travailleur téléguidé par un SIA n'est plus en capacité d'adapter librement un processus à une situation sur la base de son expérience.
- La redistribution des capacités d'action entre la machine et le travailleur, les processus et les pratiques affectent le périmètre du travail réel qui consiste à combler l'insuffisance des procédures explicites, le travail prescrit, pour atteindre les objectifs. Le travail réel décrit ce que le travailleur produit et a le sentiment de produire effectivement, tantôt en deçà, tantôt au-delà des règles et des attentes formelles. On tend même à définir la véritable compétence comme la ressource nécessaire pour combler cet écart : le professionnel démontre sa compétence qu'à partir du moment où suivre les ordres ne suffit pas pour réaliser les tâches qu'on lui demande. Le SIA peut ainsi coloniser le travail réel, soit en automatisant les pratiques antérieures, soit en éteignant son expression par l'injonction managérial à suivre l'algorithme. La qualité du travail peut alors s'en trouver « *empêchée* » si le travailleur n'est plus autorisé de mettre en œuvre ce qu'il estime nécessaire pour effectuer un travail de qualité [27]

Limité dans la possibilité de mobiliser leurs savoir-faire, les travailleurs s'exposent à une déqualification, souvent appelée « ubérisation » quand elle est occasionnée par des algorithmes. Nicholas Carr dénonce en ce sens le « *mythe de la substitution* » selon lequel à chaque fois que nous faisons appel à « *un algorithme pour nous décharger dans notre travail, nous nous émancipons pour viser un objectif plus élevé et qui exige un degré supérieur d'ingéniosité et d'intelligence* » [28]. Il constate au contraire dans les métiers qui s'automatisent un appauvrissement des tâches cognitives, une déqualification, qui altère la façon d'agir et de penser. France Stratégie perçoit un risque de « *prolétarisation des savoirs et des savoir-faire* » où *« les humains risquent de se voir déposséder de leur expertise en termes de know-how, et de perdre un ensemble de capacités et de compétences, qui non seulement peuvent être utiles à la société, mais qui contribuent aussi à alimenter le respect de soi* » (p. 6) [16].

Exemple de SIA portant une problématique de savoir-faire : Un sous-traitant du constructeur automobile développe un SIA dédié à la généralisation des connaissances empiriques mobilisées en situation de dysfonctionnement du processus. Jusqu'à présent, ces interventions extra-ordinaires sollicitaient l'expérience des travailleurs les plus qualifiés. Le SIA reçoit des notifications de dysfonctionnement des capteurs, alertent les travailleurs et guident la remédiation via des montres connectés et des écrans.

Critère 5 : Le sentiment de dépossession
4.1 Le SIA réduit-il le travail direct ? (O=1 ; N=0)
4.2 Le SIA rend-il l'activité plus facile à réaliser par tout un chacun ? (O=1 ; N=0)
4.3 Le SIA rend-il des savoir-faire obsolètes ? (O=1 ; N=0)
4.4 Le SIA génère-t-il de nouvelles tâches pour le travailleur ? (O=1 ; N=0)
Mots-clés : Identité professionnelle, savoir-faire, valeur du travail, processus, pratique, obéissance, sens du travail, identité professionnelle, Déqualification/requalification.

3.3 La responsabilité

Du latin *respondere*, répondre, la responsabilité exprime le devoir de répondre de ses actes, toutes circonstances et conséquences comprises, c'est-à-dire d'en assumer l'énonciation, l'effectuation, et par suite la réparation voire la sanction lorsque l'attendu n'est pas obtenu. La notion de responsabilité morale va plus loin : elle consiste en une capacité pour un sujet volontaire et conscient de prendre une décision sans en référer au préalable à une autorité supérieure, à pouvoir donner les motifs de ses actes, et à être jugé sur eux. La responsabilité est le contrepoids de la liberté. Il ne peut y avoir de responsabilité sans liberté : nul ne peut être tenu responsable d'actes effectués sous la contrainte.

Dans les organisations, chaque changement conduit à une nouvelle distribution des tâches et des rôles qui, consécutivement, modifie les responsabilités individuelles, ainsi que le rapport que chacun entretient avec le résultat global. Au travail, notamment dans l'industrie, l'organisation se traduit par la division d'une activité en plusieurs tâches réalisées par des travailleurs spécialisés. De façon quasi mécanique, cette division du travail a pour conséquence la fragmentation des responsabilités. L'organisation détermine qui est responsable d'une tâche. Cette responsabilité s'étend à l'exécution de la tâche, c'est la responsabilité opérationnelle ; et à la

responsabilité des effets de la part réalisée sur le résultat d'ensemble, c'est la responsabilité conséquentialiste. Plus une organisation est complexe, plus les responsabilités sont partagées entre un grand nombre d'individus et plus le sentiment de responsabilité à l'égard du tout se brouille. Pour David Graeber [29], le travail moderne prive une grande partie des travailleurs d'une motivation humaine majeure que le psychologue Karl Groos nomme la « *joie d'être cause* » : ce que le travailleur ressent lorsqu'il prend conscience que c'est lui qui provoque une action. Sans la conscience explicite d' « être cause », les travailleurs peuvent difficilement se sentir pleinement responsables de leurs activités.

L'intégration de machines intelligentes peut affecter la distribution des responsabilités suivant six mécanismes :

- Une fragmentation des responsabilités entre le travailleur et la machine intelligente : la délégation de tâche, a fortiori de tâches décisionnelles, rend le travailleur moins opérationnel *(« pourquoi assumer la responsabilité d'une tâche que je n'exécute plus* ? »). Chacun tend à sentir responsable de ce qu'il fait et à se désintéresser de ce qu'il ne fait pas ou plus. Cette nouvelle division du travail l'éloigne en outre d'une perception globale d'une séquence opératoire et donc d'un sentiment responsabilité à l'égard de l'ensemble.
- Il devient de plus en plus difficile de définir la distribution des responsabilités entre toutes les parties-prenantes (concepteur, ingénieur, programmeur, fabricant, vendeur, utilisateur), notamment dans le cas de machines apprenantes dont le comportement dépend partiellement d'une adaptation à l'environnement. Il semble exister une responsabilité « commune » ou « partagée » entre le concepteur, l'ingénieur, le programmeur, le fabricant, l'investisseur, le vendeur et l'utilisateur du robot. Sauf rares exceptions, aucun de ces acteurs ne peut être désigné comme la source ultime d'un acte. Cette approche, cependant, a pour effet de diluer tout à fait la notion de responsabilité : si tous ces acteurs ont une part de la responsabilité totale, aucun d'eux n'est entièrement responsable.
- Une emprise des processus qui réduit la liberté : la responsabilité découlant fondamentalement d'une liberté, l'emprise des algorithmes sur la pratique pourrait délier moralement le travailleur de son activité.
- Un effacement du travailleur devant l'autorité machinique : l'efficacité voire la supériorité présumée du SIA peut générer des conduites de retrait par excès de confiance, de contentement (se satisfaire d'une solution jugée correcte mais non optimale au nom de l'efficience), ou de prudence.
- L'automatisation d'une tâche peut conduire à son invisibilisation. Progressivement, la tâche automatisée échappe à l'attention, à la conscience de la situation du travailleur qui se focalise sur les tâches sur lesquelles il intervient. En outre, le robot exécute silencieusement. Basculant dans l'invisible et l'indicible, les tâches déléguées pourraient cesser d'exister socialement.
- Le problème de l'explicabilité des algorithmes d'apprentissage automatique dont l'effet de « boîte noire » est considéré un obstacle majeur à l'acceptabilité sociale des SIA dans de nombreuses activités soumises à des enjeux élevés d'imputation de la responsabilité (comme la justice et la santé, par exemple, plus généralement les SIA à haut risque tels que définis par la Commission Européenne [30]).

L'ensemble de ces facteurs peut ainsi altérer l'éthique du travail, plus précisément ce que Matthew Crawford [31] qualifie de « *vertu infra-éthique* » : une situation où le travailleur est en capacité de jugement, où, contre le « *détachement contemplatif* », il se met en jeu (moralement et physiquement), manifeste de l'intérêt, se confronte à la réalité et peut ainsi faire l'expérience directe de sa responsabilité. Cette éthique du travail est tributaire d'une expérience de l'agir humain qui découle d'une « *friction psychique* » entre l'usager et sa réalisation alors même que les outils informatiques, en minimisant ces frictions par des interfaces toujours plus intuitives, réduisent la conscience de la réalité. Crawford oppose à cette expérience désincarnée, le concept d'« *agir individuel* » comme « *expérience directe de notre responsabilité à l'égard de notre environnement matériel* », condition qu'il estime impérative pour un « *travail doté de sens* ».

Exemple de SIA portant une problématique de savoir-faire : L'intégration d'un cobot dans une activité de contrôle bactériologique du lait est bloquée par l'impossibilité d'imputer des responsabilités en cas de crise sanitaire. Aucun service n'est prêt à endosser la responsabilité d'une erreur du cobot.

Critère 6 : La déresponsabilisation

6.1. L'imputation de responsabilités en cas de problème est-elle un enjeu majeur de l'activité et de l'organisation ? (O=1 ; N=0)

6.2. Le SIA utilise des algorithmes d'apprentissage lui permettant de s'adapter dans un environnement aléatoire ? (ce qui génère de l'imprévisibilité dans le comportement du SIA) (O=1 ; N=0)

6.3. Le SIA réduit-il l'espace de jugement du travailleur (en augmentant l'emprise des procédures, en fragmentant les responsabilités, en invisibilisant des actions, en expliquant pas ses décisions recommandations)? (O=1 ; N=0)

6.4. Le SIA peut-il alimenter un effacement volontaire ou inconscient du travailleur (excès de confiance, effet de contentement, excès de prudence) ? (O=1 ; N=0)

Mots-clés : Jugement, fragmentation des responsabilités, effacement du travailleur, réduction de la liberté, conscience de la situation, explicabilité.

4 Utilisation de l'outil « MAIAT»

4.1 Quels utilisateurs pour MAIAT?

Nous envisageons plusieurs cas de figure. Le plus évident renvoie à notre propre profil, à savoir un accompagnateur-expert de transformations métiers et organisations liées à l'introduction d'une nouvelle technologie en générale, un SIA

en particulier. Dans cette configuration, l'outil sert de guide au diagnostic et l'expertise de l'utilisateur ne justifie pas une formalisation particulièrement aboutie. Cependant, dans le cadre du programme Confiance AI, nous visons la conception d'outils de construction de la confiance maniables par une plus large communauté que celles des experts. Nous ambitionnons par exemple des scénarios d'usage où l'outil est mobilisé par le fournisseur du SIA ou par son client. Dans cette perspective, le niveau de formalisation doit être plus avancé, l'outil de diagnostic doit pouvoir fonctionner sans portage et les utilisateurs rendus autonomes par un design qui facilitera son appropriation. Se pose alors la question du système d'acteurs pertinent pour réaliser de diagnostic. Nous pensons qu'il doit être réalisé de façon transparente et plurielle par un groupe représentatif de la communauté professionnelle. D'une part, cette approche ouverte est plus favorable à la construction de la confiance qui sera d'autant plus robuste que le diagnostic sera partagé. D'autre part, cela responsabilise toutes les parties-prenantes vis-à-vis d'un SIA qui devra faire l'objet de retour d'expériences et d'améliorations. MAIAT participera en ce sens à une montée en compétence collective autour de l'IA qui peut constituer un temps social fédérateur. Pour autant, cette utilisation « démocratique » peut aussi inspirer des craintes, en multipliant des questions anxiogènes qui risquent de faire passer la proposition de valeur du SIA au second plan. L'approche collégiale, ouverte et transparente sera privilégiée dans le cadre d'un climat social apaisé.

Les utilisateurs peuvent également être des représentants élus du Comité Social et Economique, qui doit légalement être informé de toute transformations organisationnelles. Or, la culture actuelle en matière de SIA est peu développée et la réception sociale d'un SIA peut se baser essentiellement sur des représentations communes de l'IA qui souffrent de deux écueils :

- La technophobie qui conduit à rejeter l'IA à partir d'idées reçues et développer à une critique non seulement peu constructive mais également éloignée de ce que sont réellement les SIA aujourd'hui. Dans ce cas de figure, notre outil permet de convertir une peur irrationnelle en vigilance pertinente.
- La technophilie qui nourrit des attentes disproportionnées vis-à-vis du SIA susceptibles 1) de générer des déceptions, les SIA étant régulièrement victimes de leur marketing ; 2) d'ignorer des points de vigilance importants. Ici, l'outil permet de convertir une confiance aveugle en confiance informée.

Or, les représentants élus au CSE peuvent réclamer le statut de « projet important » pour l'introduction d'un SIA, qualification que la direction validera ou non. La reconnaissance déclenchera le recours à une mesure d'expertise afin d'examiner les modifications des conditions de santé, de sécurité et les conditions de travail. Dans ce contexte, notre outil peut contribuer à dépassionner la concertation en l'arrimant à des critères et des indicateurs qui organiseront un diagnostic partagé et objectivé. En 2016, l'introduction au Crédit Mutuel de Watson (IBM) sur des opérations d'analyse de courriers électroniques et de réponse en temps réel aux questions des clients sur des produits techniques a justement engendré un conflit entre le CSE et la direction qui refusait de reconnaître le caractère important du projet. Le CSE (à l'époque CHSCT) « *avait fait valoir que le projet de technologie cognitive constitué par le logiciel Watson mis en place pour optimiser le travail des chargés de clientèles portait en lui-même la potentialité d'un redécoupage des missions des salariés au sein d'une agence et donc une modification notable des conditions de travail* ». Saisie, la Cours de Cassation a produit une réponse étonnante. Non seulement elle n'a pas reconnu le caractère important du projet, ce pourquoi elle était précisément saisie, mais elle a estimé en outre que les tâches seraient facilitées par Watson, considération qualitative inédite dans ce genre de jugement. Plus déroutant encore, le Tribunal de Grande Instance de Paris, sur le même usage de Watson par le CIC, avait émis un jugement très différent, en qualifiant le projet d'important non seulement en raison du nombre de salariés impactés mais aussi et surtout parce que « *l'outil dont l'implantation est discutée est déjà porteur de fonctionnalités et de capacités dont les limites sont encore mal définies, compte-tenu de son aspect "apprenant" et évolutif* ». Le TGI considèrait ainsi le caractère dynamique et évolutif de la technologie quand la CC a adopté une approche statique. Cet exemple démontre la difficulté à appréhender les effets sociaux d'un SIA et à produire une évaluation sinon objective ou moins « intersubjective » de ces nouvelles technologies. L'outil « MAIAT » peut adresser ce besoin, tant pour les CSE que pour les juges.

4.2 Comment fonctionne MAIAT?

Dans sa configuration actuelle, sommaire, MAIAT fonctionne comme un outil de notation :

- Chaque critère est composé de quatre indicateurs
- Les indicateurs se renseignent en répondant « oui » ou « non » à une question. Cette binarité laisse moins de place à la subtilité qu'une échelle graduelle qui permettrait une évaluation plus fine. En l'état actuel, nous avons pourtant maintenu cette modalité. Une échelle graduelle, en produisant des nuances rend moins lisible le diagnostic. Il s'agit de faire remonter des points de vigilance et non d'acter des nuisances. Les questions, certes formulées avec un minimum d'ambiguïté, doivent s'entendre ainsi : « *identifiez-vous un risque de…* ». L'approche binaire a également le mérite de dégager plus clairement des points d'attention et ainsi de faciliter la priorisation des actions de remédiation.
- Chaque indicateur produit un score « 0 » ou « 1 » en fonction de la réponse. Nous avons précisé l'allocation des valeurs de chaque indicateur plus haut. Ces valeurs se cumulent pour former une note entre 0 et 4 pour chaque critère.
- Ces notes formalisent un schéma radar (fig. 1) qui permet de visualiser le résultat du diagnostic. Actuellement, ce schéma ne distingue pas les deux grandes familles de critères (« Bien-être » et « Engagement »). Nous envisageons l'opportunité de les distinguer en les positionnant sur un même schéma afin d'offrir un meilleur visuel. Nous redoutons cependant que cela conduise à une différenciation excessive de critères qui demeurent très emboîtés.

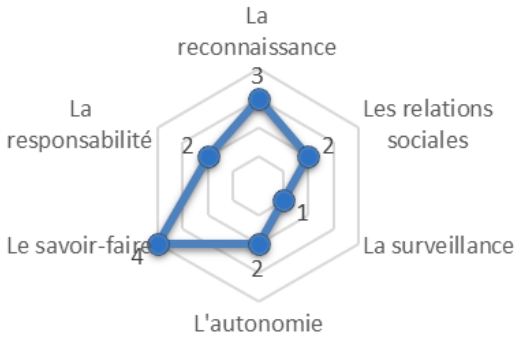


Fig.1 : Visualisation des points d'attention de MAIAT

Une fois chaque critère évalué, l'utilisateur est invité, au-delà de sa note, à en formuler une explication contextualisée, puis à évoquer les remédiations envisagées (fig.2). Actuellement, MAIAT ne produit pas de recommandations, il est seulement un outil d'aide au questionnement et à l'évaluation. Dans notre utilisation actuelle, nous pouvons être sollicités pour proposer des « solutions » à la suite du diagnostic. Dans le cas d'une formalisation sans portage par un expert, la production de recommandations est plus délicate, en raison d'un défaut de contextualisation. Cela convertirait « MAIAT» en outil d'aide à la décision ce qui 1) est très complexe quand un outil adresse des situations sociales ; 2) peut réduire l'engagement subjectif de l'utilisateur c'est-à-dire l'expression de sa perception de la situation, pour des raisons similaires à celles détaillées pour les SIA. Nous envisageons davantage MAIAT comme un outil de définition partagée de situations sociales mal structurées, sur lesquelles les acteurs manquent de ressources et de repères pour agir, plutôt qu'outil de résolution de problèmes. Si MAIAT aide à problématiser, les solutions viennent des acteurs des situations de travail. Nous avons conclu chaque série d'indicateurs par des mots-clés. Ceux-ci préfigurent une formalisation de MAIAT où les critères seront moins détaillés que dans cet article. Ils ont vocation à faciliter la compréhension des dimensions contenus dans le critère.

4.3 Quand utiliser MAIAT?

L'outil MAIAT peut être utilisé par un fournisseur pour évaluer la robustesse sociale de son SIA en vue d'améliorer le design de son système et/ou d'accompagner ses clients sur les enjeux sociaux du SIA. Nous concevons cependant MAIAT en référence au contexte du programme Confiance AI, à savoir l'expérimentation d'un assistant vocal intelligent dans un environnement industriel et plus généralement la mise à disposition d'outils génériques pour faciliter la construction de la confiance dans les SIA. A ce titre, nous envisageons l'utilisation de MAIAT dans le cadre d'un projet d'intégration d'un SIA afin d'identifier des points d'attention qu'il suscite dans un contexte de déploiement donné. Dans cette configuration d'usage, MAIAT est outil de diagnostic et de suivi de ces points d'attention en vue d'accompagner un processus de socialisation du SIA.

- Un outil de diagnostic : MAIAT permet de réaliser un pronostic d'usage à partir des résultats de l'évaluation. Celle-ci est idéalement réalisée par le collectif de travail à la suite d'un atelier commun qui fera donc ressortir des indicateurs autour des six critères. Cependant, ce diagnostic sera réalisé après un temps d'utilisation forcément limité qui ne permettra pas d'épuiser toutes les situations. Il risque notamment de faire référence à des situations de travail prescrit (*a fortiori* s'il est réalisé par les seuls managers) et non de travail réel que les travailleurs savent du reste rarement expliciter. Au final, ce premier diagnostic reposera davantage sur une représentation de l'activité que sur l'activité réelle.
- Un outil de suivi : l'usage de MAIAT ne doit dès lors pas s'arrêter à l'établissement d'un diagnostic statique mais être mobilisé dans le cadre d'un suivi et de retours d'expérience réguliers. D'une part, à l'épreuve du travail réel, les indicateurs mesurés peuvent être réévalués. D'autre part, l'usage du SIA peut faire apparaître de nouvelles situations de travail, mais aussi de nouveaux savoir-faire porteurs d'opportunités imprévisibles au moment du diagnostic.

Scardigli [32] a ainsi noté qu'il existe « *trois temps de l'insertion sociale des techniques* ». Le premier, encombré d'images, de discours et de promesses est le temps des discours prophétiques qui précèdent et accompagnent l'insertion et l'expérimentation de l'innovation technologique dans le corps social. Le deuxième, celui de la diffusion de l'innovation, voit se développer les premiers usages, l'entrée en scène des

	SUIVI DES EFFETS SOCIAUX D'UN SIA	Indicateurs	Mots-clés	Criticité (1 à 4)	JUSTIFICATION DE LA CRITICITÉ	DISPOSITIONS D'ACCOMPAGNEMENT	OBSERVATIONS	OBSERVATIONS N+ 6 MOIS	OBSERVATIONS N+ 1 AN
1- IMPACTS SUR LE BIEN-ÊTRE	La reconnaissance	1. 0/1 2. 0/1 3. 0/1 4. 0/1	Reconnaissance, singularité du travailleur, pratique, efforts, substituabilité, mérite, valorisation, inégalités justes.	3					
	Les relations sociales	1. 0/1 2. 0/1 3. 0/1 4. 0/1	Communication, coopération, travail collectif, collectif de travail, appauvrissement du langage, interactions sociales.	2					
	La surveillance	1. 0/1 2. 0/1 3. 0/1 4. 0/1	Surveillance, contrôle, traçabilité, mesure de productivité individuelle, privacy, RGPD, proportionnalité.	1					
2- IMPACTS SUR L'ENGAGEMENT	L'autonomie	1. 0/1 2. 0/1 3. 0/1 4. 0/1	Procédures, initiative, travail réel, liberté, notification, pratiques professionnelles, flexibilité cognitive.	2					
	Le savoir-faire	1. 0/1 2. 0/1 3. 0/1 4. 0/1	Identité professionnelle, savoir-faire, valeur du travail, processus, pratique, obéissance, sens du travail, identité professionnelle, Déqualification/requalification	4					
		1. 0/1 2. 0/1	Jugement, fragmentation des responsabilités, effacement du	2					

Fig.2 : Tableau de suivi de MAIAT

médiateurs et des prescripteurs. À l'enthousiasme éventuel pour une nouvelle technologie, fait souvent suite une phase de désillusion, de refus partiel ou total des promesses. Enfin, le troisième temps, celui de l'appropriation socio-culturelle de l'innovation, est celui où les usages de la technique se stabilisent : alors se produit un mouvement d'acculturation, voire de naturalisation de la technique. Rapporté au travail et à notre outil, ce séquencement de l'insertion sociale de la technique renvoie à l'étape de découverte du SIA, de son expérimentation (lors d'un PoC) et le cas échéant de son appropriation (si le SIA est mis en production). Tout au long de ce parcours, s'expriment des représentations, puis des retours d'expérience, durant lequel le SIA est transformé par le travail et le travail transformé par le SIA. MAIAT a vocation à scander ce processus de socialisation autour d'indicateurs dynamiques mesurés et renseignés collectivement.

5 Conclusion

MAIAT est un outil d'accompagnement de la construction sociale de la confiance dans un SIA dédié au travail. Son expérimentation dans le cadre du programme Confiance AI débutera en avril 2022. Si notre article est retenu, nous pourrons présenter des premiers résultats lors de la PFIA 2022 à Saint-Etienne.

