Instanciation de relations n-Aires dans des articles scientifiques guidée par une Ressource Termino-Ontologique de domaine

Martin Lentschat

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale I2S

Unité de recherche UMR IATE – UMR TETIS

Instanciation de relations n-Aires dans des articles scientifiques guidée par une Ressource Termino-Ontologique de domaine

Présentée par Martin LENTSCHAT

Le 14 Décembre 2021

Sous la direction de Patrice BUCHE,
Juliette DIBIE-BARTHELEMY,
et Mathieu ROCHE

Devant le jury composé de

Patrice BELLOT, Professeur, Université Aix-Marseille – UMR LIS	Rapporteur
Nathalie PERNELLE, Professeure, Université Sorbonne Paris Nord – UMR LIPN	Rapportrice
Nathalie AUSSENAC-GILLES, Directrice de Recherche, CNRS – UMR IRIT	Examinatrice
Konstantin TODOROV, Maître de Conférences HDR, Univ. Montpellier – UMR LIRMM	Examinateur
Patrice BUCHE, Ingénieur de Recherche HDR, INRAE – UMR IATE	Directeur
Juliette DIBIE, Directrice de Recherche, INRAE – UMR MIA Paris	Co-Directrice
Mathieu ROCHE, Directeur de Recherche, CIRAD – UMR TETIS	Co-Directeur

UNIVERSITÉ DE MONTPELLIER

Remerciements

Réaliser une thèse est une expérience unique. On apprend à prendre le temps dans l'urgence. Un temps de remises en question nombreuses, un temps de satisfactions aussi, mais surtout un temps ou un goût pour la dégustation du thé se voit remplacé par une addiction au café.

Je tiens tout d'abord à remercier Virginie, ma mère, qui il y a longtemps lors d'une conversation à propos de mon stage de licence, m'a simplement dit : *" tu as l'air d'aimer ça. Tu as déjà pensé à faire une thèse ? "*.

Un grand merci à mes encadrants, Patrice Buche, Juliette Dibie et Mathieu Roche. Leur accompagnement par une exigence bienveillante a été essentiel. Merci également à mes rapporteurs, Nathalie Pernelle et Patrice Bellot, ainsi qu'aux membres du jury, Natahlie Aussenac-Gilles et Konstantin Todorov, qui ont accepté de me consacrer de leur temps.

Merci à mes collègues, camarades de labeurs, Jacques, Lucylle, Eric, Hugo, Marie, Vincent et tous les autres.

Merci à ma famille, qui m'a offert toutes les chances possibles dans le meilleur des environnements, et à mes sœurs, Caroline, Lola et Marilou, que j'aime.

Et enfin merci à mes amis. Les anciens de Nancy, toujours là malgré la distance, Mike, Cassandra et ma filleule chérie Léa, Chloé et Jimmy, Thomas, Morgane et leur petit Alistair, Valentin "Barbu", Sarah et Mathieu. Et les nouveaux de Montpellier, Lucie "Cannelle" ainsi que tout les membres des associations du CZLR et du PUZLE.

Once you know what the question actually is, you'll know what the answer means.

DOUGLAS ADAMS
The Hitchhiker's Guide to the Galaxy

Résumé

Cette thèse s'inscrit dans le domaine de recherche des smart data, où nous recherchons des informations spécifiques au sein de documents textuels. Elle consiste à proposer de nouvelles méthodes de représentation et d'extraction de données expérimentales à partir d'articles scientifiques. Ces méthodes ont été évaluées sur un corpus d'articles dans le domaine des emballages alimentaires.

Les données expérimentales peuvent être représentées sous forme de relations n-Aires composées d'arguments symboliques et quantitatifs. Ces derniers sont constitués d'une valeur numérique et d'une unité de mesure. L'objectif de cette thèse est de peupler une base de connaissances d'instances de relations N-Aires extraites de documents scientifiques textuels. L'approche proposée s'appuie sur une Ressource Termino-Ontologique (RTO) et se décompose en deux Phases : (1) la reconnaissance et l'extraction des instances d'arguments d'intérêt et (2) la mise en relation de ces instances dans des relations n-Aires. La Phase (1) propose une représentation originale des instances d'arguments extraites, appelée SciPuRe (Scientifique Publication Representation). Celle-ci intègre des descripteurs ontologiques, lexicaux et structurels qui décrivent le contexte d'apparition des instances d'arguments et permet de les trier selon leurs pertinences. La Phase (2) s'appuie sur les informations présentes dans les tableaux des documents, extraits automatiquement, pour guider l'extraction des relations n-Aires à partir de relations partielles, les tableaux contenant une part importante des données expérimentales dans les articles scientifiques. Ces relations partielles sont ensuite complétées par les instances d'arguments reconnues lors de la Phase (1). Trois approches sont proposées et évaluées afin d'identifier les instances d'arguments qui doivent compléter les relations : l'utilisation de la structure des documents, l'analyse des cooccurrences entre les instances d'arguments dans les textes, et enfin l'utilisation de modèles de *word-embedding* permettant de mesurer les similarités entre les instances d'arguments candidates et les arguments déjà renseignés dans les relations partielles.

Nos résultats montrent l'importance du tri des instances pertinentes à l'issue de la reconnaissance des arguments lors de la Phase (1) en s'appuyant sur les descripteurs SciPuRe. Nos expérimentations montrent que les deux critères les plus importants pour déterminer la pertinence d'une instance d'argument symbolique sont la spécificité du concept associé à l'argument dans la RTO et sa fréquence dans le document. Pour les arguments quantitatifs, c'est l'appartenance de l'instance d'argument à des sections des documents qui permet de déterminer sa pertinence. Nos expérimentations sur la Phase (2) confirment l'utilité des scores de pertinence calculés lors de la Phase (1) pour discriminer les instances. L'analyse des résultats avec différents filtrages des instances d'arguments candidates selon leurs pertinences montre un net effet positif lors du filtrage de 20% des instances avec les pertinences les plus faibles. Nous avons également expérimenté la possibilité de sélectionner plusieurs candidats pour chaque instance d'argument manquante dans une relation partielle, dans une approche d'assistance aux experts du domaine qui peuvent ensuite déterminer l'instance valide. Lors de la sélection d'un seul candidat, l'approche fondée sur les analyses des cooccurrences donne les meilleurs résultats pour détecter l'instance d'argument candidate valide. Avec une sélection plus importante, de trois ou cinq candidats, l'analyse des similarités sémantiques permise par des modèles BERT de plongement lexicaux fournit de bons résultats pour la détection d'associations entre les instances d'arguments présentes dans les relations partielles et les instances d'argument candidates à la complétion des relations. Enfin, lors de la sélection de dix candidats, les expérimentations montrent que l'approche fondée sur la structure des documents est efficace pour compléter les relations n-Aires.

Instanciation de relations n-Aires dans des articles scientifiques guidée par une Ressource Termino-Ontologique de domaine

---

Mots clés : Relations n-Aires, Extraction d'Information, Extraction de Relations, Ingénierie des Connaissances, Ressource Termino-Ontologique, Représentation de données, Mesure de Pertinence

---

Abstract

This thesis belongs to the research field of smart data, where we search for specific information within textual documents. It proposes new methods of representation and extraction of experimental data from scientific articles. These methods were evaluated on a corpus of articles in the food packaging domain.

The experimental data can be represented as n-Ary relations composed of symbolic and quantitative arguments. The latter are composed of a numerical value and a unit of measurement. The objective of this thesis is to populate a knowledge base with instances of N-Ary relations extracted from scientific textual documents. The proposed approach is based on an Ontological and Terminological Resource (OTR) and is divided into two Phases : (1) the recognition and extraction of argument instances of interest and (2) the linking of these instances in n-Ary relations. Phase (1) proposes an original representation of the extracted argument instances, called SciPuRe (Scientific Publication Representation). It integrates ontological, lexical and structural descriptors that describe the context of the argument instances and allows to sort them by their relevance. Phase (2) relies on the information present in the tables of the documents, extracted automatically, to guide the extraction of partial n-Arye relations, the tables containing an important part of the experimental data in the scientific articles. These partial relations are then completed with the argument instances recognized in Phase (1). Three approaches are proposed and evaluated in order to identify the argument instances that should complete the relations : the use of document structure, the analysis of cooccurrences between the argument instances in the texts, and finally the use of *word-embedding* models allowing to measure the similarities between the candidate argument instances and the arguments already filled in the partial relations.

Our results show the importance of sorting the relevant instances after argument recognition in Phase (1) using SciPuRe features. Our experiments show that the two most important criteria for determining the relevance of a symbolic argument instance are the specificity of the concept associated with the argument

in the OTR and its frequency in the document. For quantitative arguments, it is the apparition of the argument instance in sections of the documents that determines its relevance. Our experiments on Phase (2) confirm the usefulness of the relevance scores computed in Phase (1) to discriminate the instances. The analysis of the results with different filtering of the candidate argument instances according to their relevance shows a clear positive effect when filtering 20% of the instances with the lowest relevance. We also experimented with the possibility of selecting multiple candidates for each missing argument instance in a partial relation, in an approach to assist domain experts who can then determine the valid instance. When selecting a single candidate, the approach based on co-occurrence analyses gives the best results in detecting the valid candidate argument instance. With a larger selection of three or five candidates, semantic similarity analysis enabled by BERT word embeddings model provides good results for detecting associations between the argument instances present in partial relations and the candidate argument instances for relation completion. Finally, when selecting ten candidates, the experiments show that the approach based on document structure is effective to complete the n-Ary relations.

n-Ary relations instantiation from scientific articles driven by a domain Ontological and Terminological Resource

---

Keywords : n-Ary Relations, Information Extraction, Relation Extraction, Knowledge Engineering, Ontological and Terminological Resource, Data Representation, Relevance Measure

---

Table des matières

Liste des Figures

Liste des Tables

Chapitre 1

Introduction

Sommaire

1.1 Cadre général

L'information constitue un enjeu majeur dans les sociétés du 21e siècle. L'accès à celle-ci est de plus en plus facilité par l'essor des moyens de communication. Cela passe principalement par le Web, sur lequel le nombre de sites ne cesse d'augmenter tout comme leurs contenus. Les journaux d'information en ligne, les blogs et réseaux sociaux, les bibliothèques et revues scientifiques constituent des sources d'informations qui peuvent profiter à tous. Cette quantité d'information disponible a fait émerger le besoin de gérer celle-ci en l'extrayant, en la catégorisant et en la structurant. Lorsque l'on souhaite déployer, dans ce cadre, des solutions automatiques, cela demande de passer par la donnée (i.e. le symbole brut) pour construire de l'information (i.e. de la donnée mise en contexte, le *'qui', 'quoi', 'où', 'quand'*) et ainsi permettre de créer de la connaissance (i.e. de l'information performative, ouvrant des applications, le *'comment'*) [Bellinger et al., 2004, Rowley, 2007].

Dans cette thèse nous nous intéressons aux connaissances scientifiques disponibles dans les publications des revues scientifiques en ligne. Tout comme les autres contenus sur le Web, le nombre d'articles publiés chaque jour augmente

constamment. La majorité des journaux scientifiques mettent à disposition en ligne, gratuitement ou non, les articles publiés. Cela représente une quantité de données considérable, à portée de main de la plupart des scientifiques et citoyens. Cette profusion de publications est une opportunité pour extraire et valoriser les connaissances disponibles dans les articles scientifiques.

Cette problématique concerne plusieurs domaines de recherche, dont le *Traitement Automatique du Langage* (TAL) et l'*Ingénierie des Connaissances* (IC). Le domaine du TAL s'intéresse, parmi diverses problématiques, à l'extraction des informations présentes dans des documents textuels, dont les articles scientifiques. Pour parvenir à cela, le TAL doit considérer le langage naturel (i.e. non structuré et formalisé) et identifier l'information pertinente en réussissant à identifier celle-ci sous toutes ses formes (e.g. variation terminologique, présence implicite) tout en filtrant les formes textuelles similaires ne correspondant pas à cette information (e.g. polysémie d'un terme, ambiguïté). Le TAL s'intéresse également à l'extraction d'informations dans des domaines à haute valeur ajoutée, typiquement comme les publications réalisées dans différents domaines scientifiques. En effet extraire automatiquement les connaissances présentes dans des articles permet de valoriser les connaissances qui y sont décrites. Le langage employé par les scientifiques s'éloigne du langage naturel car, sans être totalement formel, les informations y sont décrites en répondant à des normes et des impératifs propres aux publications scientifiques (e.g. précision et concision, standards de notations et de publications). En général, les données y sont décrites dans des sections spécifiques (e.g. *'Materials and Methods'*, *'Results and Discussion'*), présentées avec des termes propres au domaine étudié et rares dans le langage naturel (e.g. formule chimique d'une molécule, unités de mesure complexes) ou regroupées dans des tableaux qui structurent l'information.

Le domaine de recherche en IC a pour objectif de structurer et d'exploiter les connaissances. Les travaux en IC viennent généralement à la suite de l'extraction d'information et travaillent notamment à permettre la transformation de cette information en connaissance. Ces travaux vont par exemple structurer les informations dans des ontologies et les exploiter ensuite dans des systèmes de raisonnement. Une ontologie est une modélisation de la sémantique d'un domaine, celle-ci structure alors les concepts de ce domaine, les termes qui les représentent ainsi que les relations qu'ils entretiennent. Les domaines de recherche

du TAL et de l'IC se nourrissent mutuellement. Des techniques développées en TAL permettent par exemple d'assister la construction d'ontologies et leur maintenance. Les ontologies peuvent en retour être utiles à l'automatisation de l'extraction d'information en fournissant un formalisme sémantique des concepts d'un domaine et un vocabulaire. Les informations extraites par le TAL puis structurées dans des ontologies sont ainsi partageables, interrogeables et exploitables sur le Web. Appliquées sur des publications de recherche, ces différentes étapes permettent de valoriser les informations contenues dans les articles scientifiques.

1.2 Problématiques identifiées

Le travail de thèse présenté ici se place dans le cadre de l'extraction et de la valorisation des connaissances scientifiques. Le développement de solutions permettant d'accéder directement aux informations disponibles au sein même des publications scientifiques a en effet de nombreuses applications. Une telle application consiste par exemple à assister un expert réalisant une méta-analyse sur une problématique de recherche (i.e. analyse systématique des résultats de nombreuses études permettant de tirer des conclusions générales et faire état du consensus scientifique) en automatisant une partie du processus de recherche et d'analyse de l'information. Une autre application est d'alimenter automatiquement des bases de connaissances, elles-mêmes pouvant ensuite être consultées, interrogées ou utilisées comme ressources pour des systèmes d'aide à la décision [Guillard et al., 2015, Guillard et al., 2017, Lousteau-Cazalet et al., 2016a].

Notre travail se consacre à l'extraction de données expérimentales provenant d'articles scientifiques et leur structuration sous forme d'instances de relations n-Aires (i.e. une relation comportant n arguments). Cette représentation sous forme de relations n-Aires est utile pour représenter les informations mettant en jeu un nombre important de données. Chacun des n objets de l'information constitue alors l'un des n arguments regroupés dans la relation n-Aire. A titre d'exemple, la Figure 1 présente une information décrivant une mesure de perméabilité, contenue dans une base de connaissance, et représentée sous la forme d'une instance de relation n-Aire : *O2_Permeability_Relation*, comprenant six arguments : *Packaging*, *Partial_Pressure*, *Thickness*, *Temperature*, *Relative_Humidity* et *O2_Permeability*. La Section 3.1.1 présente plus en détails les caractéristiques

et l'utilisation des relations n-Aires.

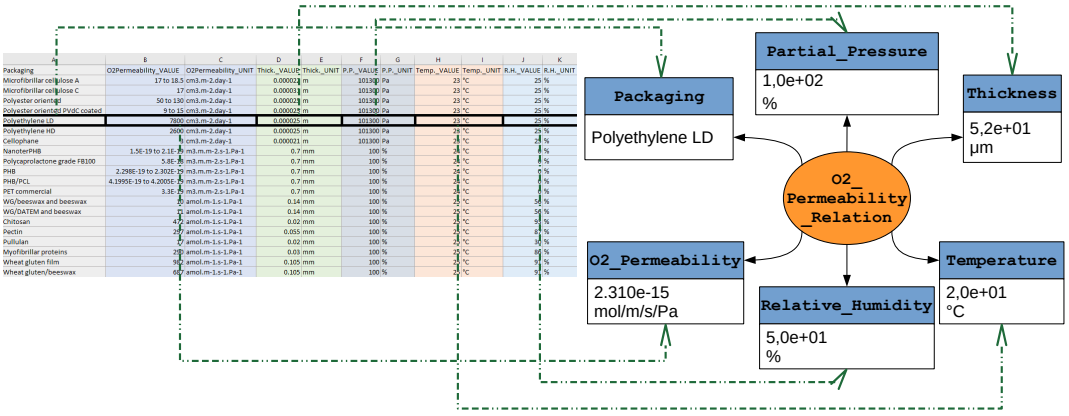


FIGURE 1 – Représentation d'une relation n-Aire depuis une base de connaissances

Plus spécifiquement, nous appliquons la méthode proposée dans cette thèse au domaine des emballages alimentaires, en portant une attention particulière aux données de perméabilité des emballages (à l'oxygène, au dioxyde de carbone et à l'eau) qui peut varier en fonction des proportions de matériaux utilisés dans la conception de l'emballage. Ces informations sont issues d'expérimentations restituées dans des articles scientifiques. Les principales difficultés d'extraction des informations dans ce domaine, partagées avec les autres domaines scientifiques, viennent de leur spécificité (i.e. vocabulaire spécifique au domaine, vocabulaire spécifique à un article, variations terminologiques inter-articles) ainsi que leur dispersion (i.e. les différentes données relatives à une même information sont présentes dans des sections différentes des articles). La multiplicité des informations (i.e. plusieurs expérimentations par article, présence d'informations et de données issues d'autres articles) ainsi que leurs différents formats (termes et valeurs numériques, présence dans le texte ou dans des tableaux) constituent également un obstacle. Malgré ces difficultés, ces informations sont importantes à extraire et à structurer afin de valoriser les connaissances scientifiques (e.g. méta-analyses, utilisations par des outils de raisonnement ou d'aide à la décision). La Figure 2 illustre cette dispersion des données dans les documents ainsi que les différents formats de données présents dans un article scientifique.

tein isolate films were lower than those of other films. The degree of resistance of various lipid films to the oxygen transmission was measured with the aim of determining the influences of polymorphic forms of the lipid and of tempering, using the method of ASTM by Kester and Fennema (1989a, 1989b, 1989c). Greener and Fennema (1989) reported oxygen permeabilities of bilayer films prepared from methyl cellulose and beeswax measured by the same method. Another work on the oxygen transmission of a methyl cellulose-palmitic acid film was reported by Rico-Pena and Torres (1990).

In a recent work from our laboratory, a simple method was proposed for the measurement of CO_2 permeability (Ayranci, Tunc, & Etci, 1999) and applied successfully to various cellulose-based edible films (Ayranci & Tunc, 2001).

In the present work, we introduce a method to measure the OP of edible films. The proposed method was applied to measure the OP of films of various composition with the aim of finding the optimum composition to minimise oxidative degradation of foods. Application of film solutions to fresh foods, such as mushrooms and cauliflower, and analysis of these foods for oxidative degradation, are also investigated.

2. Materials and methods

2.1. Materials

MC with an average molecular weight of 41 000, polyethylene glycol (PEG) with an average molecular weight of 400 and stearic acid (SA) were purchased from Sigma. Citric acid (CA), manganese(II) sulphate monohydrate, potassium iodide, sodium thiosulphate and sodium hydroxide were obtained from Merck. Oxygen gas was obtained from a local gas supplier,

PEG, the solution was homogenised with an homogeniser at 24 000 rev min^{-1} for 5 min. It was re-homogenised after the addition of fatty acid and other additives. The final solution was kept in a vacuum oven at 80 °C for about 5 h in order to remove air bubbles or dissolved air. It was then spread on 20×20 cm glass plates by adjusting the hand-operated thin layer chromatography plate coater to 0.5 mm thickness. The spread films were dried at 60 °C in an oven for 25 min and then at room temperature for 1 day.

2.3. Measurement of the OP

An edible film was sealed between two specially designed glass cups, each having a diameter of 4 cm and a depth of 5 cm (Fig. 1). Both cups have two channels. Oxygen enters to the cup on one side of the film from one channel and leaves from the other with a controlled flow rate to keep the oxygen pressure constant in that compartment. The cup on the other side of the film was purged by a stream of nitrogen entering from one channel and leaving from the other. This nitrogen acted as a carrier for oxygen permeated from the other side of the film to the wet analysis system (Fig. 1). Up to the wet analysis system, the design was mainly based on the ASTM standard (1988). A modification was made to the O_2 analysis and a classical wet analysis, based on the well-known method of iodimetry was applied for the determination of the amount of permeated O_2, instead of using an oxygen analyser as in the ASTM standard, which uses a coulometric sensor.

The mixture of N_2 and permeated O_2 was passed from the wet system for a known period of time. Then the wet system, which originally contained aqueous manganese(II) sulphate and alkaline iodide solution, was analysed for O_2 (Vogel, 1989). It is well known that manganese(II) hydroxide is rapidly and quantitatively

determined using a calibration curve prepared with gallic acid, and the results reported as mg/l.

3. Results and discussion

3.1. The effect of antioxidants on the OP of edible films containing SA

Fatty acids, such as SA, LA and PA, being edible and having hydrophobic character, are used in coating formulations as water vapour barrier materials. In previous work we had found SA to be more effective than LA and PA for decreasing WVP of cellulose-based edible films (Ayranci & Tunc, 1997, 2001). Therefore, it was of interest to see how the OP of these films was affected by the SA content. The OP values of MC-based edible films, containing varying amounts of SA in their composition, were determined by the method developed in the present work, as described earlier, and are given in Table 2, together with film thickness values.

The general trend is that the OP increases with increasing SA content of the film. This may be attributed to the formation of holes in the crystal structure of edible films as the SA content increases. These holes, which are especially formed above 15 g SA/100 g MC,

Table 2
The SA content, the thickness and the OP values of edible films at 25 °C and 0% RH

SA content g (100 g MC)$^{-1}$	Thickness 10^5 m	OP 10^9 g d^{-1} Pa^{-1} m^{-1}
0.0	1.86±0.00	6.8±0.4
5.0	1.93±0.03	5.2±0.2
15	2.10±0.01	7.7±0.9
25	1.88±0.04	8.6±0.3
40	2.00±0.00	14±1

It is clear from Table 3 that OP values of films decrease with both AA and CA contents. The only exception to this trend is at 16.7 g CA/100 g MC content. The OP values of this film were found to be slightly larger than that of the film with 3.33 g CA/100 g MC. The two antioxidants show similar effects in improving the oxygen barrier property of the films.

3.2. The effects of coating on water loss of fresh foods

The water loss of mushrooms, with coatings of varying composition, given in Table 1, and of uncoated ones, as a function of time, are shown in Fig. 2. In the coating formulations, an intermediate SA content of 20 g/100 g MC (which is equivalent to 0.6 g/3 g MC) and the highest examined CA or AA content of 16.7 g/100 g MC (which is equivalent to 0.5 g/3 g MC) were maintained according to the results presented above in Section 3.1. The % water losses of uncoated mushrooms are 3.86, 14.7 and 19.7 at the end of first, third and fifth days, respectively. Mushrooms with coatings of varying

Table 3
The antioxidant content, the thickness and the OP values of edible films containing 20 g SA/100 g MC at 25 °C and 0% RH

Antioxidant content g (100 g MC)$^{-1}$	Thickness 10^5 m	OP 10^9 g d^{-1} Pa^{-1} m^{-1}
AA		
0.33	1.9±0.2	8.3±0.2
1.67	1.87±0.03	6.5±0.1
3.33	1.8±0.0	5.8±0.2
16.67	1.80±0.02	4.5±0.2
CA		
0.33	1.68±0.0	6.4±0.3
1.67	1.57±0.03	5.39±0.03
3.33	1.49±0.01	3.9±0.2
16.67	1.62±0.02	4.7±0.2

contextes contenants des instances d'arguments

instances de relations n-Aires partielles dans des tableaux

FIGURE 2 – Extraits de Ayranci, Erol & Tunc, Sibel. (2003). *A method for the measurement of the oxygen permeability and the development of edible films to reduce the rate of oxidative reactions in fresh foods.* Food Chemistry. 80. 423-431.

Nous posons alors notre question de recherche de la manière suivante :

Quelles représentations multi-descripteurs adopter, et stratégies d'exploitation associées, pour extraire des instances de relations n-Aires à partir d'articles scientifiques ?

Dans le cadre des informations expérimentales relatives au domaine des emballages alimentaires, de précédents travaux ont déjà été réalisés sur l'extraction d'instances d'arguments corrélés de relations n-Aires [Berrahou, 2015, Berrahou et al., 2017] en utilisant l'ontologie Transmat de @Web [1] [Buche et al., 2011]. @Web est une plateforme, développée par l'équipe ICO (Ingénierie des COnnaissances - UMR IATE - INRAE - Montpellier), regroupant et structurant des connaissances issues d'articles scientifiques en domaine expérimentaux. Il regroupe aujourd'hui 332488 données, structurées dans 171 relations n-Aires différentes, elles-mêmes représentées dans 7 ontologies couvrant 8 domaines. L'ontologie Transmat est formalisée en tant que Ressource Termino-Ontologique (RTO) [Aussenac-Gilles et al., 2006, Reymonet et al., 2007, Tissaoui et al., 2013], qui modélise un domaine de connaissance à travers ses concepts et les relations entre ceux-ci tout en dissociant la composante sémantique (i.e. concepts et relations) de la composante linguistique utilisée (i.e. les termes représentant concepts et relations). Par exemple, dans la RTO Transmat, le concept H2O_Permeability possède une composante terminologique comportant les termes *'H2O permeability'* ou *'water permeability'*. Les données expérimentales sont alors représentées dans cette RTO sous la forme d'instances de relations n-Aires, comme illustrées dans la Figure 1.

Dans les articles des domaines expérimentaux, les auteurs décrivent les objets de leurs études (e.g. emballages, biomasses), la méthode expérimentale, ses conditions et enfin les résultats des mesures effectuées. Ces informations sont présentes dans les textes des articles sous forme de termes (i.e. mot ou suite de

1. `https://ico.iate.inra.fr/atWeb/-Septembre2021`

mots) ainsi que de valeurs numériques et leurs unités de mesure. Ces informations textuelles sont majoritairement non structurées, exprimées dans les phrases et servant le discours déployé dans l'article. Les articles présentent également des tableaux regroupant et/ou structurant certaines informations importantes ou spécifiques. Cette dispersion des informations implique que les instances d'arguments (e.g. données telles que le nom d'un emballage ou son épaisseur) d'une même relation n-Aire (e.g. l'information présentant un résultat de mesure de perméabilité) ne seront pas présents dans la même phrase ou paragraphe. De la même manière, les informations structurées dans les tableaux ne concernent que quelques arguments de la relation (e.g. une valeur de perméabilité est donnée conjointement à son emballage mais sans les paramètres de contrôle de la mesure). Les articles scientifiques présentent également différentes relations n-Aires (e.g. des mesures de perméabilités à différents gaz, oxygène, vapeur, dioxyde de carbone) et chacune de ces relations peut posséder plusieurs instances (e.g. des mesures de perméabilité à l'oxygène de plusieurs emballages différents). Un grand nombre d'instances d'arguments peut donc être détecté sans que les liens de relations entre celles-ci ne soient nécessairement explicites, dû à la dispersion des instances d'arguments dans les documents et à la multiplicité des instances d'arguments et de relations.

La forme d'apparition des instances d'arguments dans les textes des documents nécessite également d'être prise en compte. En effet, l'utilisation d'une RTO fournit pour chacun des concepts pouvant constituer des instances d'arguments de relations n-Aires (i.e. concepts d'intérêts) un vocabulaire pouvant être employé pour détecter ses apparitions dans les textes. Cependant la couverture terminologique de la RTO ne peut pas être complète, car chaque article présente ses propres sujets de recherche (e.g. le nom d'un nouvel emballage) et ne respecte pas toujours les conventions du domaine (e.g. l'exposant et l'ordre des unités dans les unités de mesure varient). Il est important d'être capable de capturer l'ensemble de ces variations, car celles-ci représentent des instances des arguments pour les relations n-Aires recherchées. Il existe également un grand nombre d'informations pouvant être confondues avec des instances d'arguments d'intérêts. Ces confusions peuvent apparaître dans plusieurs cas : une graphie similaire à celles d'arguments recherchés (e.g. toutes les températures trouvées dans l'article, comme $25°C$, ne sont pas nécessairement celles utilisées lors de la mesure de perméabilité d'un emballage), une information issue d'une autre publication (e.g. des résultats issus

d'une autre publication cités à titre de comparaison) ou encore simplement des informations si peu spécifiques que peu dignes d'intérêt (e.g. le terme *'packaging'* est bien une instance d'un des concepts sous l'argument Packaging, mais celle-ci est si peu spécifique qu'elle ne présente pas de valeur informative). Écarter ces données ne constituant pas des instances d'arguments de relations n-Aires est essentiel afin de ne pas risquer d'extraire des informations sans réelle valeur. La Figure 2 illustre les formes et contextes d'apparition des instances d'arguments de relations n-Aires ainsi que les instances de relations n-Aires partielles présentent dans les tableaux des documents.

Les principaux verrous identifiés dans l'extraction de relations n-Aires dans les articles scientifiques sont donc :

Verrou 1. La *dispersion* des instances d'arguments des relations n-Aires implique un travail de mise en relation de celles-ci dans des instances de relations. En effet la distance entre les instances d'arguments pouvant appartenir à une même instance de relation rend inadaptée l'utilisation de certaines méthodes classiques employées pour l'extraction de relations textuelles [Bach and Badaskar, 2007, Chan and Roth, 2011, Bunescu and Mooney, 2005, Pawar et al., 2017] qui opèrent au niveau de la phrase ou du paragraphe. Nous observons par exemple dans La Figure 2 que les composants des emballages sont décrits dans *2.1 Materials* tandis que leurs différentes proportions le sont dans *3.1 The effect of antioxidants ...* et *3.2 The effects of coating ....* Les résultats des mesures de perméabilité à l'oxygène sont donnés dans la *Table 3* et incluent également les mesures d'épaisseur des emballages tandis que les paramètres de contrôle sont présents dans l'entête du tableau et structurées différemment. La méthode de mesure de ces perméabilités est nommée et présentée en *2.3 Measurement of OP*.

Verrou 2. L'*hétérogénéité* des formats de données dans les articles scientifiques représente un verrou à considérer. Les informations d'intérêt sont présentes dans les textes des documents mais également structurées dans les tableaux. Cela correspond à plusieurs contextes d'apparition de l'information, nécessitant l'utilisation de méthodes d'extraction adaptées ainsi que la capacité de comparer et relier les données apparaissant dans des contextes différents. Il est également nécessaire de détecter plusieurs types d'informations : symboliques (e.g. noms des emballages, des composants

et de la méthode utilisée) et quantitatives (e.g. valeur de perméabilité, de proportions et différents paramètres de contrôle).

Verrou 3. Les *variations* lexicales des instances d'argument et des unités de mesures dans les articles sont également importantes. Lever ce verrou est essentiel pour garantir une meilleure exhaustivité des informations extraites.

Verrou 4. La *pertinence* des instances d'arguments reconnues est importante pour limiter le bruit contenu dans les résultats. En effet reconnaître, sans analyse de pertinence, une instance d'argument constituant un résultat faux-positif risque de transmettre celle-ci dans une instance de relation n-Aire.

Verrou 5. La *multiplicité* des instances de relations présentes dans les documents implique enfin de devoir associer chaque instance d'argument à l'instance de relation appropriée. Associer une instance d'argument à une autre relation serait en effet une erreur.

1.3 Définitions et hypothèses de travail

L'accès à des sources de données de plus en plus volumineuses a donné naissance au domaine du *big data* [De Mauro et al., 2016]. Ce domaine de recherche désigne les approches s'attaquant à la gestion de données en grande quantité, variées, non structurées et présentant potentiellement des redondances. Les travaux dans ce domaine se consacrent majoritairement à l'extraction de grandes tendances obtenues par le recoupement des quantités de données analysées. Au contraire de ces approches, notre travail se concentre principalement sur la valeur individuelle de chacune des données. Nos travaux relèvent alors du domaine des *smart data* [Marcia, 2017, Duong et al., 2017], où l'analyse du contexte dans lequel apparaît chacune des données est nécessaire pour en tirer l'information recherchée. Cette approche a pour but de considérer la valeur individuelle de chaque donnée, de manière indépendante tout comme dans l'ensemble des données extraites. En nous appuyant sur les critères caractérisant les smart data [Marcia, 2017, Duong et al., 2017], nous définissons ce domaine de recherche comme répondant à la Définition 1. Dans ce contexte, l'Hypothèse 1 stipule que la bonne manipulation des données détectées dans les textes repose sur la détection d'un ensemble de descripteurs permettant de faire ressortir la valeur individuelle de chacune.

Définition 1 *Le smart data est une façon d'organiser et de sémantiser les données de différentes natures en renseignant leurs contextes et pertinence afin d'être en capacité de les manipuler à toutes les échelles, de la donnée unique et particulière aux ensembles de données hétérogènes.*

Hypothèse 1 *Dans une approche smart data, un ensemble de descripteurs permet de représenter le contexte d'apparition des instances d'arguments et des instances de relations n-Aires. La manipulation des instances d'arguments dans le texte des documents, tout comme leurs manipulations afin de former des instances de relations n-Aires partielles ou complètes, est rendue possible par un ensemble de descripteurs reposant sur différents critères et rassemblés dans une unique représentation des instances d'arguments.*

La méthode proposée pour l'extraction des instances d'argument produit un grand nombre de résultats et un tri des données pertinentes est alors nécessaire. Dans les travaux d'extraction de l'information, la pertinence est usuellement entendue comme la capacité d'un résultat à satisfaire les besoins en information d'un utilisateur [Cooper, 1971]. Dans notre travail, nous entendons la satisfaction de ce besoin d'information comme la capacité d'une instance d'argument à faire partie d'une instance de relation n-Aire d'intérêt, relation qui contient un argument résultat (e.g. la relation de perméabilité doit contenir la valeur de perméabilité), selon notre Définition 2. L'Hypothèse 2 stipule que l'exploitation des descripteurs d'une représentation des instances d'argument permet l'estimation de leur pertinence.

Définition 2 *Une instance d'argument est considérée comme pertinente dès lors qu'elle est en mesure de prendre part à une instance de relation n-Aire recherchée par l'utilisateur.*

Hypothèse 2 *Il est possible de déterminer un ensemble de descripteurs du contexte d'apparition des instances d'arguments et des instances de relations n-Aires qui facilite l'extraction des instances pertinentes*

Notre travail de thèse s'appuie ainsi sur les Définitions 1 et 2 afin de proposer une méthode générique d'extraction des connaissances depuis des articles

scientifiques. Suivant les Hypothèses 1 et 2, notre approche se fonde principalement sur l'identification et l'exploitation de descripteurs aptes à décrire chacune des données extraites et permettre de les manipuler afin d'identifier les données pertinentes et reconstituer les instances de relations n-Aires recherchées.

1.4 Approche et Contributions de la thèse

Nous avons conçu l'ensemble de notre méthode d'extraction des connaissances depuis des articles scientifiques en domaine expérimental dans une démarche générique. Notre méthode exploite la définition des relations n-Aires ainsi que la composante terminologique d'une RTO afin de guider le processus d'extraction des connaissances. Une RTO modélise les relations n-Aires dans une *core-ontology* et les concepts, ainsi que leurs terminologies, propres à chaque domaine expérimental dans des *domain-ontology*. Cette core-ontology, et donc la modélisation des relations n-Aires, étant partagée par les différentes domain-ontology, cela garantit la généricité de notre approche. Notre processus d'extraction est ici appliqué au domaine des emballages alimentaires, mais se veut donc applicable à tout domaine expérimental pour lequel une RTO de domaine existe. Nous utilisons ici la RTO Transmat, disponible sur la plateforme @Web [Buche et al., 2011], qui représente les informations dans le domaine du transfert de matières dans les emballages sous forme de relations n-Aires. Plus de détails sur les RTO sont disponibles dans la Section 2.1 et un exemple d'extraction dans un autre domaine est donné en Section 2.6.4.

Dans le cadre du smart data, il est nécessaire de concevoir une représentation riche des données, multi-descripteurs et contextualisée. Pour cela, les représentations utilisées en recherche d'information reposent généralement sur des critères lexicaux (e.g. représentations en sac de mots) et ne sont pas assez riches. Au contraire, d'autres approches déployées en utilisant des techniques de *deep learning* résultent en des représentations riches mais difficilement interprétables ou explicables [**?**]. Dans le cadre de cette thèse, nous proposons une représentation, nommée SciPuRe (Scientific Publication Representation) [Lentschat et al., 2020, Lentschat et al., 2021b], des instances d'arguments reconnues dans les documents ainsi qu'une représentation, appelée STaRe (Scientific Table Representation), des instances de relations n-Aires partielles détectées dans les tableaux des documents.

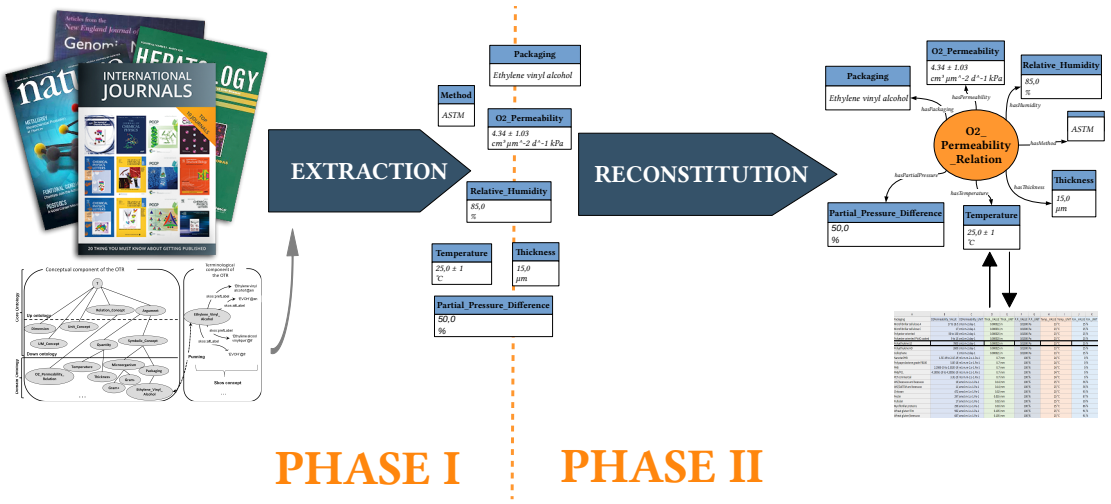


FIGURE 3 – Extraction en deux phases de relations n-Aires

Celles-ci décrivent les instances d'arguments et de relations n-Aires selon différents critères qui permettent ensuite d'analyser chaque instance d'argument individuellement. Les descripteurs de la représentation SciPuRe sont également utiles afin d'estimer la pertinence des instances d'arguments reconnues. Les descripteurs de SciPuRe et de STaRe offrent également un ensemble de critères riches et utiles pour la reconstitution des instances de relations n-Aires.

Pour lever le premier verrou (i.e. la dispersion des instances d'arguments dans un texte), nous adoptons une stratégie en deux phases, tel que illustré dans la Figure 3. La Phase I sera consacrée à l'identification des instances d'arguments dans les documents et la Phase II les mettra en relation dans des instances de relations n-Aires. Cette stratégie est une approche classique dans les travaux d'extraction des relations n-Aires [Zhou et al., 2014, Berrahou et al., 2017].

Pour lever le deuxième verrou (i.e. l'hétérogénéité des formats de données), nous proposons la représentation SciPuRe des instances d'arguments et la représentation STaRe qui représente les relations n-Aires partielles présentes dans les tableaux des articles. Ces représentations permettent de traiter l'hétérogénéité des formats de données en élaborant des descripteurs communs aux instances d'arguments et aux instances de relations partielles afin de procéder à la reconstitution des instances des relations n-Aires lors de la Phase II.

Pour lever le troisième verrou (i.e. les variations des instances d'arguments dans le texte), nous utilisons des méthodes existantes de reconnaissance des variations

terminologiques [Bourigault and Jacquemin, 1999], de reconnaissance de nouvelles unités de mesure [Berrahou, 2015] et une méthode proposée dans cette thèse pour l'extraction des acronymes. Nous intégrons l'ensemble de ces méthodes dans la construction de nos représentations SciPuRe et STaRe.

Afin de participer à la levée des quatrième et cinquième verrous identifiés (i.e. besoin de mesurer la pertinence des instances d'arguments reconnues puis associer chaque instance d'argument à l'instance de relation qui lui correspond), cette thèse propose trois contributions principales et originales au regard de l'état de l'art de la reconnaissance des instances de relations n-Aires à l'échelle des documents.

La première contribution exploite une représentation multi-descripteurs des instances d'arguments pour les relations n-Aires. Nommée *Scientific Publication Representation* (SciPuRe), celle-ci emploie des descripteurs ontologiques, lexicaux et structurels afin de caractériser les instances d'arguments extraites. La représentation SciPuRe [Lentschat et al., 2020, Lentschat et al., 2021b] exploite les informations pouvant être récoltées tout au long des étapes de l'extraction des instances d'arguments. Cette représentation SciPuRe nous permet de proposer des scores de pertinence des instances d'arguments extraites durant la Phase I. Ces mesures exploitent les descripteurs de la représentation SciPuRe afin de concevoir des scores de pertinence lexicaux et sémantiques. Ceux-ci permettent d'ordonner les instances d'arguments extraites et écarter par seuillage une partie des faux-positifs de la collection des instances d'arguments extraites à l'issue de la Phase I. Ces scores de pertinence sont adaptés aux différents types d'instances d'arguments recherchées (e.g. instances symboliques ou quantitatives) et combinés dans une approche multi-critères.

Notre seconde contribution porte sur la reconstitution des relations n-Aires. Notre Phase II débute par l'extraction automatique, guidée par la RTO, des relations n-Aires partielles présentes dans les tableaux des articles afin de prendre en considération ces informations structurées. En effet, les auteurs des publications scientifiques présentent de manière synthétique, sous forme de tableaux, les résultats d'expérimentation présentés dans l'article. Les informations contenues dans ces tableaux sont donc généralement importantes, concernant l'objet principal des articles. La représentation SciPuRe des instances d'arguments est étendue en Phase II pour représenter ces instances de relations n-Aires partielles extraites des tableaux présents dans les articles. Cette représentation originale, nommée

Scientific Table Representation (STaRe), spécifie la nature et la structure de la relation n-Aire et situe l'instance de relation partielle dans son document. STaRe comprend également les descripteurs ontologiques et lexicaux des instances d'argument composant la relation. Nous recherchons alors à compléter les instances de relations partielles issues des tableaux avec des instances pertinentes d'arguments identifiées dans le texte. Nous avons pour cela évalué les apports de méthodes fondées sur l'utilisation de la structure des documents, les associations de cooccurrences fréquentes entre instances d'arguments et leurs liens de similarités donnés par des modèles de langage de type word-embedding. Dans une démarche d'assistance à l'annotation d'instances de relations n-Aires par des experts, nous évaluons également l'impact de proposer plusieurs candidats pour chaque argument manquant dans les relations à compléter. L'ensemble de nos travaux se situe ainsi dans une démarche globale proposant un pipeline complet, du corpus d'articles scientifiques possédant une RTO du domaine aux des instances de relations n-Aires extraites dans une optique d'ajout de celles-ci à des bases de connaissances.