Remerciements

Nous remercions nos partenaires du programme Confiance AI pour leur…confiance, en particulier Renault Group et Spix Industry. Nous remercions Maître Oustin Astorg pour le partage de ses travaux et analyse du cas Watson au Crédit Mutuel et au CIC. Nous remercions également le Future of Work Working Group du PMIA pour son travail de recensement de cas d'usage réels de l'IA au travail. Nous remercions enfin l'équipe du LaborIA pour notre prometteuse coopération. Nous remercions enfin Alexandre Blanche, Samuel Crespy et Augustin Debray pour avoir relu et commenté la première version de ce texte. Les auteurs assument l'entière responsabilité de son contenu.

6 Références

[1] C . Benedikt, M. Osborne, « The Future of Employment : How Susceptible are Jobs to Computerization ? », *Technological Forecasting and Social Change*, Vol. 11. 4, issue C, pp.254-280, 2013.
[2] D. Autor, « Why Are There Still So Many Jobs? The History and Future of Workplace Automation », *Journal of Economic Perspectives*, vol. 29, n°3, pp. 3-30, 2015.
[3] M. Arntz, T. Gregory, U. Zierahn, « The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis », *OECD Social, Employment and Migration Working Papers*, No. 189 ; Paris, OECD Publishing, 2016.
[4] J. Suzman, *Travailler, La grande affaire de l'humanité*, Flammarion, 2021.
[5] GPAI, Future of Work, AI Observatory at the Workplace: https://gpai.ai/projects/future-of-work/ai-at-work-observation-platform/ai-observatory-at-the-workplace.pdf
[6] MTEI : https://travail-emploi.gouv.fr/actualites/l-actualite-du-ministere/article/laboria-centre-de-ressources-et-d-experimentations-sur-l-intelligence
[7] Confiance AI: https://www.confiance.ai/
[8] J.-P. Brun, N. Dugas, « La reconnaissance au travail : analyse d'un concept riche de sens », *Gestion*, volume 30, numéro 2, pp. 79-88, été 2005.
[9] M.-E. Bobillier Chaumon, M.E., « Acceptation située des TIC dans et par l'activité : Premiers étayages pour une clinique de l'usage », *Psychologie du Travail et des Organisations*, Vol. 22(1), pp. 4-21, 2016.
[10] D. Wolton, *Penser la communication*, Flammarion, 1997.
[11] J. Perry, *La procrastination*, Autrement, 2020.
[12] Y. Ferguson, *Politiser l'action publique. Une approche par les instruments, le cas du programme Constellation*, Sociologie. Université Toulouse le Mirail - Toulouse II, 2014.
[13] S., Caroly, Y., Clot, « Du travail collectif au collectif de travail : développer des stratégies d'expérience », *Formation Emploi*, Vol.88, pp. 43-55, 2004.
[14] Y. Clot, *Le prix du travail bien fait. La coopération conflictuelle dans les organisations*, La Découverte, 2021.
[15] S. Tisseron, *Le jour où mon robot m'aimera. Vers l'empathie artificielle*, Albin Michel, 2015.
[16] V. Scardigli, *Un anthropologue chez les automates*, PUF, 2001.
[17] France Stratégie, *Anticiper les impacts économiques et sociaux de l'Intelligence Artificielle. Annexe 1 : L'intelligence Artificielle en quête d'acceptabilité et de confort*, 2016.
[18] Cnil, La surveillance, video-protection au travail, 2019 : https://www.cnil.fr/sites/default/files/atoms/files/travail-vie_privee.pdf
[19] S. Head, *Mindless: Why Smarter Machines are Making Dumber Humans?*, Basic Books, 2014.
[20] J.-P, Meyer, J.P, N.J. Allen, (1991) A Three-Component Conceptualization of Organizational Commitment. *Human Resource Management Review*, 1, pp. 61-89.
[21] F. Dubet, *Injustices. L'expérience des inégalités au travail*, Seuil, 2006.
[22] Freyssenet, M., *Le processus de déqualification-surqualification de la force de travail*, Paris, CSU, 1974.
[23] Zacklad, M. « *Intelligence Artificielle : représentations et impacts sociétaux* », CNAM, 2017.
[24] Mission Villani, *Donner un sens à l'intelligence artificielle*, rapport mission parlementaire, 2018.
[25] Y. Clot, *Le travail sans l'homme*, La Découverte, 1995.
[26] R. Gelin, O. Guilhem, *L'intelligence artificielle, avec ou contre nous ?* La documentation française, 2021.
[27] Y. Clot, *Le travail à cœur. Pour en finir avec les risques psychosociaux*, La Découverte, 2015.
[28] N. Carr, *Remplacer l'humain. Critique de l'automatisation de la société*, L'Echappée, 2017.
[29] D. Graeber, *Bullshit Jobs*, Les Liens qui Libèrent, 2018.
[30] Commission Européenne, *Proposition de règlement du parlement européen et du conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'union*, 2021.
[31] M., Crawford, M., *Éloge du carburateur*, La Découverte, 2016.
[32] P. Musso, « Usages et imaginaires des TIC », in *L'évolution des cultures numériques*, FYP éditions, pp.201-210, 2009.

XAI et information géographique: application aux reconstructions paléoenvironnementales

Bastien Zimmermann[1], Matthieu Boussard[1], Nicolas Boulbes[2], Sophie Grégoire[3]
[1] craft.ai
[2] Institut de Paléontologie Humaine, Fondation Albert Ier, Paris, UMR 7194 HNHP, EPCC-CERP Tautavel
[3] UMR 7194 HNHP, EPCC-CERP Tautavel

bastien.zimmermann@craft.ai, matthieu.boussard@craft.ai, nicolas.boulbes@cerptautavel.com, sophie.gregoire@cerptautavel.com

Résumé

Ce travail illustre les apports de l'explicabilité à la reconstruction des paléoenvironnements. Il s'agit de construire un modèle permettant, à partir des données de fouilles, de déterminer l'environnement correspondant à un niveau archéologique et une période du Paléolithique donnée. Dans ce contexte, la prédiction seule du modèle a moins de valeur que les explications sous-jacentes qui permettent aux archéologues de remettre en cause leurs hypothèses. Dû aux incertitudes sur les données, ce travail porte sur les explications orienté données, comme data-Shapley. Enfin, cet article propose l'utilisation d'un système d'information géographique permettant d'exploiter au maximum les informations issues des outils d'explicabilité.

Mots-clés

XAI, explicabilité, Système d'information géographique, apprentissage automatique, communautés animales, paléoenvironnements.

Abstract

This work shows the contribution of explicability to the reconstruction of paleoenvironments. It aims at building a model allowing, from excavation data, to infer the environment corresponding to a given layer and a given paleolithic period. In this context, the prediction of the model alone has less value than the underlying explanations that allow archaeologists to question their assumptions. Due to the uncertainties in the data, this work focuses more on data-oriented explanations tools, such as data-Shapley. Finally, a contribution of this article is the use of a Geographic Information System allowing us to exploit to the maximum the information we can obtain from the explainability tools.

Keywords

XAI, explainable AI, Geographic Information System, machine learning, animal communities, palaeoenvironments.

1 Introduction

Rendre les systèmes à base d'Intelligence Artificielle (IA) fiables est l'un des enjeux majeur de la recherche et du développement autour des techniques d'apprentissage machine. La fiabilité d'un système est aisément remise en cause lorsque celui-ci fonctionne de façon obscure ou bien s'il repose sur de mauvaises fondations par exemple de mauvaises données. L'intelligence artificielle explicable (XAI) est l'un des piliers permettant d'aboutir à une IA de confiance. En particulier, les méthodes explicatives centrées sur les données s'inscrivent dans une démarche visant à recentrer l'attention sur les données qui ont autant, voire plus d'importance que le modèle dans le processus d'apprentissage automatique. En effet des données corrompues ou utilisées dans le mauvais contexte mènent à des conclusions erronées. Dans ce contexte, les algorithmes évaluant la qualité des points de données, permettent d'orienter plus efficacement le travail de nettoyage, d'augmentation, d'amélioration de la qualité de celles-ci et ainsi de construire des modèles plus performants. Nous avons ici utilisé **Beta-Shapley** sur des données de fouilles archéologiques, travaillées dans le cadre du programme ANR SCHOPPER [6], afin d'attribuer une valeur à nos différents points de données.

L'objectif est d'utiliser un algorithme d'apprentissage machine pour prédire un biome (une unité écologique positionnée géographiquement) en fonction des espèces animales caractéristiques de ce biome ou des écorégions qui le compose. Les restes fossiles d'animaux retrouvés dans un niveau archéologique sont utilisés pour estimer, en se basant sur le principe de l'actualisme, le biome associé et par conséquent les conditions climatiques qui avaient cours au moment de la mise en place du niveau.

Beta-Shapley permet d'avoir une analyse en amont de ce processus et d'analyser ce qui sert à entraîner notre modèle. Dans un premier temps, la pertinence de l'outil a été évaluée au travers de l'identification de données de haute et faible qualité. Nous montrons la cohérence entre la valeur du point et son impact sur la performance d'un modèle. Ensuite une analyse plus poussée met en valeur les différentes informations que peut apporter cet outil. Finalement, les explications apportées par ces outils permettent de voir nos données sous un nouveau jour et ouvrent des perspectives inédites, notamment via l'exploitation de la di-

mension spatiale. L'utilisation croissante des systèmes d'information géographique couplée à des modèles de distribution d'espèces (Ecological Niche Modelling) pour l'étude des paléoenvironnements[12],justifie le développement de méthodes explicatives d'apprentissage machine sur ce type d'outil.

Modèles prédictifs

1.1 Contexte

Afin d'apporter une illustration concrète des apports des outils d'XAI et plus spécifiquement les outils XAI orientés data, nous avons choisi l'exemple de la reconstruction paléoenvironnementale à partir d'un site archéologique. Ainsi, il s'agit de déterminer l'environnement et le climat correspondant à une époque donnée en considérant certains indices biologiques identifiés lors de fouilles archéologiques. Le site utilisé est la grotte paléolithique de la Caune de l'Arago à Tautavel, dans le sud de la France, qui a bénéficié de 54 ans de fouilles et d'études multidisciplinaires d'une séquence stratigraphique de 15 mètres d'épaisseur, développée entre 690 000 ans et 90 000 ans BP. [1]. Ce site a livré près de 600 000 objets répartis dans 55 niveaux archéologiques. La richesse de cet enregistrement paléolithique, la qualité de la conservation des vestiges et le système d'enregistrement standardisé des données dont il bénéficie, en font l'un des meilleurs terrains d'application des outils présentés ici pour les reconstructions paléoenvironnementales. Les reconstructions paléoenvironnementales se basent sur le principe de l'actualisme. Il se base sur l'hypothèse que les systèmes biologiques du passé fonctionnent de la même manière que ceux que l'on peut observer actuellement. Par exemple s'il est possible d'observer la répartition actuelle des différentes espèces animales pour différentes régions du globe, il est théoriquement possible de déterminer le climat du passé à partir d'un assemblage faunique fossile grâce aux affinités écologiques connues des espèces animales actuelles en les transposant aux communautés animales fossiles. Ce principe permet d'identifier le climat correspondant à une couche archéologique donnée, à partir des restes d'ossements retrouvés dans cette dernière. Toutefois, de nombreux facteurs de risques pouvant conduire à une conclusion erronée sont à prendre en considération, tels :

- la qualité du jeu de données actuelles
- la représentativité des taxons retrouvés lors des fouilles (état de conservation et biais dûs à la prédation)
- l'évolution des espèces (adaptation, migration, disparition)
- le principe même d'actualisme.

Des approches d'apprentissage machine pour la reconstruction des paleoenvironnements ont été proposées dans [10]à partir de données polliniques. Nous proposons ici d'utiliser des modèles basés sur des données fauniques. Plus fondamentalement, là ou les auteurs se focalisent sur l'aspect prédictif des modèles, nous nous intéressons ici aux explications des prédictions que l'apprentissage machine peut fournir.

2 Présentation des Données

Les données utilisées dans cet article sont de deux natures. Les unes actuelles, utilisées comme référence pour la constitution des modèles et les autres, archéologiques avec l'objectif de les classer en utilisant des modèles d'apprentissage machine et de définir ainsi les conditions environnementales associées.

2.1 Les jeux de données sur les environnements actuels

Le dataset actualiste utilisé est le *wildfinder dataset* [2], c'est une représentation biogéographique de la biodiversité terrestre. L'unité de base est l'écorégion comme représenté Fig. 1. Les écorégions sont établies en fonction de critères biogéographiques, définies comme "une unité étendue de terre ou d'eau qui contient un assemblage d'espèces, de communautés naturelles et de conditions environnementales, qui se distingue au plan géographique" Chaque éco-

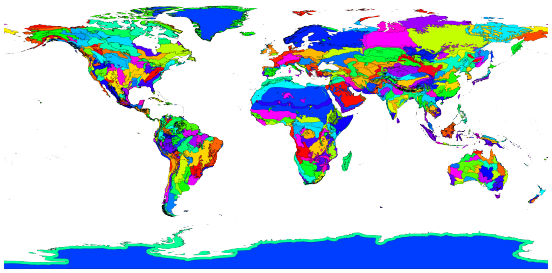


FIGURE 1 – Les différentes écorégions du monde

région contient une liste des espèces présentes sur les plus de 26000 répertoriées, ainsi qu'un type de biome associé. Ce dernier est un ensemble d'écosystèmes caractéristique d'une aire biogéographique. Chacun des 14 biomes est représentatif d'un certain climat (Fig. 2). Un niveau encore plus général de description existe, celui des écozones qui, au nombre de 8, représentent bien la répartition de la faune actuelle sur la planète. Les deux que nous considérons dans ce travail, sont le Paléarctique correspondant à l'Europe, l'Afrique du nord, les deux-tiers nord de l'Asie ainsi que le Moyen-Orient (sauf l'Arabie), et le Néarctique correspondant à l'essentiel de l'Amérique du Nord, c'est à dire les écozones de l'hémisphère nord. La richesse des données est présentée par la Fig. 3. Les différents biomes ne sont pas représentés de manière égale, il y a plus d'écorégions de *Temperate Broadleaf and Mixed Forests* que de *Montane Grasslands and Shrublands*. De plus certains biomes ont une diversité d'espèces plus faible, un nombre moins grand d'espèces est présent en moyenne dans les écorégions de la *Tundra* que ceux de *Temperate Coniferous Forests*.

Le premier outil d'explicabilité sur les données utilisable est une visualisation de celles-ci. La figure Fig.4 permet d'avoir des représentations alternatives. L'utilisation de la PCA (Principal Conponent Analysis) et t-SNE (t-distributed Stochastic Neighbor Embedding) permettent d'obtenir en deux dimensions une représentation pertinente

des données. En effet la PCA permet de transformer les variables entre elles et de ne garder que les deux composantes principales, décorrellées des autres et expliquant au mieux la variance. L'algorithme t-SNE, lui, a pour caractéristique de conserver la proximité des points pendant la transformation réduisant la dimension.
Ces deux représentations haut niveau permettent de constater que certain clusters existent et d'apprécier la dimension Vapnik–Chervonenkis de nos données (la théorie du même nom vise à expliquer l'apprentissage d'un point de vue statistique). En effet il semble à priori possible de distinguer les différents groupements de biomes.

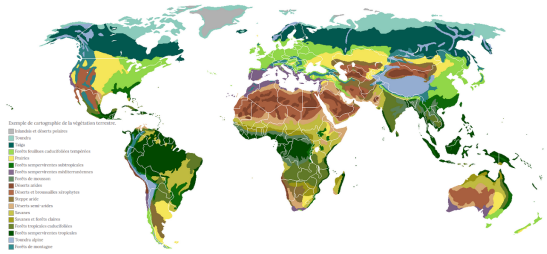


FIGURE 2 – Les différents biomes du monde

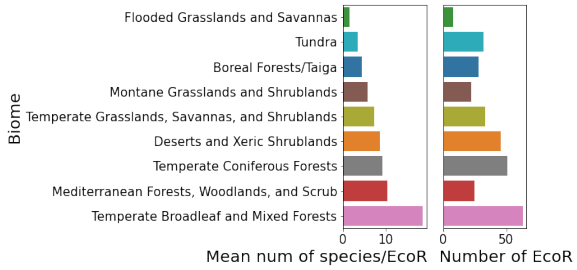


FIGURE 3 – Nombre d'espèces par écorégion par biome et nombre d'écorégions par biome. (seulement les espèces de la Caune de l'Arago)

2.2 Le jeu de données paléolithique

Un dataset est constitué avec les espèces de faunes identifiées dans tous les niveaux archéostratigraphiques de la Caune de l'Arago. Il rassemble les espèces de grands et petits mammifères, amphibiens, reptiles et oiseaux, déterminées à partir des restes fossiles (ossements, dents) et correspond à l'inventaire taxonomique de la communauté de vertébrés présents dans chaque couche archéologique de la Caune de l'Arago. Au total, le nombre d'espèces qui représentent les variables dans le dataset s'élève à 144. Le but de ce dataset est d'identifier les biomes et écorégions représentées dans chaque niveau archéologique afin de reconstituer l'environnement de la grotte et d'identifier le type de paysage et le climat dominant à chaque période d'occupation du site par des groupes humains.
Une fois les deux datasets constitués, nous restreignons le jeu de données WWF [2]. Géographiquement seules les écozones pertinentes pour le site de la Caune de l'Arago

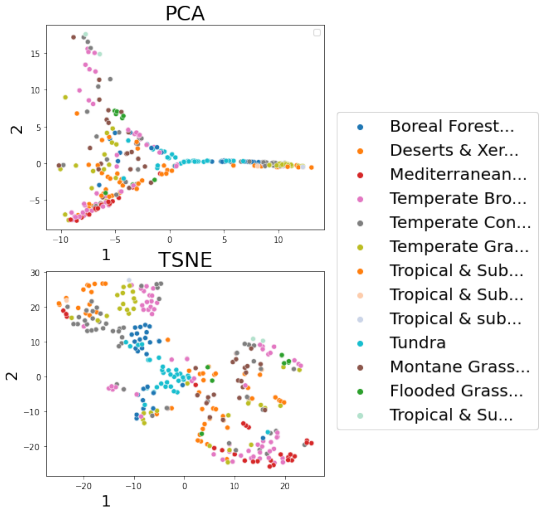


FIGURE 4 – Le jeu de donnée WWF représenté en deux dimensions

sont conservées, soit les écozones Paléarctique et Néarctique. Comme l'ensemble des taxons retrouvés lors des fouilles ne représentent qu'une petite partie des taxons existant de nos jours, le dataset WWF a dû être adapté. Les postulats de départ sont les suivants :

- Considération des espèces en présence/absence afin que l'écart entre les quantifications actuelles naturelles et celles du corpus archéologique (nécessairement plus limité) ne biaise pas les résultats des prédictions.
- Le choix de ne pas utiliser le critère d'abondance des taxons permet de plus d'éviter les biais liés à la conservation archéologique différentielle et à ceux de la sélection des espèces par leurs prédateurs (Hommes, carnivores, rapaces).
- Considération exclusive des espèces ayant été retrouvées au moins une fois dans au moins un des niveaux archéologiques, afin de ne pas prendre en compte des espèces trop éloignées des assemblages fossiles.
- Remplacement de certaines espèces éteintes par l'espèce actuelle la plus proche au plan écologique. Les espèces sans équivalent actuel (ex :le rhinocéros de prairie) sont écartés du dataset.

La figure 5 représente le dataset ainsi restreint. On peut y voir la répartition des 127 espèces restantes selon les écorégions auxquelles elles appartiennent de nos jours. Celles-ci sont plus nombreuses dans deux écorégions de l'Europe de l'Ouest (Northeastern Spain and Southern France Mediterranean forests, englobant la Caune de l'Arago, et Western European broadleaf forests) et leur densité s'affaiblit en s'en éloignant.

2.3 Objectif : Prédiction de types de Biome

Ces 127 taxons nous permettent donc d'entraîner un modèle dans l'objectif de prédire un biome à partir de la liste d'espèces (présence/absence) présentes dans l'écorégion. L'ob-

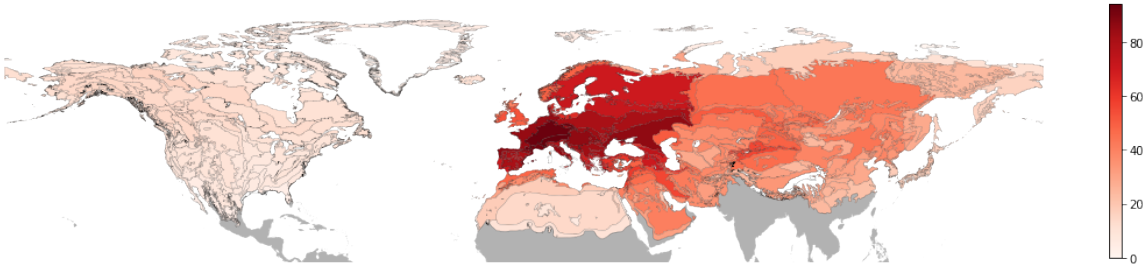


FIGURE 5 – Nombre d'espèces trouvées dans la grotte de l'Arago vivant dans chaque éco-region

jectif principal est d'exploiter ce modèle en inférence sur ces données archéologiques et prédire les conditions climatiques qui régnaient au cours de la mise en place des couches archéologiques en fonction des espèces retrouvées. Le modèle multi-classe de la figure 6 identifie les probabilités de présence des différents biomes pour chaque couche archéologique.
Cette approche prends une nouvelle dimension au travers d'outils d'explicabilité, l'inférence seule du modèle ne présentant en effet que peu de valeur pour un expert. Accompagner une prédiction d'explications permet de l'enrichir et d'apporter de nouvelles informations. Un des exemples est l'utilisation de Shap [8] pour fournir des explications accompagnant les inférences d'un modèle comme présenté figure 7.

3 Un outil d'explicabilité orienté données : beta-Shapley

Dans l'objectif d'attribuer une valeur aux points de donnés, A. Ghorbani et al. ont introduit le concept de data-Shapley [4]. Cette méthode se base sur le concept de théorie des jeux des valeurs de Shapley. Introduites initialement par Lloyd Shapley, elle propose une méthode équitable de répartition de gain. Ainsi, à partir d'un algorithme d'apprentissage et d'un jeu de donnée d'entraînement, data-Shapley est une métrique quantifiant la valeur de chaque point du jeu d'entraînement par rapport à la performance du prédicteur. Cette approche a de nombreux avantages, notamment que les points de faible valeur capturent les valeurs aberrantes et points corrompus, les points de haute valeur peuvent nous informer sur quel type de nouvelles données pourrait profiter à notre étude [5].
On définit la contribution marginale Δ_j comme suit [7] :

Definition 3.1 (Contribution Marginale)**.** Pour une fonction $h ; j \in [\![1 \ ; \ n]\!], n = |D|$ avec D notre dataset, on définit la contribution marginale d'un point $z^* \in D$ par rapport à $j-1$ points comme :

$$\Delta_j(z^*; h; D) = \frac{1}{\binom{n-1}{j-1}} \sum_{S \in D_j^{\setminus\{z^*\}}} h(S \cup z^*) - h(S)$$

avec $D_j^{\setminus z^*} = \{S \subseteq D \setminus \{z^*\} : |S| = j-1\}$

Le calcul de la valeur de data-Shapley pour un point de donnée est défini par [7] :

Definition 3.2 (Data Shapley)**.** La data-Shapley du point $z^* \in D$

$$\psi_{shap}(z^*; U; D) := \frac{1}{n} \sum_{j=1}^{n} \Delta_j(z^*; U; D)$$

avec $|D| = n ; U : \cup_{j=0}^{k} z^j \to \mathbb{R}$ une fonction d'utilité représentant la performance d'un modèle entraîné sur un dataset $\cup_{j=0}^{k} z^j ; k \in \mathbb{N}$

Les data-Shapley satisfont de manière unique les propriétés suivantes [7] :

1. **Efficacité :** Les allocations s'additionnent à la valeur de l'utilité du dataset complet.

$$\forall U, \sum_{z \in D} \psi(z; U; D) = U(D) \tag{1}$$

2. **Symétrie :** Pour tout U et toute permutation π sur D

$$\forall S \subseteq D, \psi(U(\pi(S))) = U(\pi(\psi U(S)))$$

3. **Joueur nul :** Un point z_i qui apporte une contribution marginale nulle reçoit une allocation nulle.

$$U(S \cup \{z*\}) = U(S) \ \forall S \subseteq D \setminus z^* \ \psi(z*; U; D) = 0$$

4. **Linéarité :** $\forall U_1, U_2$ fonctions d'utilité, $\forall \alpha_1, \in \mathbb{R}$,

$$\begin{aligned} \psi(z^*; \alpha_1 U_1 + U_2; D) =& \alpha_1 \psi(z^*; U_1; D) + \\ & \psi(z^*; U_2; D) \end{aligned} \tag{2}$$

Une généralisation de cet outil est présentée au travers de beta-shapley définies dans [7].
Les valeurs dites beta-Shapley sont dérivées des data-Shapley en relaxant l'axiome d'efficacité (1) et ajoutant deux hyperparamètres (α, β) décidant de poids ajoutés en fonction de la cardinalité. Une valeur élevée de α mettra une importance accrue sur les ensembles de petite cardinalité et réciproquement β élevé mettra l'importance sur les ensembles de grande cardinalité. Une illustration des paramètres est présentée Fig. 8.

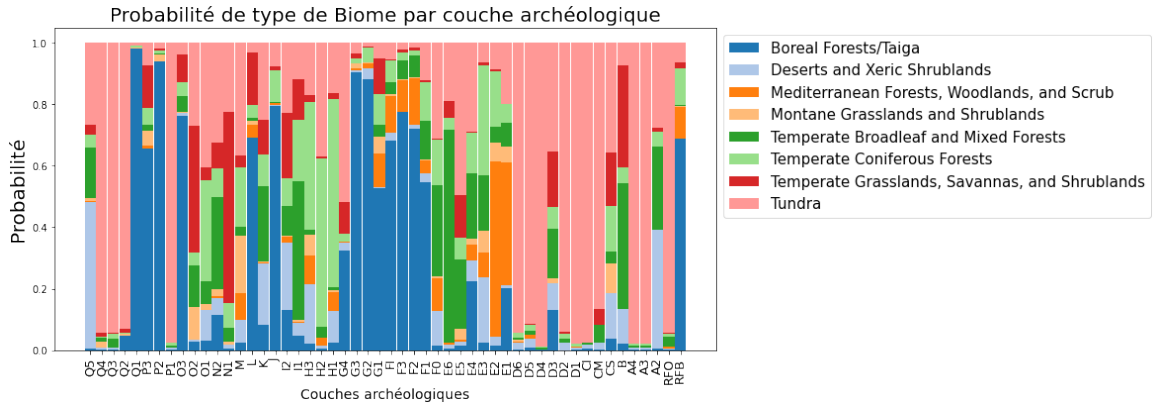


FIGURE 6 – Distribution de la représentation des biomes dans chaque couche archéologique

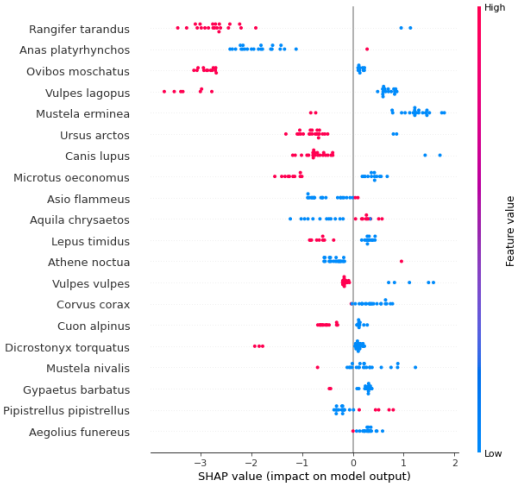


FIGURE 7 – Visualisation de l'importance des taxons via les shapley values pour le modèle de prédiction des biomes

Definition 3.3 (Beta Shapley)**.** La Beta Shapley du point $z^* \in D$

$$\psi_{beta}(z^*; U; D; w^{(n)}) := \frac{1}{n} \sum_{j=1}^{n} w^{(n)}(j) \Delta_j(z^*; U; D) \tag{3}$$

avec $w^{(n)} : [n] \to \mathbf{R}$ tel que :

$$n = \sum_{j=1}^{n} \binom{n-1}{j-1} w^{(n)}(j) \Delta_j(z^*; U; D)$$

En particulier nous utiliserons un schéma de poids défini à partir des paramètres α, β défini comme :

$$w^{(n)}_{\alpha,\beta}(j) = n \frac{Beta(j+\beta-1, n-j+\alpha)}{Beta(\alpha,\beta)}$$

avec : $Beta(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$; Γ la fonction gamma.

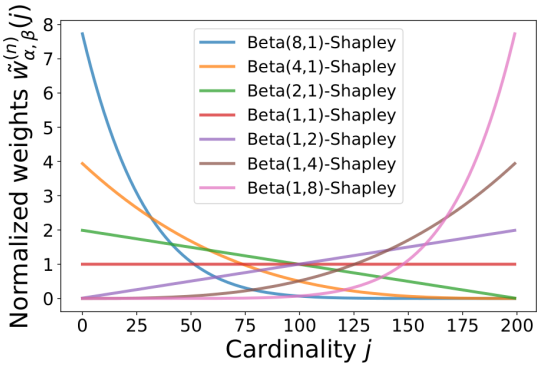


FIGURE 8 – Illustration des rôles de α , β

Pour notre étude nous avons calculé les valeurs beta-Shapley de chaque écorégion. Pour ce faire nous avons calculé la moyenne des valeurs de beta-Shapley au travers d'une stratégie de *stratified shuffle split*, nous le pouvons grâce à la propriété de linéarité (2) des valeurs beta-Shapley. Cette méthode de split permet une validation croisée en conservant le pourcentage de points de chaque classe. Cela nous permet de garder une évaluation des performances non biaisées tout en ayant une valeur pour chaque point de notre dataset.
Ainsi pour chaque split, les valeurs de beta-Shapley sont approximées par une méthode de Monte-Carlo dont la convergence est supervisée via la statistique de Gelman-Rubin [3]. Le modèle utilisé est un modèle de gradient boosting, LightGBM, et la métrique d'utilité est l'exactitude multi-classe sur l'ensemble du test correspondant à l'ensemble d'entraînement fourni par le *stratified shuffle split*.
La figure 9 donne un aperçu des résultats pour différents paramètres α, β. Plus la valeur attribuée à une écorégion est faible plus sa couleur sera rouge, plus sa valeur sera importante plus elle sera verte, la couleur blanche correspond à 0.
On distingue aisément que β élevé permet d'isoler des points précis tandis qu'un α élevé met en valeur des groupes de points. En effet, plus le α est élevé plus on se rapproche

d'une approche du type leave-one-out qui correspond à enlever un point et mesurer la différence de performance.

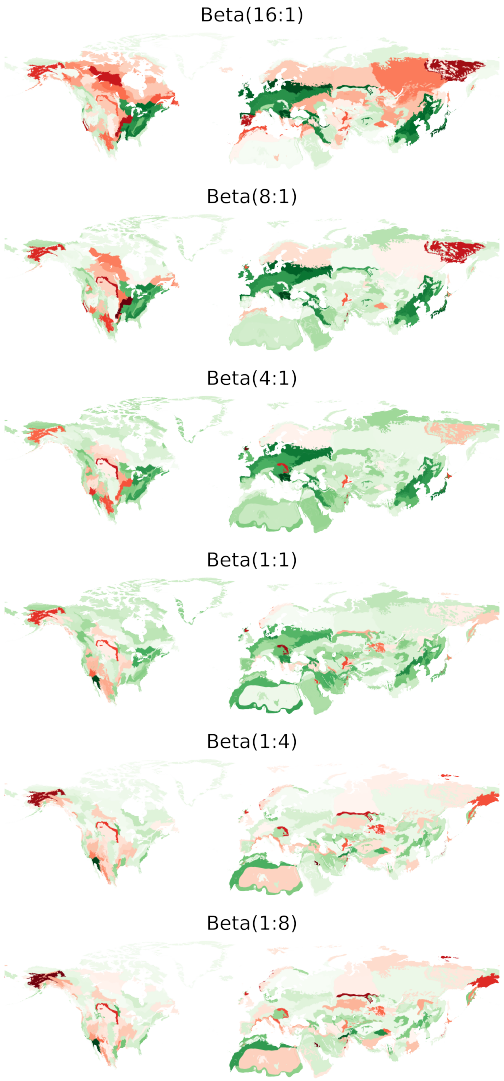


FIGURE 9 – Illustration des rôles de α , β sur les data-Shapley values

3.1 Relation entre qualité des données et Valeurs de Shapley

Notre jeu de données n'est pas parfait, en effet comme décrit précédemment nous n'avons gardé que certaines espèces, ce qui entraîne que certaines écorégions apparaissent identiques de par leur assemblage d'espèces animales ; ceci est problématique car elles sont cependant caractérisées par des biomes différents. Ces écorégions qualifiées d'ambiguës sont représentées figure 10.

Les beta-Shapley values des ces écorégions sont faibles comme présenté Fig. 11. En effet celles-ci, de par leur défauts ne permettent pas un bon entraînement du modèle, voire lui nuisent et donc reçoivent une valeur faible en conséquence.

Une façon alternative de constater le fonctionnement de cet

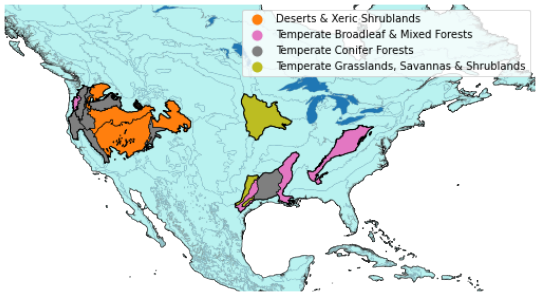


FIGURE 10 – Écorégions ambiguës

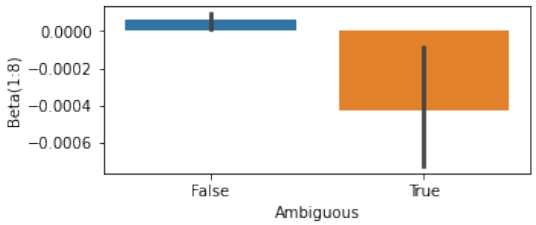


FIGURE 11 – Distribution of $Beta(1:8)$ en fonction de l'ambiguïté

outil est une analyse de la correspondance performance et valeur de Beta-Shapley comme présentée Fig. 12.

Plusieurs heuristiques de valorisation de points de données sont évaluées. Chaque heuristique propose un classement des écorégions à partir duquel les points sont enlevés un à un par ordre d'importance. À chaque étape le modèle est réentraîné et sa performance est évaluée. L'on constate donc que beta-Shapley est significativement plus efficace qu'une méthode de sélection aléatoire. En effet, les heuristiques basées sur beta-Shapley permettent même un gain de performance en enlevant des point néfastes.

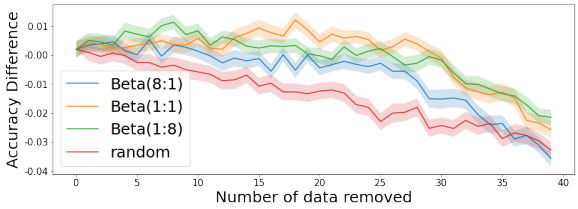


FIGURE 12 – Suppression successive de points du jeu d'entraînement et impact sur la performance

3.2 Inspection de tendances groupées

Les beta-Shapley paramétrées telles que $\alpha \gg \beta$ par exemple $Beta(8:1)$ mettent de larges poids sur les petites cardinalités et dé-bruite les grandes. Par conséquent les attribution sont plus homogènes et il est possible de distinguer l'impact de groupes de points sur la performance du modèle.

Par exemple, si on groupe ces beta-Shapley values par biome, il apparait une immédiate distinction entre le biome *Temperate Broadleaf and mixed Forests* et les autres (Fig.

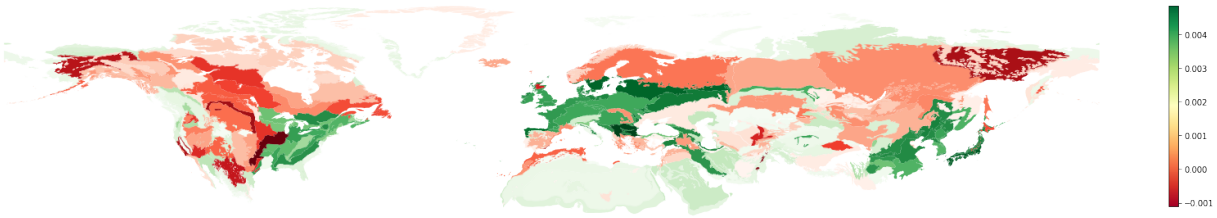


FIGURE 13 – Distribution des Beta(8 :1) par ecoregion

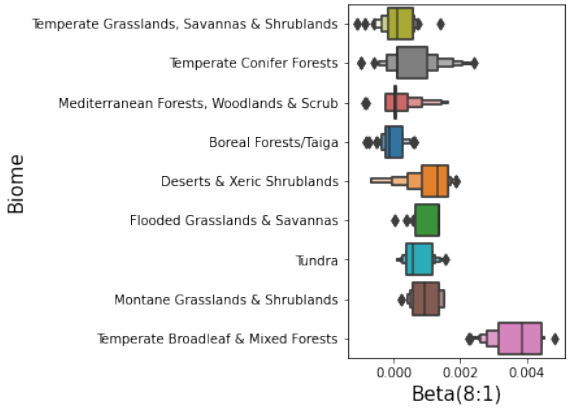


FIGURE 14 – Distribution des Beta(8 :1) par Biome

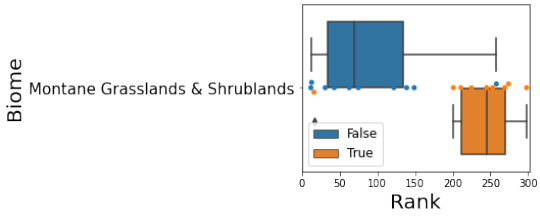


FIGURE 15 – Distribution du rang du Buteo buteo en fonction de sa présence dans le *Biome Montane Grasslands & Shrublands*

14). Les écoregions appartenant à ce Biome contribuent plus à augmenter la performance du modèle. Ce biome est à la fois le plus représenté dans le domaine de nos données (Fig 3) mais aussi, celui qui présente la plus grande variété d'espèces par écorégion. Il est possible que l'exactitude multi-classes, fonction d'utilité utilisée, ait augmenté ces valeurs en faveur du biome susmentionné. En effet celle-ci est sensible au déséquilibre des différentes classes, elle mesure mal les performances sur celles les moins représentées.

3.3 Inspection d'espèces influentes

Beta-Shapley nous donne accès à un classement d'écorégions reflétant leur utilité marginale en tant que point d'un ensemble d'entraînement. On considère ainsi le rang d'une écorégion le classement de sa valeur de *Beta(1 :8)* ordonnées dans l'ordre croissant. Ce rang permet de mettre en valeur des éléments utiles à l'entraînement du modèle. Par exemple en regardant pour les écoregions du Biome *Montane Grasslands & Shrublands* et en conditionnant sur la présence de "Buteo buteo" (Buse variable) une distinction claire apparaît (Fig 15). Pour une écorégion donnée, la présence de cet animal permet de mieux déterminer si le biome est *Montane Grasslands & Shrublands* ou non. Les écorégions contenant cet animal ont été plus informatives pour la détermination de ce biome et pour notre modèle que celles qui n'en avaient pas.

Cette information prend une nouvelle dimension une fois replacée dans son contexte. En effet une visualisation sur une carte, ou dans un système d'information géographique, donne à un expert plus d'éléments pour arriver à des conclusions. Par exemple pour le cas présent, le Buteo buteo n'est pas présent dans les plateaux Tibétains au sud du désert du Taklamakan tandis que le biome s'étend sur une zone continue. Dans le cadre de notre étude, l'écorégion du Karakoram ouest dans laquelle vit la buse nous permet d'identifier plus facilement le biome *Montane Grasslands & Shrublands* comparé aux écorégions du même biome dans lesquelles cet animal ne vit pas.

L'on peut avoir des informations de manière similaire avec le Vulpes lagopus (renard polaire) et le biome *Temperate broadleaf & mixed forests*. Les écoregions semblant pertinentes étant : *Sarmatic mixed forests* et *Baltic mixed forests*.

4 Intégration à un système d'information géographique

Afin de rentre les outils d'explicabilité plus facilement utilisables, nous avons intégré les valeurs de beta-Shapley et les autres attributs disponibles dans l'outil Kepler.gl [11]. Kepler.gl est un outil de visualisation de données géospatiales open source permettant une vue interactive et en trois dimensions de nombreux calques. Ces deux éléments permettent de remettre les données dans leur contexte ainsi qu'une exploration personnalisée des données.

Les valeurs Beta(8 :1) shapley sont présentées figure 18. Chaque zone segmentée de couleur uniforme corresponds à une écorégion, la couleur ainsi que la hauteur correspondent à la valeur de beta Shapley. Le minimum correspondant au bleu le plus sombre et à une hauteur nulle, le maximum au rouge le plus intense. Il est possible de constater que l'écorégion du Sahara du Nord apparaît en sur-brillance, c'est le résultat d'un clic de l'utilisateur qui peut de cette ma-

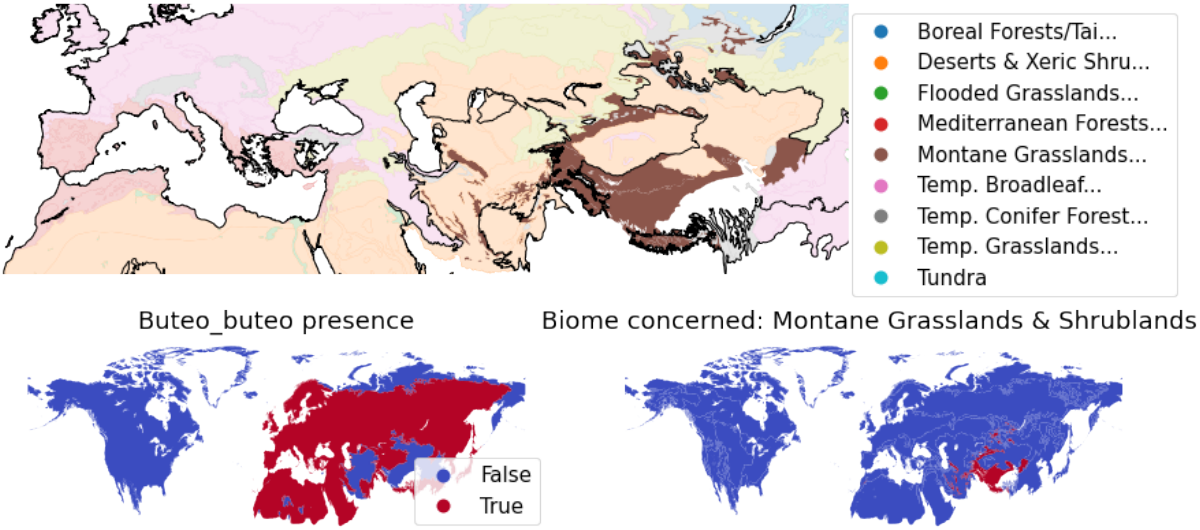


FIGURE 16 – Cartes de répartition du Buteo buteo

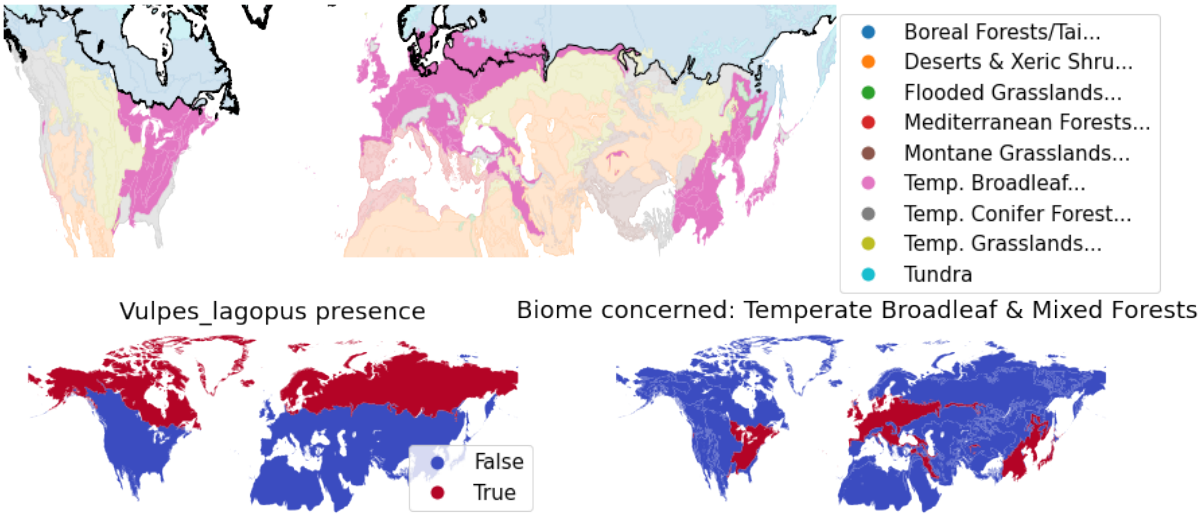


FIGURE 17 – Cartes de répartition du Vulpes lagopus

nière accéder au détail de ce point de données. Au delà des multiples personnalisations possibles (coloration, variable affichées, ...) l'utilisateur est libre de naviguer sur la carte et de choisir le point de vue qui lui convient le mieux dans cet espace en trois dimensions.

Les figures 19, 20 présentent les mêmes informations décrites section 3.3. Ainsi en couleur orangée apparaît le biome nous intéressant : *Montane Grasslands & Shrublands* pour la détermination duquel la présence de la buse est informative. Grâce à ces visualisations il est aisé de déterminer que l'écorégion du *North Tibetan Plateau-Kunlun Mountains alpine desert* est la seule changeant significativement de hauteur parmi les régions *Montane Grasslands & Shrublands*. Le Buteo buteo apparait dans cette ecorégion cependant elle possède une valeur de beta-Shapley faible ce qui contraste avec ses voisines. La forte informativité de cette écorégion peut être due à la présence ou l'absence d'un autre animal. Grâce à cette visualisation un expert du domaine disposerait à la fois d'un point de départ pour avoir un regard critique sur les données et de la nouvelle dimension que constitue les valeurs de beta-Shapley, celui de l'informativité d'un point de donnée vis-à-vis d'un modèle d'apprentissage machine.