Enfin, la troisième contribution consiste en l'élaboration de trois corpus, et des guides d'annotation associés, utilisés pour évaluer les méthodes d'extraction proposées dans cette thèse, et mis à disposition des communautés de recherche en TAL et en IC. L'un des corpus a été obtenu par un processus d'annotation manuelle des instances d'arguments reconnues par trois annotateurs dans 50 articles scientifiques du domaine des emballages alimentaires [Lentschat et al., 2021a]. Il constitue le Gold Standard [Lentschat et al., 2021c] utilisé pour l'évaluation de l'extraction des instances d'arguments symboliques et quantitatifs concernant la composition des emballages alimentaires et leurs perméabilités au gaz. Un corpus présente les relations n-Aires partielles issues des tableaux des articles [Lentschat et al., 2021d]. Il comporte les résultats d'une annotation manuelle de 31 tableaux issus de 10 articles ainsi que les résultats d'une annotation automatique suivant une méthode de l'état de l'art [Buche et al., 2011, Hignette et al., 2009]. Un dernier corpus, utilisé comme Gold Standard [Lentschat, 2021] pour évaluer nos approches de reconstitution des relations n-Aires, est composé des relations n-Aires partielles issues des tableaux complétées manuellement avec les instances d'arguments présents dans les textes des documents. Ces corpus permettent de contribuer à la création de ressources pour l'extraction de relations n-Aires dans les domaines scientifiques expérimentaux et l'évaluation des méthodes déployées. Il existe peu de corpus annotés avec des relations n-Aires à disposition, et cela est

particulièrement marqué lorsque l'on recherche des articles scientifiques annotés au niveau du document entier [Zhou et al., 2014]. Le domaine des emballages alimentaires n'est pas un domaine déjà couvert par d'autres Gold Standards. De manière générale, des corpus riches pouvant être employés comme standards pour l'évaluation et la comparaison des approches conçues pour l'extraction de relations n-Aires en domaine de spécialité doivent encore être construits. De plus, les corpus créés dans le cadre de cette thèse peuvent permettre l'évaluation de méthodes proposées dans différents travaux (e.g. reconnaissance d'unités de mesure, extraction automatique de données dans les tableaux, identification d'entités nommées en domaine de spécialité).

1.5 Organisation du manuscrit

Ce manuscrit de thèse est organisé autour de deux chapitres reprenant les deux phases d'extraction des relations n-Aires :

I. Le Chapitre 2 présente la Phase I de notre méthode consistant en l'extraction des instances d'arguments dans les documents. La Section 2.1 présente les ressources termino-ontologiques, leurs caractéristiques et utilisations, ainsi que la RTO Transmat que nous utilisons dans notre domaine d'application. La Section 2.2 dresse un état de l'art des méthodes utilisées en extraction d'informations et plus généralement celles pouvant être transposées à l'extraction des instances d'arguments. La Section 2.3 présente notre méthode d'extraction d'instances d'arguments de relations n-Aires guidée par une RTO. Cela comprend les prétraitements étendant la couverture de la composante terminologique de la RTO en Section 2.3.1 et les étapes d'identification et de désambiguïsation des instances d'arguments en Section 2.3.2. La représentation SciPuRe est présentée en Section 2.3.3. Cette représentation permet de regrouper, durant le processus d'extraction, des descripteurs ontologiques, lexicaux et structurels pour les instances d'arguments extraites afin de les différencier, d'identifier leurs similarités et minimiser leurs ambiguïtés. La Section 2.4 propose des scores estimant la pertinence des instances d'arguments extraites. Ceux-ci sont fondés sur les descripteurs de la représentation SciPuRe et des méthodes combinant ces scores sont également étudiées. La Section 2.5 présente le Gold Standard

conçu pour évaluer notre travail d'extraction des instances d'arguments. Enfin, les résultats de notre méthode d'extraction des instances d'arguments sont présentés en Section 2.6, notamment par l'évaluation des scores de pertinence lexicaux et sémantiques, seuls et en combinaison.

II. Le Chapitre 3 présente la Phase II de notre méthode consistant à compléter les instances de relations partielles issues des tableaux des articles avec les instances d'arguments extraites précédemment. La Section 3.1 dresse un état de l'art de l'extraction des relations n-Aires et des familles de méthodes utilisées. La Section 3.2 décrit notre approche de complétion des relations n-Aires basée sur différents critères. Celle-ci débute par une sélection des instances candidates pour un argument manquant dans une relation (Section 3.2.3) ainsi qu'un dédoublonnage fusionnant les instances redondantes (Section 3.2.4). Les méthodes de discrimination des instances candidates sont ensuite présentées en Section 3.2.5, celles-ci s'appuyant sur des critères Structurels, Fréquentistes et par Plongements Lexicaux. La Section 3.3 décrit la méthode à l'état de l'art utilisée pour l'extraction automatique des relations partielles présentes dans les tableaux (Section 3.3.1) et le Gold Standard des relations n-Aires (Section 3.3.2). Enfin, les résultats de notre méthode de complétion des relations n-Aires partielles sont examinés en Section 3.4. Cette dernière partie analyse les apports des approches Structurelles, Fréquentistes et par Plongements Lexicaux (Sections 3.4.1, 3.4.2 et 3.4.3) conjointement aux effets d'une sélection plus large du nombre d'instances candidates proposées pour compléter un argument manquant dans une instance de relation n-Aire.

Chapitre 2

Extraction d'Instances d'Arguments dans les Articles Scientifiques

Sommaire

Nous présentons ici la première Phase de notre processus d'extraction de connaissances expérimentales depuis des articles scientifiques et leur structuration sous forme de relations n-Aires. Cette Phase I se consacre à la reconnaissance et à l'extraction des instances d'arguments présentes dans les textes. Pour cela, notre méthode effectue différents pré-traitements dans les documents avant de rechercher les variations terminologiques permettant de reconnaître le plus grand nombre possible d'instances d'arguments. Ces instances d'arguments sont ensuite extraites et une représentation, *Scientific Publication Representation* (*SciPuRe*) incluant des descripteurs ontologiques, lexicaux et structurels, est associée à chacune. Enfin, nous avons conçu une méthode pour évaluer la pertinence des instances d'arguments extraites. Cette méthode repose sur l'utilisation de leurs représentations SciPuRe. La méthode propose des scores utilisés pour ordonner les instances d'arguments selon leurs pertinences, afin de trier les faux positifs obtenus.

2.1 Ressource Termino-ontologique

Dans cette sous-section nous présentons l'usage des Ressources Termino-Ontologiques (RTO) dans les domaines de l'ingénierie des connaissances et de l'extraction d'information. Une RTO est une ressource qui permet de représenter conjointement les concepts d'un domaine et les phénomènes linguistiques qui permettent de les repérer dans un texte [Reymonet et al., 2007]. Plus spécifiquement, nous détaillons la RTO que nous utilisons, ses spécificités et son utilité dans notre processus d'extraction des relations n-Aires.

2.1.1 Les Ressources Termino-Ontologiques

En intelligence artificielle, et plus précisément en ingénierie des connaissances, les ontologies sont des modèles proposés pour représenter les connaissances d'un domaine. Cela permet la conception de systèmes exploitant ces connaissances préalablement formalisées grâce à l'ontologie : *'Une ontologie est une spécification explicite d'une conceptualisation'* [Gruber, 1993]. Les ontologies se positionnent comme une représentation sémantique de concepts et de relations entre ceux-ci dans le domaine considéré. Une ontologie est généralement définie non seulement dans un domaine, mais dans un contexte et pour une tâche donnée [Bachimont, 2000]. Cela influence le sens des termes et expressions linguistiques utilisés dans cette ontologie pour formaliser les concepts et relations.

Une ontologie repose sur un langage et sa syntaxe formelle. Le langage d'une ontologie doit permettre de définir de manière formelle les concepts et les relations qui les lient. Cette formalisation repose le plus souvent sur le langage *Web Ontology Language* (OWL) définissant les formules de base [Guarino et al., 2009] suivantes :

owl :Class : les concepts représentés dans l'ontologie.

owl :DatatypeProperty : les attributs de ces concepts.

owl :ObjectProperty : les relations entre deux concepts.

owl :subClassOf : la relation de subsomption (i.e. hiérarchie) entre deux concepts.

Le langage OWL, standardisé par le W3C, est conçu pour faciliter la mise en œuvre et l'utilisation des ontologies pour enrichir les ressources du Web. Celui-ci est une extension des langages *Ressource Description Framework* (RDF), et RDFS (i.e. *'S' = 'Schema'*), qui utilisent `rdfs:label` pour modéliser l'association entre les termes et les concepts qu'ils désignent. Cependant cette association ne dissocie pas le terme, composante linguistique, et le concept qu'il représente, composante sémantique. Cette non-différentiation peut poser un problème lors de travaux cherchant à extraire ou annoter des données textuelles, les termes de l'ontologie n'ayant par exemple pas de catégories grammaticales. Cette problématique a mené à la naissance de la notion de Ressource Termino-Ontologique (RTO) [Aussenac-Gilles et al., 2006, Reymonet et al., 2007, Tissaoui et al., 2013]. Cette nouvelle modélisation propose que le terme, la composante linguistique, soit dissocié du concept, composante sémantique, qu'il représente. Cela est fait par

exemple en utilisant le langage *Simple Knowledge Organisation System* (SKOS) qui s'appuie sur RDF afin de structurer des vocabulaires [Touhami et al., 2011]. Les triplets RDF s'établissent alors entre une instance de concept de type $skos : Concept$, et un terme selon une propriété $skos : Label$. Plusieurs types de termes sont classiquement définis comme $skos : prefLabel$, un terme préféré pour représenter le concept et $skos : altLabel$, un terme alternatif (e.g. variation terminologique, acronyme ou expression lexicale sémantiquement équivalente). D'autres propriétés existent également afin de s'adapter à d'autres types d'utilisation, comme $skos : hiddenLabel$ qui permet de couvrir les fautes orthographiques. D'autres propriétés peuvent également être définies afin de définir des usages particuliers lorsque nécessaire.

En utilisant ces formalismes, les Ressources Termino-Ontologiques (RTO) peuvent ainsi représenter les connaissances d'un domaine. Plusieurs ont déjà été élaborées, comme celles présentées sur la plateforme @Web [1]. Les RTO créées avec @Web, nommées naRyQ (n-ary Relations between Quantitative and Qualitative experimental data), permettent de représenter des relations n-aires entre des données expérimentales quantitatives et qualitatives [Buche et al., 2013a, Buche et al., 2013b]. Elles sont divisées en deux parties, une partie nommée *core-ontology* et une partie nommée *domain-ontology*. Un exemple d'une RTO spécifique au domaine d'application et utilisée dans le cadre de cette thèse est donné en Figure 4.

Core-Ontology La core-ontology, dans sa partie supérieure up-core, comprend une représentation générique de la structure des relations n-Aires et de leurs arguments. Sa partie inférieure, down-core-ontology, prend en compte deux types d'arguments pouvant composer les relations, les arguments symboliques et les arguments quantitatifs. Les arguments symboliques représentent notamment les objets de l'expérimentation (e.g. Packaging) et les arguments quantitatifs représentent les données possédant une valeur numérique et une unité de mesure (e.g. Temperature).

Domain-Ontology La domain-ontology représente les concepts spécifiques au domaine étudié, ainsi que leurs vocabulaires. Chaque concept symbolique ou

1. `https://ico.iate.inra.fr/atWeb/`

quantitatif est joint à une composante terminologique permettant de lui associer des labels (*PrefLabel* ou *AltLabel*). Les concepts quantitatifs sont également liés à des instances de concepts d'unités de mesure. Chaque instance de concept d'unité de mesure possède également un ensemble de labels. L'ensemble de ces labels est utilisé par les tâches d'extraction et d'annotation de données textuelles.

Cette structure en core-ontology et domain-ontology est propre aux RTO n-AryQ [Touhami et al., 2011, Buche et al., 2012] définissant la notion de relations n-Aires, dans la core-ontology, et les utilisant, dans la domain-ontology, afin de représenter les connaissances. Les RTO considèrent également, à travers leurs représentations, les termes comme des ressources particulières. Cela permet de les utiliser dans des tâches d'extraction ou d'annotation de données textuelles. Notre thèse repose sur l'utilisation de ces termes et leur place dans la structure de l'ontologie, ainsi que sur la conceptualisation en relation n-Aire des informations.

2.1.2 La RTO de domaine Transmat

Dans cette thèse, nos travaux ont été appliqués à des articles scientifiques dans le domaine des emballages alimentaires. La RTO N-aryQ utilisée, Transmat [Guillard et al., 2018] [2], permet de représenter des données expérimentales de ce domaine, structurées sous forme de relations n-Aires. La RTO Transmat permet en particulier de représenter les connaissances dans le domaine des transferts de matières et est illustrée en Figure 4. Comme toute RTO N-aryQ, Transmat possède une core-ontology et une domain-ontology. Sa core-ontology est commune avec d'autre RTO conceptualisant les connaissances de domaines expérimentaux sous forme de relations n-Aires et sa domain-ontology définit les concepts spécifiques à son domaine.

Pour notre travail, nous nous intéressons plus particulièrement aux connaissances relatives à la perméabilité aux gaz des emballages alimentaires et l'impact de leurs compositions sur celle-ci. Ces connaissances peuvent être représentées à travers des relations n-Aires, regroupant un nombre n d'arguments en les reliant à un concept de relation. Par exemple la relation H2O_Permeability_Relation regroupe comme arguments la mesure de

2. `https://ico.iate.inra.fr/atWeb/`

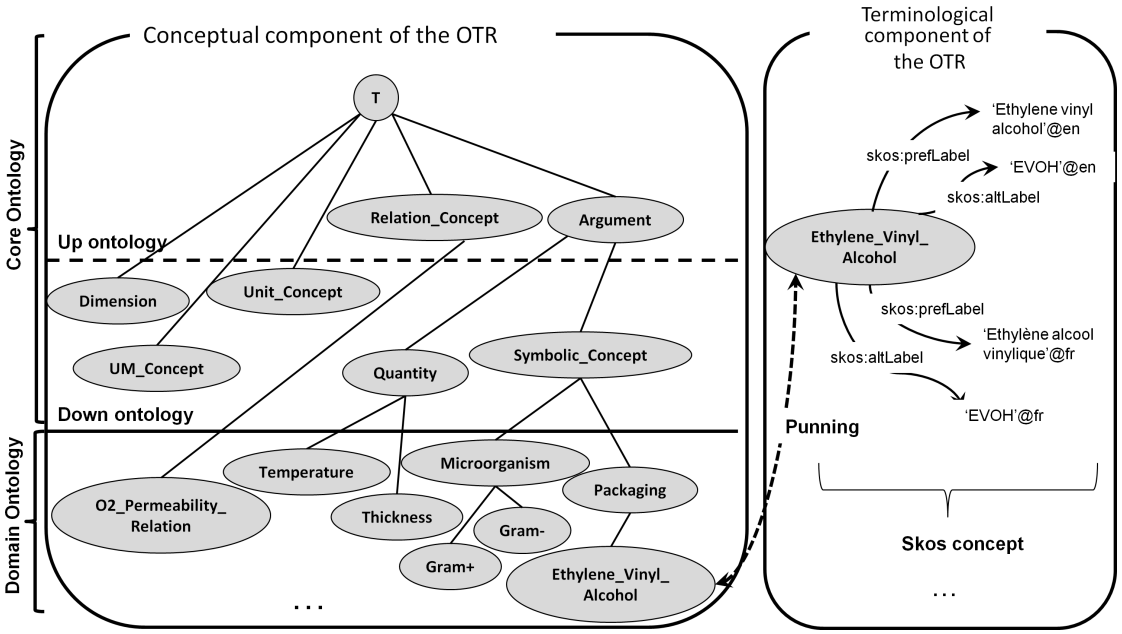


FIGURE 4 – Illustration de la RTO Transmat

perméabilité (H2O_Permeability), l'emballage étudié (Packaging) et les conditions expérimentales (Method, Temperature, Relative_Humidity et Partial_Pressure). La notion de relation n-Aire, ainsi que son utilisation dans nos travaux, est décrite plus en détail dans la Section 3.1.1.

Les Figures 5 et 6 présentent des extraits des hiérarchies de concepts sous les concepts d'arguments symboliques et quantitatifs. Nous observons que les concepts Symbolic_Concept et Quantity_Concept se divisent en différents sous-concepts. Les arguments des relations n-Aires recherchées appartiennent à ces sous-concepts. Une relation n-Aires est en effet définie par sa signature, c.a.d le type de son argument résultat (e.g. O2_Permeability) et de ses arguments d'entrée (e.g. Packaging, Temperature, Relative_Humidity). Nous parlons alors du *top-concept* d'un argument pour désigner le concept le plus générique, le plus haut dans la hiérarchie, présent dans la signature de la relation n-Aire (i.e. correspondant au type de l'argument présent dans la relation n-Aire). C'est par exemple le cas des concepts Packaging ou Temperature. Mais une instance d'une relation n-Aire peut contenir une instance d'argument appartenant à l'un des sous-concepts du top-concept. Ceux-ci sont nommés concepts spécifiques, ils dépendent d'un top-concept et héritent de ses propriétés (e.g. Low_Density_Polyethylene est un Polyethylene, qui est lui-même un Packaging). Une instance d'argument

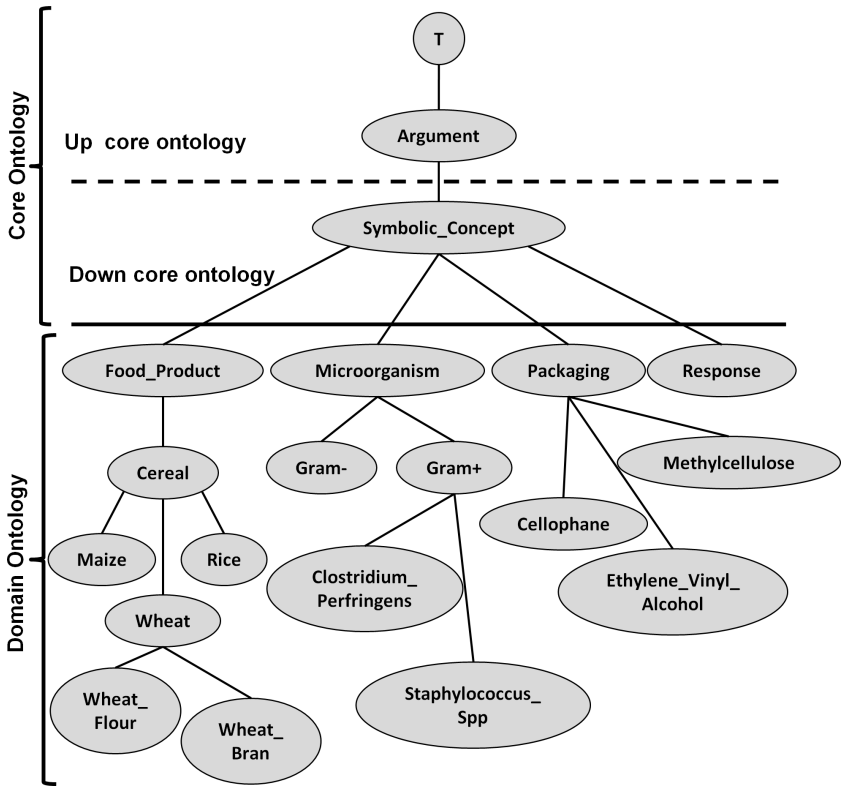


FIGURE 5 – Extrait de la hiérarchie de concepts symboliques dans la RTO

appartenant à une relation n-Aire peut être l'instance d'un top-concept ou de l'un de ses concepts spécifiques, selon cette relation de subsomption. Généralement les instances d'arguments quantitatifs correspondent à des top-concept de la RTO (e.g. *temperature*, *thickness*) et les instances d'arguments symboliques à des concepts spécifiques (e.g. *low density polyethylene*, *method ASTM-96*). Nous remarquons également que la hiérarchie des concepts possède un nombre de niveaux bien plus important parmi les concepts symboliques (Figure 5) que pour les concepts quantitatifs (Figure 6). Cela constitue une différence notable entre arguments symboliques et arguments quantitatifs qui sera exploitée dans la suite de cette thèse.

L'ontologie Transmat décrit 62 concepts de relation basés sur l'utilisation de 2432 concepts symboliques, 82 concepts de quantité et 62 concepts d'unité. Nous nous intéressons dans cette thèse à quatre relations, trois concernant les propriétés de perméabilité des emballages (CO2_Permeability_Relation, H2O_Permeability_Relation, O2_Permeability_Relation) et une concernant leurs compositions (Impact_Factor_composition_Relation).

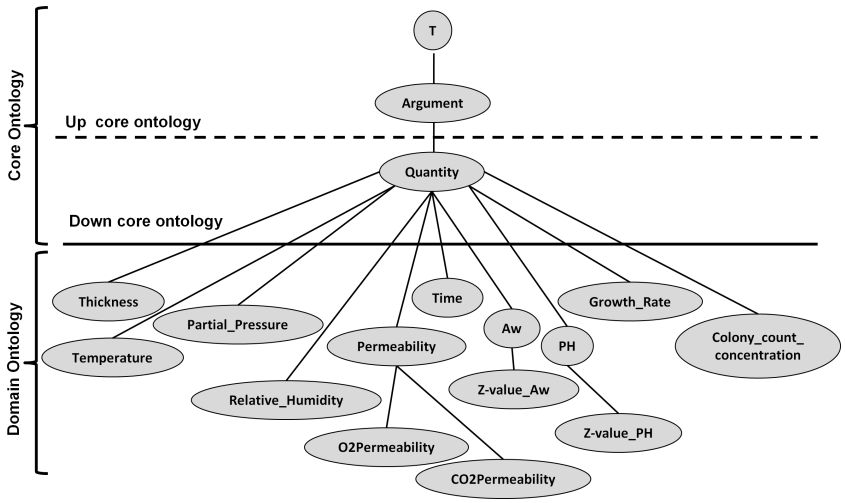


FIGURE 6 – Extrait de la hiérarchie de concepts quantitatifs dans la RTO

2.2 Etat de l'Art - extraction d'information en domaine spécialisé

Nos travaux se consacrent à l'extraction de connaissances expérimentales depuis des articles scientifiques sous forme de relations n-Aires. Cela nous place dans le domaine de l'extraction d'information, plus précisément l'extraction de relations complexes en domaine de spécialité. Comme indiqué dans le Chapitre 1, nos hypothèses de travail consistent à décomposer le processus d'extraction des instances de relations n-Aires en deux Phases. La Phase I se concentre sur l'identification et l'extraction des instances d'arguments présents dans les articles scientifiques relatifs aux relations n-Aires d'intérêt. La Phase II cherche à relier les arguments extraits précédemment pour constituer des instances de relations n-Aires. Cette section étant dédiée à la Phase I, nous nous concentrerons dans cet état de l'art sur les travaux en rapport avec l'extraction des instances d'arguments. Ces instances d'arguments sont présentes dans les articles scientifiques sous forme d'entités textuelles, de termes (mots ou syntagmes) ou de valeurs numériques et leurs unités de mesure.