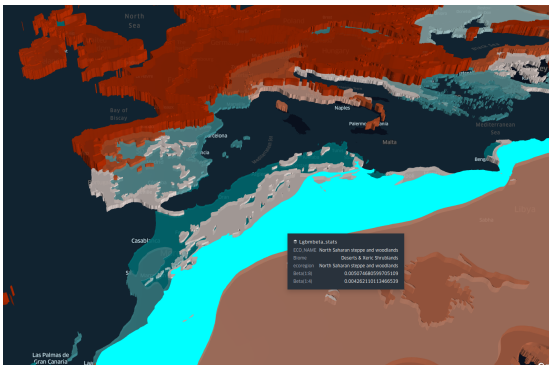


FIGURE 18 – Exemple de l'interface de Kepler.gl

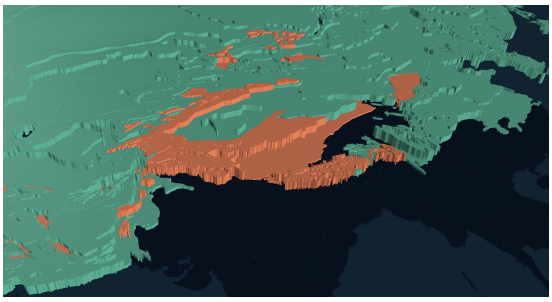


FIGURE 19 – Les information sur Buteo buteo vus sur Kepler - la hauteur représente présence de l'animal

5 Discussion et conclusion

Le but de cet article est de présenter l'utilisation et l'utilité d'un outil d'explicabilité centré sur les données dans le contexte concret qu'est la reconstruction des paléoenvironnements. Nous avons montré comment l'exploitation des informations géographiques contenues dans les données permettent de mieux visualiser les explication des modèles d'apprentissage machine. Les éléments mis en valeurs ne sont pas des conclusions, il est nécessaire qu'un paléontologue, expert du domaine, s'approprie cet outil et y apporte son savoir et ses connaissances. La fiabilité et l'utilité de cet outil qui a été décrite au fil des expérimentations reste à nuancer. En effet, l'information délivrée est complexe, l'utilité d'un point de donnée pour un modèle d'apprentissage machine par rapport à une fonction de score est pour le moins nuancée. En dépit des bonne propriétés (section 3) beaucoup de dimensions différentes impactent les valeurs de beta-Shapley. Pour tirer des conclusions valables il est à minima nécessaire de connaître à la fois les caractéristiques du type de modèle apprenant utilisé ainsi que ses limitations mais aussi de connaître le dataset sur lequel on l'entraîne.

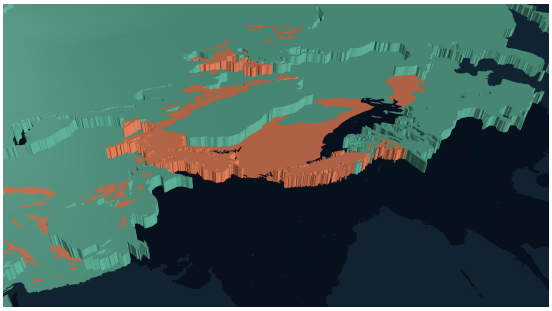


FIGURE 20 – Les information sur Buteo buteo vu sur Kepler - la hauteur représente le rang

L'une des limitations de beta-Shapley est sa complexité. En effet le calcul des valeur exactes est de complexité exponentielle en nombre de points. Bien que certaines méthodes permettent des approximations moins coûteuses comme les méthodes de Monte-Carlo [9], son utilisation sur des jeux de données volumineux devient quasiment impossible. Il reste cependant possible de l'appliquer à un échantillon de points représentatifs, l'utilité de cet outil serait néanmoins diminuée. Par exemple pour la détection d'anomalies, qui elle serait appropriée dans un contexte de données massives, il ne serait pas possible d'utiliser ce procédé. Il serait néanmoins possible de l'utiliser à des fins de vérification.

Références

[1] Henry de Lumley. *La Caune de l'Arago Tome I, Tautavel-en-Roussillon, Pyrénées-Orientales, France.* Éd. du CNRS, Paris, 2014.

[2] World Wildlife Fund. Wildfinder : online database of species distributions, 2006.

[3] Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4) :457 – 472, 1992.

[4] Amirata Ghorbani and James Zou. Data Shapley : Equitable Valuation of Data for Machine Learning. *arXiv :1904.02868 [cs, stat]*, June 2019. arXiv : 1904.02868.

[5] Amirata Ghorbani, James Zou, and Andre Esteva. Data Shapley Valuation for Efficient Batch Active Learning. *arXiv :2104.08312 [cs, stat]*, April 2021. arXiv : 2104.08312.

[6] Sophie Grégoire, Nicolas Boulbes, Bernard Quinio, Matthieu Boussard, Caroline Chopinaud, Anne-Marie Moigne, Agnès Testu, Vincenzo Celiberti, Cédric Fontaneil, Christian Perrenoud, Anne-Sophie Lartigot Campin, Thibaud Saos, Tony Chevalier, Véronique Pois, Henry de Lumley, Marie-Antoinette de Lumley, Antoine Harfouche, Rolande Marciniack, Philippe Carrez, and Thierry Hervé. Innovative multidisciplinary method using Machine Learning to define human behaviors and environments during the Caune de l'Arago (Tautavel, France) Middle Pleistocene occupations. In Archaeopress, editor, *Big Data and Archaeology : Proceedings of the XVIII UISPP World Congress (4-9 June 2018, Paris, France), Sessions III-1 François Djindjian, (éd.) ; Paola Moscati, (éd.)*, Proceedings of the XVIII UISPP World Congress (4-9 June 2018, Paris, France), pages 28–47. 2021.

[7] Yongchan Kwon and James Zou. Beta shapley : a unified and noise-reduced data valuation framework for machine learning. *CoRR*, abs/2110.14049, 2021.

[8] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[9] Tomasz Pawel Michalak, Aadithya V. Karthik, Piotr L. Szczepanski, Balaraman Ravindran, and Nicholas R. Jennings. Efficient computation of the shapley value for game-theoretic network centrality. *CoRR*, abs/1402.0567, 2014.

[10] Magdalena Sobol and S. Finkelstein. Predictive pollen-based biome modeling using machine learning. *PLOS ONE*, 13 :e0202214, 08 2018.

[11] Open source. Kepler.gl, a powerful open source geospatial analysis tool for large-scale data sets. Online. Accessed : 2022-04-12.

[12] Christian Willmes, Kamil Niedziółka, Benjamin Serbe, Sonja Grimm, Daniel Groß, Andrea Miebach, Michael Maerker, Felix Henselowsky, Alexander Gamisch, Masoud Rostami, Ana Mateos, Jesús Rodríguez, Heiko Limberg, Isabell Schmidt, Martin Müller, Ericson Hölzchen, Michael Holthausen, Konstantin Klein, Christian Wegener, and Georg Bareth. State of the art in paleoenvironment mapping for modeling applications in archeology-summary, conclusions, and future directions from the paleomaps workshop. *Quaternary*, 3 :13, 05 2020.

Représentation explicable du comportement de systèmes complexes : automates pour les séries temporelles multivariées

I. Chraibi Kaadoud[1*] L. Fahed[1] T. Tian[1] Y. Haralambous[1] P. Lenca[1]
[1]IMT Atlantique, Lab-STICC, UMR CNRS 6285, Brest, France

{ikram.chraibi-kaadoud, lina.fahed, tian.tian, yannis.haralambous, philippe.lenca}@imt-atlantique.fr

Résumé

La compréhension des états de systèmes complexes représentés par des séries temporelles peut être une tâche ardue, notamment à cause des changements permanents des événements contextuels internes et externes aux systèmes. Pour faire face à ce défi, nous proposons la méthode XR-CSB (eXplainable Representation of Complex System Behavior) dont l'objectif est de représenter l'évolution de tels systèmes d'une manière intelligible et explicable reposant sur : (i) un clustering vertical afin de détecter les états du système, (ii) une représentation visuelle explicable utilisant des automates dépliés, (iii) une pré-modélisation explicable fondée sur des métriques explicatives. Quatre représentations, évaluées par des experts du domaine applicatif de notre travail, montrent que XR-CSB passe à l'échelle et répond à leurs attentes en termes d'explicabilité et d'intelligibilité.

Mots-clés

Séries temporelles multivariées, représentation, explicabilité, automate, système complexe

Abstract

Understanding the states of complex systems represented by time series can be a difficult task, especially because of the constant changes in contextual events internal and external to the systems. To face this challenge, we propose the method XR-CSB, eXplainable Representation of Complex System Behavior, whose objective is to represent the evolution of such systems in an intelligible and explainable way based on : (i) a vertical clustering in order to detect the states of the system, (ii) an explainable visual representation using unfolded automata (iii) an explainable pre-modeling based on explainable metrics. Four representations, evaluated by experts in the application domain of our work, show that XR-CSB scales up and meets their expectations in terms of explainability and intelligibility.

Keywords

Multivariate time series, representation, explainability, finite-state automata, complex system

*Contact author

1 Introduction

Dans de nombreux cas d'applications, telles que l'énergie, le trafic urbain, etc., des systèmes de plus en plus complexes sont développés et utilisés [4]. Un système complexe peut être décrit par un ensemble important d'entités/variables interagissant dans le temps [11, 8]. Il est ainsi généralement représenté par des séries temporelles multivariées, où chaque série représente une variable du système dont les intervalles de valeurs correspondent à des états pour l'intervalle de temps associé [12].

Extraire des connaissances de ces systèmes et notamment leur évolution, est un défi lié aux changements permanents des événements contextuels internes et externes aux systèmes. Il est alors difficile de déterminer quelles sont les caractéristiques pertinentes qui contribuent à un changement d'état, et au-delà de cela de comprendre le comportement du système. Dans ce contexte, des techniques de visualisation de données ont été proposées [11, 21, 19], telles que les techniques de coordonnées parallèles, les matrices de nuage de points, etc. Des approches fondées sur les automates à états finis (FSA pour *Finite State Automata*), considérées comme représentations à faible dimension [1] impliquant une discrétisation des séries temporelles, ont également été proposées [26, 20]. Ces techniques réduisent la complexité algorithmique et offrent un haut niveau d'explicabilité – nous nous en inspirons pour proposer une représentation explicable d'un système de production d'énergie. Une telle représentation capture les états de tout ou partie du système en s'inscrivant dans le domaine de l'Intelligence Artificielle (IA) explicable (XAI) et intelligible [23].

Nous traitons le cas d'une centrale thermique qui brûle du charbon et du gaz afin de produire de la vapeur pour générer de l'électricité. Cette centrale dispose d'équipements (et d'ensemble d'équipements dont des chaudières, équipé(e)s de plusieurs centaines de capteurs) ayant des vocations différentes : transformer, consommer ou produire de l'énergie. La centrale est ainsi assimilée à un système complexe composé de séries temporelles multivariées non étiquetées, au comportement «chaotique» (i.e. irrégulier) et *a priori* non déterministe. Les experts du domaine industriel [2] contri-

1. I.e. résultant d'un processus de réduction des données dimensionnelles contenant autant d'informations que possible que les données d'origine.

2. Sociétés informatiques, éditeurs de logiciels, qui gèrent des données

buent en donnant leur avis sur le système. L'objectif est d'améliorer leur compréhension du fonctionnement de la centrale afin d'optimiser la consommation d'énergie.
Nos principales hypothèses sont les suivantes : (i) Les valeurs des séries temporelles à un instant donné représentent l'état du système en cet instant et permettent de détecter des états, en particulier les états «rares» ; (ii) les états du système peuvent être caractérisés par différentes métriques explicatives liées à l'évolution des séries temporelles : vitesse, vélocité et accélération des valeurs au sein d'un état ; (iii) les états du système, ainsi que leurs caractéristiques (métriques explicatives), peuvent être considérés comme faisant partie d'un FSA pour lequel il existe des représentations visuelles efficaces ; (iv) le FSA est une représentation synthétique, intelligible et compréhensible du comportement du système dans le temps et donc une aide à la décision.
Nous proposons une méthodologie de représentation explicite du comportement des systèmes complexes (XR-CSB pour *eXplainable Representation of Complex System Behavior*) utilisant les FSA pour les séries temporelles multivariées, fondée sur : **(i) L'extraction de connaissances** au travers de l'utilisation de *clustering vertical* des séries temporelles afin de détecter les états du système. Cette approche originale est indépendante de la taille des séries temporelles et applicable à des séries uni/multivariées, ce qui permet un contrôle de la complexité ; **(ii) la représentation de la connaissance explicable** par le biais de l'utilisation de FSA pour représenter le comportement du système, et **(iii) la prémodélisation de l'explicabilité** via l'utilisation de métriques explicatives pour enrichir les FSA, approche d'ingénierie des fonctionnalités explicables spécifiques au domaine (*pre-modeling explainability domain-specific explainable feature-engineering approaches*) [18].
La section 2 présente des travaux connexes sur l'XAI et la représentation des systèmes complexes. La méthode proposée est décrite dans la section 3, les résultats expérimentaux en section 4 – nous concluons en section 5.

2 État de l'art

Disposer de données étiquetées est très coûteux et parfois-même impossible, ce qui a été le cas pour notre application – des méthodes non-supervisées doivent alors être appliquées [3]. Nous présentons ici des travaux liés à la compréhension et à la représentation du fonctionnement d'un système complexe dans ce contexte.

• **Clustering de séries temporelles multivariées :** les méthodes de clustering de séries temporelles multivariées [6, 1, 26] nécessitent que toutes les séries aient la même taille afin de garantir le bon fonctionnement du clustering et la fiabilité des résultats. Elles permettent de détecter des comportements communs entre les séries temporelles afin par exemple, d'étiqueter automatiquement celles-ci par la suite ou de détecter des motifs fréquents. Cependant, elles ne permettent pas de représenter le comportement caractérisé par les valeurs de différentes séries temporelles à un moment donné et ne sont donc pas adaptées à notre objectif.

• **IA explicable (XAI) :** des modèles, très efficaces, de systèmes complexes existent [9]. Ils sont cependant souvent qualifiés de «boîtes noires» et l'XAI vise a les rendre plus intelligibles, plus transparents et plus accessibles ou encore à concevoir directement des modèles explicables [9, 10]. L'XAI a pour objectif de fournir une *explication* des mécanismes internes et/ou du comportement d'un système, sous la forme d'un ensemble d'énoncés construits pour décrire un ensemble de faits qui clarifie les causes, le contexte et les conséquences de ces faits [7]. Une explication est auto-suffisante et adressée au public cible selon sa connaissance *a priori* du domaine/système, de ses attentes et du contexte [22].
Une explication est donc une interface entre le système et le public cible (dans notre cas, les experts du domaine) [10, 14]. Les explications et l'intelligibilité peuvent, en effet, être beaucoup plus importantes que la *performance pure* dans les systèmes d'IA [10, 9], par exemple pour faciliter le contrôle et l'acceptation par l'utilisateur, ainsi que parfois pour des questions légales. Cela nécessite des compétences interdisciplinaires telles que l'interaction centrée sur l'homme ou une expertise en IA [17, 16, 2]. Les représentations visuelles jouent dans ce contexte un rôle important [24].
La majorité des méthodes de visualisation récents traite de la visualisation du modèle de «boîte noire» à travers l'analyse de sensibilité [5], et ignore parfois les aspects liés à la multidimensionnalité des données. Il est important de noter que nous cherchons à représenter les résultats d'un modèle transparent (clustering) sur un système.
Soulignons les travaux de [13, 26] dans lesquels les séries temporelles sont segmentées au moyen de la détection des points de changement [3]. L'inconvénient de cette approche est qu'elle ne prend pas en compte les informations dynamiques et temporelles ainsi que l'interaction et la dépendance entre les séries temporelles.

• **Représentation de séries temporelles multivariées à l'aide d'automates :** l'utilisation de FSA – pour représenter, surveiller, estimer ou même prédire les états des systèmes – est d'une grande importance pour la prise de décisions émergeant de l'interaction experts-système.
Dans les travaux associés [6, 13, 26], des variables explicatives sont extraites de chaque série temporelle. Elles constituent les états du système et la succession de ces variables représente une séquence d'états. Un clustering est ensuite effectué pour détecter des modèles similaires (successions d'états) entre les séries temporelles. Ceci permet d'extraire un comportement commun partagé entre les séries temporelles, mais n'est applicable ni sur un très petit nombre de séries, ni à l'extraction du comportement global du système car l'interaction et la dépendance (ou non) entre les séries temporelles sont ignorées. Une telle approche n'est pas adaptée aux systèmes complexes car elle ne permet pas de

représentant le type d'installation que nous étudions.

3. Un clustering destiné à traiter des données numériques "simple" (k-means par exemple pour les vecteurs numériques) est appliqué sur les variables explicatives et les fenêtres de données d'entrée.

simplification intelligible du comportement global du système.

3 Modèle XR-CSB : représentations explicables pour le comportement de systèmes complexes

Notre travail se concentre sur (i) l'étude des changements de comportement d'un système complexe représenté par des séries temporelles multivariées à différents pas de temps, (ii) la détection et la compréhension de l'évolution de l'état du système vers d'autres états, et (iii) la représentation de son évolution d'une manière intelligible et explicable.

Pour cela, nous proposons XR-CSB, un modèle en 3 étapes, inspiré de travaux de génération de FSA [15]. La figure 1 présente un schéma général du modèle.

3.1 Etape 1 : clustering vertical de séries temporelles

Afin d'effectuer ce que nous appelons un clustering vertical, nous utilisons l'algorithme k-means [25]. Notre point d'entrée est la série temporelle X représentée par un ensemble de vecteurs $X = [X_1, \ldots, X_m]$ et une fenêtre temporelle fixe W de longueur w (mesurée en minutes). Nous appliquons k-means, avec la distance euclidienne, de manière à partitionner X en k clusters et ce indépendamment des pas de temps associés [4]. Les clusters obtenus, qui émergent des valeurs des différentes séries temporelles regroupées, représentent les états du système. Le choix de la valeur de k est fonction de la valeur de silhouette score calculée.

3.2 Etape 2 : explicabilité via des automates

Processus de génération d'automates (figure 4(a)) : Nous générons ensuite un automate où chaque nœud représente un état du système/cluster. La génération de l'automate suit les étapes suivantes décrite dans [15] : pour chaque vecteur X_t, un nœud est crée dans le FSA dont l'identifiant (id) est celui du cluster auquel le k-means l'a associé. Une arête directe est créée avec un poids de 1 entre le nœud en cours et celui créé. Si le nœud existe déjà dans l'automate, et l'arête également, le poids de celle-ci est incrémenté de 1. Sinon une nouvelle arête avec un poids de 1 est créée entre les deux nœuds existants. Dans le cas où deux vecteurs consécutifs X_t et X_{t+1} appartiennent au même cluster, une boucle récurrente est ajoutée au nœud associé. Ce processus aboutit à la génération d'un automate aux transitions pondérées représentant l'arrangement des clusters/états. La figure 4(a) représente l'automate résultant. Le poids des transitions est indiqué par une couleur, elle indique la durée pendant laquelle le système reste dans un état : plus elle est foncée, plus le poids de la transition est important.

Processus de génération de chemin : automate déplié (figures 4(b,d)) : Un *automate déplié* correspond à un chemin temporel le long des états du système. Nous proposons deux représentations visuelles d'un tel chemin et de l'évolution des états du système dans le temps : (1) la figure 4(b) représente la durée de chaque état via des informations numériques (durée des états en minutes, date UTC et UTC+2), et les mêmes nœuds/états possèdent le même identifiant ; (2) la figure 4(d) représente chaque état par un rectangle dont la taille est proportionnelle à la durée de l'état, et les mêmes états ont le même code-couleur.

3.3 Étape 3 : pré-modélisation du processus d'explicabilité, utilisation de métriques explicatives

Un processus d'explicabilité préalable à la modélisation consiste en l'extraction de métriques explicatives afin d'enrichir l'automate déplié. Afin de caractériser chaque état $S_{(t_{\text{start}}, t_{\text{end}})i}$, nous proposons trois métriques à calculer pour chaque ensemble de vecteurs associés à la fenêtre $W \in [t_{\text{start}}, t_{\text{end}}]$:

• **la vitesse moyenne** : notée Sp, elle correspond à la valeur moyenne des vitesses de changement des valeurs entre les temps t et $t-1$.

• **la vélocité moyenne** : notée Vl, elle correspond à la valeur moyenne des vélocités calculées pour chaque vitesse (la dynamique de l'évolution des valeurs)

• **l'accélération moyenne** : notée Ac, elle correspond à la valeur moyenne des accélérations et représente la rapidité de changement de la vitesse moyenne Sp des valeurs des états sur la fenêtre W.

À noter que si un état S_i se produit plusieurs fois dans une fenêtre W, alors les métriques explicatives seront calculées à chaque fois sur les vecteurs liés à l'état en question pendant cette fenêtre temporelle associée. La figure 4(c) représente un automate déplié enrichi de métriques explicatives.

4 Expérimentations

Nous présentons le jeu de données, puis une évaluation de la scalabilité de XR-CSB, ainsi qu'une évaluation qualitative visant à mesurer son pouvoir explicatif [5].

4.1 Description des données industrielles

Notre jeu de données provient des capteurs des équipements d'une centrale thermique qui mesurent des quantités physiques différentes (tonne/h, °C, bar, etc). Les relevés de capteurs peuvent être renseignés ou non, sans forcément que cela soit une anomalie. Le jeu de données contient les enregistrements de 377 capteurs couvrant une période de trois ans, avec des valeurs collectées toutes les 10 minutes. Au niveau du pré-traitement, en fonction d'une fenêtre temporelle de largeur w (nombre de pas de temps)

4. La distance euclidienne ne tenant pas compte des différences d'échelles ou des possibles corrélations entre les séries temporelles permet de proposer une méthodologie générique transposable sur différents contexte d'applications et cela indépendamment des caractéristiques des séries temporelles.

5. Les expériences sont réalisées sur un MacBook Pro, Apple M1, 16GB de RAM. Le language de programmation Python est utilisée, ainsi que les librairies Numpy (https://numpy.org/), Scipy (https://scipy.org/), Matplotlib (https://matplotlib.org/), Networkx (https://networkx.org/documentation/stable/index.html), Scikit-learn (https://scikit-learn.org/dev/index.html) et Pandas (https://pandas.pydata.org/docs/index.html).

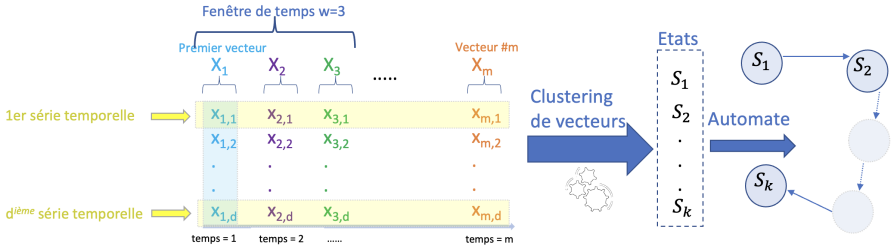


FIGURE 1 – Schéma général du modèle XR-CSB : séries temporelles multivariées, clustering et automates.

et d'un nombre de séries temporelles à analyser, le jeu de données est ainsi transformé en une matrice de dimension $(m \times d)$ où $m \in \{1, \ldots, w\}$ et d est le nombre de séries temporelles. Le traitement peut être incrémental puisque la matrice est dynamique : pour chaque nouveau pas de temps, une colonne est ajoutée avec les valeurs des variables correspondantes. Toutes les valeurs sont normalisées (par la moyenne et l'écart-type). Lorsque $d > 1$, une moyenne est calculée à partir des valeurs normalisées.

4.2 Evaluation de la scalabilité

Nous avons évalué la scalabilité de notre approche en calculant le temps d'exécution pour 92 capteurs lorsque : (1) le nombre de clusters k varie sur une fenêtre de temps fixe de $w = 144$ pas de temps (figure 2), (2) lorsque la fenêtre de temps w varie pour $k = 7$ clusters (figure 3). Pour ces deux analyses, le temps d'exécution augmente presque linéairement et l'algorithme présente une bonne scalabilité pour les valeurs de w et k. Au niveau technique, notre évaluation de la scalabilité montre que ce modèle est non-coûteux en terme de temps d'execution ou de puissance computationnelle. Notons que pour l'analyse de plus de deux ans d'enregistrements, il faut compter 20 minutes de temps de traitement, temps relativement raisonnable compte tenu du contexte industiel.

4.3 Évaluation de l'explicabilité

Les résultats de XR-CSB sont destinés aux experts du domaine de nos partenaires industriels. Ainsi, une évaluation *qualitative* a été faite via un questionnaire sur la qualité de la représentation et l'explicabilité du comportement d'un ensemble d'équipements. Composé de questions à choix multiples, le questionnaire porte sur trois études de cas et sur le profil de l'expert du domaine [6].

4.4 Résultats

Nous proposons 3 études de cas : **(i) Cas d'étude A.1** : analyse d'un capteur C_1 dont l'unité physique est «tonnes par heure» (t/h) lorsque celui-ci présente un comportement monotone, **(ii) Cas d'étude A.2** lorsqu'il présente un comportement dynamique, **(iii) Cas d'étude B** : analyse des 92 capteurs d'un équipement B. Pour chaque cas d'utilisation, quatre représentations sont proposées (figure 4).

6. Les questions sur le genre et l'âge ont été exclues du questionnaire car jugées non pertinentes pour cette étude.

4.4.1 Profil des répondants

Profil professionnel : Parmi les six répondants, quatre sont des *data scientists*, un est ingénieur en apprentissage machine et un est directeur d'exploitation. Tous travaillent sur des systèmes complexes : trois depuis moins d'un an, un entre 1 et 2 ans et deux depuis plus de 5 ans. Enfin, tous utilisent des représentations visuelles dans leur travail quotidien pour expliquer ou transmettre des informations.
Confiance et explications du comportement de systèmes IA : quatre préfèrent les explications multi-modales (combinant plusieurs formes), un les représentations visuelles et un a répondu que seuls les résultats comptent (notamment en apprentissage machine supervisé). Notons toutefois qu'aucun ne préfère des explications unimodales (textuelles ou tabulaires) et que leur confiance dans les résultats des systèmes d'IA est fonction des «enjeux» traités.
Expérience professionnelle liée aux systèmes complexes : deux rencontrent «occasionnellement» le problème de la représentation du comportement des systèmes complexes dans le temps (50% des projets), trois en rencontrent «parfois» (60% des projets) et un rencontre ce problème «régulièrement» (70% des projets).

4.4.2 Évaluation du pouvoir explicatif des représentations

Nous évaluons dans cette section l'acceptabilité de la représentation pour chacune des 3 études de cas. La figure 5 présente les résultats de l'évaluation par les experts industriels de l'explicabilité de chaque représentation de notre modèle pour chaque cas d'utilisation. Globalement, les représentations (b), (c) et (d) sont considérées comme «peu représentatives» ou «totalement représentatives». La représentation (a) n'a pas convaincu les experts. Sur les 3 études de cas, la représentation (d) s'est avérée la plus intéressante pour les experts. Enfin, les commentaires ajoutés par les experts montrent une préconisation à fusionner les représentations (b) et (d) décrites comme *intuitives et intéressantes car les états sont identifiables ainsi que les distributions temporelles*.
En résumé, les représentations dépliées ont reçu une bonne évaluation sur leur capacité à représenter le comportement d'un système dans le temps. La représentation (d) était particulièrement convaincante grâce aux systèmes de couleurs et à la simplicité de la visualisation, tout comme la représentation (b) (malgré une évaluation moins positive). La représentation (c), jugée intéressante (retour d'un répon-

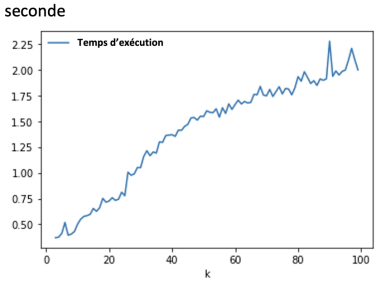


FIGURE 2 – Analyse de 92 capteurs sur 24 heures (144 enregistrements) : Temps d'exécution calculé selon un nombre de clusters $k \in \{3, \ldots, 100\}$.

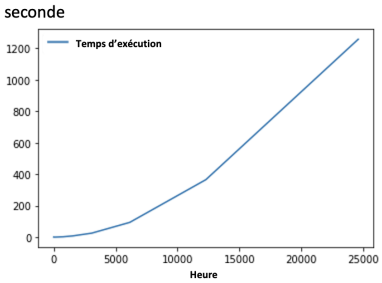


FIGURE 3 – Analyse de 92 capteurs pour $k = 7$: Temps d'exécution calculée selon la longueur de la fenêtre w avec $h \in \{8, 16, 24, 48, 96, 192, 384, 768, 1536, 3072, 6144, 12288, 24576\}$ heures.

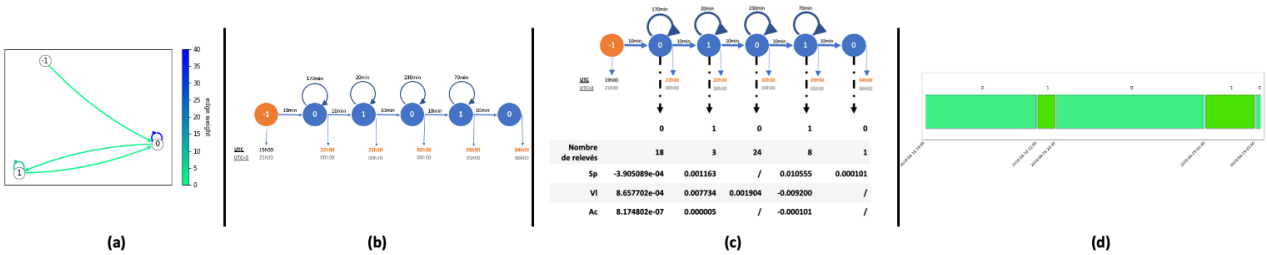


FIGURE 4 – Présentation des quatre représentations évaluées pour l'étude qualitative, résultant de l'analyse d'un seul capteur au comportement monotone (sans variation significative des valeurs) clusterisé avec $k = 2$. (a) un automate dont les couleurs des transitions indiquent leur poids : plus elle est foncée, plus le poids de la transition est élevé (le nœud -1 désignant le début de l'analyse, ne fait pas partie du groupe de données) ; (b) un automate déplié avec une représentation visuelle du temps pour chaque état par des valeurs numériques ; (c) un automate déplié avec des métriques explicatives ; (d) un automate déplié avec une représentation visuelle du temps pour chaque état.

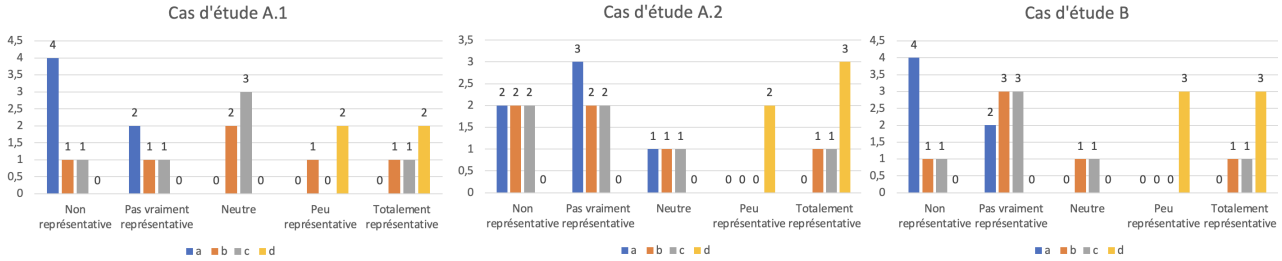


FIGURE 5 – Évaluation de l'explicabilité de chaque représentation pour chaque cas d'utilisation. Les valeurs correspondent au nombre de répondants.

dant) mais limitée en termes d'explicabilité du comportement, n'a suscité aucun rejet mais aucune adoption majeure non plus. Plusieurs facteurs peuvent expliquer ce résultat : (i) le choix des métriques et la création des représentations ont été faits par une approche orientée données (approche de *data science*), (ii) sans une forte implication des experts (notre public cible), (iii) sans un besoin explicite d'explicabilité de leur part, et (iv) sans contexte sur l'objectif du présent travail. Enfin, le retour des experts sur les seuils d'intelligibilité des automates (entre 6 et 10 nœuds maximum) et le nombre de capteurs au delà duquel l'analyse du système devient complexe (entre 10 et 15 selon un expert senior) confirme que XR-CSB est en adéquation avec les besoins métier en terme de représentation visuelle explicable du comportement d'un système complexe.

5 Conclusion et travaux futurs

XR-CSB est un travail original basé sur le clustering vertical et la génération d'automates dépliés pour décrire et comprendre le comportement d'un système complexe dans le temps. Notre étude montre que les représentations dépliées sont intéressantes pour les experts car elles donnent des informations sur le comportement du système, qu'il soit simple (une série temporelle) ou complexe (une centaine de séries temporelles), sur de courtes ou longues fenêtres de temps, que le comportement soit ponctuel (un pas de temps), ou non. Dans nos travaux futurs, nous souhaitons étudier les représentations que notre modèle permet d'ex-

traire afin de détecter les capteurs qui jouent un rôle discriminant dans le changement d'état du système complexe et intégrer cela dans notre explication du système. Cela pourrait aider les gestionnaires à identifier les équipements importants pour la gestion de l'énergie de l'installation, et plus globalement contribuer à la question de l'explicabilité des algorithmes de clustering appliqués aux séries temporelles.

Remerciements

Nous remercions le programme FEDER (Conseil régional de Bretagne et Union européenne) pour le financement du projet, les entreprises Energiency et Script&Go, et Raphaël Charbey (Energiency) pour son aide au sujet des visualisations.

Références

[1] Saeed Reza Aghabozorgi, Ali Seyed Shirkhorshidi, and Ying Wah Teh. Time-series clustering - A decade review. *Inf. Syst.*, 53 :16–38, 2015.

[2] Jean-Pierre Barthélemy, Gilles Coppin, and Philippe Lenca. Cognitive approach to decision making and practical tools. *IFAC*, 39(4) :123–128, 2006.

[3] Agnès Braud, Pierre Gançarski, Corinne Grac, Agnès Herrmann, Florence Le Ber, and Harrison Vernier. Classification de séries temporelles hétérogènes pour le suivi de l'état des cours d'eau. In *EGC*, volume E-37 of *RNTI*, pages 71–82, 2021.

[4] Leo Carlos-Sandberg and Christopher D Clack. Incorporation of causality structures to complex network analysis of time-varying behaviour of multivariate time series. *Scientific Reports*, 11(1) :18880, 2021.

[5] Paulo Cortez and Mark J. Embrechts. Using sensitivity analysis and visualization techniques to open black box data mining models. *Inf. Sci.*, 225 :1–17, 2013.

[6] Michel C. Desmarais and François Lemieux. Clustering and visualizing study state sequences. In *EDM*, pages 224–227, 2013.

[7] Jess Drake. *Introduction to Logic*. Scientific e-Resources, 2018.

[8] Janine Guespin-Michel. La science des systèmes complexes. https://urlz.fr/hIxs, 2016. Consulté : 2022-5-4.

[9] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5) :93 :1–93 :42, 2019.

[10] David Gunning. Darpa's explainable artificial intelligence (XAI) program. In *IUI*. ACM, 2019.

[11] David Harel. Statecharts : A visual formalism for complex systems. *Sci. Comput. Program*, 8(3) :231–274, 1987.

[12] Forough Hassanibesheli, Niklas Boers, and Jürgen Kurths. Reconstructing complex system dynamics from time series : a method comparison. *New J. Phys*, 22(7) :073053, 2020.

[13] Bryan Higgs and Montasir Abbas. Segmentation and clustering of car-following behavior : Recognition of driving patterns. *IEEE Transactions on Intelligent Transportation Systems*, 16(1) :81–90, 2014.

[14] Ikram Chraibi Kaadoud, Lina Fahed, and Philippe Lenca. Explainable AI : a narrative review at the crossroad of knowledge discovery, knowledge representation and representation learning. In *MRC@IJCAI*, vol 2995 of CEUR Workshop Proceedings, pages 28–40, 2021.

[15] Ikram Chraibi Kaadoud, Nicolas P. Rougier, and Frédéric Alexandre. Knowledge extraction from the learning of sequences in a long short term memory (LSTM) architecture. *Knowl. Based Syst.*, 235 :107657, 2022.

[16] Elisabeth Le Saux, Philippe Lenca, and Philippe Picouet. Dynamic adaptation of rules bases under cognitive constraints. *Eur. J. Oper. Res*, 136(2) :299–309, 2002.

[17] Philippe Lenca. Human centered processes. *Eur. J. Oper. Res*, 136(2) :231–232, 2002.

[18] Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. Explainable k-means and k-medians clustering. In *Proceedings of ICML*, pages 7055–7065, 2020.

[19] Vung Pham, Ngan Nguyen, Jie Li, Jon Hass, Yong Chen, and Tommy Dang. MTSAD : Multivariate time series abnormality detection and visualization. In *IEEE Big Data*, pages 3267–3276, 2019.

[20] Miriam García Soto, Thomas A Henzinger, and Christian Schilling. Synthesis of hybrid automata with affine dynamics from time-series data. In *HSCC*, pages 1–11, 2021.

[21] Andreas Theissler. *Detecting anomalies in multivariate time series from automotive systems*. PhD thesis, Brunel Univ. School of Engineering and Design, 2013.

[22] Bas C. van Fraassen. The pragmatic theory of explanation. *Theories of Explanation*, 8 :136–155, 1988.

[23] Daniel S Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Commun. ACM*, 62(6) :70–79, 2019.

[24] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. How do visual explanations foster end users' appropriate trust in machine learning ? In *IUI*, pages 189–201, 2020.

[25] Zheng Zeng, Rodney M. Goodman, and Padhraic Smyth. Learning finite-state machines with self-clustering recurrent networks. *Neural Comput.*, 5(6) :976–990, 1993.

[26] Yihuan Zhang, Qin Lin, Jun Wang, and Sicco Verwer. Car-following behavior model learning using timed automata. *IFAC-PapersOnLine*, 50(1) :2353–2358, 2017.

Session 5 : Planification / Raisonnement

L'IA au service de l'engagement des secours

Guy A. Narboni[1], Commandant Nicolas Mathieu[2]

[1] implexe - ToMCo
[2] SDIS de Seine-et-Marne - Agence du Numérique de la Sécurité Civile

r-d@implexe.eu

Résumé

Cette présentation du moteur de mobilisation développé pour le projet NexSIS décortique la mécanique d'engagement des secours qui fait suite à un appel d'urgence aux pompiers. En suivant le fil d'une modélisation en Prolog, elle illustre le rôle toujours actuel de l'« IA symbolique » dans une conception sur mesure de systèmes d'aide à la décision, au plus proche du métier.

Mots-clés

Automatisation du raisonnement, Modélisation par Contraintes, Programmation en logique, Programmation mathématique, Systèmes d'aide à la décision.

Abstract

This overview of the 'mobilisation engine' developed for the French NexSIS project details the mechanics of the dispatch process that follows an emergency call for Fire & Rescue Services. Using Prolog modeling as guiding thread, it illustrates the long-lasting contribution of 'Symbolic AI' in the design and fine-tuning of custom-built Decision Support Systems.

Keywords

Automation of reasoning, Constraint-based modeling, Logic programming, Mathematical programming, Decision-Support Systems.

1 Introduction

Porté par l'Agence du numérique de la sécurité civile, « NexSIS 18-112 » est un grand projet national [1] d'unification des systèmes d'information et de commandement des services d'incendie et de secours (IS). Son objectif est d'améliorer la gestion des interventions d'urgence à travers une nouvelle génération de Systèmes de Gestion « souverains » dont les principaux modules sont dédiés :

— à la cartographie (SIG)
— à la réception et à la qualification des Appels d'urgence (SGA)
— à la création et au suivi des Opérations (SGO)
— aux Echanges d'informations (SGE) avec les autres partenaires de l'urgence comme le SAMU ou les forces de l'ordre.

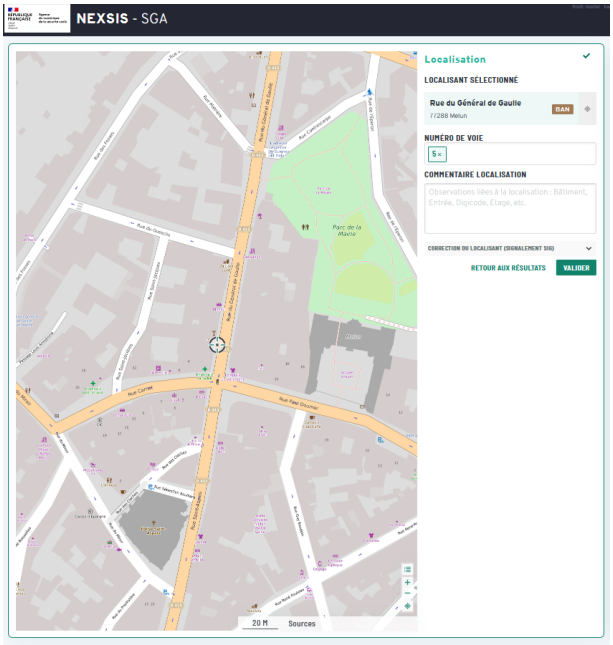


FIGURE 1 – Alerte notifiée par le SGA

FIGURE 2 – Prise en compte par le SGO

La création d'une opération est la réponse apportée par le Centre Opérationnel Départemental d'Incendie et de-Secours (CODIS) à une situation d'urgence, localisée sur son territoire et concernant la protection des personnes, des biens et de l'environnement.
Les services territoriaux d'IS disposent sur leur secteur de moyens d'intervention dont la répartition géographique découle d'une analyse de risques et d'une stratégie de « couverture opérationnelle » de ces derniers.
Traiter une alerte, c'est répondre vite et bien à l'urgence en envoyant rapidement sur les lieux de l'incident les secours appropriés. C'est en deux mots mobiliser les bonnes ressources, au bon endroit et au bon moment.

Le « moteur de mobilisation » est la pièce maîtresse de NexSIS dans ce processus. Il doit aider l'opérateur dans l'évaluation de la situation présentée et dans la prise de décision d'engagement des secours.
L'apport principal du moteur concerne la phase « réflexe » de la mobilisation où il s'agit, en temps réel, de planifier les mesures à prendre en réaction immédiate, de manière systématique, avec des moyens structurants (dimensionnés sur la base de scénarios « majorants »). Par exemple : un signalement de feu ordinaire déclenchera une mission de lutte contre l'incendie qui se traduira réglementairement par un départ de fourgon pompe-tonne avec 6 pompiers à bord.

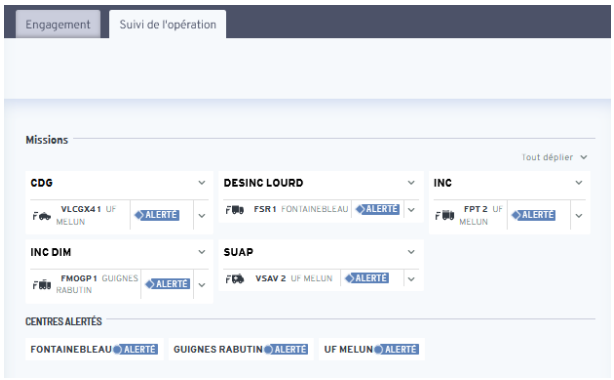


FIGURE 3 – Opération engagée

FIGURE 4 – Vue détaillée de l'équipage du fourgon INC

FIGURE 5 – Ordre de départ des moyens

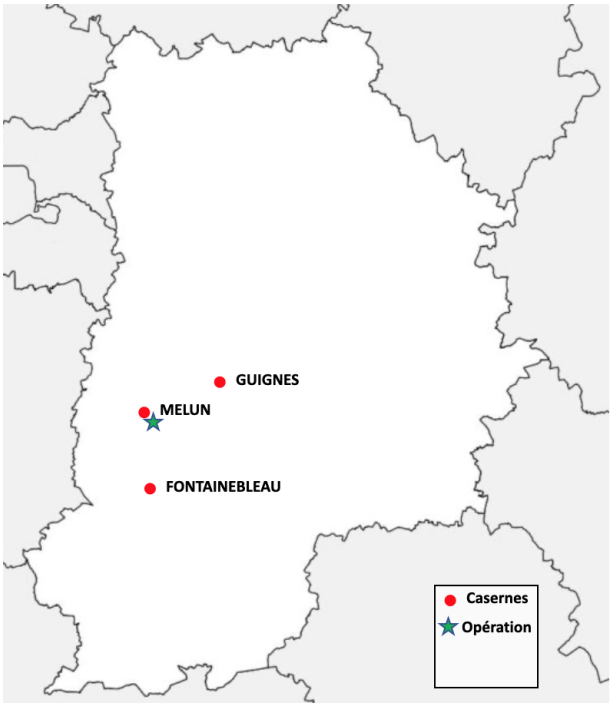


FIGURE 6 – Casernes mobilisées pour l'opération

La réponse du moteur se construit en deux temps [2][3][4] :

1. définition d'une réponse tactique — en termes de « missions » à accomplir (c'est-à-dire, en vue d'effets à obtenir)
2. définition sur cette base d'une réponse opérationnelle — par une affectation concrète de ressources (matérielles et humaines) aux missions.

Cette « traduction » de la situation d'urgence en moyens à engager est un challenge technique où, nous allons le voir, l'IA a toute sa part.

2 Le moteur de planification et la réponse tactique

La situation remontée exige-t-elle d'intervenir ou non ? La décision d'intervention est du ressort du service d'IS territorialement compétent. S'il y a intervention, il faut la planifier comme une opération.
La mécanisation de ce processus de décision fait appel à un moteur de règles.

2.1 Adaptabilité au territoire

Au niveau national, la plate-forme NexSIS fournit un moteur générique. Pour fonctionner, ce système doit être complété par une base de connaissances spécifique qui formalise les pratiques opérationnelles propres au département concerné.
Le découpage thématique de cette base en grandes familles (feux, accidents, secours d'urgence à personne, etc.) ramène l'examen d'une situation d'urgence à l'application d'un jeu de règles plus limité qui précise la doctrine d'engagement dans ce cas.

2.2 Règles métier

Les règles d'engagement sont définies dans leur grandes lignes par le Règlement Opérationnel local. Elles s'énoncent en pratique sous la forme : « si la situation d'urgence présente telles et telles caractéristiques, alors il faut telle mission pour y faire face ».
Les conclusions des règles font intervenir non seulement la qualité de la réponse (le type de mission à effecter) mais aussi sa quantité (le nombre de missions étant un indicateur du « volume » de secours à engager).
Les conditions des règles portent sur les attributs qui qualifient la situation comme dans l'exemple de la figure 7, mais pas seulement. Elles peuvent également faire référence aux conclusions d'autres règles. En particulier, une condition peut porter sur le nombre total de missions, comme dans la figure 8 où la règle se lit : « si le nombre de missions engagées est supérieur à 3, alors il faut prévoir *a minima* une mission de commandement ».