2.2.1 Extraction d'information - cadre général

La naissance de l'extraction d'information en tant que domaine de recherche est située dans les années 1970 avec la conférence *Message Understanding Conferences* (MUC). Les MUC successifs ont encouragé ce domaine de recherche, proposé un cadre d'évaluation standard (e.g. fondée sur les mesures du rappel et de la précision des résultats) ainsi que des corpus textuels annotés. D'autres événements tels que les conférences *Automatic Content Extraction* (ACE) [Doddington et al., 2004] et *Knowledge Base Population* (KBP) [Ji and Grishman, 2011] ont également contribué à augmenter la diversité et la complexité des tâches à réaliser ainsi que la quantité de données textuelles annotées disponibles pour l'entraînement de méthodes d'apprentissage supervisées. Ces travaux concernent par exemple l'extraction d'entités nommées (e.g. noms de personnes, d'organisations, de lieux), de coréférences (i.e. entités textuellement différentes, mais de sens équivalents) et des relations existant entre les différentes entités. D'autres conférences se concentrent sur des thématiques particulières dans le domaine de l'extraction d'information. SemEval [3] [4] (Semantic Evaluation) est consacrée à des tâches analysant la sémantique de l'information (e.g. désambiguïsation de termes, mesure de similarité entre textes, étiquetage des rôles sémantiques). TREC (Text REtrieval Conference) [Harman, 1996] cherche elle à encourager le développement de méthodes pour l'extraction d'information au sein de corpus larges couvrant différents domaines (e.g. web et blogs, journaux, articles de chimie et de biomédecine, textes légaux, spams). CLEF (Conference on Labs of the Evaluation Forum) [Bellot et al., 2019] se concentre sur l'extraction d'information dans des contextes multi-langues et multi-modaux (i.e. différents contextes et formats d'apparition de l'information).

Dans le cadre de la Phase I de notre processus, nous nous concentrons dans cette section sur les tâches d'extraction des entités textuelles et aborderons ensuite les points particuliers à considérer lors de l'extraction d'information en domaine de spécialité dans la Section 2.2.2.

L'importance des prétraitements en extraction d'information. L'extraction d'information vise à localiser les termes, mots ou syntagmes,

3. `https://semeval.github.io/`
4. `https://en.wikipedia.org/wiki/SemEval`

pouvant être catégorisés dans des classes d'intérêt, extraits afin être stockées (i.e. généralement dans une base de données) et interprétés [Grishman, 2019]. Les catégories d'entités pertinentes pouvant varier, il est généralement considéré comme d'intérêt toute entité satisfaisant le besoin d'information de l'utilisateur [Cooper, 1971]. La Figure 7 illustre l'architecture générale adoptée par les tâches d'extraction d'information. Cela débute par un ensemble de prétraitements afin de faire ressortir les caractéristiques du texte pouvant ensuite venir enrichir l'identification des entités : segmentation des phrases (Sentence Segmentation) puis des tokens (Tokenization), couramment les mots composants le vocabulaire du texte (`['le', 'président', 'des', 'États-Unis', 'a', 'été', 'élu', 'le', '3 novembre 2020']`). Les tokens peuvent également être décomposés en préfix-radical-suffixe ou en caractères. Vient ensuite l'identification des catégories grammaticales des mots (i.e. *Part of Speech Tagging - POS*) (e.g. `['le|DET', 'président|NOUN', 'des|ADP', 'États-Unis|PROPN', 'a|AUX', 'été|AUX', élu|VERB, le|DET, 3|NUM, novembre|NOUN, 2020|NUM]`). Ces catégories permettront par exemple d'obtenir l'arbre syntaxique de chaque phrase [Grishman, 2015]. Les entités composées d'un ou plusieurs tokens sont ensuite réorganisées (Entity Reorganization) afin d'identifier les unités de sens (i.e. syntagmes) de la phrase (e.g. `'le président des États-Unis|PERSON', 'États-Unis|COUNTRY', 'le 3 novembre 2020|DATE'`). Cela concerne typiquement les entités représentants des personnes, lieux, dates, entreprises ou organisations, et permet de conserver le sens porté par une entité nommée afin de la considérer comme une unité du discours, au lieu de la décomposer en tokens. Vient ensuite une étape de désambiguïsation des entités (Entity Disambiguation) permettant de catégoriser les entités dont la nature serait ambiguë. Par exemple l'entité nommée *'Paris'* peut correspondre à la ville de Paris en France, la ville de Paris au Texas, ou encore à l'actrice Paris Hilton (e.g. `'Paris'` $\rightarrow$ `Paris(France)|CITY, Paris(Texas)|CITY, Paris Hilton|PERSON` ?). Cette étape consiste généralement à lier une entité nommée à une entrée dans une base de connaissances structurée [Dou et al., 2015, Shah and Jain, 2014]. Cela est particulièrement important à réaliser lorsque l'information que l'on désire extraire doit être intégrée à de l'information provenant de sources différentes. Ces étapes sont essentielles, car leurs résultats constituent les fondations sur lesquels les étapes suivantes (i.e. l'extraction des entités d'intérêt, leurs catégorisations et l'identification des liens

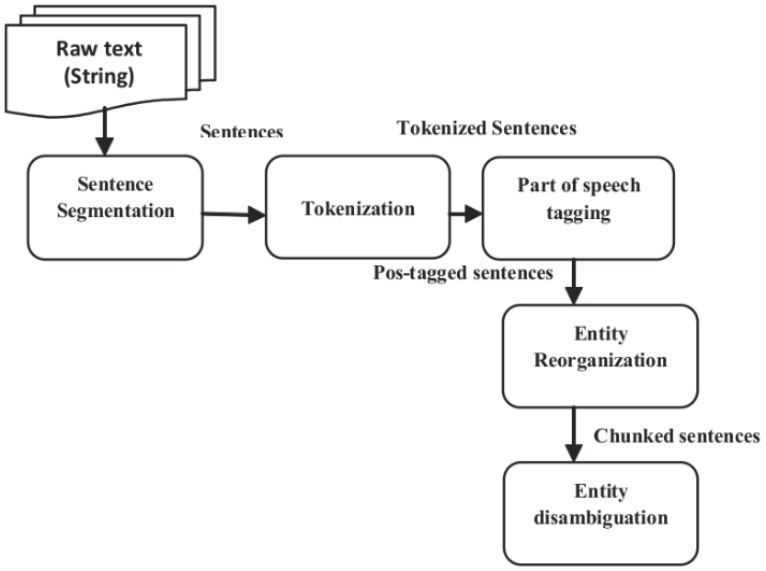


FIGURE 7 – Architecture générale en extraction d'information [Singh, 2018]

qu'elles entretiennent) vont se construire.

Ces prétraitements ont classiquement été réalisés par des modèles de langages obtenus selon deux approches : les approches par règles linguistiques et les approches statistiques. Les approches par règles linguistiques exploitent des règles, exprimées par exemple sous forme d'expressions régulières, afin de reconnaître phrases et mots puis de procéder à leur analyse grammaticale et syntaxique [Kapetanios et al., 2013]. Concevoir ces règles manuellement est très coûteux en temps, car ce travail doit être réalisé par des experts linguistes. Il a donc très tôt été conçu des méthodes permettant d'apprendre ces règles automatiquement sur des corpus annotés. Par exemple, [Brill, 1995] a proposé une méthode pour apprendre à une machine à reconnaître les catégories grammaticales des tokens. Sa méthode se fonde sur un ensemble de règles initiales, conçues manuellement ou aléatoirement, annotant automatiquement un corpus. Les résultats sont ensuite comparés à un corpus annoté manuellement servant de référence afin d'apprendre de nouvelles règles permettant de rapprocher le résultat de la première annotation au corpus de référence. Ces règles sont composées de deux éléments, la transformation à faire (e.g. transformer les tags MOD en NOUN) et une condition de déclenchement de cette règle (e.g. le token précédent est un DET). Plusieurs itérations sont ainsi effectuées, apprenant une nouvelle règle dès lors que celle-ci permet d'améliorer les résultats. L'ensemble final des règles ainsi obtenues peut ensuite être directement

utilisé sur d'autres corpus sans nécessiter de phase d'apprentissage.

Dans une autre approche, les méthodes statistiques ont exploité un ensemble varié de techniques pour parvenir à créer des modèles de langage à partir de corpus annotés. On compte parmi ces techniques l'utilisation d'arbres de décision [Magerman, 1995] et de modèles probabilistes tels que les modèles de Markov, de classifieurs et de machine à vecteur de support [Nadeau and Sekine, 2007]. Aujourd'hui, l'utilisation de réseaux de neurones profonds domine et fournit à l'ensemble de la communauté du TAL de nombreux modèles de langages prêt à l'emploi. Parmi les plus connus, nous citerons par exemple les modèles BERT [Vaswani et al., 2017] de l'entreprise Google, Stanza [Qi et al., 2020] de l'université Stanford, spaCy [5] et FLAIR [Akbik et al., 2019] de l'université Humboldt de Berlin. Ces modèles sont appris sur de larges corpus (e.g. OntoNote 5.0 [6], CoNLL-2003 [Sang and De Meulder, 2003]), servant également de Gold Standard afin de mesurer les performances de ces modèles dans différentes tâches. Les principales différences entre les modèles de langages cités proviennent de leurs méthodes de construction et d'apprentissage. BERT utilise par exemple des mécanismes d'attention [Vaswani et al., 2017] permettant au réseau de neurones de choisir les exemples sur lesquels s'entraîner alors que FLAIR effectue des représentations contextuelles des mots [Akbik et al., 2019]. Un modèle de langage ainsi entraîné peut être directement utilisé dans différentes tâches de traitement automatique du langage essentielles dans les travaux d'extraction d'information. Grâce à des bibliothèques d'algorithmes préexistantes, ces modèles vont alors pouvoir effectuer les tâches de segmentations du texte en phrases et en tokens, reconnaissance des catégories grammaticales des tokens, lemmatization, identification des entités nommées, analyse de l'arbre de dépendance syntaxique. Certains modèles incluent également des fonctions spécifiques telles que l'analyse de sentiments (Stanza) ou la création automatique d'une base de connaissances pour désambiguïser les entités (SpaCy).

D'autres tâches plus spécifiques peuvent venir enrichir le traitement de l'information. La détection de co-références, c'est-à-dire de termes référant à la même instance d'une entité, permet par exemple de lier le nom employé dans une phrase au pronom de la phrase suivante [Ng, 2017]. La reconnaissance

5. `https://spacy.io/`
6. `https://catalog.ldc.upenn.edu/LDC2013T19`

d'information temporelle (e.g. dates spécifiques, intervalle de temps ou fréquence) [Amami et al., 2012] apporte elle de l'information pouvant être utile pour ordonner les informations extraites, particulièrement lorsque l'on travaille sur des corpus biographiques ou journalistiques. Aujourd'hui, l'utilisation de modèles de langages obtenus par l'entraînement de réseaux de neurones sur de larges corpus est généralisée pour réaliser les prétraitements à l'extraction d'information qui enrichiront les textes avec de l'information lexicale, grammaticale, syntaxique et sémantique. L'objectif est ensuite d'extraire les informations d'intérêt pour le travail à réaliser. Là encore les modèles de langages peuvent être d'une grande aide, car ceux-ci incluent des fonctions permettant de procéder directement à la reconnaissance de l'information désirée. [Shin et al., 2020] emploie par exemple le modèle BERT afin de détecter automatiquement les informations spatiales (i.e. entités et relations les liants). Lorsque l'information recherchée est plus spécifique, le travail est généralement centré autour de deux approches : l'utilisation de motifs linguistiques et l'utilisation de ressources terminologiques.

Utilisation de motifs linguistiques. La conception de motifs linguistiques a pour but de concevoir des expressions, incluant des critères lexicaux, grammaticaux, syntaxiques et/ou sémantiques, qui une fois reconnus dans un texte permettent d'extraire l'information qu'ils contiennent. Par exemple [Makarov, 2018] utilise, sur des articles de journaux présentant des mots-clés relatifs aux événements politiques (e.g. *'protest'*, *'campaign'*), des motifs reconnaissant les entités nommées au pluriel et possédant une relation de dépendance syntaxique à un verbe. Ce verbe est ensuite utilisé afin de former un motif candidat composé de l'ensemble de l'arbre de dépendance sous ce verbe. Cela résulte en des extractions (e.g. *'Protesters gathered on the street chanting slogans.'*) en grand nombre. Les auteurs de [Makarov, 2018] utilisent plusieurs méthodes de filtrage classiques afin de sélectionner les motifs extrayant l'information d'intérêt : règles statistiques ne conservant que les motifs fréquents (i.e. > 15) au nom fréquent également (i.e. > 5) et règles syntaxiques exigeant que l'élément identifié comme nom ait une relation de type *nsubj* ou *agent* avec le verbe. Les motifs restants sont ainsi considérés comme pertinents et employés comme base afin de générer de nouveaux motifs permettant d'extraire de l'information pertinente, mais présentant des variations qui ne seraient pas couvertes par les motifs existants. Pour cela [Makarov, 2018] étend l'application des ses motifs

aux entités nommées sémantiquement similaires à celles déjà extraites, par le calcul d'une distance cosinus sur leurs représentations vectorielles, puis mesurant l'association des nouveaux motifs obtenus avec les motifs existants en utilisant une mesure de *positive pointwise mutual information*.

Utilisation de ressources terminologiques. L'utilisation de ressources terminologiques externes permet de définir la liste des termes à extraire. Cette ressource fournit a minima un lexique, comme dans un simple dictionnaire, indiquant les termes pertinents à extraire. D'autres ressources, thésaurus et ontologies [Dou et al., 2015, Shah and Jain, 2014], apportent de l'information sémantique additionnelle pouvant servir à classifier les termes extraits ou à les aligner avec cette base de connaissance. L'emploi d'une telle ressource est généralement requis dès que l'on souhaite extraire de l'information spécifique, comme développé dans la Section suivante, afin par exemple de déclencher l'extraction dans des contextes linguistiques favorables à la reconnaissance de l'information désirée [Makarov, 2018], ou au contraire classer les résultats en sortie de l'extraction [Uban et al., 2021].

Une ressource terminologique peut également être employée afin d'annoter automatiquement un corpus en vue de l'utiliser dans une méthode d'apprentissage. On parle alors de supervision distante [Mintz et al., 2009]. Cette technique se heurte au bruit qu'elle crée dans les données textuelles, en annotant des éléments non pertinents du texte, et qui affecte négativement les résultats de l'extraction [Roth et al., 2013]. Ce bruit peut être réduit, par exemple en utilisant une source de données textuelles annotées manuellement comme première base d'entraînement [Pershina et al., 2014]. Les informations extraites par ces techniques peuvent également être utilisées pour construire des bases de connaissances comme FreeBase [Bollacker et al., 2008], DBPedia [Auer et al., 2007] ou Google Knowledge Graph [Steiner et al., 2012]. Ces bases de connaissances peuvent être interrogées et utilisées pour de la recherche de documents, des systèmes d'aide à la décision ou par des agents conversationnels.

Extraction ouverte d'information. L'extraction d'information concerne également des questions plus complexes à aborder. L'extraction ouverte d'information (*open information extraction*) consiste à extraire toutes les

informations reconnaissables dans un texte, sans spécifiquement cibler certaines entités ou relations [Etzioni et al., 2008]. Ce cadre à l'avantage de ne pas se limiter à un ensemble défini d'entités ou de relations et ainsi être plus adaptables aux différents domaines, à la taille, ou à l'hétérogénéité du corpus [Niklaus et al., 2018a]. Il a été relevé que les approches déployées dans ce domaine se concentrent sur l'utilisation de règles linguistiques, apprises automatiquement ou non, reposant sur l'analyse grammaticale et syntaxique des phrases [Niklaus et al., 2018a]. Représentée sous forme de n-uplets, généralement de triplets, la méthode d'évaluation de la validité des informations extraites est toujours discutée [Niklaus et al., 2018a]. Sans standard d'évaluation ni de corpus important partagé dans ce but, la mesure des résultats dans ce domaine de recherche se concentre sur la précision et accorde moins d'importance au rappel des approches déployées. Une autre question de recherche importante est l'extraction d'information dans des domaines spécialisés, posant des difficultés spécifiques, abordée dans la prochaine section.

2.2.2 Extraction d'information - domaines scientifiques et de spécialités

Les travaux en extraction d'information dans des articles scientifiques s'attaquent à des problématiques spécifiques. En effet, les publications produites en science sont par nature complexes et spécialisées. L'extraction d'information en domaine de spécialité couvre ainsi de nombreux champs de recherche différents, les plus courants étant les domaines de la médecine, chimie et biologie, mais concerne également l'agriculture ou les sciences des matériaux. Les tâches réalisées en extraction d'information dans ces domaines expérimentaux sont variées [Nasar et al., 2018] (les questions de recherche spécifiques à l'extraction d'information dans les documents de sciences sociales ne seront pas abordées ici). Par exemple, l'extraction de la terminologie des documents permet de représenter le vocabulaire d'un domaine ou d'indexer les documents [Roche et al., 2015]. L'extraction des instances d'entités spécifiques permet elle d'assister la création d'ontologies de domaine et d'alimenter des bases de connaissances [Dou et al., 2015, Shah and Jain, 2014]. Nous nous concentrerons ici sur ce dernier point.

Les méthodes d'extraction d'information en domaine de spécialité concernent majoritairement le domaine de la médecine [Marsi and Öztürk, 2015] et les domaines biomédicaux [Andrade and Bork, 2000, Perera et al., 2020]. Des corpus annotés de bonne qualité sont disponibles en grande quantité dans ces domaines, obtenus à partir de plateformes telles que PubMed ils permettent de déployer des méthodes d'apprentissage automatique sur ces corpus annotés [Jonnalagadda et al., 2015]. Dans d'autres domaines, présentant un nombre plus faible de corpus annotés, des approches non supervisées, ou sans apprentissage, sont nécessaires. L'utilisation de règles construites manuellement permet par exemple d'extraire des informations générales, [Klink et al., 2000] par exemple identifier la structure des documents (e.g. les sections telles qu'Abstract ou Conclusion) puis utiliser des motifs et la fréquence des termes dans ces sections afin d'en extraire mots-clés, noms d'auteurs ou métadonnées. Ce type d'approche est générique, applicable à tous les articles scientifiques, mais n'extrait que des informations génériques et ne permet pas de cibler des entités propres au domaine ou répondant à des besoins spécifiques.

Les domaines spécialisés possèdent certes un vocabulaire et une syntaxe différente, mais les catégories d'entités recherchées sont également généralement différentes [Krallinger et al., 2013]. La recherche de noms de gènes est par exemple une tâche courante dans le domaine de la médecine. Par exemple des noms de gènes tels que *'RAG-1'* ou *'APOL3'* ont une forme spécifique et ne seraient pas capturés par des approches génériques de reconnaissance d'entités nommées. Le ciblage d'information spécifique, non seulement à un domaine mais également à une requête de l'utilisateur (i.e. relative à une information ciblée), est toujours un problème récurrent qui se heurte non seulement à la difficulté de reconnaître ces informations, mais également aux nombreux résultats pouvant constituer des faux positifs. [Avram et al., 2021] a par exemple obtenu un f-score moyen de .37 sur l'extraction de quantité, leurs unités de mesures et propriétés ainsi que les relations qu'elles entretiennent en employant des méthodes d'apprentissages reposant sur l'utilisation de *Conditional Random Fields* et de réseaux de neurones *bi-Long Short Term Memory*. Le meilleur score sur ces tâches (i.e. SemEval-2021 Task 8 [Harper et al., 2021]) est de .51 [Davletov et al., 2021] et a été obtenu en utilisant le modèle de langage pré-entrainé RoBERTa [Liu et al., 2019].

Dans notre travail, nous ciblons des informations spécifiques (i.e. mesures

de perméabilités, noms des emballages alimentaires ou paramètres de contrôle) et avons pour cela besoin d'une méthode les recherchant dans l'ensemble des documents. Ce n'est pas le cas de l'ensemble des méthodes procédant à de l'extraction d'information dans des articles scientifiques, une revue de littérature [Jonnalagadda et al., 2015] a par exemple montré que l'extraction d'information dans les articles scientifiques se consacre généralement uniquement aux *abstract* des documents. En effet l'information essentielle des articles y est généralement regroupée par les auteurs, cette information essentielle correspondant souvent à l'information considérée comme pertinente dans une tâche d'extraction de l'information.

L'apprentissage en extraction d'information spécialisée et ses limites. L'utilisation de méthodes d'apprentissage supervisé est permise lorsque des corpus annotés en qualité et quantité suffisante sont disponibles. Cette approche consiste généralement à utiliser des modèles d'extraction d'information, construisant par exemple des motifs [Groth et al., 2018] ou des modèles de word-embedding [Malmasi et al., 2015]. Cependant le résultat de l'apprentissage peut être limité au domaine spécifique du corpus et n'est pas nécessairement transférable à un autre domaine.

Une manière de contourner cette question du transfert de l'apprentissage à d'autres domaines peut être de considérer plusieurs niveaux de règles [Nédellec, 2004], génériques et spécifiques. Les règles génériques permettent de sélectionner les contextes (e.g. phrases) dans lesquelles l'information est susceptible d'apparaître. La détection de termes spécifiques (i.e. *trigger words*) permet ensuite de sélectionner la ou les règles spécifiques devant ensuite être utilisées pour extraire l'information pertinente.

Ces dernières années, l'utilisation de réseaux de neurones profonds est devenue la technique dominante pour l'extraction de l'information, y compris en domaine de spécialité. La principale raison est que les différents modèles de réseaux de neurones ne nécessitent pas, ou peu, de sélectionner les caractéristiques importantes à considérer pour l'apprentissage [Hahn and Oleynik, 2020]. Cela mène alors à la création de nombreuses études à partir des corpus d'apprentissages disponibles. Ces systèmes obtiennent également en général de meilleures performances, et résultent en des modèles de langage pouvant être réutilisés sur d'autres documents.

Cependant il ne s'agit pas uniquement de trouver le bon réseau de neurones, ou modèle de langage, à appliquer à un corpus. En effet l'enrichissement des documents avec différentes caractéristiques, comme l'identification de termes spécifiques utilisés comme trigger-words, permet en effet d'améliorer les résultats obtenus [Lin et al., 2020, Khetan et al., 2021].

Des modèles de word-embedding ont également étés entraînés spécifiquement sur des corpus d'articles scientifiques afin d'être plus efficaces dans les prétraitements du texte (e.g. tokenization, reconnaissance des catégories grammaticales, analyse syntaxique) et dans l'extraction d'informations spécifiques. SciSpaCy [7] [Neumann et al., 2019a] et BioBERT [8] [Lee et al., 2020] fournissent par exemple des modèles entraînés sur des corpus d'articles dans le domaine biomédical. Ces modèles peuvent être employés afin d'améliorer la reconnaissance des informations d'intérêt, des relations qu'elles entretiennent et leurs classifications [Mondal, 2020]. Un rapide ré-entraînement d'un tel modèle permet également de l'adapter aux spécificités du corpus d'étude [Poerner et al., 2020].

Cependant, lorsque l'on cherche à reconnaître des entités textuelles dans un domaine de spécialité, des corpus pour l'apprentissage ne sont pas toujours disponibles. De plus, la spécificité des entités recherchées (e.g. nom de gènes ou de molécules ayant une morphologie particulière) nécessite une approche adaptée à tous les cas rencontrés. C'est pourquoi l'utilisation de motifs génériques et créés manuellement est parfois préférée, particulièrement dans le cadre d'extraction ouverte d'information. Les informations extraites sont ensuite triées selon différents critères (e.g. lexicaux, sémantiques, statistiques) afin d'en conserver l'essentiel. Il s'agit alors de reconnaître des instances de tuples (e.g. $drug :: has\_effect\_on :: gene$) [Niklaus et al., 2018b] ciblant les informations d'intérêt. Celles-ci sont obtenues à partir de motifs utilisant principalement les dépendances syntaxiques. L'avantage de cette approche est que celle-ci ne requière la création que d'un nombre réduit de motifs. Les instances de tuples reconnus sont ensuite sélectionnées afin d'extraire l'information d'intérêt en se basant sur des informations lexicales (e.g. presence de trigger-words) [Mesquita et al., 2013], sémantiques (e.g. détection de polarité ou quantité) [Gashteovski et al., 2017] ou de fréquences dans le cadre de l'utilisation de supervision distance [Mintz et al., 2009]. Une autre possibilité est

7. `https://allenai.github.io/scispacy/`
8. `https://github.com/dmis-lab/biobert`

d'utiliser un modèle issu de domaines disposant de corpus pour l'apprentissage et de les appliquer au domaine étudié. Cependant, il a été démontré que cela dégrade les résultats de l'extraction d'information [Peng et al., 2019], particulièrement lorsque les domaines du modèle et de l'étude sont éloignés. Ainsi, dans les domaines spécialisés, l'utilisation de ressources externes est généralement encore nécessaire.

Utilisation de ressources externes. L'utilisation de ressources externes permet de représenter l'information d'intérêt à extraire. Celle-ci peut être un thésaurus, dictionnaire des termes d'un domaine, [Kim et al., 2017], fournissant une liste de trigger-words permettant de déclencher l'extraction d'information dans un contexte favorable (e.g. phrases contenant un nom de gène) et de sélectionner les instances pertinentes (e.g. instances de motifs incluant un verbe d'interaction). Les ontologies sont des ressources plus complexes, qui structurent les concepts d'un domaine dans une représentation sémantique. Ces ressources sont très utiles, et utilisées, dans l'extraction d'information en domaine spécialisé [Dou et al., 2015, Shah and Jain, 2014] car les connaissances du domaine qu'elles fournissent permettent de guider ou de contraindre le processus d'extraction d'information et de représenter formellement les instances extraites en les alignant avec les concepts du domaine [Konys, 2018]. Les ontologies peuvent ainsi être utiles dans différentes approches. Les informations sémantiques représentées peuvent également guider un processus d'apprentissage en filtrant les motifs sélectionnés à la suite d'un apprentissage [Bellandi et al., 2007]. Dans une autre optique, [Liu and El-Gohary, 2017] a par exemple démontré que l'utilisation d'une ontologie pour sélectionner les données d'entraînement d'un *conditional random field* permet d'obtenir de bonnes performances tout en réduisant la quantité d'annotations manuelle nécessaire à l'apprentissage. Les ontologies peuvent également être utilisées afin d'enrichir les textes en information sémantique avant de construire des motifs pour l'extraction d'information [McDowell and Cafarella, 2006a, Berrahou et al., 2017]. L'utilisation d'une ontologie nécessite cependant que celle-ci couvre l'ensemble de l'information recherchée, afin d'être capable d'aligner les représentations de ses concepts avec les instances contenues dans les documents. La première préoccupation lorsque l'on utilise le vocabulaire d'une ressource externe pour guider le processus d'extraction d'entités est donc la couverture terminologique du domaine d'intérêt [Pazienza et al., 2005, Cram and Daille, 2016].

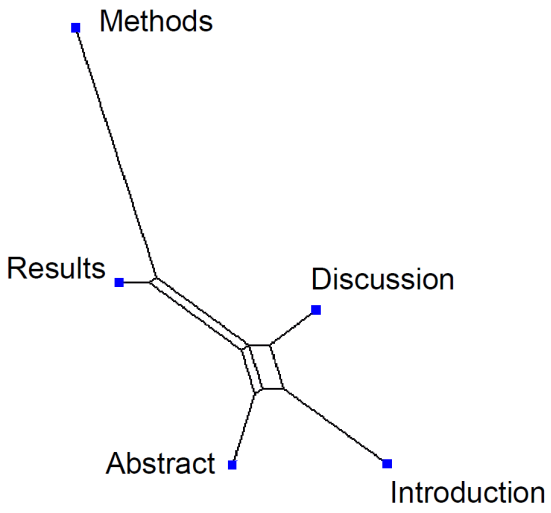


FIGURE 8 – Similarité des sections d'articles scientifiques selon leurs mots-clés fréquents [Shah et al., 2003]

Particularités des articles scientifiques. Lorsque l'intégralité du contenu des articles est considérée, la division en section des articles scientifiques peut être exploitée. Des travaux précédents ont expérimentalement démontré que des sections différentes des articles contiennent des informations significativement différentes [Mao et al., 2003, Cohen et al., 2010] en extrayant automatiquement les mots-clés les plus courants dans les sections. La Figure 8 illustre par exemple la similarité des sections des articles scientifiques selon les mots-clés qui y sont détectés [Shah et al., 2003]. Ainsi [Hofmann et al., 2009] a déterminé que les sections *Abstract* et *Discussion* étaient les plus propices à la découverte de mots-clés. Notre travail ne se consacre pas spécifiquement à cette tâche, mais la structure des articles scientifiques a un impact certain sur les tâches d'extraction d'information et sa prise en compte sera un point important pour notre représentation des instances d'arguments et notre méthode d'évaluation de la pertinence des instances d'arguments extraites (Section 2.4).

Les articles scientifiques font un usage intensif des acronymes. Il est important de reconnaître les acronymes présents dans une étude, car ceux-ci peuvent constituer des instances d'entités pertinentes à être extraites. Certains de ces acronymes sont courants, partagés par tous les auteurs d'un domaine (e.g. *NLP = Natural Language Processing*), mais d'autres peuvent être plus spécialisés et concerner un sujet particulier (e.g. *RTO = Ressource Termino-Ontologique*), ou encore n'être définis que dans un document pour les besoins particuliers de

l'étude (e.g. *MAM = Mon Acronyme à Moi / SciPuRe = Scientific Publication Representation*). L'utilisation d'une ressource existante (e.g. Acronym Finder [9]) ne peut donc pas couvrir l'ensemble des cas présents dans les articles d'un corpus. De plus, une telle ressource n'étant pas nécessairement spécifique au domaine des articles étudié, des acronymes d'autres domaines pourraient venir parasiter la reconnaissance des acronymes du corpus. La principale manière d'aborder la problématique est de reconnaître de potentiels acronymes, principalement à travers un ensemble d'heuristiques comme l'identifiant des termes composés de majuscules et n'appartenant pas à un dictionnaire standard [Wren and Garner, 2002, Ehrmann et al., 2013a, Dannélls, 2006]. Une fois ces potentiels acronymes identifiés, l'objectif est de trouver leurs formes étendues dans les textes, souvent présentes une fois à côté de leurs acronymes. ACROMED [Pustejovsky et al., 2001] par exemple effectue cet alignement entre acronymes et formes étendues en utilisant une analyse syntaxique superficielle contrôlée par quelques règles linguistiques. D'autres études exploitent des mesures statistiques [Jain et al., 2007, Okazaki and Ananiadou, 2006] recherchant les termes fréquemment rencontrés autour des acronymes. De l'apprentissage [Xu and Huang, 2005] à partir de corpus contenant un nombre important d'acronymes est également possible, mais nécessite un corpus annoté suffisant. Nous avons développé une méthode d'extraction des acronymes, présentée en Section 2.3.1, qui se veut simple, directe et adaptée aux spécificités observées dans notre domaine d'étude.

Les domaines expérimentaux font également un grand usage des unités de mesure. Certaines unités sont communes, utilisées également en dehors des domaines spécialisés (e.g. $°C$, km/h) ou obéissent à des définitions de formations formelles [Thompson and Taylor, 2008], mais d'autres sont beaucoup plus complexes, multi-dimensionnelles, et majoritairement employées pour les besoins du domaine étudié (e.g. $cm^3\mu mm^{-2}d^{-1}kPa$). Leurs reconnaissances, ainsi que de celle de leurs variations (e.g. $cm^3\mu mm^{-2}d^{-1}kPa \rightarrow cm^3.m^{-2}.s^{-1}.bar.\mu m$), est donc une question importante [Foppiano et al., 2019]. En effet, les unités de mesure étant au cœur de la formulation des données expérimentales, leur reconnaissance est essentielle pour extraire l'information dans ces domaines [Jessop et al., 2011a, Jessop et al., 2011b]. L'utilisation de ressources recensant les

9. `https://www.acronymfinder.com/`

unités de mesure est courante [Rijgersberg et al., 2013, Grau et al., 2009] mais peu de travaux s'attaquent à l'extraction de variants d'unités de mesure. Les quelques travaux existants utilisent principalement des mesures de similarité [Cohen et al., 2003, Van Assem et al., 2010] pour associer unités de mesure existantes et variations. Dans cette thèse, nous utilisons pour cette tâche une méthode existante [Berrahou et al., 2015], décrite en Section 2.3.1

2.3 Méthodologie - extraction et représentation des instances d'arguments

Dans cette section, nous présentons le processus d'extraction des instances d'arguments des relations n-Aires guidé par une Ressource Termino-Ontologique (RTO). Cette méthode est générique et repose sur l'utilisation d'une RTO décrivant un domaine expérimental qui définit les types d'arguments de relations n-Aires à extraire et guide cette extraction, notamment via sa composante terminologique. Le choix d'une méthode d'extraction n'utilisant pas de méthode d'apprentissage provient de l'absence de corpus d'articles annotés en quantité et en qualité suffisante. Nous avons choisi d'utiliser une RTO n-AryQ [Buche et al., 2013a, Touhami et al., 2011, Buche et al., 2012] de domaine afin de guider l'extraction des instances d'arguments de relations n-Aires. Cette RTO, détaillée en Section 2.1, permet de définir les relations n-Aires et les arguments les composants, qui représentent les connaissances d'un domaine ainsi que les informations qui les forment. Elle présente également une composante terminologique permettant de détecter les termes relatant la présence d'instances d'arguments dans les documents. Elle est ici appliquée au domaine des emballages alimentaires, en utilisant la RTO Transmat [Guillard et al., 2018] [10], et les exemples utilisés dans cette section appartiennent donc à ce domaine.

L'objectif de cette Phase I consiste à utiliser une RTO pour extraire les instances d'arguments des relations n-Aires d'intérêt présentes dans le corps de texte des documents. Nous avons décomposé le processus d'extraction, présenté en Figure 9, en quatre étapes. Les deux premières étapes (cf. Section 2.3.1) effectuent les prétraitements nécessaires à l'extraction d'un maximum d'instances

10. `https://ico.iate.inra.fr/atWeb/`

d'arguments des relations n-Aires.

La RTO ne représentant pas l'intégralité de la terminologie utilisée dans le domaine considéré, l'étape *Variation Extraction* va tout d'abord augmenter la couverture de la RTO utilisée en recherchant les variations terminologiques des labels de la RTO, les acronymes ainsi que de nouvelles unités de mesure. Cela permet d'augmenter la couverture du vocabulaire de la RTO et ainsi identifier un maximum de termes pouvant constituer des instances d'arguments symboliques ou permettant de désambiguïser des instances d'arguments quantitatifs lors des étapes suivantes.

L'étape *Text Processing* procède ensuite à des prétraitements textuels nécessaires à notre travail d'extraction (e.g. tokenisation, identification des sections). La Phase I procède ensuite à l'identification et l'extraction des instances d'arguments (cf. Section 2.3.2).

L'étape *Entity Extraction* identifie les termes des documents relatifs aux instances d'arguments symboliques (i.e. un mot ou une suite de mots) et quantitatifs (i.e. valeur numérique et unité de mesure) en se basant sur la composante terminologique de la RTO précédemment étendue.

L'étape *Disambiguisation* utilise ensuite l'analyse des liens de dépendances syntaxiques afin de désambiguïser les instances d'arguments quantitatifs grâce à la détection de liens entre unités de mesure et termes de la RTO relatifs à ces arguments (e.g. *'oxygen permeability'*, *'thickness'*). Une fois ces étapes d'extraction des instances d'arguments terminées, nous créons pour chacune une représentation baptisée *Scientific Publication Representation* (SciPuRe).

Cette représentation [Lentschat et al., 2020] exploite des informations récoltées tout au long des étapes de la Phase I, comme indiqué dans la Figure 9, afin de décrire chaque instance d'argument extraite. La représentation SciPuRe (cf. Section 2.3.3) est composée de trois types de descripteurs, ontologiques, lexicaux et structurels. Elle est ensuite utilisée pour déterminer la pertinence des instances d'arguments extraites en Phase I, mais également pour fournir des informations essentielles à la mise en relations des instances d'arguments dans des instances de relations n-Aires dans la Phase II.

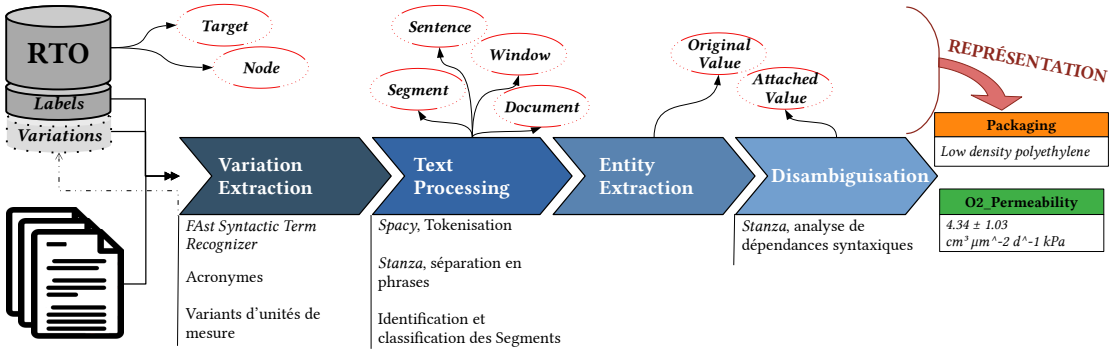


FIGURE 9 – Extraction d'instances d'arguments de relations n-Aires en domaine expérimental guidée par une RTO

2.3.1 Prétraitements pour l'extraction des instances d'arguments

Le processus d'extraction d'instances présenté en Figure 9 débute par deux opérations de prétraitement : l'augmentation de la couverture de la composante terminologique de la RTO ainsi que le prétraitement textuel des articles scientifiques. Les exemples présentés dans cette section s'appuient sur la RTO Transmat [Guillard et al., 2018], présentée en Section 2.1, et qui permet de représenter des données expérimentales dans le domaine du transfert de matière dans les emballages alimentaires. Cette RTO comprend une composante terminologique pour chacun de ses concepts qui sera utilisée pour guider le processus d'extraction des instances d'arguments.

Variation Extraction. La RTO guide le processus d'extraction d'instances sur la base de ses concepts et de la composante terminologique qui leur est associée. Les unités de mesure et les *labels* des concepts impliqués dans les relations n-Aires d'intérêt définissent les tokens formant les instances des arguments d'intérêt dans les documents, tels qu'illustrés dans l'Exemple 1. La composante terminologique de la RTO ne contenant pas toutes les variations terminologiques qu'il est possible de rencontrer dans les articles, il est nécessaire de faire un certain nombre de prétraitements pour l'enrichir. Cela concerne les formes alternatives des termes représentant les concepts de la RTO présents dans les textes tels que la forme plurielle ou acronymique d'un label de la RTO (e.g. *relative humidity* → *relative humidities* / *RH*). La recherche de formes variantes des labels de la RTO est

également appliquée aux unités de mesure. Augmenter la couverture du vocabulaire de la RTO est crucial, car l'identification des labels de la RTO présents dans le texte est essentielle pour la reconnaissance des instances d'arguments. Cette étape de la Phase I comprend trois types de prétraitement présentés ci-dessous :

— **Recherche des variants terminologiques :** Pour améliorer l'extraction d'instances d'arguments, le vocabulaire de la RTO a été enrichi de variations terminologiques en utilisant *FASTR* [Bourigault and Jacquemin, 1999]. FASTR utilise une analyse linguistique des phrases, ce qui lui permet de ne pas être dépendant de fréquences d'associations. Il peut ainsi être utilisé sur des corpus réduits et dans des domaines spécialisés, où les termes spécialisés du domaine présentent des variations terminologiques pertinentes à être reconnues. FASTR extrait les variations terminologiques d'une liste de termes dans un document via l'analyse des caractéristiques morphologiques et syntaxiques.

Cet outil débute par la recherche des variations entre les mots de la liste donnée en entrée et les mots du corpus en recherchant les lemmes et stems communs. L'analyse des catégories grammaticales permet de contrôler la lemmatisation des mots en s'assurant de détecter le bon lemme racine et ainsi arbitrer les cas d'homographie. FASTR peut reconnaître les labels des concepts de la RTO présents au pluriel (e.g. *temperature* → *temperatures*) ou sous forme adjectivale (e.g. *thickness* → *thick*). Cet outil peut traiter des termes composés d'un ou plusieurs mots en exploitant des règles comme l'insertion de modificateurs, de déterminants et de prépositions (e.g. *linear polyethylene* → *linear low density polyethylene*). Nous avons adapté l'algorithme de FASTR aux formes de la langue anglaise pour pouvoir capturer davantage de termes à mots multiples en levant la restriction conservant l'ordre des mots (e.g. *oxygen permeability* → *permeability to oxygen*). L'apport de l'utilisation de FASTR est évalué en Section 2.6.1.

— **Recherche des acronymes :** Nous avons également développé une tâche de reconnaissance des acronymes représentant les concepts de la RTO présents dans les textes. Notre méthode de reconnaissance d'acronymes débute par l'identification des labels de la RTO présents dans les textes. Une liste des potentiels acronymes est sélectionnée à l'aide d'une heuristique consistant à rechercher les séquences constituées majoritairement

de caractères majuscules, chiffres et caractères spéciaux (e.g. *(*, *)*, -). Une autre liste est également constituée en reconnaissant les labels de la RTO présents dans le texte. Un score de similarité est ensuite calculé entre chacun des potentiels acronymes et des labels de la RTO reconnus dans une même phrase en utilisant un coefficient de Dice [Dice, 1945].

Cette mesure nous permet de considérer la similarité entre un label de la RTO et un potentiel acronyme sans être trop restrictifs quant à l'ordre de ses caractères. En effet les méthodes existantes [Ehrmann et al., 2013b, Veyseh et al., 2021] ne sont pas toujours adaptées à la reconnaissance de cas uniques à un document, car exploitant la fréquence d'apparition et d'association des acronymes, ou à des formes de termes spécialisés décomposés en plusieurs caractères dans un acronyme (e.g. $Dice(\text{"}low\ density\ polyethylene\text{"}, \text{"}LDPE\text{"}) = .86$). Les associations dont le score est supérieur à un seuil déterminé sont ajoutées aux labels alternatifs de la RTO.

— **Recherche de variants d'unité de mesure :** L'identification des unités de mesure dans les domaines expérimentaux est également une préoccupation fréquente dans certains domaines [Foppiano et al., 2019]. En effet les auteurs ne respectent pas, ou ne peuvent pas toujours respecter, les conventions de nommage standard des unités de mesure apparaissant dans un domaine. Nous avons utilisé la méthode décrite dans [Berrahou et al., 2015] pour reconnaître les variations d'unités de mesure.

Cette méthode a été développée sur un corpus similaire à celui utilisé dans cette thèse et a été réécrite en langage Python afin d'être compatible avec nos autres codes d'extraction. Elle commence par rechercher les candidats pouvant constituer de nouvelles unités de mesure. La recherche commence par sélectionner toutes les suites de tokens situées entre deux termes d'un dictionnaire standard (i.e. ici le dictionnaire de la librairie Python `nltk` comprenant $236k+$ mots de l'anglais). Les suites contenant au moins un token déjà référencé dans une unité de mesure de la RTO sont conservées comme candidats. Les candidats sont ensuite filtrés en deux temps, par une mesure d'indice de Jaccard ($> .50$) puis une mesure de Damerau-Levenshtein ($. > 70$) entre candidats et unités de mesure existantes. Ces scores permettent de mesurer la proximité de chaque candidat avec les unités de mesure déjà existantes dans la RTO et sélectionner les candidats possédant des scores de proximité au-dessus d'un seuil défini (i.e. $> .70$). Ils permettent également

d'associer les candidats retenus à une unité de mesure existante afin de lier les nouvelles unités de mesure détectées aux concepts quantités de la RTO. Par exemple l'unité de mesure inconnue $u_1 = kgmPa^{-1}s^{-1}m^{-2}$ est associée à l'unité de la RTO $u_{RTO} = lb.m.m^{-2}.s^{-1}.Pa^{-1}$ en deux temps : $Jaccard(u_1, u_{RTO}) = \frac{4}{6} = .70$ puis $Dammerau - Levenshtein(u_1, u_{RTO}) = max[0; \frac{5-1}{5}] = .80$.

Exemple 1 *Reconnaissance d'instances d'arguments dans les textes :*
The permeability of (*low density* (*polyethylene*)) *films* ((*LDPE*)) *was measured with the* (*ASTM D95-96*) *method at* $\underline{25 \pm 1}$ [$°C$]. *The film had a* (*thickness*) *of* $\underline{15}$ [μm] *and showed optimal barrier properties with a* (*permeability to oxygen*) *of* $\underline{4.34 * 10^{-3}}$ [$cm^3\mu mm^{-2}d^{-1}kPa$]. *This measurement was obtained at a constant* (*RH*) *of* $\underline{85.0}$ [%].