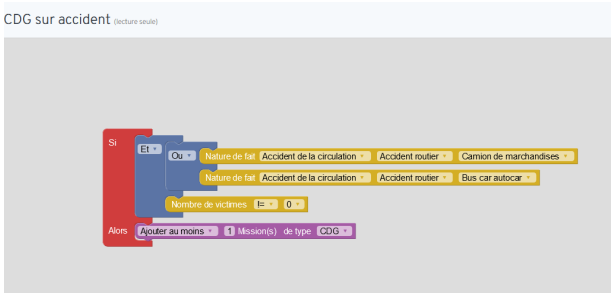


FIGURE 7 – Vue de la structure logique d'une règle

2.3 Modélisation

Les règles d'engagement doivent être pensées dans une logique monotone de montée en puissance de l'engagement : « plus j'ai ceci, plus je dois avoir cela ». En optimisation

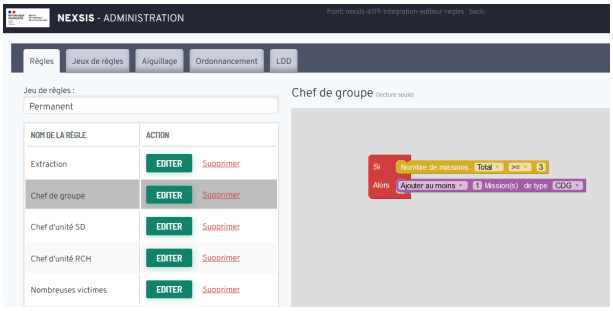


FIGURE 8 – Règle dépendant des résultats d'autres règles

multi-critères, on retrouve des notions similaires sous l'appellation de « règles de décision graduées » [5].
Ces règles s'éloignent du cadre normatif [6] retenu par la plupart des systèmes de gestion de règles métier. Nous avons donc conçu un « mini langage » déclaratif spécialement dédié à l'expression de ces connaissances [7]. Il est essentiellement de niveau propositionnel, ou plus exactement « pseudo-booléen » puisqu'il doit permettre de sommer des volumes.

Soit M la taille du catalogue de missions. Nous avons au plus M variables de décision $(x_1, \ldots, x_M)$ représentant des entiers positifs ou nuls qui correspondent au nombre des différents types de missions à engager.
Chaque règle est une implication logique : *conditions* $\rightarrow$ *conclusion* où la conclusion impose à l'une des variables une valeur entière minimale ($x_j \geq c_j$) et où chacune des conditions fait également intervenir un seuil. Ce seuil de déclenchement peut porter :

- soit sur une fonction booléenne y des paramètres d'entrée, à savoir les attributs booléens qui qualifient l'affaire (selon une nomenclature [8][9] désormais commune à tous les services d'urgence). Dans ce cas, il est égal à 1 :
  $y \geq 1$
- soit sur une variable de décision :
  $x_k \geq c_k$
- soit sur le nombre total des missions :
  $x_1 + \cdots + x_M \geq c_k$

Une fois la situation connue, les paramètres d'entrée s'effacent. Les règles qui ne s'appliquent pas sont tout simplement ignorées. Si bien que, techniquement, les règles restant à satisfaire sont des disjonctions d'inégalités linéaires que l'on peut écrire sous la forme :

$$(\vec{a}_{1_r}\vec{x} \geq c_{1_r}) \wedge \ldots \wedge (\vec{a}_{n_r}\vec{x} \geq c_{n_r}) \Rightarrow (x_r \geq c_r) \quad (R_r)$$

où $\vec{x}$ désigne le m-uplet des variables de décision et où les coefficients des produits cartésiens sont positifs ou nuls (puisqu'ils sont soit tous égaux à 1, soit tous égaux à 0 à l'exception d'un seul 1).
Cette syntaxe particulière généralise les clauses de Horn propositionnelles [10]. Il est à noter que si les inégalités

n'étaient pas toutes dans le même sens (dans le membre gauche de la règle), on pourrait poser avec ce langage n'importe quel problème SAT en variables binaires, et donc encourir le risque d'une explosion combinatoire dans la recherche d'une solution.

2.4 Résolution

Sur le plan technique, l'automatisation de la déduction repose sur la traduction des règles à satisfaire en un système de contraintes.

Le problème de la décision tactique se présente en fin de compte comme un problème d'optimisation sous contraintes : trouver le plus petit volume de secours satisfaisant à l'ensemble des règles énoncées. Autrement dit :

$$\begin{gathered}\text{Minimiser } f \\ \text{sous les contraintes } (R_r) \text{ du jeu de règles} \\ \text{et de la fonction de coût } f \geq x_1 + \cdots + x_M \\ \text{avec les } x_i \geq 0 \text{ et entiers.}\end{gathered}$$

Or, de la même façon que la résolution unitaire est complète pour les formules Horn SAT, on montre que la propagation est complète avec les jeux de règles exhibant cette structure [10]. Le résultat découle de l'observation que toutes les contraintes du système sont min-closes [11]. Le problème est donc décidable en temps polynomial.

Avec un solveur CLP(FD), la décision s'effectue sans avoir à recourir à l'énumération : si le produit cartésien des domaines des variables ne se réduit pas au vide, son plus petit élément est la solution optimale recherchée.

L'implantation du moteur se résume donc à l'écriture de quelques lignes de code. Avec des variables entières, la traduction des règles se fait naturellement [1]. Et pour résoudre un cas, une seule inférence de la machine Prolog suffit.

2.5 Innovation

Le langage proposé a été validé sur une base représentative de cas d'usage, constituée au départ de 200 règles métier à conclusions multiples (soit l'équivalent d'un millier de clauses à conclusion unique ou proprement « Horn »).

Ses premiers utilisateurs cobayes semblent s'approprier sans trop de difficulté sa logique combinatoire où il n'y a pas d'ordre entre les règles et où toutes les règles s'appliquent de concert.

Comment le montrent les précédentes captures d'écrans, la rédaction de ces règles logiques par un « paramétreur » s'effectue au travers d'une interface graphique, sans recourir à l'écriture d'une quelconque formule mathématique.

Avec une version libre de Prolog, les temps de réponse du moteur de règles sont de l'ordre du centième de seconde sur une machine standard, autrement dit, sans latence pour une application déployée d'emblée « dans le Cloud » [2].

L'innovation consiste ici à remplacer par des réponses métier beaucoup plus affinées et contextualisées un catalogue de réponses codifiées qui ne couvre pas toutes les situations d'urgence — seulement les cas les plus courants et, qui plus est, avec des volumes prédéfinis.

A l'aide d'un petit nombre de règles, on peut en effet couvrir un très large éventail de situations. A travers ces règles, toutes les informations communiquées sur l'incident sont susceptibles de moduler la réponse opérationnelle : aussi bien la description structurée de l'alerte transmise par le SGA que son enrichissement automatique par le Système d'Information Géographique (notamment sur les risques propres à la zone d'intervention).

Mais l'innovation doit aller de pair avec la conduite du changement. Jusqu'où la doctrine doit-elle être poussée ? La contrepartie en effet est que la combinatoire des configurations de départ va grossir très vite. On ne pourra plus raisonner en « extension ». Il faudra raisonner en « compréhension ». Nous avons rappelé pour cela l'importance accordée à l'intelligibilité des jeux de règles. Leur cycle de vie devra être soumis à un processus de validation de cohérence rigoureux.

3 Le moteur d'affectation et la réponse opérationnelle

Une fois la réponse tactique définie, il s'agit de trouver dans les centres d'IS les ressources correspondantes pour lancer l'opération. Il faut pouvoir détacher les moyens prévus pour les dépêcher sur place. Est-ce possible au regard de l'inventaire des matériels et personnels disponibles ? Et si oui, comment assurer au mieux :

- la rapidité d'intervention
- la connaissance du terrain
- la qualité de service (quitte à la dégrader si l'on ne peut faire autrement)
- la capacité de répondre ultérieurement à une urgence potentiellement plus grave (par une préservation de la couverture opérationnelle des risques) ?

L'aide à la décision conduit clairement ici à la formulation d'un problème d'optimisation multi-critères. Sa mécanisation fait appel à un solveur de contraintes.

3.1 Déclinaison préalable

Que demande une mission ? Essentiellement un engin avec son armement au complet, constitué notamment de personnel formé. Là encore, la doctrine de la tutelle départementale s'applique. Elle s'exprime en Prolog par des règles qui indiquent quel engin retenir de façon préférentielle pour chaque mission. Elle précise également le nombre de postes à pourvoir, avec pour chacun d'eux les compétences idoines :

- compétences liées à la mission
- compétences liées à l'engin pour certaines manipulations
- compétences liées à son acheminement (typiquement, un permis pour le conducteur du véhicule).

Cette déclinaison donne à l'objectif fonctionnel une structuration organique adaptée au territoire.

1. Une forme propositionnelle serait nettement moins intuitive et l'on s'épargne aussi l'obstacle de la traduction des contraintes de sommation [12]

2. Sur la plate-forme NexSIS, le moteur Prolog est virtualisé dans un conteneur Docker.

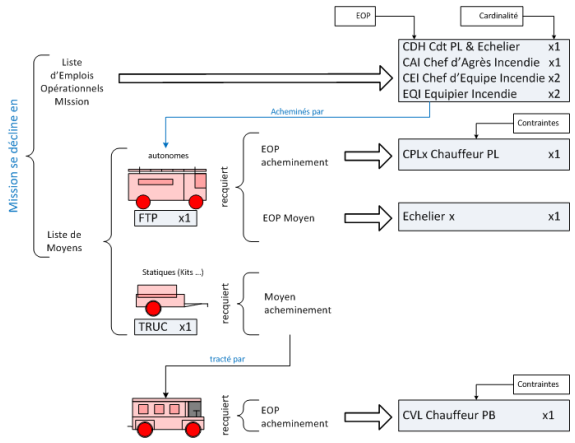


FIGURE 9 – Illustration de la déclinaison d'une mission (source : Document support au design SGO, ANSC, 2019)

3.2 Modélisation

La déclinaison des missions débouche sur la construction d'un modèle d'allocation des ressources en matériel et en personnel. Les casernes étant les centres de ressources, le problème principal consiste à affecter chacune des missions à une caserne.

Classiquement (cf. graphe de la figure 10), cela se modélise comme un problème « de transport » mettant en relation l'offre (les casernes) et la demande (les missions). Bien entendu, il faut que la caserne dispose de l'engin approprié avec un équipage mobilisable.

La résolution du problème principal dépend donc de la résolution de deux problèmes subordonnés :

- l'un relatif au choix des engins (nouveau problème de transport)
- l'autre, au choix des agents (problème d'affectation ou à nouveau de transport, selon que l'on individualise ou non les ressources en personnel).

Si l'on met tout cela à plat, des contraintes dites de liaison apparaissent entre ces 3 problèmes qui prennent la forme de contraintes d'intégrité : pour une mission m se décomposant en un besoin d'engin e, il ne peut y avoir d'affectation de e à un véhicule v de la caserne c que si c est affectée à m. Autrement dit, l'affectation de e à v implique l'affectation de m à c. Avec des variables binaires explicitant ces choix, cette implication se traduit par l'inégalité : $b_{e,v} \leq b_{m,c}$.

Il en va de même pour les besoins de postes.

Le système de contraintes à satisfaire pour que l'opération soit réalisable est donc formé par la réunion de 5 sous-systèmes :

- les contraintes du problème principal
- celles des deux problèmes subordonnés
- les contraintes des deux systèmes de liaison.

Le modèle peut être formulé avec des contraintes linéaires en Programmation en Nombres Entiers ou avec des contraintes globales en Programmation par Contraintes — option retenue ici pour rester dans l'univers CLP(FD).

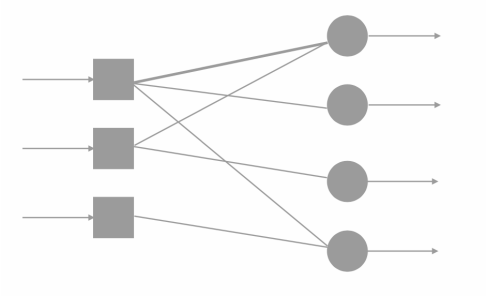


FIGURE 10 – Modélisation de la satisfaction des besoins (à gauche) par des ressources (à droite) en capacités limitées

Sur le plan « graphique », la capacité des arcs dans les problèmes subordonnés (graphes bipartis engins-véhicules et postes-agents) est conditionnée par la valeur des variables de décision du problème principal (les arcs missions-casernes devenant des coefficients annulateurs s'ils ne sont pas sélectionnés). Du point de vue de la complexité malheureusement, le problème résultant n'est plus un simple problème de flot.

3.3 Définition des objectifs

Si les contraintes à respecter se conçoivent aisément, il n'en va pas de même pour les critères d'optimisation. Prenons celui de la célérité : lorsque le dispositif se limite à une mission, la réponse la plus rapide vient de la caserne la plus proche. Mais si le dispositif est constitué de plusieurs missions, comment comparer deux réponses ? Une solution peut être plus rapide pour une mission et plus lente pour une autre. Faut-il prendre le min, le max, la moyenne des temps d'intervention ? La réponse n'est pas évidente.

La modélisation en contraintes proposée ouvre des perspectives auxquelles le métier ne sait pas encore répondre. Elle n'interdit pas de reproduire les heuristiques gloutonnes qu'emploient les logiciels de gestion actuels. Ces derniers effectuent une recherche séquentielle des moyens de secours selon un ordonnancement des casernes préalablement défini sur chaque portion de territoire. Mais ils ne peuvent produire ainsi qu'une seule proposition d'engagement.

Cette sorte d'optimisation « au niveau de la mission » a des effets pervers bien connus. Pour éviter de se voir préempter les compétences rares, elle incite à placer en premier les missions qui les exigent. Ceci a pour effet de dénaturer l'ordre métier qui ne correspond plus à la séquence d'arrivée des engins attendus sur les lieux de l'urgence.

En permettant de réserver à l'avance un volant de compétences rares, la pose des contraintes nous fait avancer vers une optimisation « au niveau de l'ensemble des missions ».

En termes de priorité opérationnelle, la modélisation actuelle permet déjà de faire cohabiter les « listes de défense » (pour la connaissance du terrain) avec les « temps de transit » (pour la célérité), ce qui en soi est une nouveauté.

A l'avenir, la prise en compte d'autres axes comme la qualité et la préservation de la couverture ira dans le sens d'une optimisation plus globale.

3.4 Sélection finale

Dire qu'une mission peut être affectée à une caserne, c'est dire que cette caserne est en capacité d'armer un engin pour l'assurer. Or, sur le plan des ressources humaines, on trouve souvent des solutions équivalentes du point de vue de l'optimalité. C'est donc à ce moment qu'il convient de tenir compte des multiples règles en vigueur pour « prioriser » l'affectation des personnels dans les casernes en fonction :

— des plannings journaliers des feuilles de garde,
— du cumul des temps d'engagement,
— de l'ancienneté,
— ...

On touche ici à des questions d'organisation de l'activité dans les casernes qui sont au cœur de l'acceptabilité sociale de la solution proposée. Là encore, Prolog permet d'exprimer avec une relative aisance les heuristiques métier de sélection des agents.

3.5 Innovation

A défaut de savoir définir « la » meilleure solution, on pourra, en faisant varier les critères d'optimisation, proposer à l'opérateur du CODIS un bouquet de solutions, toutes correctes par construction. D'un côté, le système d'aide à la décision le libèrera de tâches répétitives qui ont peu de valeur ajoutée. De l'autre, il lui demandera une plus grande acuité de jugement dans le choix de l'ordre de départ. Loin de le « robotiser », l'application devrait au contraire le faire monter en expertise.

4 Conclusion

Nous avons présenté dans cet article les rouages essentiels de la mécanique d'engagement des secours. La réponse « théorique » définit en quelque sorte le cadre de l'opération et la réponse « instanciée » complète ce tableau.
Pour les pompiers, le moteur de mobilisation est un outil de logistique opérationnelle. A l'image d'un système expert, il est par conception « paramétrable » par des doctrines, afin d'apporter des réponses respectueuses des spécificités locales.
Pour le développement et le déploiement de ce composant à forte valeur ajoutée, l'Agence du numérique de la sécurité civile a fait le pari de Prolog. L'une des raisons est que ce langage permet de « traduire le métier » de manière visible et compréhensible, et par là de modéliser beaucoup plus finement les besoins. Nous espérons avoir montré qu'il se prêtait aussi admirablement à la résolution des problèmes posés.

Remerciements

A l'Agence, à nos collègues de la BSPP, du SDMIS, des sociétés ToMCo et Cosytec, sans qui ce beau projet ne se serait pas devenu réalité.

A Alain Colmerauer, pour avoir inventé Prolog, il y a tout juste 50 ans.

Références

[1] Préfet Guillaume Lambert. Création d'un système unifié de traitement des appels, des alertes et des opérations des services d'incendie et de secours. Ministère de l'Intérieur, 2016.

[2] J. Rohmer and M. Carruel. Moteurs de règles et de contraintes dans NexSIS : Besoins et recommandations. Note d'orientation, ToMCo, 2020.

[3] Agence du Numérique de la Sécurité Civile. Rapport d'activité, 2020.

[4] N. Mathieu and N. Mortada. Moteur de mobilisation. Présentation au 127 congrès des Pompiers de France, 2021.

[5] D. Dubois and H. Prade. Towards a reconciliation between reasoning and learning - a position paper. In *Procs. SUM 2019*, LNAI 11940, 2019.

[6] Object Management Group. Decision model and notation. https ://www.omg.org/spec/DMN, 2015.

[7] G. Narboni. From deontic logic to mathematical optimization : a rule formalism for emergency decision-making (extended abstract). In *Procs. Int. Conf. on the Integration of Constraint Programming, Artificial Intelligence and Operations Research*, CPAIOR 22, Los Angeles, 2022.

[8] OASIS Emergency Management TC. Emergency data exchange language (EDXL). Organization for the Advancement of Structured Information Standards, 2009.

[9] GT CISU. Cadre d'interopérabilité des services d'urgence. Groupe de travail interministériel, 2019.

[10] G. Narboni. On rule systems whose consistency can be locally maintained. *AI Communications*, 26, 2013.

[11] P. Jeavons and M. Cooper. Tractable constraints on ordered domains. *Artificial Intelligence*, 79(2), 1995.

[12] O. Bailleux and Y. Boufkhad. Efficient CNF encoding of boolean cardinality constraints. In *Procs. Int. Conf. on Principles and Practice of Constraint Programming*, CP 03, LNCS 2933, 2003.

DOCaMEx, un outil web pédagogique qui propose une structuration de la connaissance inédite à base de cartes conceptuelles et d'arborescences de raisonnement technologique

C. Baudrit[1], P. Buche[2], J. Couteaux[1], J. Cufi[2], C. Fernandez[1], A Oudot[2]

[1] INRAE, Univ.Bordeaux-I2M, 33400, Talence, France
[2] INRAE, Univ. Montpellier-IATE, 34060, Montpellier, France

Cedric.baudrit@inrae.fr,patrice.buche@inrae.fr,julien.couteaux@u-bordeaux.fr,
julien.cufi@inrae.fr,christophe.fernandez@inrae.fr,Alrick.oudot@inrae.fr

Résumé

Une plateforme web DOCaMEx a été développée pour structurer la filière fromagère qui doit faire face à une perte des savoirs et des savoir-faire. Cette plateforme offre un support d'aide au raisonnement technologique et à la transmission des savoirs et savoir-faire fromagers. Elle est composée (1) d'un moteur de raisonnement (CAPEX), qui s'appuie sur des arborescences, capable de proposer des leviers d'action à mettre en œuvre pour corriger ou maintenir une qualité et (2) un livre électronique de connaissances (LdC), qui s'appuie sur des cartes conceptuelles, capable de donner accès à l'ensemble des connaissances capitalisées dans les filières.

Mots-clés

Ingénierie de la connaissance, cartes conceptuelles, arbres de décision.

Abstract

The cheese Industry is currently confronted with a loss of knowledge and know-how, the web platform DOCaMEx has been developed to help limit this loss. The platform offers support with technological reasoning along with the transmission of cheese-making knowledge and know-how. It is composed of (1) a reasoning engine (CAPEX), based on tree structures that is capable of proposing action levers whose implementation can correct or maintain quality and (2) an electronic knowledge book (LdC), based on conceptual maps, capable of giving the user access to the entirety of the knowledge gathered from all sectors of the industry.

Keywords

Knowledge engineering, conceptual maps, decision tree

1 Introduction

Le secteur de l'agro-alimentaire peine à recruter (ex: >3000 postes en CDI sont à pourvoir dans la filière laitière chaque année), ce qui engendre une perte des savoirs et des savoir-faire, et un turnover important de personnel. Fragilisée par ce contexte, la filière doit également faire face à des normes de plus en plus contraignantes, tout en peinant à s'appuyer sur une quantité croissante de connaissances génériques, et de données massiques, dues aux avancées scientifiques et technologiques. La capitalisation des savoirs technologiques, scientifiques, et des savoir-faire empiriques, est un enjeu important pour la pérennité et le développement des filières agroalimentaires. La filière fromagère n'échappe pas à ce constat et doit se doter d'outils numériques permettant de structurer leurs domaines de connaissances puis de développer et d'exploiter les bases de connaissances. Les procédés traditionnels de fabrication fromagère bénéficiant d'une Indication Géographique (AOP/IGP) reposent sur une multitude de savoirs, de savoir-faire et d'expériences (connaissances), forgés au cours du temps. Au niveau de la transformation, les fromagers-affineurs, par leurs pratiques et leur savoir-faire, adaptent leurs pratiques aux variations des caractéristiques des laits et des fromages. Ce positionnement engendre une richesse de savoir-faire des acteurs issue de leur propre expérience dont la transmission est essentiellement assurée par la voie de l'apprentissage sur site (faire pour savoir). Des évolutions internes aux appellations, en particulier en termes de renouvellement et de formation des opérateurs, fragilisent fortement la préservation et la transmission de ces savoir-faire. Les manières de transmettre les connaissances varient d'un individu à l'autre, également en fonction de la période et de l'activité. Organiser la transmission des savoirs et savoir-faire dans l'entreprise, c'est permettre aux collaborateurs d'acquérir une méthodologie afin de rendre plus efficace et pérenne le transfert des expertises au sein des équipes. Dans un contexte de transition numérique DOCaMEx[1] (résultat d'un collectif de 20 partenaires techniques et scientifiques) propose un outil web innovant permettant de recueillir, structurer et remobiliser des savoirs et savoir-faire fromagers d'une filière ou d'une entreprise. Il est composé de 2 outils interconnectés :

- un moteur de raisonnement CAPEX qui vise à apporter une aide à la recommandation d'actions technologiques permettant de corriger un défaut ou de maintenir une qualité d'un produit alimentaire dans un procédé de fabrication.
- un livre électronique de connaissances (LdC [2]) qui vise à transférer la connaissance et expliciter les

[1] https://www.youtube.com/watch?v=lBT3T-rsJBQ

recommandations proposées par CAPEX

Son but est (1) d'éviter la perte des savoirs et savoir-faire lors des départs à la retraite, (2) de favoriser le partage des expertises, (3) d'optimiser et fiabiliser le passage de relais entre deux ou plusieurs personnes, (4) de formaliser les bonnes pratiques et capitaliser sur le transfert pour pérenniser la démarche et (5) de préserver le patrimoine connaissances et compétences des filières. DOCaMEx a l'avantage d'être utilisable par tout type d'utilisateurs dans tout type d'environnement et avec différents niveaux d'accessibilité. Il propose une interface ergonomique intégrée et administrable dans l'entreprise qui permet de minimiser des erreurs de saisies, de construire et mettre à jour automatiquement le livre de connaissances et d'enrichir CAPEX avec les retours d'expériences.