Legend : (*label de concept symbolique*) [*unité de mesure*] *valeur numérique*

Text Processing. Après l'extension de la couverture de la composante terminologique de la RTO, les textes sont découpés en tokens en utilisant la librairie Python *spaCy 3*[11] tandis que *Stanza* [Qi et al., 2020, Zhang et al., 2020] est utilisé pour la délimitation des phrases. Ces outils ont été choisis pour leurs performances à l'état de l'art sur ces traitements linguistiques[12]. Notre corpus de documents étant constitués d'articles scientifiques dans un domaine spécialisé, nous avons observé que spaCy présentait un meilleur comportement sur la tokenization tandis que Stanza était mieux adapté à la segmentation des phrases. Cela provient de la possibilité de définir un dictionnaire personnalisé de termes pour la tokenization (i.e. le vocabulaire de la RTO).

La structure des documents, à travers les noms des sections, sous-sections et tables ou figures, est également identifiée. Cela permet de conserver cette information structurelle, utilisée ensuite pour décrire les instances d'arguments. Les noms de sections et sous-sections sont regroupés automatiquement en classes en se basant sur une distance de Levenshtein [Levenshtein, 1966] faible (i.e. ≤ 3), afin de principalement englober les formes plurielles et ne pas être trop permissif à d'autres variations, et en regroupant les sections dont les noms sont inclus dans

11. `https://spacy.io/`
12. `https://spacy.io/usage/facts-figures#benchmarks`

les noms d'autres sections. E.g. $classe(R\&D) = \{Results\ and\ Discussions\} \cup \{Result and Discussion\} \cup \{Results\} \cup \{Discusion\}$). Cette étape *Text Processing* est la deuxième étape de la Phase I, telle que présentée dans la Figure 9.

2.3.2 Extraction et désambiguïsation des instances d'arguments

Entity Extraction Une fois les étapes de prétraitements terminées, le vocabulaire étendu de la RTO est employé pour reconnaître les termes et les unités de mesure qui constitueront les instances d'arguments. L'utilisation d'une ressource présentant une composante terminologique est courante en recherche d'information dans des domaines spécialisés [McDowell and Cafarella, 2006a, Kim et al., 2017, Berrahou et al., 2017] et permet ainsi de guider le processus d'extraction en ciblant les entités d'intérêt. Les valeurs numériques sont également identifiées. L'Exemple 1 présente différentes entités identifiées avec ce processus d'extraction.

Ces identifications constituent la troisième étape, *Entity Extraction*, de la Figure 9. L'ensemble des labels appartenant aux instances d'arguments des relations n-Aires d'intérêts sont recherchés. Cela concerne les termes faisant référence à des instances d'arguments symboliques (e.g. *'LDPE* dans l'Exemple 1), à des arguments quantitatifs (e.g. *'thickness'* dans l'Exemple 1), à des unités de mesure référencées dans la RTO (e.g. $cm^3 \mu m m^{-2} d^{-1} kPa$ dans l'Exemple 1) ainsi que les valeurs numériques. Les suites de tokens reconnus appartenant à une même catégorie sont regroupées en un même terme (e.g. $'28, 12', .10^{-8'} \rightarrow' 28, 12.10^{-8'}$, $'agar - agar', films' \rightarrow' agar - agar films'$). Les associations entre valeurs numériques et unités de mesure sont ensuite recherchées afin de lier celles-ci et former de potentielles instances d'arguments quantitatifs. Cela est réalisé en analysant l'arbre de dépendances syntaxiques des phrases, une relation de dépendance entre une valeur numérique et une unité de mesure permettant de lier celles-ci. Cette analyse est une approche courante [Mohit and Hwa, 2005] en extraction d'entités. Cela est fait à l'aide de la librairie stanza dont l'outil d'analyse de dépendance syntaxique obtient d'excellentes performances à l'état de l'art sur les domaines biomédicaux [Qi et al., 2020, Zhang et al., 2020].

Si cette analyse de la structure des phrases ne permet pas d'associer une unité de mesure à chaque valeur numérique, alors nous recherchons l'unité de mesure la

plus proche. Cette recherche par proximité privilégie les unités placées à la suite de la valeur numérique, et avec une distance limite arbitraire de 15 tokens, afin de couvrir les cas où valeur numérique et unité de meure sont éloignés (e.g. *'the permeability values were respectively of* $4.34 * 10^{-3}$ *for the packaging (A1) and of* $8.54 * 10^{-3}$ *"$cm^3 \mu m m^{-2} d^{-1} kPa$' for the packaging (B1)*).

Désambiguïsation. Afin d'extraire les termes reconnus en tant qu'instances d'arguments de relations n-Aires, il est parfois nécessaire de procéder à une étape de désambiguïsation. En effet si les instances liées à des arguments symboliques ne se réfèrent qu'à un seul concept de l'ontologie et ne sont donc pas ambiguës, ce n'est pas le cas des instances d'arguments quantitatifs. Cela correspond à la quatrième étape, *Disambiguation*, dans la Figure 9. Cette étape de désambiguïsation concerne principalement les instances d'arguments quantitatifs, l'unité de mesure pouvant se référer à plusieurs arguments différents. Un exemple simple est l'unité de mesure %, pouvant se référer à l'argument Relative_Humidity, mais étant aussi souvent utilisé pour décrire d'autres informations de quantités ne constituant pas des instances des arguments recherchés. Pour cette désambiguïsation nous analysons les associations pouvant être faites entre les instances de $['numerical\_value', 'measure\_unit']$ et les termes identifiés précédemment et se référant à un concept de quantité de la RTO.

Dans l'Exemple 1, $['4.34 * 10^{-3}', 'cm^3 \mu m m^{-2} d^{-1} kPa']$ peut être désambiguïsé en une instance de l'argument O2_Permeability par son association avec *'permeability to oxygen'*. Cela est fait de manière similaire à l'association d'une unité de mesure à une valeur numérique, en analysant les liens de dépendances syntaxiques ou la proximité des termes dans la phrase. Les instances de $['numerical\_value', 'measure\_unit']$ présentant une unité de mesure non ambiguë peuvent être extraites comme instance d'arguments quantitatifs sans nécessiter de désambiguïsation. Cela concerne principalement les unités de mesure fondamentales ne pouvant représenter qu'un seul concept de quantité, tel que $°C$ qui ne peut être qu'une température.

L'ensemble des étapes constituant la Phase I n'est supervisé par aucun corpus d'entraînement et s'appuie uniquement sur des techniques de reconnaissance directes guidées par la structure et la composante terminologique d'une RTO.

Ce procédé a été conçu pour être générique, pouvant être utilisé dans différents domaines pour l'extraction de données expérimentales en étant guidé par une RTO de ce domaine. La Section 2.6 présente une évaluation de la méthode d'extraction proposée, incluant les scores de rappel, précision et f-score pour chacun des arguments d'intérêts dans le domaine des emballages alimentaires.

2.3.3 SciPuRe : Scientific Publication Representation

Nos effectuions ensuite une représentation, baptisée *Scientific Publication Representation* (SciPuRe), associé à chaque instance d'arguments extraite afin de décrire ses caractéristiques et son contexte d'apparition. Nous avons proposé cette représentation dans [Lentschat et al., 2020]. La représentation SciPuRe intègre un ensemble de descripteurs ontologiques, lexicaux et structurels respectant les critères suivants de représentation de l'information [Boyce et al., 2017] : discriminer les différences, identifier les similarités, décrire précisément et minimiser l'ambiguïté. Les représentations des données extraites en extraction d'information sont parfois centrées sur les besoins spécifiques de l'étude (e.g. représentation de données quantitatives [Hao et al., 2017]) ou exploitent principalement des descripteurs lexicaux [Wood et al., 2021, Zaenen et al., 2010] ou d'une ressource externe (e.g. une base de connaissances [Franco-Salvador et al., 2014]). Notre représentation offre elle une richesse de description des instances élevée permettant de les décrire selon plusieurs critères tout en étant générique et adaptées à différents formats de données, symboliques et quantitatifs.

Cette approche se situe dans le domaine de recherche du smart data et permet de sémantiser les données extraites par une description de leurs contextes et particularités [Marcia, 2017, Duong et al., 2017]. La représentation SciPuRe est employée pour l'évaluation de la pertinence des instances d'arguments et fournit des descripteurs utiles à la reconstitution des instances de relations durant de la Phase II décrite en Section 3.2.

SciPuRe contient trois catégories de descripteurs, reposant sur les informations communes à toutes les instances d'arguments :

— **Descripteurs Ontologiques** : Le descripteur *Target* indique le top-concept de l'argument d'une relation n-Aire auquel l'instance est associée. Le

top-concept d'un argument de relation n-Aire étant le concept le plus haut dans la hiérarchie de l'ontologie pouvant faire partie de la signature de la relation (e.g. les concepts Packaging ou Temperature dans O2_Permeability_Relation). Le descripteur *Node* spécifie le sous-concept spécifique que l'instance représente, c'est-à-dire le concept de la RTO associé au label utilisé pour la reconnaissance ou la désambiguïsation de l'instance (i.e. respectivement pour les arguments symboliques ou quantitatifs).

Par exemple, l'instance extraite *'LDPE'* de l'Exemple 1 correspond à une étiquette alternative du concept Low_Density_Polyethylene qui est lui-même un sous-concept de Packaging. Son descripteur Target est donc *'Packaging'* et son descripteur Node est *'Low_Density_Polyethylene'*.

— **Descripteurs Lexicaux** : Le descripteur *Original_Value* contient le texte correspondant à l'instance d'argument extraite. Le descripteur *Attached_Value* correspond aux termes utilisés lors de l'étape de désambiguïsation. Pour un argument quantitatif, ce sont les termes relatifs à une quantité de la RTO reconnus dans la phrase qui ont été utilisés pour désambiguïser l'unité de mesure. Lorsqu'aucune désambiguïsation n'est nécessaire, l'unité de mesure n'étant pas ambiguë, Attached_Value correspond au PrefLabel du concept spécifique de l'argument (i.e. le descripteur Node). Un argument symbolique n'étant pas ambigu, nous considérons que leurs descripteurs Attached_Value et Original_Value sont les mêmes.

Dans l'Exemple 1, *'RH'* est un label alternatif du concept Relative_Humidity et permet la désambiguïsation de *'50 %'*. Le descripteur Original_Value est donc *'50 %'* et le descripteur Attached_Value de l'instance est *'RH'*.

— **Descripteurs Structurels** : Les descripteurs *Sentence*, la phrase, et *Window*, c'est-à-dire la phrase précédente, actuelle et suivante, indiquent le contexte lexical dans lequel l'instance d'argument apparaît. Cette taille de fenêtre contextuelle a été estimée comme la plus favorable à la reconnaissance d'instances d'arguments liés dans des instances de relations par [Berrahou et al., 2017]. Le descripteur *Segment* permet de prendre en compte la structure d'un article scientifique en indiquant les sections et sous-sections dans lesquelles les instances d'arguments ont été extraites. Ce descripteur donne un contexte spécifique aux informations extraites dans les

articles scientifiques en situant l'information dans un contexte large au sein du document Enfin le descripteur *Document* fournit les références de l'article (i.e. titre, auteurs, année) et le DOI permet de l'identifier de manière unique.

Deux exemples de représentations SciPuRe extraites de la phrase de l'Exemple 1 sont présentés dans les Tableaux 1 et 2, respectivement pour une instance d'argument symbolique et une instance d'argument quantitatif.

	Descripteur	Valeur
LEX.ONT.	Target	`Packaging`
	Node	`Low_Density_Polyethylene`
LEX.ONT.	Original Value	*'LDPE'*
	Attached Value	*'LDPE'*
STRUCTURELS	Sentence	*'The permeability of ... at* $25 \pm 1\ °C$*'*
	Window	[∅, *'The permeability ...* $25 \pm 1\ °C$*'*, *'The film ...* $d^{-1}kPa$*'*]
	Segment	*'Results and Discussion'*
	Document	*Barrier properties of chitosan coated polyethylene*
	DOI	*10.1016/j.memsci.2012.02.037*

Table 1 – *SciPuRe* d'une instance symbolique

	Descripteur	Valeur
LEX.ONT.	Target	`O2_Permeability`
	Node	`O2_Permeability`
LEX.ONT.	Original Value	[*'*$4.34 * 10^{-3}$*'*, *'*$cm^3 \mu m m^{-2} d^{-1} kPa$*'*]
	Attached Value	*'permeability', 'to', 'oxygen'*
STRUCTURELS	Sentence	*'The film had ...* $d^{-1}kPa$*'*
	Window	[*'The permeability ...* $25 \pm 1\ °C$*'*, *'The film ...* $d^{-1}kPa$*'*, ∅]
	Segment	*'Results and Discussion'*
	Document	*Barrier properties of chitosan coated polyethylene*
	DOI	*10.1016/j.memsci.2012.02.037*

Table 2 – *SciPuRe* d'une instance quantitative

2.4 Calcul de pertinences des instances d'arguments

À la suite de l'extraction des instances d'argument et la construction de leurs représentations SciPuRes, décrites en Section 2.3, nous proposons des indicateurs permettant de mesurer la pertinence des instances extraites. Nos méthodes décrites ici visent à distinguer les instances d'arguments pertinentes des instances non pertinentes au regard des relations n-Aires recherchées dans le texte. Par exemple, dans le domaine des emballages alimentaires, les instances de l'emballage faisant l'objet des expérimentations décrites dans l'article sont pertinentes, car faisant partie d'instances de relations n-Aires d'intérêts, alors que les noms d'emballages cités à des fins de comparaisons ne le sont pas. Un autre exemple est la valeur de la température contrôlée pendant l'expérience qui ne doit pas être confondue avec celles impliquées dans la préparation de l'emballage. Cette question de la pertinence des informations est une préoccupation récurrente en recherche d'information [Cooper, 1971, Mizzaro, 1998]. La pertinence d'une information est habituellement définie par la manière dont celle-ci satisfait le besoin d'information d'un utilisateur [Cooper, 1971]. Dans notre application, une instance d'argument est considérée comme pertinente si celle-ci appartient à une instance d'une relation n-Aire d'intérêt (i.e. relations de perméabilité et de composition de l'emballage).

Nous avons décidé d'aborder cette question par la mesure de scores de pertinence pour chacune des instances extraites à l'issue de la Phase I. Ces scores de pertinence sont calculés à partir des descripteurs de SciPuRe. L'objectif est d'attribuer des scores de pertinence plus élevés aux instances valides, afin de pouvoir les distinguer des instances non pertinentes. Cette différentiation est faite à travers un ordonnancement des instances d'arguments selon leurs scores de pertinence. Cela permet de choisir un seuil pour filtrer les résultats de l'extraction. Une évaluation par Precision@N [Craswell, 2009] (également connue sous le nom de Precision@K [Manning et al., 2008]) est présentée dans la section 2.6 pour évaluer les effets des scores de pertinence proposés ci-dessous.

2.4.1 Score de pertinence lexicale

Afin de différencier les instances d'arguments pertinentes, des scores de pertinence lexicale basés sur la notion de l'importance et de la discrimination des termes des instances par le calcul de l'indicateur *Tf-idf* (*tf-idf*) [Gerard Salton, 1983] sont proposés. *Tf* (Term Frequency) est basé sur l'hypothèse que les termes les plus fréquents dans un document sont les plus importants. *idf* (Inverse Document Frequency) vise à refléter la nature discriminante des termes, tout en donnant plus d'importance à ceux qui ne sont spécifiques qu'à quelques documents. Ces scores classiques en recherche et extraction d'information ont été choisis pour leur capacité à considérer à la fois le caractère fréquent et discriminant d'un terme. Il est en effet attendu que les instances d'arguments pertinentes soient plus fréquentes (e.g. le terme dénotant l'emballage étudié dans un article) ou spécifiques à un document ou à une section (e.g. les paramètres de contrôle sont plus souvent présents dans les sections *Materials and Methods*). Les descripteurs de SciPuRe possèdent toutes les caractéristiques permettant de déployer ces mesures.

Le tableau 3 liste différents scores de pertinence lexicale proposés. Les descripteurs SciPuRe fournissent les éléments permettant de calculer des scores lexicaux à différents niveaux pour une instance d'argument extraite. Nous avons choisi de proposer les descripteurs de la représentation SciPuRe pour faire cela et tirer parti de la richesse de description des instances fournies en considérant différents descripteurs textuels des instances d'arguments ainsi que différents contextes.

Les indicateurs $tf-idf$ utilisent un terme et comparent sa fréquence (tf) ou sa présence (idf) par rapport à un certain contexte. Le descripteur Attached_Value est l'élément indiquant la manifestation de l'instance dans le texte. Celui-ci correspond au terme associé à une instance d'argument symbolique ou au terme ayant permis de désambiguïser une instance d'argument quantitatif. Le descripteur Target peut être employé pour considérer un terme plus générique, celui correspondant au PrefLabel du top-concept de l'argument dans la RTO. Dans le cas d'un terme relatif à une instance d'emballage alimentaire, Attached_Value serait par exemple le terme *'low density polyethylene'* alors que Target serait *'packaging'*. Ce terme est pris dans un certain contexte. Le contexte considéré est généralement le Document dans l'utilisation du score $tf-idf$. Cependant les articles scientifiques

présentent une structure spécifique, représentée dans SciPuRe, et qui peut être considérée. Le Segment (i.e. la classe de section contenant l'instance d'argument) permet d'étudier la présence des instances dans ces contextes spécifiques. Puisque les Segment sont regroupés en classes, nous nous situons dans le cadre plus générique de l'indicateur $tf - icf$ (*Term Frequency - Inverse Category Frequency*) proposé dans [Wang and Zhang, 2010].

Nom	terme	contexte	Equation
$TF^{term}_{document}$	Attached Value t	Document d	$\frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$
$TF^{term}_{segment}$	Attached Value t	Segment s	$\frac{f_{t,s}}{\sum_{t' \in s} f_{t',s}}$
$TF^{target}_{segment}$	Target a	Segment s	$\frac{f_{a,s}}{\sum_{a' \in s} f_{a',s}}$
$IDF^{term}_{document}$	Attached Value t	Document d	$log \frac{\lvert D \rvert}{\lvert d \in D : t \in d \rvert}$
$ICF^{term}_{segment}$	Attached Value t	Segment s	$log \frac{\lvert S \rvert}{\lvert s \in S : t \in s \rvert}$
$ICF^{target}_{segment}$	Target a	Segment s	$log \frac{\lvert S \rvert}{\lvert s \in S : a \in s \rvert}$

TABLE 3 – Définition des scores de pertinence lexicale selon les descripteurs de SciPuRe

Les effets des ordonnancements des instances d'arguments selon ces scores de pertinence lexicale sont évalués en Section 2.6.2.

2.4.2 Score de pertinence sémantique

Les instances d'arguments de relations n-Aires extraites des documents scientifiques étant destinées à être utilisés par des experts ou des systèmes avancés d'aide à la décision, la mesure de leur pertinence doit également refléter leur pouvoir informatif. Ce pouvoir informatif peut être considéré à travers la spécificité du concept [Harispe, 2014]. Par exemple, si l'instance du concept représentée par *'multilayer film'* est une instance d'argument de Packaging, l'instance du concept *'PE films coated with chitosan'*, plus spécifique, lui serait préféré. Cette notion de la pertinence repose sur la relation de subsomption des concepts fournis par la RTO que nous utilisons.

SciPuRe inclut le concept spécifique de la RTO associé à l'instance (Node) et son top-concept (Target). La distance, c'est-à-dire le nombre d'arêtes dans le graphe de la RTO entre les concepts correspondant aux Node et Target, est calculée en utilisant la hiérarchie des concepts de la RTO. Elle exprime la mesure de spécificité de l'instance, inspirée de [Norman et al., 1965], dans le score de pertinence sémantique *Conceptual Distance* CD^{node}_{target}, illustré dans l'Équation 1). La pertinence de chaque instance correspond à la distance entre le Node n et la Target a notée $dist(n, a)$. Cette distance est comparée à la distance maximale entre le concept signature de l'argument considéré, a, et tous ses sous-concepts, n', notée $max(dist(n', a) : n' \sqsubseteq a)$, où $\sqsubseteq$ désigne la relation de subsomption dans la RTO. La mesure de pertinence de CD^{node}_{target} est attendue comme plus discriminante pour les instances d'arguments symboliques, car celles-ci sont décrites à des niveaux de spécialisation plus élevés dans la RTO.

$$Conceptual\ Distance \quad CD^{node}_{target} = \frac{1 + dist(n, a)}{1 + max(dist(n', a) : n' \sqsubseteq a)} \tag{1}$$

2.4.3 Combinaison des scores

Les scores de pertinence présentés ci-dessus peuvent être utilisés seuls ou en combinaison. Par exemple, afin de considérer conjointement la fréquence relative et le caractère discriminant d'un terme dans un document, les scores *tf* et *icf* sont souvent combinés par multiplication. Nous avons déterminé d'autres moyens de combiner nos scores de pertinence afin de tirer parti des spécificités de chacun et combiner leurs effets. Des combinaisons linéaires et séquentielles des scores sont proposées afin de bénéficier des propriétés associées à chaque type de score de pertinence. Par exemple, combiner un score lexical de type *tf* avec le score sémantique CD^{node}_{target} permet de prendre en compte à la fois la fréquence de l'instance extraite dans les textes (i.e. critère lexical) et la spécificité de son concept associé (i.e. critère sémantique). Notons qu'avant de combiner les scores de pertinence, il faut d'abord les normaliser sur une échelle $[0, 1]$. Combiner des scores de pertinence de nature différente permet également de combiner l'usage des descripteurs de SciPuRe afin de discriminer les instances d'arguments.

Combinaison linéaire. La combinaison linéaire, Équation 2, additionne les différents scores après avoir attribué un poids à chacun d'eux. La somme totale de tous les poids α_i est toujours égale à 1.

$$Linear(Score_i) = \sum_{i=1} \alpha_i.Score_i \quad : \quad \sum_{i=1} \alpha_i = 1 \tag{2}$$

Cette combinaison possède certaines limites dans l'attribution des scores α_i, une trop grande différence entre ceux-ci revenant à trop privilégier un score au détriment des autres. Au contraire, des valeurs similaires des α_i rapprochent la combinaison linéaire de la moyenne des scores. Cette attribution des α_i est également dépendante de l'utilisateur. La possibilité d'utiliser des fonctions d'apprentissage pour fixer les scores de pertinence à combiner linéairement, ainsi que les valeurs à attribuer à α_i, est discutée en Section 4.3. Afin de combiner plus efficacement les effets des différents scores, nous avons établi une façon de les combiner séquentiellement.

Combinaison séquentielle. L'objectif de la combinaison séquentielle est d'utiliser un score de pertinence sur un ensemble d'instances d'arguments ayant été pré-filtrées par un autre score. Par exemple, le score CD^{node}_{target} peut être utilisé pour éliminer d'abord les instances les moins spécifiques dans l'ontologie (i.e. aspect sémantique). Ensuite, un score lexical tel que $TF^{term}_{segment}$ permettra de sélectionner les instances extraites les plus fréquentes dans l'ensemble des résultats restants.

La combinaison séquentielle consiste donc à ordonner les instances extraites en fonction d'un premier score $Score_1$. Un sous-ensemble constitué de la proportion θ (%) des premiers résultats est ensuite réordonné selon un $Score_2$. Ce processus peut être répliqué i fois jusqu'au dernier score à considérer $Score_i$ et θ_i afin de bénéficier des effets spécifiques de chaque score de manière séquentielle. Naturellement, le choix de l'ordre de combinaison est important. Cet ordre, tout comme le choix de la proportion θ des instances sélectionnées à chaque étape, doit être déterminé par l'utilisateur de notre méthode d'extraction. La possibilité d'utiliser des fonctions d'apprentissage pour fixer les scores de pertinence à combiner séquentiellement, leur ordre ainsi que les valeurs de θ, est discutée en Section 4.3.

2.5 Instances d'arguments - présentation du corpus

Afin d'évaluer les résultats de notre méthode d'extraction d'instances d'arguments de relations n-Aires guidée par une RTO, présentée en Section 2.3, nous avons conçu un Gold Standard des instances d'arguments. Les effets de l'attribution de scores de pertinence lexicale et sémantique aux instances d'arguments, présentés en Section 2.4 sont également évalués. Ce Gold Standard a été obtenu par l'annotation manuelle d'un corpus de 50 articles scientifiques par trois annotateurs sur un serveur WebAnno [Eckart de Castilho et al., 2016]. L'objectif a été d'annoter les entités textuelles constituant les instances d'arguments relatives aux relations n-Aires d'intérêt. Il en résulte un Gold Standard constitué de 1772 instances d'arguments, détaillé ci-dessous ainsi que dans [Lentschat et al., 2021a] et disponible sur [Lentschat et al., 2021c]. Concevoir des corpus pour l'évaluation des résultats est une nécessité lorsque le domaine d'application est très spécialisé. Ces corpus peuvent ensuite être mis à disposition et ainsi constituer pour d'autres des ressources exploitables lors du déploiement de méthode utilisant de l'apprentissage. Il n'existe pas de ressources textuelles annotées sur le domaine considéré dans cette étude, et utilisable pour l'évaluation des tâches abordées dans le cadre de notre travail. Ce domaine étant spécifique, cela réduit l'intérêt pour conception de corpus annotées. Dans notre domaine d'étude, les publications scientifiques sur les emballages alimentaires, la quantité de ressources textuelles brutes existantes est plus faible (e.g. 200 documents sont présents sur la plateforme @Web [13]). De plus l'annotation est particulièrement difficile, les annotateurs familiers du domaine étant plus rares.

2.5.1 Conception du Corpus annoté

Le jeu de données a été obtenu à partir de 50 articles, publiés dans plusieurs journaux et accessibles sur la plateforme ScienceDirect, rassemblés manuellement au format HTML. Les documents ont été transformés depuis leur format HTML en format texte, les codes pour cette tâche sont disponibles avec l'ensemble de données sur une archive ouverte Dataverse [Lentschat et al., 2021c]. Le but était

13. `https://ico.iate.inra.fr/atWeb/`

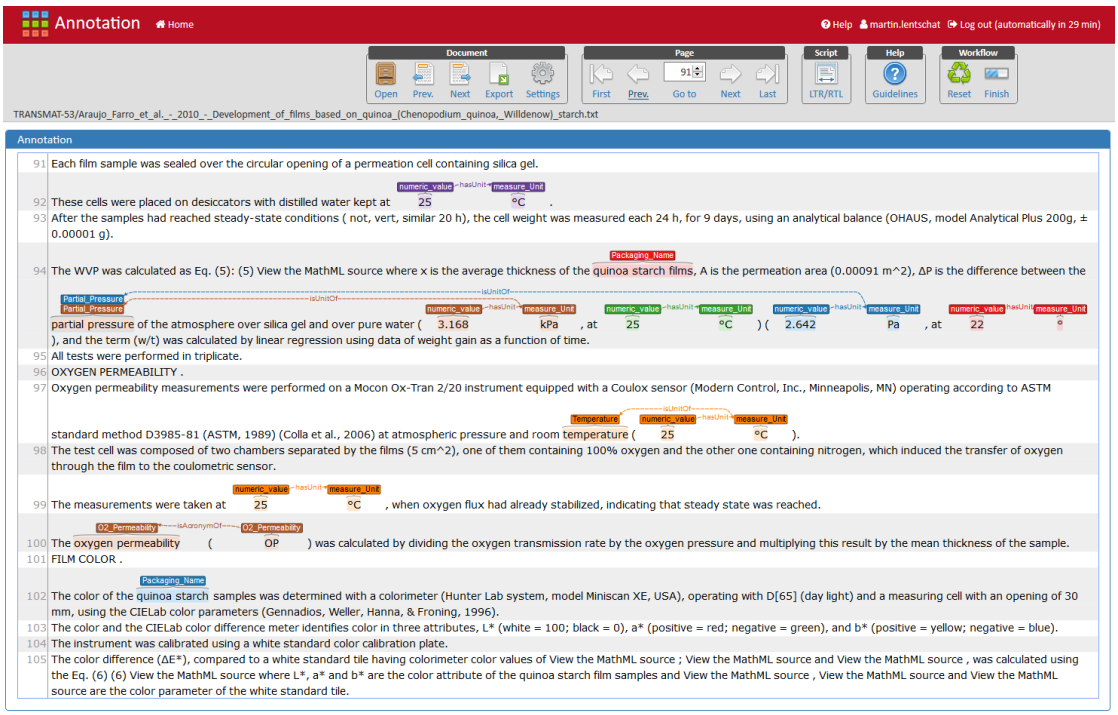


FIGURE 10 – Annotation manuelle sur WebAnno

d'obtenir des fichiers adaptés à l'annotation manuelle sur un serveur WebAnno [Eckart de Castilho et al., 2016]. Les fichiers sont ainsi nettoyés de toutes les informations liées au format HTML pour ne conserver que les informations textuelles, c'est-à-dire le corps du texte, et structurelles (noms des sections et des sous-sections). Dans les figures, comme les graphiques ou les images, la légende et le titre ont été conservés et inclus dans le corps du texte. Le traitement et la segmentation du texte appliqués aux phrases ont été réalisés à l'aide de la librairie Python Stanza [Qi et al., 2020, Zhang et al., 2020]. Les tableaux ont été identifiés et re structurés dans un format adapté à WebAnno. L'annotation et l'extraction des données présentes dans les tableaux ont été traitées comme des tâches séparées, ce travail étant présent en Section 3.3.2. Les titres des tableaux et des figures, ainsi que leurs légendes, ont été considérés comme du texte et inclus dans le processus d'annotation. Toutes les informations spécifiques aux articles, telles que le copyright, la liste des auteurs ou les références, ont été supprimées. L'annotation a été effectuée sur un service WebAnno, déployé sur un serveur docker de l'équipe ICO de l'UMR IATE (Montpellier, France). Une illustration de la tâche d'annotation est présentée dans la Figure 10.

Quatre relations n-Aires d'intérêt ont été sélectionnées dans la RTO pour être annotées : les relations de perméabilité (la perméabilité à l'oxygène, O2_Permeability_Relation, au dioxyde de carbone, CO2_Permeability_Relation, et à l'eau, H2O_Permeability_Relation) et la composition des emballages alimentaires (Impact_Factor_Component_Relation). Ces relations ont été sélectionnées car elles constituent le sujet principal des articles scientifiques du corpus. Les publications scientifiques de ce domaine contiennent en effet toujours a minima une relation de perméabilité et une relation renseignant sur la composition des emballages. Généralement ces articles étudient plusieurs formulations d'un emballage, en faisant varier les proportions des différents constituants, et mesurent les perméabilités des différentes formulations.

Seules les instances d'arguments liées à ces relations ont été annotées. Les informations similaires à l'un des arguments recherchés, mais non liées à l'une des relations d'intérêt n'ont pas été annotées. Cela concerne par exemple une température liée aux conditions de stockage des emballages, une information distincte des températures utilisées comme paramètre de contrôle pour les mesures de perméabilités.

Trois annotateurs ont effectué le travail d'annotation manuelle en parallèle et de manière indépendante. L'annotateur 1 est l'annotateur principal, responsable de la construction du Gold Standard. Les annotateurs 2 et 3 sont exercés au travail d'annotation dans le domaine étudié, leurs rôles ont été dans un premier temps de faire remonter les points importants à considérer dans le travail d'annotation sur ce domaine ainsi que de valider le schéma d'annotation. La tâche d'annotation a d'abord été effectuée sur un sous-corpus de 10 documents par l'annotateur 1, qui a conçu un schéma d'annotation préliminaire. L'annotation consiste à identifier toutes les informations qui constituent une instance d'argument d'une des relations n-Aires d'intérêt. Chacune des relations comprenant différents arguments représentés dans les textes par des informations symboliques et quantitatives, une instance peut être constituée d'un mot unique ou d'une suite de mots, pouvant eux-mêmes être distant les uns des autres. La version finale du schéma d'annotation a été obtenue après un processus itératif incluant l'ensemble des annotateurs. Ce schéma est disponible en ligne au côté des textes annotés sur le Dataverse [Lentschat et al., 2021c]. Il a été décidé de ne pas annoter les données contenues dans les tableaux avec l'outil WebAnno, en raison de la difficulté de la tâche et

du risque de saturation du serveur dû à une forte concentration d'annotations que cela entraîne. Notons d'ailleurs que l'interface de WebAnno n'est pas conçue pour annoter des données dans des tableaux.

Une fois le schéma d'annotation établi, l'annotation a été effectuée par les trois annotateurs. L'annotateur 1 a annoté l'intégralité des documents du corpus et les annotateurs 2 et 3 ont annoté cinq documents différents chacun. Il s'agit respectivement des cinq premiers et des cinq derniers documents, par ordre alphabétique. Les instructions données aux annotateurs étaient d'identifier uniquement les instances d'arguments liées aux relations de perméabilité des emballages en se référant aux arguments des quatre relations n-Aires définies dans l'ontologie Transmat. Le schéma d'annotation définit les étiquettes permettant d'annoter les éléments des textes composant les instances d'arguments, tel qu'illustré dans les Figures 10, 11, 12 et 13. Ces étiquettes permettent de signaler les éléments relatifs aux arguments des relations n-Aires (e.g. Packaging, O2_Permeability) ainsi que les valeurs numériques et les unités de mesure. L'identification des instances symboliques est simple : un mot ou une séquence de mots (e.g. l'instance Method dans la Figure 11).

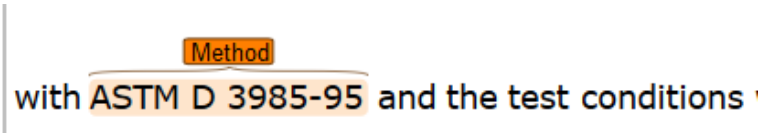


FIGURE 11 – Annotation d'une méthode de mesure sur WebAnno

Les instances quantitatives nécessitent de lier la valeur numérique identifiée et l'unité de mesure, ainsi que parfois le terme utilisé pour désambiguïser l'unité. Des relations binaires orientées sont utilisées pour lier les annotations, comme illustré dans la Figure 12. Deux relations binaires existent pour les instances quantitatives : *hasUnit* relie une étiquette *numeric_value* à une étiquette *measure_Unit* et *isUnitOf* relie une étiquette *measure_Unit* à une étiquette relative aux paramètres expérimentaux (e.g. *"°C"* à *"température"*).

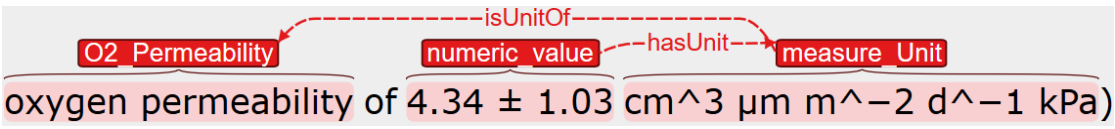


FIGURE 12 – Annotation d'une valeur de perméabilité sur WebAnno

En outre, une relation *isCompositionValueOf* permet de relier une valeur numérique ou une unité de mesure à une étiquette *Packaging_Component* afin d'annoter une donnée de composition de l'emballage, qui peut être adimensionnelle (e.g. comme les proportions illustrées dans la Figure 13).

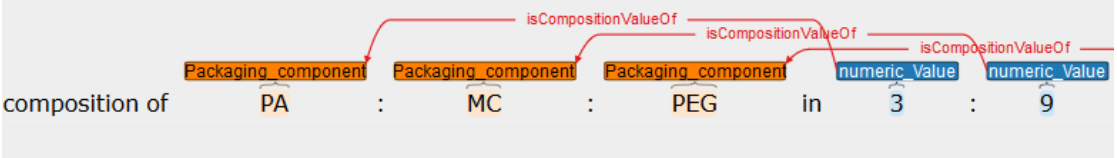


FIGURE 13 – Annotation de proportions sur WebAnno

Afin de faciliter la tâche d'annotation, les informations avec des séquences de caractères exactement similaires présentes à différents endroits dans les documents n'ont pas été annotées, à moins que cela ne soit nécessaire (par exemple, si le texte dupliqué est lié à une autre instance d'intérêt). Ce choix a été fait, car de nombreux doublons étaient présents dans les documents, mais l'annotation de toutes ces occurrences ne génère pas d'informations nouvelles pour la tâche à accomplir.

Un résumé du travail d'annotation est présenté dans la Table 4. Un peu moins de deux mille annotations ont u total été effectuées par les trois annotateurs. Celles-ci sont légèrement plus nombreuses pour les arguments symboliques ($\approx$ 60%) et concerne en grande majorité les noms d'emballages et leurs composants. Les arguments quantitatifs relevés concernent principalement les valeurs de composition des emballages alimentaires et les valeurs de perméabilités mesurées. Les paramètres de contrôle sont présents en moindres quantités. En effet un même paramètre de contrôle est généralement partagé par les différentes expérimentations décrites dans un article (i.e. une même valeur de température est utilisée dans toutes les mesures de perméabilité à l'oxygène de différents emballages). Nous remarquons également que certains paramètres de contrôle ne sont pas présents dans tous les documents. Par exemple la donnée de pression partielle, Partial_Pressure, a rarement été relevée car ce paramètre est inclus dans l'unité de mesure de la perméabilité (e.g. *'Pa'*, *'bar'*).

Faire réaliser un travail d'annotation par plusieurs annotateurs permet de calculer des scores d'accord inter-annotateurs afin d'évaluer le niveau de consensus entre les annotateurs. Deux scores ont été calculés sur les documents communs annotés de manière indépendante, d'une part par les annotateurs 1 et 2 et d'autre

Target	annot. 1 docs 1-50	annot. 2 docs 1-5	annot. 3 docs 46-50	**Final** 50 docs
SYMBOLIC	988	127	42	**1050**
packaging	431	60	30	**476**
method	43	7	5	**46**
impact_factor_component	514	60	7	**528**
QUANTITATIVE	686	30	81	**722**
component_qty_value	365	16	13	**379**
permeability	150	6	42	**165**
relative_humidity	58	3	9	**61**
temperature	54	4	9	**56**
thickness	44	1	5	**45**
partial_pressure	15	0	3	**16**
TOTAL	1676	157	123	**1772**

TABLE 4 – Distribution des instances d'arguments dans le Gold Standard.

part par les annotateurs 1 et 3. L'outil intégré à WebAnno pour le calcul de l'accord inter-annotateur, DKPro Statistics [Eckart de Castilho and Gurevych, 2014], a mesuré un Kappa de Cohen moyen de $\kappa_C = .98$. Cependant, cet outil compare les annotations ayant des positions similaires dans les documents, mais pas les étiquettes attribuées à des mots identiques présents dans différents endroits des documents (e.g. si l'annotateur 1 a annoté un nom d'emballage dans l'abstract du document, et que l'annotateur 2 a annoté ce même nom, avec la même forme textuelle, dans l'introduction du document, alors ces deux annotations ne seront pas comparées). Un score de *Gwet's Kappa* [Gwet, 2014] a donc été calculé, en comparant les étiquettes annotées sur les termes similaires (c'est-à-dire un mot ou une séquence de mots) sans tenir compte de leurs positions dans les textes. Le code utilisé pour le calculer est disponible dans le jeu de données [Lentschat et al., 2021c]. Nous avons obtenu un score d'accord moyen de $\kappa_G = .62$. Cela souligne la difficulté pour les annotateurs de déterminer la pertinence de l'ensemble des informations en raison des variations terminologiques et de la dispersion de l'information dans les documents.

Le processus d'agrégation des annotations dans un fichier unique pour constituer le Gold Standard doit ensuite tenir compte des différents choix d'étiquettes faits par les annotateurs. Les distributions des annotations présentées dans le tableau 4 montrent que le Gold Standard final est principalement constitué

d'annotations faites par l'annotateur 1. Les annotations des annotateurs 2 et 3 viennent compléter le Gold Standard. Pour remédier aux différences entre les étiquettes attribuées aux informations, le premier annotateur s'est vu confier le rôle de *curateur* et a décidé de l'annotation finale. Le nombre et la répartition des informations dans l'annotation finale sont résumés dans la dernière colonne du tableau 4. Un programme d'extraction, disponible dans le jeu de données [Lentschat et al., 2021c], permet de passer du format de WebAnno à un format .csv contenant uniquement les instances d'arguments.

2.5.2 Composition du Gold Standard

Le Gold Standard des instances d'arguments [Lentschat et al., 2021a] est composé d'instances d'arguments symboliques et quantitatifs concernant la composition des emballages alimentaires et leurs perméabilités au gaz. Ce Gold Standard a été utilisé pour évaluer la méthode d'extraction décrite en Section 2.3, ainsi que l'effet d'un ordonnancement des instances d'arguments selon des scores de pertinence de la Section 2.4. Cette méthode d'extraction est guidée par l'ontologie Transmat, décrite en Section 2.1, et a été appliquée sur un corpus de 50 documents (environ 258000 mots et 9400 phrases après nettoyage). Ces 50 documents sont considérés comme représentatifs du domaine de la perméabilité des emballages. À titre de comparaison, la plateforme *@Web*[14] qui stocke ce type de données dans le dossier *PackPermeability* héberge actuellement environ 200 documents dans ce domaine contenant les informations requises et ont été manuellement extraites.

Il est à noter que d'autres travaux d'extraction automatique de données scientifiques ont également été menés sur de petits corpus. Elles reflètent la quantité limitée de ressources textuelles disponibles dans les domaines spécialisés, au-delà du domaine médical. Une étude récente [Brack et al., 2020] utilise un corpus de 110 documents comprenant 110 résumés provenant de divers domaines, de l'agriculture aux sciences informatiques. Les auteurs comparent leurs résultats avec d'autres travaux utilisant des corpus plus importants et des méthodes de *machine learning* : [Beltagy et al., 2019, Luan et al., 2018]. Ils ont obtenu des résultats similaires et ont conclu qu'un corpus de 110 résumés est suffisant pour de telles tâches. Une autre étude [Minard et al., 2010] a utilisé 300 rapports de

14. `https://ico.iate.inra.fr/atWeb/`

radiologie pour l'extraction d'instances quantitatives et symboliques liées à la recherche sur les reins. Notre corpus est constitué d'un nombre de documents plus faible, mais prend en considération l'intégralité du contenu des articles scientifiques.

Le Gold Standard obtenu comprend les instances d'arguments annotées dans des articles ayant des séquences de caractères différentes, la position des instances dans les documents n'a pas été prise en compte. Ce choix a été fait, car de nombreux doublons étaient présents dans les documents, et l'annotation de toutes les occurrences n'aurait donc pas généré les informations nécessaires à la tâche à accomplir. Le Gold Standard des instances d'arguments organise les données dans un tableau (un exemple de ligne de données est présenté dans la Table 5). Les données sont décrites par un ensemble de caractéristiques :

Doc : le titre de l'article à partir duquel les données ont été annotées.

DOI : le *Digital Object Identifier* de chaque article.

Target : le concept générique représenté par les données dans la RTO Transmat.

Original_Value : le texte annoté constituant la donnée : une liste de tokens annotés pour les données symboliques, deux listes de tokens annotés pour les données quantitatives (une liste de valeurs numériques et une liste d'unités de mesure).

Attached_Value : la liste des tokens annotés pour désambiguïser une unité de mesure lorsque cela est nécessaire pour les données quantitatives. Aucune pour les données symboliques.

Annotator : l'identifiant de l'annotateur.

Type : la catégorie de concept de l'ontologie, symbolique, quantitative ou adimensionnelle.

Ce Gold Standard [Lentschat et al., 2021a, Lentschat et al., 2021c] a été utilisé dans une étude sur l'extraction et la pertinence des données expérimentales [Lentschat et al., 2020, Lentschat et al., 2021b]. Le nombre et la répartition des informations annotées dans le corpus sont indiqués dans la Table 4.

Features	Values
Document	Barrier and surface properties of chitosan-coated greaseproof paper
DOI	https ://doi.org/10.1016/j.carbpol.2006.02.005
Target	permeability
Original_Value	*(['3400'], ['cm', '^', '3', 'mm', '/', '(', 'm', '^', '2', 'atm', 'day', ')'])*
Attached_Value	*['carbon', 'dioxide']*
Type	QUANTITY
Annotator	1

TABLE 5 – Exemple d'une ligne du Gold Standard des instances d'arguments

2.6 Résultats et discussions

Nous présentons ici les résultats de nos méthodes d'extraction et d'ordonnancement des instances d'arguments de relations n-Aires, guidées par la Ressource Termino-Ontologique (RTO) Transmat, décrite en Section 2.3. L'utilité des différents scores de pertinence (Section 2.4) conçus à partir de SciPuRe (Section 2.3.3) pour trier les instances d'arguments est également évaluée. Dans le cadre de cette évaluation, nous utilisons le Gold Standard décrit en Section 2.5 et employé à la fois pour évaluer les résultats de l'extraction des instances d'arguments et évaluer l'effet de leurs ordonnancements selon les scores de pertinence [Lentschat et al., 2020, Lentschat et al., 2021b].