2 Capitalisation et structuration de la connaissance

Les informations collectées, intégrant des mécanismes explicatifs, ont été structurées sous la forme d'un espace contenant des arborescences de raisonnement. Tous les autres types et sources d'information ont été structurés dans le livre électronique de connaissances (LdC) dont le squelette s'appuie sur le formalisme des cartes conceptuelles (Cmaps) [1,4].

2.1 Arborescences de raisonnement

Les raisonnements techniques (issus d'un recueil auprès des détenteurs du savoir et des savoir-faire) sont structurés sous la forme d'arborescences. La méthode consiste à interroger l'expertise des professionnels ou techniciens sur les éléments de raisonnement pouvant expliquer l'apparition de défauts ou l'élaboration de certains critères de qualité des produits. La structuration des données sous la forme d'arborescences de raisonnement intègre ensuite les mécanismes explicatifs tout en représentant des relations de cause à effet potentielles entre les défauts/qualités, appelés descripteurs par la suite, et les leviers d'intervention correctifs (correction de défauts) ou recommandés (maintien d'une qualité) [2]. La définition du descripteur permet d'établir le champ couvert par l'arborescence mise en place. Il s'agit de décrire le plus finement possible le défaut ou la qualité sur laquelle va porter l'arborescence. Un guide méthodologique a été mis en place permettant de structurer les situations à l'origine du défaut ou de la qualité en regroupant les situations proches à l'aide d'opérateurs logiques.

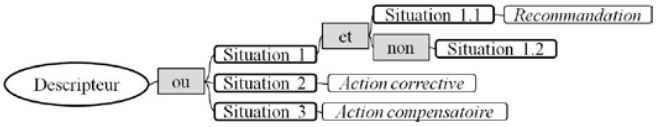


Figure 1: Arborescence explicative

Les situations à l'origine d'un défaut ou d'une qualité sont liées par des relations telles que « ou », « et », « sauf si », « non », *etc*. qui peuvent êtres couplées pour modéliser les interactions complexes entre des situations ou de la présence d'un contexte particulier dont il faut tenir compte. Pour chaque situation, des actions correctives, compensatoires ou des recommandations sont proposées. Par exemple, la figure 1 exprime le fait que le descripteur (défaut ou qualité du produit) peut s'expliquer par la situation 1, 2 ou 3 sur lesquelles il existe une action corrective (*resp*. une action compensatoire) pour corriger (*resp*. compenser) la situation 2 *(resp*. 3).

2.2 Cartes conceptuelles (Cmaps)

Les cartes conceptuelles constituent le squelette du LdC et s'appuient sur des graphes sémantiques composés d'un concept principal relié par des relations ontologiques aux concepts capables de décrire le concept principal. Il a été montré qu'une carte conceptuelle structurée hiérarchiquement facilitait la compréhension du lecteur en minorant sa désorientation et sa charge cognitive [1]. La relation ontologique de type taxonomique « est-un » permet de positionner le concept principal dans un groupe bien défini. La relation de synonymie permet de fournir des synonymes (professionnels) du concept principal. La relation ontologique de type méréologique « a-pour-partie » permet de lier une entité à ses parties. Cette relation peut être spécifiée selon les thèmes abordés ; par exemple la relation « a-pour-composant » (*resp*. a-pour-étape) permet de relier un objet (*resp*. un procédé) à ses composants (*resp*. ses étapes). Les relations du domaine servent à faire émerger des connaissances du domaine ; elles permettent d'indiquer avec quelles méthodes et outils un concept est mesuré, observé, caractérisé, contrôlé,

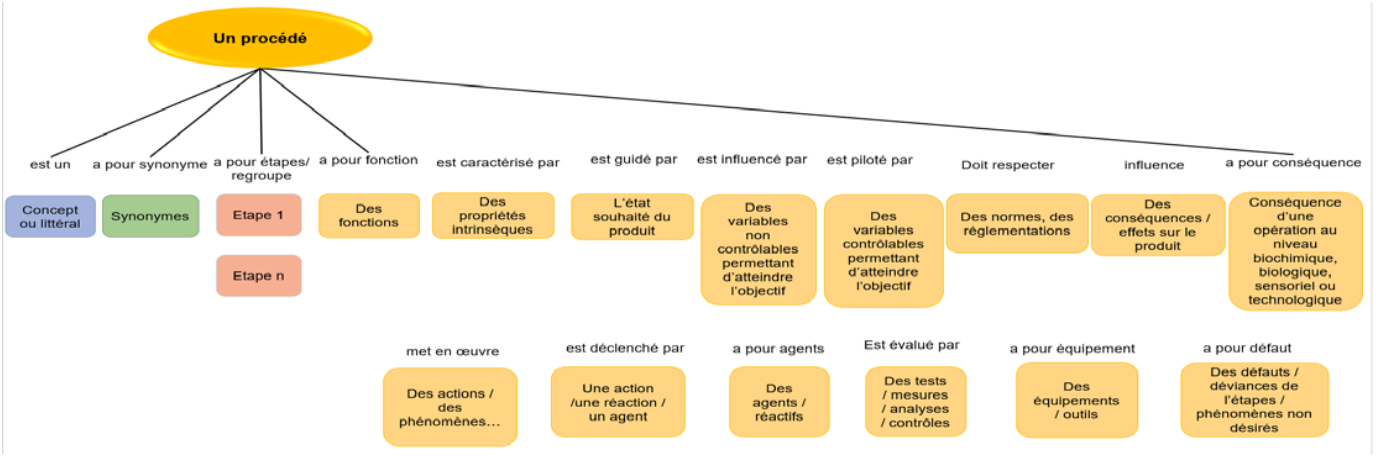


Figure 2: Cmap canonique capable de décrire tout type de procédé

etc. Différentes cartes canoniques ont été développées permettant de décrire tout type de procédé, de produit ou encore de phénomène. Par exemple, la Figure 2 représente la cmap canonique d'un procédé, i.e le modèle qui est capable de décrire tout type de procédé.

3 Portail web DOCAMEX

Le portail web DOCaMEx est un dispositif numérique qui permet de recueillir, structurer et remobiliser des savoirs et savoir-faire fromagers d'une filière ou d'une entreprise. Il donne accès à un livre électronique de connaissances (LdC) et au moteur de raisonnement (CAPEX) structurés et organisés comme des documents hypermédia et fortement interconnectés. Les interconnexions ont été développées entre ces deux éléments pour explorer les connaissances de manière interactive ou pour faciliter la compréhension en aidant l'utilisateur à trouver le sens des informations qui lui sont proposées en respectant une iconologie commune et une homogénéisation des termes employés. La base de connaissance pour instancier le portail DOCAMEX s'est construite à partir d'un collectif d'expert (à ce jour 15 experts en technologie fromagère issus des ENIL[2], Actalia[3], INRAE) et alimenté par un retour constant des filières utilisant l'outil.

3.1 Livre électronique de connaissances (LdC)

Les livres électroniques de connaissances (LdCs) sont des réseaux hypertextes dotés d'une sémantique ouverts sur l'internet qui permettent de structurer la connaissance d'un domaine à l'aide de cartes conceptuelles (voir section 2.2), de graphes de processus, de graphes d'influence, de documents téléchargeables, de pages Web et de fiches de connaissances hypermédias. Le LdC s'appuie sur un vocabulaire du domaine standardisé qui est disponible dans un glossaire. Les fiches de connaissance sont composées d'un ensemble prédéfini de champs descriptifs cliquables : titre, illustration, explications, date de création, auteurs, mots clés, voir aussi et références bibliographiques. L'illustration peut être une vidéo, un son, une photo, un dessin, un graphique, un tableau, une équation, etc., un lien vers un document consultable en ligne. Les graphes de processus permettant de représenter un processus multi-étapes. Les étapes sont des opérations unitaires ordonnées, caractérisées par des variables d'état, ayant chacune des données en entrée et un résultat en sortie, les variables d'état. Les graphes d'influence dans les LdCs proposent une représentation verticale des influences entre les concepts avec en amont les concepts causaux qui influencent un concept principal central qui influence en conséquence des concepts en aval. Les cartes conceptuelles, les graphes de processus, les graphes d'influence et les fiches de connaissances sont interconnectées par des liens hypertextes. La représentation granulaire progressive des cartes conceptuelles permet de situer à un niveau de détail relatif au concept principal et les liens qu'elles contiennent pointent vers des connaissances à un niveau de détail supérieur (plus détaillé). Cette structuration de la connaissance permet au LdC d'être évolutif et minimise la désorientation de l'utilisateur. L'outil est utilisable par tout type d'utilisateurs dans tout type d'environnement et avec différents niveaux d'accessibilité et peut être intégré et administrable dans l'entreprise. Un logiciel Cheeser_MBK© [3], permettant de gérer le paramétrage de la création et de l'exploitation d'un livre de connaissances, a été développé pour s'adapter au contexte des filières fromagères et des écoles de laiterie.

3.2 CAPEX

Le cœur du moteur de raisonnement CAPEX est basé sur des raisonnements techniques recueillis après des experts sous la forme d'arborescences de raisonnement (voir section 2.1). L'arborescence associée à un descripteur pourra être reliée à différentes ressources du LdC par 3 types de liens donnant accès à des fiches de connaissance, des cartes conceptuelles ou des ressources bibliographiques (voir section 3). L'utilisation en réseau de CAPEX a nécessité la construction d'une base de données spécifique (Triple Store), d'un hébergement sur un serveur web (JavaEE, Kotlin, GraphQ server), la création d'une interface web (SPA, TypeScript, Angular, GraphQL client), et d'un serveur d'authentification (CAS). GraphQL a été choisi afin d'uniformiser le dialogue entre l'interface et le moteur de raisonnement. La base de donnée orientée graphe choisie étant un triplestore celle-ci est interrogée au travers du langage standardisé SPARQL qui, en s'appuyant sur l'ontologie générique[4], permet d'exploiter le graphe RDF représentant les connaissances. A partir d'un défaut ou d'une qualité d'un produit, l'outil final de raisonnement technologique se présente sous la forme d'arborescences avec des liens de type « s'explique par » et en fin d'arborescence, par des propositions d'actions correctives ou compensatoires. L'objectif étant d'accompagner l'utilisateur dans sa réflexion (aide à la décision). L'expertise du technicien ou du formateur réside dans sa capacité à adapter la réponse proposée au contexte professionnel concerné.

4 Utilisabilité et scénario d'utilisation dans la filière des formages IGP-AOP

L'évaluation de l'outil dans sa capacité à transférer des connaissances de façon efficace comprend trois composantes évaluées (1) le degré de compréhension, qu'on peut expliquer comme la capacité d'un individu à acquérir, assimiler, transformer et exploiter de nouvelles connaissances ; (2) la désorientation qui mesure la tendance à perdre le sens de l'emplacement et de la direction dans un hypertexte et (3) la charge cognitive qui mesure l'effort utilisé dans un processus d'apprentissage. L'agrégation de ces trois composantes va fournir la note globale de la capacité de transfert de l'outil. Le degré de compréhension de l'outil a été mesuré par l'intermédiaire de deux tests réalisés avant (pré-test) et après (post-test) la navigation dans l'outil DOCaMEx [7]. Enfin, trois sessions de tests finaux d'évaluation du transfert de connaissance et de l'accompagnement au raisonnement technologique impliquant des utilisateurs de filières extérieures au projet (15 participants) ont été organisées. Les résultats montrent qu'au cours de la navigation dans la plateforme web DOCaMEx, les participants arrivent à accéder

[2] https://www.enil.fr/

[3] https://www.actalia.eu/

[4] CAPEX ontology https://doi.org/10.15454/9Z4PS3

aux différents contenus sans avoir une charge cognitive importante, avec une désorientation négligeable, et beaucoup de facilité pour aller chercher des contenus techniques précis. La charge cognitive et la désorientation ont été mesurées par la méthode d'auto-évaluation de l'effort mental investi au cours de l'utilisation du livre de connaissances sur une échelle de Likert à sept-points (très faible jusqu'à très fort) [5]. Il a aussi été constaté une nette amélioration des performances du raisonnement technologique lors de l'utilisation du moteur de raisonnement CAPEX (58% de bonnes réponses supplémentaires avec l'outil). Ce bilan confirme que les performances attendues par ce nouveau dispositif d'exploration des connaissances fromagères sont satisfaites.

4.1 Scénario d'aide aux raisonnements technologiques

Une entreprise de fromagerie doit faire face à un problème qui se caractérise par la présence d'un duvet noirâtre en surface de ses fromages communément appelé « poil de chat ». Cet accident en fromagerie est un développement en surface du fromage d'une moisissure du genre Mucor. La fromagerie souhaiterait corriger mais surtout comprendre l'origine du problème. Dans un premier temps l'utilisation de l'outil CAPEX va aider à identifier la source du problème et proposer des recommandations ou des actions correctives pour corriger ce défaut. L'entreprise choisira la recommandation la plus adaptée en fonction de ses contraintes. Par exemple, CAPEX identifie que ce défaut est souvent lié à des modifications de pratiques ou d'ambiance, remettant en cause l'équilibre « flore technologique – flore indésirable ». Il suffira par exemple pour l'entreprise de vérifier si le nouvel employé respecte toutes les phases de nettoyage.

4.2 Scénario de transfert et apprentissage

L'entreprise, faute de temps ou par manque d'employés expérimentés, pourra alors s'appuyer sur le contenu et l'ergonomie du livre électronique de connaissances pour rappeler les bonnes pratiques et former le nouvel arrivant en lui donnant accès par exemple à une fiche de connaissance détaillant la « gestion de l'hygiène des planches d'affinage ». L'interconnexion des deux outils Capex – LdC permet d'obtenir des recommandations, des explications et des connaissances de manière intuitive. Par exemple, pour un défaut d'aspect de type « poils de chat », Capex propose qu'une des causes possibles soit le mauvais nettoyage de planche, à ce moment-là l'utilisateur peut cliquer sur nettoyage des planches d'affinage et cela ouvre une fiche de protocole de nettoyage, une vidéo de bonne pratique et même vers Capex pour d'autres défauts liés à l'affinage. L'outil illustre les défauts par des photos permettant à l'opérateur de mieux identifier le type de défaut en cas de doute ou d'incapacité à le reconnaitre.

5 Conclusion

L'outil développé constitue un formidable système informatique de capitalisation et de transmission des compétences professionnelles. Ce dispositif numérique innovant est configuré sur une plateforme hébergeant un socle commun mis à jour régulièrement consultable par tous ou de manière privative en naviguant dans des partitions confidentielles propres à chaque filière. Les outils développés s'adressent à des techniciens experts ou débutants (service technique d'une filière ou d'une entreprise), des fromagers, des animateurs de filière ou des responsables d'usine, ainsi qu'à des formateurs et apprenants (formation initiale et continue). Cette plateforme fonctionnelle et structurée incrémentée de données dites « génériques » rendues publiques permet d'envisager le déploiement de l'outil vers d'autres filières ou entreprises laitières. Le moteur de raisonnement CAPEX permet aux utilisateurs de naviguer dans des arbres de raisonnement technologique. Une nouvelle version de CAPEX est actuellement en cours d'élaboration. Elle permettra d'une de prendre en considération les retours des utilisateurs sur la mise en œuvre des actions recommandes en terme d'efficacité technologique) et d'autre part de trier les actions correctives proposées selon plusieurs critères (coût, risque sanitaire, efficacité technologique, …).

Remerciements

Ce travail a été soutenu par le financement du projet CASDAR DoCaMEx (AAP IP 2016 n° 5624 de CAS DAR 2016 coordonné par le CTFC) et par la Direction Interministérielle du Numérique (DINUM) du Ministère de la Transformation et de la Fonction Publiques dans le cadre du Plan France Relance.

6 Références

[1] Amadieu, F., Van Gog, T., Paas, F., Tricot, A., & Mariné, C. (2009). Effects of prior knowledge and concept-map structure on disorientation, cognitive load, and learning. *Learning and Instruction*, 19(5), 376-386.

[2] Buche, P., Cuq, B., Fortin, J., & Sipieter, C. (2019). Expertise-based decision support for managing food quality in agri-food companies. *Computers and Electronics in Agriculture*, *163*, 104843.

[3] Fernandez Christophe, Ndiaye Amadou, Baudrit C. (2021) Cheeser_MBK: an Electronic Knowledge Books Designer for French Cheese Sector (3.9.0) [PHP7, Mariadb, HTML5]. DDN.FR.001.080046.000.S.P.2021.000.20900

[4] Krieglstein, F., Schneider, S., Beege, M., & Rey, G. D. (2022). How the design and complexity of concept maps influence cognitive learning processes. Educational technology research and development, 1-20.

[5] Pass, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory : *Educational Psychologist*, 38,(1), pp.63-71.

[6] Suciu, I., Ndiaye, A., Baudrit, C., Fernandez, C., Kondjoyan, A., Mirade, P. S., ... & Della Valle, G. (2021). A digital learning tool based on models and simulators for food engineering (MESTRAL). Journal of Food Engineering, 293, 110375.

[7] Vakilifard, A., & Armand, F. (2011). Les effets de la carte conceptuelle hiérarchique sur la compréhension littérale et inférentielle de textes informatifs en langue seconde. *Canadian modern language review*, *67*(2), 217-245.

Vers une ingénierie des exigences dirigée par les données : analyse automatique d'avis d'utilisateurs

Jialiang Wei*, Anne-Lise Courbis*, Thomas Lambolais*, Binbin Xu*,
Pierre Louis Bernard** and Gérard Dray*

*: EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France
**: EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Montpellier, France
*: firstname.lastname@mines-ales.fr **: firstname.lastname@umontpellier.fr

Résumé

Ces travaux s'inscrivent dans le cadre de l'ingénierie des exigences dirigée par les données et notamment les avis d'utilisation. Les commentaires en ligne d'utilisateurs d'applications sont une source importante d'informations pour en améliorer leur fonctionnement et en extraire de nouveaux besoins. Nous proposons une analyse automatisée en utilisant CamemBERT, un modèle de langue en français qui fait aujourd'hui état de l'art. Afin d'affiner ce modèle, nous avons créé un jeu de données de classification multi-labels de 6000 commentaires issus de trois applications du domaine de l'activité physique et la santé[1]*. Les résultats sont encourageants et permettent d'identifier les avis concernant des demandes de nouvelles fonctionnalités.*

Mots-clés

Ingéniérie des exigences, Ingénierie guidée par les données, TALN, CamemBERT, Apprentissage profond

Abstract

We are concerned by Data Driven Requirements Engineering, and in particular the consideration of user's reviews. These online reviews are an important source of information for extracting new needs and improvement requests. In this paper, we provide an automated analysis using CamemBERT, which is a state-of-the-art language model in French. In order to fine tune the model, we created a multi-label classification dataset of 6000 user reviews from three applications in the Health & Fitness field. The results are encouraging and make it possible to identify automatically the reviews concerning requests for new features.

Keywords

Requirements Engineering, Data Driven Requirements Engineering, NLP, CamemBERT, Deep Learning

1. Données disponible en ligne : https://github.com/Jl-wei/APIA2022-French-user-reviews-classification-dataset

1 Introduction

L'ingénierie des exigences (IE) est l'une des phases du cycle de développement des systèmes, visant à élaborer un cahier des charges aussi clair et complet que possible, à partir des interviews de différentes parties prenantes (clients, utilisateurs, donneurs d'ordre). Il s'agit d'une des phases les plus critiques [1], jouant un rôle fondamental pour obtenir un logiciel de qualité. Les activités principales de l'IE sont : la conduite des interviews, l'élicitation, la modélisation du domaine et l'analyse de faisabilité. L'étape d'élicitation a pour objectif de mettre en exergue les besoins réels des parties prenantes [2]. Les approches traditionnelles visant à établir des modèles explicites d'exigences à partir d'interviews, brainstorming, observations, etc. sont enrichies par de nouvelles approches, visant à exploiter les retours des utilisateurs mis en ligne sur les stores d'applications ou les réseaux sociaux. Ainsi, la communauté d'IE propose de relever de nouveaux défis [3] consistant à développer une ingénierie des exigences dirigée par les données permettant de traiter un volume important d'avis d'utilisateurs.

Les stores d'applications, comme Google Play® et App Store®, sont devenus des plate-formes de communication entre utilisateurs et développeurs assurant la collecte d'avis sur chacune des applications en téléchargement. Ces avis sont riches en information car ils contiennent des critiques et des préférences, des retours d'expérience, des rapports d'erreur et des demandes de nouvelles fonctionnalités. Par exemple, l'application Facebook reçoit plus de 4000 commentaires journaliers dont 30% peuvent être considérés comme une source pour identifier de nouveaux besoins [4]. Exploiter manuellement de telles sources serait laborieux. Nous proposons d'effectuer un Traitement Automatique du Langage Naturel (TALN) pour filtrer les avis et cibler les plus pertinents en vue d'identifier de nouvelles exigences. Plus précisément, nous nous sommes intéressés à BERT (*Bidirectional Encoder Representations from Transformers*) [5], qui est un modèle d'apprentissage profond basé sur une architecture de type *Transformer*. BERT peut être affiné pour effectuer différentes tâches telles que la classification de textes, la réponse aux questions, l'inférence en langage naturel, etc. Ce sont les modèles basés

sur BERT qui ont obtenu les meilleures résultats sur la plupart des tâches de la communauté TALN. BERT étant entraîné en anglais, nous nous sommes tournés vers CamemBERT [6], un modèle de type BERT qui a été entraîné sur le corpus OSCAR en français afin d'avoir de meilleures performances sur les commentaires écrits en français.
Pour affiner le modèle BERT, il est nécessaire de l'entraîner avec un ensemble de données annoté. Il existe plusieurs jeux de données de commentaires annotés pour effectuer la classification [7–13]. La plupart de ces jeux de données sont en anglais. Certains sont multi-langues [8, 13]. Mais aucun d'eux n'intègre suffisamment de commentaires rédigés en français. Nous proposons donc de créer un jeu de données en français et d'explorer les résultats obtenus. Notre sujet d'étude étant à terme le développement selon une approche dirigée par les données d'applications de suivi de séniors pour « *bien vieillir en santé* », nous nous sommes intéressés aux commentaires de trois applications de la catégorie « Health & Fitness » de Google Play.
Dans la partie suivante nous présentons le jeu de données constitué ainsi que la méthode mise en œuvre pour conduire notre expérimentation. Les résultats de classification obtenus sont discutés dans la troisième partie. Enfin, dans la dernière partie, nous exposons les perspectives envisagées sur ces travaux.

2 Méthode mise en œuvre pour l'apprentissage

2.1 Jeu de données et son annotation

L'usage de CamemBERT pour classifier les avis d'utilisateurs écrits en français nécessite un jeu de données annoté. En l'état actuel de nos connaissances, il n'en existe pas. Nous en avons donc créé un. Nous avons tout d'abord collecté sur la plateforme Google Play des avis rédigés en français concernant les applications Garmin Connect, Huawei Health et Samsung Health. Le nombre de commentaires collectés et la taille choisie pour constituer le jeu de données pour la phase d'apprentissage sont présentés dans le Tableau 1. Nous avons ensuite associé manuellement des labels aux avis. Nous avons choisi d'utiliser quatre labels, comme cela est proposé dans [12] : *Évaluation*, *Rapport d'erreur*, *Demande de nouvelles fonctionnalités* et *Expérience utilisateur*. L'*évaluation* est généralement un texte simple qui exprime le sentiment général de l'utilisateur sous forme d'éloge, de critique ou de dissuasion. Le *rapport d'erreur* est relatif aux problèmes rencontrés lors de l'utilisation de l'application : perte de données, arrêt brutal de l'application, problème de connexion, etc. La *demande de nouvelles fonctionnalités* concerne de nouveaux services, de nouveaux contenus ou encore de nouvelles interfaces. Enfin, les avis de type *expérience utilisateur* sont des expériences relatées par les utilisateurs qui sont décrites par exemple comme des conseils d'utilisation, des fonctions ou des usages perçus comme très utiles ou faciles d'utilisation. Il est possible d'associer plusieurs labels à un même avis. Par exemple, l'avis : *"c'est une très bonne application. Elle nous aide beaucoup à rester toujours actif."* sera annoté dans les catégories *Évaluation* et *Expérience Utilisateur*. L'avis *"j'aimais bien cette application mais elle ne fonctionne plus : message d'erreur téléphone rooté ! ! faux et j'ai vérifié."* sera annoté dans les catégories *Évaluation* et *Rapport d'erreur*. Les avis sont annotés par quatre auteurs de l'article, et révisés par Jialiang Wei. Le Tableau 2 indique le nombre d'avis classés dans chacune des catégories pour les trois applications cibles. La somme des avis classés n'est pas égale au total des avis puisque certains avis ont été affectés à plusieurs classes.

App	#Avis	Échantillon
Garmin Connect	22880	2000
Huawei Health	10304	2000
Samsung Health	18400	2000

TABLEAU 1 – Applications et avis collectés pour la phase d'entraînement

App	Total	(Ev)	(R)	(D)	(Exp)
Garmin Connect	2000	1260	757	170	493
Huawei Health	2000	1068	819	384	289
Samsung Health	2000	1324	491	486	349

TABLEAU 2 – Classification manuelle des avis en quatre familles : (**Ev**)aluation, (**R**)apport d'erreur, (**D**)emande de fonctions, (**Exp**)érience utilisateur

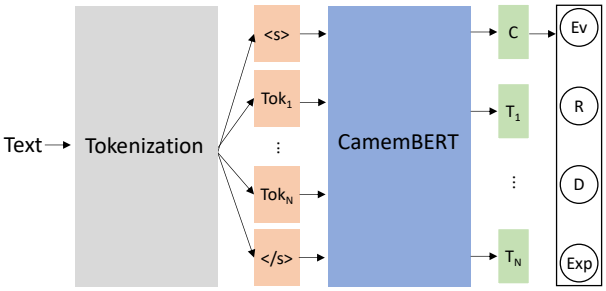


FIGURE 1 – Vue d'ensemble du modèle CamemBERT pour la classification multilabels

2.2 Modèle de classification

La classification des avis d'utilisateurs est un problème de classification de type TALN [14]. Ces dernières années, de nombreuses méthodes d'apprentissage statistique ont été appliquées à la classification d'avis. Maalej et al. [12] utilise des réseaux bayésiens, des arbres de décision et régression logistique multinomiale pour effectuer la classification des avis dans ces mêmes quatre catégories : *évaluation*, *rapport d'erreur*, *demande de fonctionnalités* et *expérience utilisateur*. Restrepo Henao et al. [15] applique BERT pour faire la classification des avis dans les catégories : *rapport de problèmes*, *demande de fonctionnalités*, et *avis non pertinent* ; ils comparent les performances des modèles d'apprentissage BERT et d'un réseau de neurones convolutif à apprentissage profond et concluent que BERT obtient les

meilleures taux de précision et de rappel. Mekala et al. [16] comparent la performance dans le domaine du *crowdsourcing* (production participative) de BERT avec des modèles de Machines à Vecteurs de Support, des classifieurs de type FastText et ELMo sur des avis d'utilisateurs classés en deux catégories : *inutile* et *utile*. BERT obtient également les meilleurs résultats. Les bonnes performances de BERT ont guidé notre choix sur sa version française CamemBERT.
CamemBERT ne peut utiliser du texte brut en entrée. Une phase de pré-traitement est nécessaire : chaque avis est découpé en un ensemble de termes, ou jetons [17]. Les balises `<s>` et `</s>` sont ajoutées pour marquer respectivement le début et la fin de l'avis. Les avis sont ensuite formatés à la longueur 512 en complétant ceux qui sont de taille inférieure avec la balise `<PAD>`. Ceux qui sont de taille supérieure sont tronqués. Un masque est alors associé à chaque avis afin de repérer les symboles `<PAD>` (valeur 0) et les termes (valeur 1). Puis, on attribue à chaque terme un identifiant numérique. Enfin, pour adapter CamemBERT à la tâche de classification souhaitée, nous modifions son architecture en ajoutant à son dernier niveau une couche linéaire à quatre sorties, comme le montre la Figure 1.

3 Expérimentations et résultats

Nous avons utilisé la bibliothèque PyTorch pour la mise en œuvre CamemBERT. Le modèle a été entraîné avec trois epochs, le paramètre `batch_size` et l'optimiseur AdamW avec un taux d'apprentissage fixé à $2e^{-5}$. Le modèle a été entraîné sur un ordinateur de 48 Go de mémoire RAM, muni d'un processeur Intel i7-7820HQ et d'une carte graphique NVIDIA Quadro M2200 de 4 Go de VRAM. L'évaluation de la performance de CamemBERT sur la classification des avis s'est faite sur deux expérimentations. Dans la première, les modèles sont entraînés sur une partie des avis des trois applications et testés sur l'autre partie. Dans la seconde expérimentation, les modèles sont entraînés sur les avis de deux applications et évalués sur ces applications et sur la troisième application.

3.1 Apprentissage à partir des trois applications

Pour cette expérimentation, nous avons sélectionné 60% des avis des trois applications comme ensemble d'apprentissage, 20% comme ensemble de validation et 20% comme ensemble de test en utilisant un échantillonnage stratifié. Cette opération a été répétée 10 fois. Les critères d'évaluation de performance des modèles sont : la précision, le rappel et la F1-Mesure. Les résultats moyens obtenus sur les 10 apprentissages sont présentés dans le Tableau 3.
Comme le montre le Tableau 3, les résultats obtenus sont corrects dans l'ensemble : la moyenne de la F1-Mesure est de 0,89. Les résultats concernant le critère *Expérience Utilisateur* sont inférieurs aux trois autres. Ceci s'explique de deux façons. Tout d'abord les avis concernant ce critère sont plus variés, de nature moins homogène que les autres critères, ce qui rend plus difficile l'apprentissage des caractéristiques. Aussi, il serait nécessaire de faire un apprentissage sur un ensemble plus important d'avis pour améliorer les résultats. La seconde raison réside vraisemblablement dans le fait que quatre personnes ont participé à l'annotation des avis et leur interprétation du critère *Expérience Utilisateur* peut influencer les performances du modèle.

	Précision	Rappel	F1
Évaluation	0,88	0,93	0,91
Rapport d'erreur	0,92	0,93	0,93
Demande de fonctions	0,85	0,83	0,84
Expérience utilisateur	0,81	0,73	0,77
Moyenne pondérée	0,88	0,89	0,89

TABLEAU 3 – Résultats de la classification des avis des trois applications

3.2 Apprentissage à partir de deux applications

Dans cette expérimentation, nous avons utilisé la même stratégie d'échantillonage stratifié : 60% des avis de deux applications comme base d'apprentissage, 20% comme validation. Les tests s'appliquent sur les 20% restants et également sur tous les avis de la troisième application. Nous avons fait ceci pour les trois combinaisons possibles d'applications sélectionnées pour l'apprentissage versus le test. Par exemple, une de ces combinaisons a consisté à réaliser l'apprentissage / validation sur 1600 avis de Garmin Connect et 1600 de Huawei Health et le test sur 800 avis de ces applications et les 2000 avis de Samsung Health. Pour chaque combinaison, 10 expériences ont été réalisées. Les résultats moyens des 30 apprentissages (3×10) sont présentés dans les Tableaux 4 et 5.

	Précision	Rappel	F1
Évaluation	0,88	0,93	0,90
Rapport d'erreur	0,91	0,93	0,92
Demande de fonctions	0,85	0,81	0,83
Expérience utilisateur	0,79	0,72	0,76
Moyenne pondérée	0,87	0,88	0,88

TABLEAU 4 – Résultats de classification des avis pour les 20% restants de deux applications

	Précision	Rappel	F1
Évaluation	0,88	0,92	0,90
Rapport d'erreur	0,85	0,92	0,88
Demande de fonctions	0,80	0,74	0,75
Expérience utilisateur	0,77	0,69	0,73
Moyenne pondérée	0,86	0,86	0,85

TABLEAU 5 – Résultats de classification des avis : apprentissage sur deux applications et test sur la troisième application

Les résultats présentés dans le Tableau 4 sont légèrement inférieurs à ceux du Tableau 3. Ils montrent que l'on peut obtenir de bonnes performances de classification avec un jeu de données restreint. Dans le Tableau 5, on peut obser-

ver que la moyenne de la précision et du rappel pour lesquels les modèles n'ont pas été entraînés sont inférieurs de 1-2% par rapport aux modèles ayant servi pour l'entraînement. Cette expérimentation montre que les résultats restent de très bonne qualité en entraînant le modèle sur un sous-ensemble d'applications.

4 Discussion et travaux futurs

Dans ce travail, nous avons créé un jeu de données pour la classification multi-labels d'avis rédigés en français d'utilisateurs d'applications de suivi d'activité physique. Nous avons utilisé le modèle CamemBERT pour classifier ces avis et les résultats obtenus montrent de bonnes performances. Les expérimentations d'entraînement et de test sur des applications distinctes montrent qu'il est possible de généraliser le modèle sur des applications de même champ applicatif. Ces résultats nous encouragent à poursuivre nos travaux relatifs à l'ingénierie des exigences dirigée par les données pour le développement de différentes versions d'une application de prévention des effets du vieillissement par le suivi de l'activité physique des seniors pour bien vieillir en santé. Nous envisageons d'affiner la classification des demandes de nouvelles fonctionnalités en permettant d'identifier, par une approche non supervisée, les concepts du domaine de l'application cible, comme par exemple, les vitesses et temps de marche, les déséquilibres ou le suivi du sommeil. Cette classification pourrait être présentée visuellement pour indiquer au concepteur quelles sont les fonctionnalités principalement demandées par les utilisateurs.
L'objectif global sera de (i) regrouper les demandes et propositions par catégorie de besoins, (ii) mettre en correspondance ces besoins avec les modèles présents dans le cahier des charges, afin de pouvoir les prendre en compte plus facilement dans la conception.

Références

[1] A. van Lamsweerde, *Requirements Engineering : From System Goals to UML Models to Software Specifications*. Wiley, January 2009.

[2] A. Bennaceur, T. Than Tun, Y. Yu, and B. Nuseibeh, "Requirements Engineering," in *Handbook of Software Engineering*. Springer Verlag, 2019, pp. 1–44.

[3] X. Franch, "Data-Driven Requirements Engineering : A Guided Tour," in *Communications in Computer and Information Science*, vol. 1375. Springer, 2021, pp. 83–105.

[4] D. Pagano and W. Maalej, "User feedback in the appstore : An empirical study," in *2013 21st IEEE International Requirements Engineering Conference (RE)*, 2013, pp. 125–134.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT : Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 — 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies — Proceedings of the Conference*, vol. 1, October 2019, pp. 4171–4186.

[6] L. Martin, B. Muller, and P. J. e. a. Ortiz Suárez, "CamemBERT : a Tasty French Language Model," in *58th Annual Meeting of the Association for Computational Linguistics*, Seattle, United States, July 2020.

[7] G. Williams and A. Mahmoud, "Mining Twitter Feeds for Software User Requirements," in *2017 IEEE 25th International Requirements Engineering Conference (RE)*, 2017, pp. 1–10.

[8] S. Scalabrino, G. Bavota, B. Russo, M. D. Penta, and R. Oliveto, "Listening to the Crowd for the Release Planning of Mobile Apps," *IEEE Transactions on Software Engineering*, vol. 45, no. 1, pp. 68–86, 2019.

[9] E. Guzman, M. El-Haliby, and B. Bruegge, "Ensemble Methods for App Review Classification : An Approach for Software Evolution," in *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2015, pp. 771–776.

[10] F. Palomba, M. Linares-Vásquez, and G. e. a. Bavota, "Crowdsourcing user reviews to support the evolution of mobile apps," *Journal of Systems and Software*, vol. 137, pp. 143–162, 2018.

[11] A. Ciurumelea, A. Schaufelbühl, S. Panichella, and H. C. Gall, "Analyzing reviews and code of mobile apps for better release planning," in *2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2017, pp. 91–102.

[12] W. Maalej, Z. Kurtanović, H. Nabil, and C. Stanik, "On the automatic classification of app reviews," *Requirements Engineering*, vol. 21, no. 3, pp. 311–331, 2016.

[13] C. Stanik, M. Haering, and W. Maalej, "Classifying Multilingual User Feedback using Traditional Machine Learning and Deep Learning," in *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*, 2019, pp. 220–226.

[14] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning–Based Text Classification : A Comprehensive Review," *ACM Comput. Surv.*, vol. 54, no. 3, April 2021.

[15] P. R. Henao, J. Fischbach, D. Spies, J. Frattini, and A. Vogelsang, "Transfer Learning for Mining Feature Requests and Bug Reports from Tweets and App Store Reviews," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 2021, pp. 80–86.

[16] R. R. Mekala, A. Irfan, E. C. Groen, A. Porter, and M. Lindvall, "Classifying User Requirements from Online Feedback in Small Dataset Environments using Deep Learning," in *2021 IEEE 29th International Requirements Engineering Conference (RE)*, 2021, pp. 139–149.

[17] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Berlin, Germany : Association for Computational Linguistics, August 2016, pp. 1715–1725.

Poster

Analyse automatique de documentation technique – application sur des retours d'essais en développement

C. Berthou - Safran Aircraft Engines - celine.berthou@safrangroup.com

Résumé

Des essais sont réalisés lors de la phase de développement du moteur et un rapport d'essai est émis lorsqu'un évènement se produit. L'opérateur réalisant l'essai décrit l'évènement en langage naturel. Cette source d'information, sous format textuel, n'est pas ou peu capitalisée à ce jour.

La finalité des travaux entrepris est la mise en place d'un modèle d'apprentissage automatique de la cause de l'évènement, à partir de ces rapports d'évènements, dans un cadre non supervisé faute de label systématiquement renseigné et viable. L'objectif est de capitaliser sur les évènements produits en développement afin d'améliorer l'aide au diagnostic pour les éventuels futurs évènements en service.

Cet article présente les modèles testés pour répondre à la problématique. Nous avons testé un premier modèle de topic modeling LDA (Latent Dirichlet Allocation) puis un modèle de langage neuronal BERT (Bidirectional Encoder Representations from Transformers). L'application de ces deux modèles nous a permis de réaliser une classification automatique des descriptifs d'évènements par cause d'évènement selon deux approches distinctes.

Mots-clés

NLP, apprentissage non supervisé, topic modeling, LDA, transformer, BERT.

Abstract

Tests are carried out during the engine development phase and a test report is issued when an event occurs. The operator performing the test describes the event in natural language. This source of information, in textual format, is not or little capitalized until now.

The purpose of this work is to set up an automatic learning model of the cause of the event, based on these event reports, in an unsupervised learning for lack of a systematically informed and viable label. The objective is to capitalize on the events produced in development in order to improve diagnostic assistance for any future events in service.

This article presents the models tested to solve the problem. We tested a first LDA (Latent Dirichlet Allocation) topic modeling model and then a BERT (Bidirectional Encoder Representations from Transformers) neural language model. The application of these two models allowed us to carry out an automatic classification of the descriptions of events by cause of event according to two distinct approaches.

Keywords

NLP, unsupervised learning, topic modeling, LDA, transformer, BERT.

1 Introduction

De manière générale, nous disposons d'une grande quantité de données texte dans différents domaines d'application : description d'essais, description d'évènements, descriptif de maintenance etc… L'ensemble de ces données est peu exploité alors qu'elles détiennent des informations essentielles à une meilleure connaissance du moteur, son développement, son bon fonctionnement et sa maintenance.

Un besoin est donc identifié d'exploiter cette importante source d'information et de développer des méthodes et des outils de *text-mining* permettant de traiter ces données texte de manière automatique. Plus particulièrement, nous disposons de rapports d'essais d'évènements en développement détaillant, selon le cas, l'essai réalisé, le contexte de l'évènement et sa description.

Nous souhaitons labelliser de manière automatique l'ensemble des rapports d'évènements en développement selon le type d'évènement produit, en utilisant des techniques de *text-mining* (exploitation de données non structurées telles que du texte écrit en langage naturel).

Cet article présente la classification automatique obtenue des rapports d'évènements par typologie d'évènement. Les sections §2 et §3 présentent les deux approches utilisées pour aboutir à ce résultat : en premier lieu, le modèle de topic modeling LDA (*Latent Dirichlet Allocation*) et ensuite, le modèle de réseaux de neurones transformer BERT (*Bidirectional Encoder Representations from Transformers*). La discussion en §4 résume les acquis et les constats afin de proposer des pistes de travaux futurs.

2 Modèle de topic modeling LDA

2.1 Chaine de traitement

Le topic modeling (identification de topics ou sujets) est une méthode de classification non supervisée de documents, équivalente au clustering pour des données numériques [1].

Pour mettre en œuvre ce type de modèle, nous allons appliquer la chaine de traitement présentée en Figure 1.

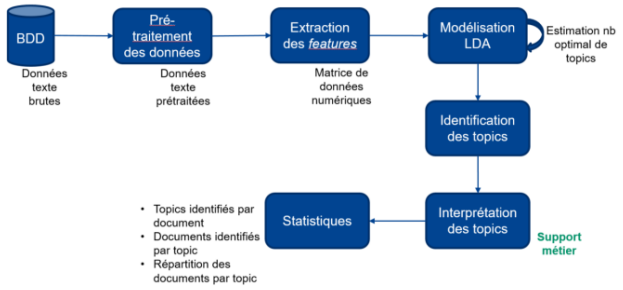


Figure 1. Chaine de traitement pour l'application du modèle LDA.

Les rapports d'évènements d'essais en développement sont extraits de la base de données sous format textuel brut. Les données texte sont ensuite nettoyées et les caractéristiques (*features*) sont extraites. On passe ainsi de données texte à une matrice de données numériques sur laquelle on peut appliquer le modèle LDA, après avoir estimé le nombre optimal de topics d'évènements à extraire de l'ensemble des rapports d'évènements considérés. Une fois les topics identifiés par le modèle LDA, on cherche à les interpréter avec l'aide du métier, et on en déduit un ensemble de statistiques.

2.2 Résultats

Le modèle LDA retenu permet d'obtenir 65 topics. Chaque topic est caractérisé par ses douze termes les plus représentatifs qui sont classés par ordre décroissant d'importance. On peut, à partir de ces termes représentatifs, inférer le topic d'évènement avec le support des métiers. On représente graphiquement ces 65 topics d'évènements en deux dimensions via une ACP (Analyse en Composantes Principales), à partir de la librairie pyLDAvis de Python [2].

Cette approche va nous permettre, après l'interprétation des topics d'évènements identifiés, de constituer une liste de classes d'évènements la plus exhaustive possible, que l'on va utiliser dans l'approche neuronale que l'on a souhaitée aussi tester.

3 Modèle transformer BERT

3.1 Chaine de traitement

BERT est l'acronyme de Bidirectionnal Encoder Representations from Transformers. C'est un modèle de langage pré-entraîné, développé en 2018 par Google, qui repose sur des réseaux neuronaux et est très utilisé en classification de textes. Pour la mise en place de ce modèle, nous allons suivre la chaine de traitement présentée en Figure 2 :

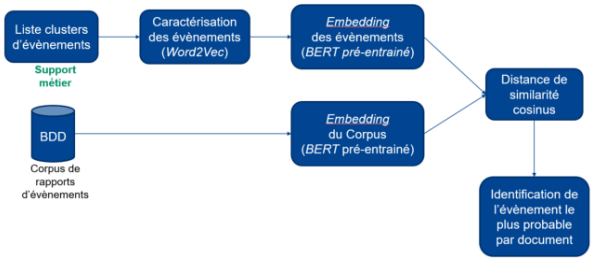


Figure 2. Chaine de traitement pour l'application du modèle BERT.

Après avoir consolidé avec les métiers la liste exhaustive des évènements, nous les caractérisons avec le modèle de word embedding Word2Vec. Puis nous réalisons l'embedding d'une part des évènements précédemment caractérisés, et d'autre part, du corpus de rapports d'évènements d'essais en développement. Nous allons procéder ensuite par distance de similarité cosinus entre un rapport d'évènement donné et chaque cluster d'évènement, pour identifier à chaque rapport d'évènement la combinaison d'évènements associée et ainsi le cluster d'évènement le plus probable.

3.2 Résultats

Maintenant que l'embedding de chaque rapport d'évènement et de chaque cluster d'évènement est réalisé, il reste à mesurer la distance entre chaque rapport d'évènement et chaque cluster d'évènement. On utilise pour cela la mesure de similarité cosinus entre les deux vecteurs correspondants. Pour un rapport d'évènement donné, le cluster d'évènement le plus probable est celui dont la probabilité de similarité est la plus élevée.

4 Discussion et perspectives

Nous disposons d'une base de données importante de rapports d'essais en développement où lorsqu'un évènement se produit en essai, l'évènement est décrit en langage naturel par l'opérateur. L'objectif fixé est de labelliser de manière automatique l'ensemble de ces rapports selon une typologie d'évènement.

Pour cela, plusieurs méthodes ont été testées. Comme les données considérées ne sont labellisées que dans 8% des cas et que la fiabilité de cette labellisation n'est pas assurée, nous avons testé des méthodes d'apprentissage non supervisé.

Nous avons utilisé le modèle LDA de topic modeling. Ce modèle permet d'estimer, par score de cohérence, le nombre optimal de topics d'évènements, définis par les termes les plus représentatifs identifiés par le modèle. Nous avons testé ensuite le modèle neuronal de type transformer BERT. On fournit en entrée du modèle une liste exhaustive d'évènements, et le modèle prédit à partir de la description de l'évènement les clusters d'évènements les plus probables par pourcentages de similarité. On constate, dans tous les cas, la nécessité d'être guidé par la connaissance métier.

Ce projet a permis d'appliquer différentes méthodes de NLP (*Natural Language Processing*) pour des données réelles d'essais, et de développer des outils exploitables et réutilisables pour d'autres types de données texte, symboliques, liées aux connaissances de l'ingénierie : les données de maintenance, les données des questions posées par les clients à Safran avec les réponses associées. Ce premier travail pourra donc être adapté à ces nouvelles données et complété en termes de méthodologies et d'outils.

Références

[1] D. M. Blei, *Probabilistic topic models*, 2012.
[2] Carson Sievert, Kenneth E. Shirley, *LDAvis: A method for visualizing and interpreting topics*, 2014.