2.6.1 Résultats de l'extraction des instances d'arguments

Le tableau 6 présente, par type d'instance d'arguments, le nombre d'instances distinctes annotées dans le Gold Standard, le nombre d'instances reconnues par l'extraction des instances d'arguments de relations n-Aires et ses résultats selon le rappel (cf. Équation 3), la précision (cf. Équation 4) et le f-score (cf. Équation 5). Ces mesures sont données en micro, signifiant que nous considérons l'ensemble des résultats sans considérer les distributions des résultats (i.e. instances) à l'intérieur de leurs classes (i.e. arguments).

$$Rappel = \frac{vrais\ positifs}{vrais\ positifs + faux\ négatifs} \tag{3}$$

$$Précision = \frac{vrais\ positifs}{vrais\ positifs + faux\ positifs} \tag{4}$$

$$F - Score = 2 * \frac{rappel * précision}{rappel + précision} \tag{5}$$

Les résultats de l'extraction des instances d'arguments sont présentés dans la Table 6. Le rappel est de .85, avec quelques variations selon les catégories d'instances considérées. Cela est bon, considérant que l'on cherche lors de cette Phase I à identifier un maximum d'instance d'arguments afin d'être en mesure de reconstituer des relations n-Aires complètes lors de la phase suivante.

La précision est de .41, et est sujette à davantage de variations, avec une moyenne de .47 pour les instances symboliques et de .14 pour les instances quantitatives. Ceci est dû au plus grand nombre de faux positifs dans l'extraction d'instances incluant des valeurs numériques. Par exemple, de nombreuses températures ont été identifiées (1925) par rapport au nombre de températures distinctes annotées (54), à savoir celles associées aux mesures de perméabilité. Cette disparité entre le nombre d'instances annotées dans le Gold Standard et le nombre d'instances extraites a également été constatée pour les instances symboliques. La précision dépend également du type de concept considéré : la précision associée à Method (.16) étant par exemple beaucoup plus faible que celle de Component (.56). Ceci est dû au nombre élevé d'occurrences du terme générique *'method'*, un faux positif, car ne portant pas d'informations, par rapport aux désignations spécifiques de la méthode telle que *'ASMT D95-96'*. Les instances de l'argument Partial_Pressure étant rares dans le corpus (i.e. 15 instances identifiées dans le Gold Standard de 50 documents), nous les avons exclues de cette présentation des résultats.

Les valeurs de rappel obtenues montrent que la Phase I a permis l'extraction d'un grand nombre d'instances valides. La précision étant plus inégale, les instances extraites ont dû être filtrées pour obtenir les informations pertinentes. Ce point sera abordé lors des expérimentations décrites en Section 2.6.2, basées sur les scores de pertinence décrits dans la Section 2.4.

Target	#distinct	#reconnus	rappel	précision	f-score
SYMBOLIQUE	988	16665	.85	.47	.61
Packaging	431	6940	.86	.37	.51
Component	514	9506	.84	.56	.67
Method	43	219	.77	.16	.26
QUANTITATIF	303	3994	.86	.14	.24
Permeability	150	832	.83	.16	.27
Relative_Humidity	55	696	.88	.28	.43
Thickness	44	541	.100	.14	.24
Temperature	54	1925	.83	.08	.15
GENERAL	1291	20659	.85	.41	.55

#distinct : nombre d'instances distinctes présentes dans le Gold Standard.
#reconnus : nombre d'instances reconnues lors de l'extraction.

TABLE 6 – Résultats de l'extraction des instances d'arguments

Apport de la recherche de variations terminologiques et d'acronymes. Nous avons évalué les effets de l'augmentation de la couverture de la composante terminologique de notre RTO sur notre méthode d'extraction d'instances d'arguments de relations n-Aires. Pour cela, nous avons mesuré les valeurs de rappel, précision et f-score obtenues lors d'une extraction sans augmentation de la couverture (cf. Baseline dans la Table 7), en utilisant la recherche de variants terminologiques de notre version modifiée de FASTR [Bourigault and Jacquemin, 1999] (cf. +Fastr dans la Table 7) et de recherche des acronymes (cf +Acronymes dans la Table 7) avec la méthode utilisant l'ensemble des augmentations disponibles (GENERAL dans la Table 7).

Approche	rappel	précision	f-score	#instances reconnues
Baseline	.74	.38	.50	14603
+Fastr	.80	.37	.51	18749
+Acronymes	.79	.36	.49	15664
GENERAL	.85	.41	.55	20669

TABLE 7 – Effet de l'augmentation de la couverture de la RTO

Nous observons dans la Table 7 que les méthodes d'augmentation de la couverture de la RTO reposant sur la détection de variants terminologiques et de formes acronymiques ont pour effet d'augmenter le rappel en dégradant très légèrement la précision. +Fastr a augmenté de .06 le rappel de Baseline en perdant

.01 de précision. +Acronymes a augmenté de .05 le rappel de Baseline en perdant .02 de précision. Ces résultats sont attendus, l'augmentation de la couverture de la RTO augmente mécaniquement le nombre de termes permettant de détecter la présence d'instances d'arguments. Des résultats intéressants se produisent lorsque l'outil FASTR et la recherche d'acronymes sont utilisés conjointement. Nous observons alors dans notre approche GENERAL, une augmentation du rappel de .11 et de la précision de .03 comparé à Baseline. Cela provient de la détection d'éléments obtenus grâce aux utilisations successives de FASTR et de la recherche d'acronymes. Par exemple le texte *'... ethylene-vinyl alcohol copolymer with 44% ethylene molar content (EVOH44) ...'* contient une forme longue d'un nom d'emballage ainsi que sa forme acronymique. La forme longue étant un terme variant, il est nécessaire de l'identifier afin de pouvoir ensuite y associer son acronyme. Cet acronyme peut ensuite être reconnu dans de nombreux autres endroits du document, résultant en l'extraction de nouvelles instances d'arguments pertinentes.

Comme montré ci-dessus, la quantité d'instances d'arguments extraites lors de la Phase I est importante. Avec un rappel de .85, cela indique que la majorité des instances d'arguments pertinentes sont reconnues. En revanche la précision générale de .41, qui varie selon l'argument considéré, indique une grande part de résultats faux-positifs. Les sections suivantes de ce chapitre sont dédiées au calcul de la pertinence des instances d'arguments, afin de les ordonner et discriminer un maximum des instances pertinentes.

2.6.2 Évaluation des scores de pertinence

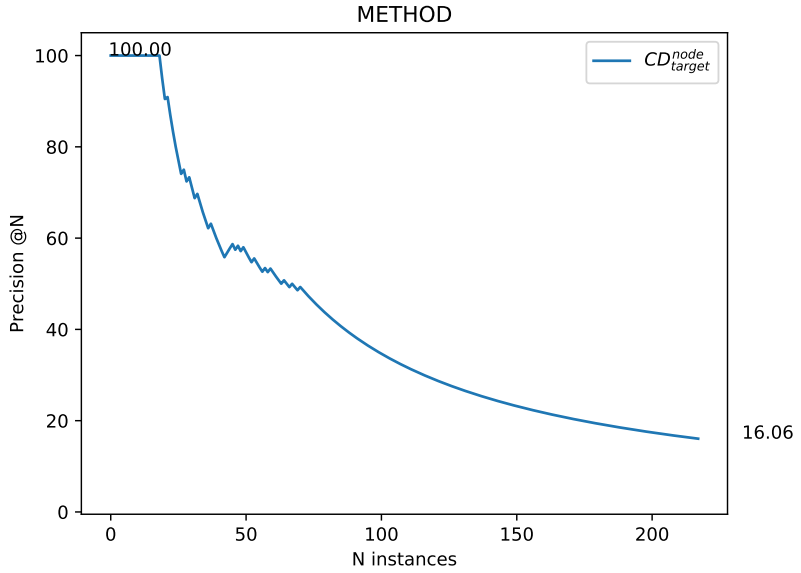
Nous avons utilisé la mesure de *Precision@N* [Craswell, 2009] (également connue sous le nom de Precision@K [Manning et al., 2008]) pour évaluer l'utilité des scores de pertinence dans le cadre d'un ordonnancement des résultats. Cela consiste pour un ensemble d'instances d'arguments classées en fonction d'un score donné, à calculer la valeur de précision des N premiers résultats. Les variations de N, de $N = 1$ à $N = all$, permettent de représenter la variation de la précision en fonction de N sur une courbe. Cette procédure d'évaluation met en évidence la précision obtenue avec un score de pertinence en fonction du nombre d'instances d'arguments sélectionnées. Les graphiques de précision@N ont facilité la sélection

d'un seuil pour filtrer les résultats en fonction d'un score de pertinence et ont permis de décider de son utilisation en combinaison avec d'autres scores.

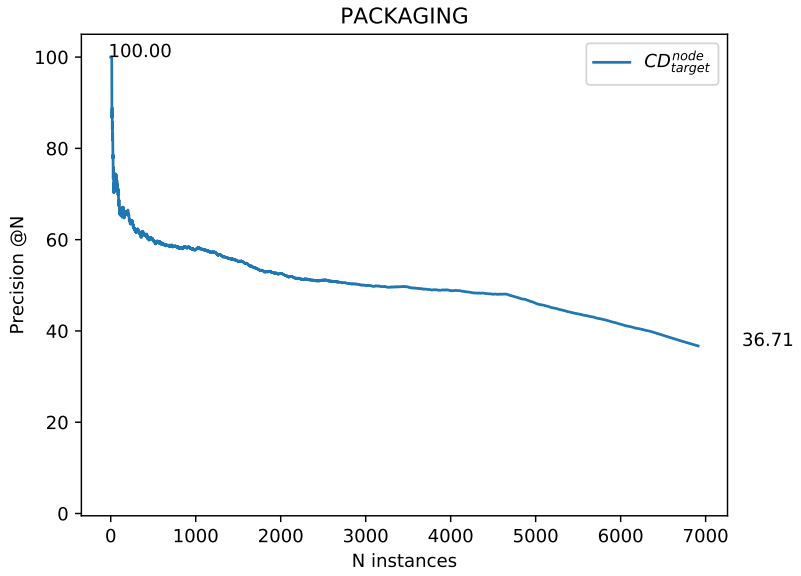
La figure 14 présente la précision@N des instances des arguments Method et Packaging ordonnées en fonction de leurs scores CD^{node}_{target}. L'axe des abscisses indique le nombre des N *meilleures* instances sélectionnées selon CD^{node}_{target}, tandis que l'axe des ordonnées indique les valeurs de Précision@N associées. Dans la figure 14a, la Précision@N est de 100% jusqu'à $N = 24$ instances d'arguments, puis elle diminue progressivement et atteint la précision moyenne de Method à $N = all$. Au-delà de $N = 70$ ($\approx 35\%$ des instances), la courbe est monotone et décroissante, c'est-à-dire que les instances sélectionnées au-delà de ce seuil ne sont que des faux positifs. La figure 14b montre un comportement similaire, bien que moins prononcé, pour Packaging. Nous avons observé le même comportement pour les instances associées à Component. Ces résultats indiquent que le fait de trier les instances des concepts symboliques avec le score CD^{node}_{target} afin de ne retenir que celles ayant les meilleures valeurs est un moyen efficace de filtrer les résultats valides.

L'impact des scores de pertinence lexicale de type tf pour un classement des instances d'arguments symboliques est présenté dans la Figure 15a avec les instances d'arguments de Packaging. En effet les noms des emballages sur lesquels portent les études sont souvent répétés dans chaque document. Nous avons observé des résultats similaires pour les instances d'arguments de Component. $TF^{term}_{segment}$ a également donné des résultats exploitables, et s'est avéré meilleur, pour les instances de Method (voir tableaux 8 et 9). Par conséquent, les noms des méthodes recherchées semblent également être plus fréquemment présents dans des sections spécifiques (par exemple, *Materials and Methods*). Il serait possible de décider de filtrer une partie des résultats avec des scores de pertinence lexicale de type tf en retirant, par exemple, les 25 derniers % tout en acceptant le fait que certaines instances valides seraient perdues (risque réduit par la présence de doublons).

Les scores tels que idf et icf ont donné de bons résultats pour mesurer la pertinence des instances quantitatives (voir Figure 15b). L'utilisation des Segment dans le score $ICF^{term}_{segment}$ a donné les meilleurs résultats. Les données expérimentales quantitatives pertinentes sont en effet présentes dans des sections spécifiques (par exemple, *Materials and Methods*), comme le reflètent les scores de pertinence lexicale. La courbe de Precision@N décline rapidement pour les



(a)



(b)

FIGURE 14 – *Precision@N* mesurées dans l'ordonnancement d'arguments symboliques par le score CD^{node}_{target}

instances quantitatives dont la précision globale est faible. Les scores de pertinence de type icf pourraient donc être utilisés pour filtrer grossièrement les résultats (en éliminant environ 75% de la population) sans risquer d'exclure trop d'informations pertinentes.

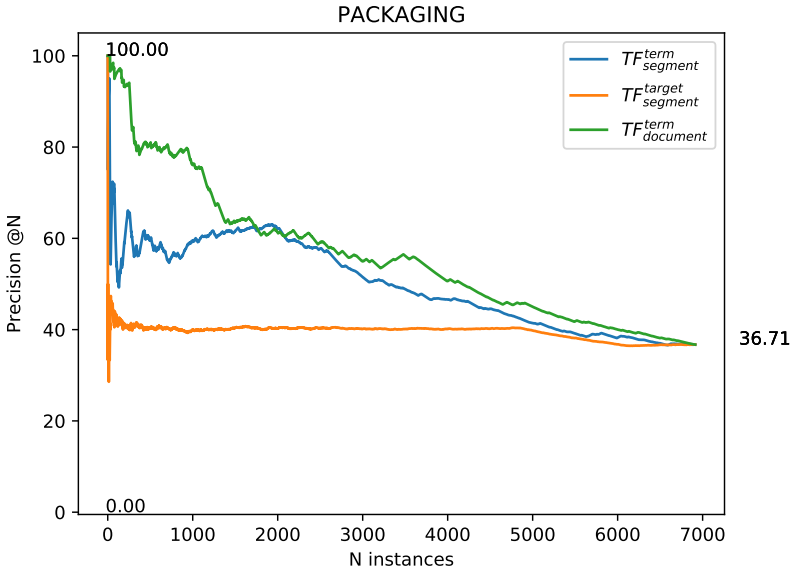
La Précision Moyenne et la R-Précision sont des standards utilisés pour représenter l'efficacité d'un système d'ordonnancement par une valeur unique. La Précision Moyenne est un standard dans la communauté TREC [Voorhees et al., 2005], et fournit la moyenne des précisions à différents niveaux de rappel. La R-Precision est la valeur de précision sur les n premiers résultats ordonnés, où n est le nombre connu de résultats pertinents. La R-Précision est considérée comme plus adaptée lorsque la proportion de faux positifs extraits est important [Manning et al., 2008]. C'est pour cela que cet indicateur sera privilégié pour évaluer l'utilité des scores de pertinence. Les tableaux 8 et 9 donnent les valeurs de précision des instances d'arguments ordonnées à l'aide des scores de pertinence. Par exemple, sur un ensemble de 100 résultats contenant 20 vrais positifs et 80 faux positifs, une méthode ordonnant parfaitement ces résultats selon leur pertinence obtiendrait une précision moyenne de .51 alors que sa R-Précision serait de 1.00. Ainsi, la R-Précision s'adapte mieux à la proportion d'informations pertinentes parmi les résultats extraits.

Target	p*	CD_t^n	TF_d^t	TF_s^t	TF_s^a	IDF_d^t	ICF_s^t	ICF_s^a
SYMBOLIC	.47	.55	**.64**	.51	.52	.53	.49	.47
Packaging	.37	.49	**.56**	.50	.40	.31	.30	.36
Component	.56	.60	**.71**	.52	.61	.70	.63	.56
Method	.16	**.41**	.18	.28	.25	.20	.25	.18
QUANTITATIVE	.14	.13	.13	.13	.14	.12	.13	.14
Permeability	.16	.16	.13	.15	.17	.14	**.21**	.15
Relative_Humidity	.28	.27	.28	.27	.33	.22	.33	**.35**
Thickness	.14	.14	.14	.14	.13	.11	.19	**.20**
Temperature	.08	.07	.08	.08	.08	.06	.05	.05

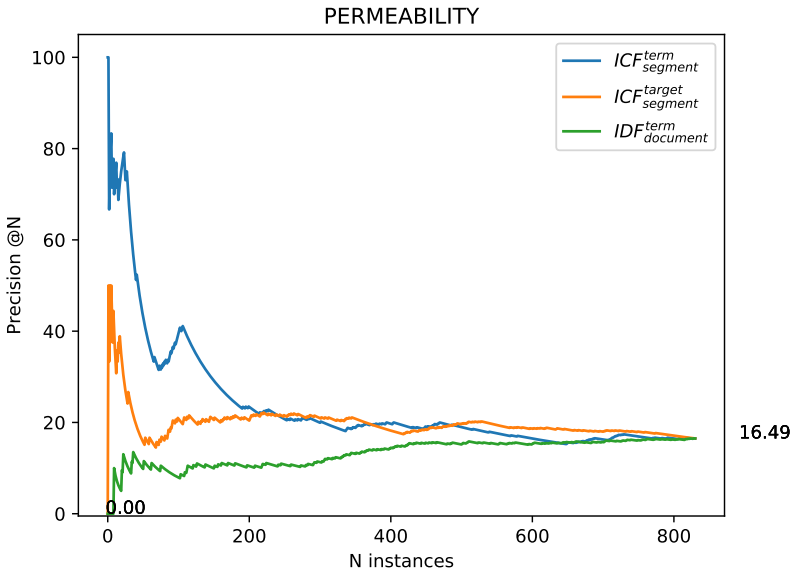
p* : baseline précision

TABLE 8 – Valeurs de Précision Moyenne des instances d'arguments ordonnées selon les scores de pertinence

Dans l'ensemble, le scores CD_{target}^{node} a révélé des améliorations des valeurs de précision pour les instances symboliques, mais pas pour les instances quantitatives. La valeur de R-Précision montre une nette amélioration de la précision des



(a)



(b)

FIGURE 15 – *Precision@N* mesurées dans l'ordonnancement d'arguments symboliques par les scores de pertinence lexicale

Target	p*	CD_a^n	TF_d^t	TF_s^t	TF_s^a	IDF_d^t	ICF_s^t	ICF_s^a
SYMBOLIC	.47	.55	**.66**	.55	.52	.45	.46	.47
Packaging	.37	.50	**.59**	.57	.41	.22	.24	.36
Component	.56	.58	**.73**	.53	.61	.63	.63	.56
Method	.16	**.64**	.17	.44	.28	.25	.31	.19
QUANTITATIVE	.14	.13	.13	.12	.17	.12	.12	.15
Permeability	.16	.17	.9	.17	.29	.07	**.30**	.10
Relative_Humidity	.28	.27	.28	.26	.31	.19	.35	**.49**
Thickness	.14	.14	.13	.12	.12	.11	**.24**	.22
Temperature	.08	.06	.08	.04	.06	.02	.04	.03

p* : baseline precision

TABLE 9 – Valeurs de R-Precision des instances d'arguments ordonnées selon les scores de pertinence

instances symboliques. La précision des instances d'arguments symboliques est égale à .47 et s'est améliorée jusqu'à .55 avec CD_{target}^{node} et jusqu'à .66 avec $TF_{document}^{term}$. Cela varie selon l'argument considéré, et est plus visible avec les instances de Method, CD_{target}^{node} ayant amélioré la précision de .16 jusqu'à .64. Cela confirme notre intuition selon laquelle CD_{target}^{node} dépend fortement de la structure de l'ontologie et n'est donc pas applicable aux instances quantitatives. Nous avons observé que les scores lexicaux employant la fréquence (i.e. tf) sont appropriés pour mesurer le score de pertinence des instances symboliques. $TF_{document}^{term}$ est bien adapté pour Packaging et Component, la raison étant que les termes liés aux instances symboliques sont le sujet principal des articles et sont donc très fréquents. Les instances de Method sont plus spécifiques à certaines sections des documents, donc $TF_{segment}^{term}$ est plus approprié.

Les instances quantitatives recherchées sont spécifiques à certaines sections des articles, mais elles ne sont pas présentes en grand nombre. L'amélioration obtenue en comparant la R-Précision à la précision précédente des instances quantitatives dépend de l'argument considéré. À l'exception des instances Température, la R-Précision a montré une amélioration significative : de .16 à .30 pour les Perméabilité et de .14 à .24 pour Thickness avec le score $ICF_{segment}^{term}$. Pour les instances de Relative_Humidity, $ICF_{segment}^{target}$ a montré une meilleure amélioration avec une précision de .28 et une R-Précision de .49 après ordonnancement. Le score de pertinence lexicale utilisant le modèle icf est donc approprié pour trier les instances quantitatives valides. Les instances de Temperature constituent

une exception remarquable, aucun des scores de pertinence ne présentant de résultats intéressants. Cela est dû à la grande proportion de faux positifs dans les résultats de l'extraction (1925 instances reconnues pour 54 valeurs de températures recherchées), mais provient également de l'absence de termes explicites sur lesquels baser le calcul d'un score : le terme *"température"* est très générique, se retrouve uniformément dans toutes les sections des documents, et n'est pas systématiquement présent aux côtés de l'unité de mesure ($°C$ n'étant pas ambigu).

2.6.3 Évaluation des combinaisons des scores

Comme les scores de pertinence lexicale sont bien adaptés aux instances symboliques, nous avons mené des expérimentations sur leur combinaison avec le score CD^{node}_{target} pour améliorer les effets de l'ordonnancement. Les combinaisons linéaires et séquentielles de $TF^{term}_{document}$ avec le score sémantique CD^{node}_{target} ont été évaluées afin de comparer leurs effets respectifs. Les exemples de ces combinaisons sont donnés avec des instances de Packaging. Cela nous a permis de combiner des scores utilisant différents types d'informations (lexicales et sémantiques) afin d'affiner les évaluations de la pertinence des instances d'arguments.

Combinaison linéaire. La combinaison linéaire de la Figure 16a est une combinaison des scores CD^{node}_{target} et $TF^{term}_{document}$ des instances de l'argument symbolique Packaging. Cette combinaison renforce ainsi la spécificité sémantique des instances par rapport à leurs fréquences dans les documents. Nous n'avons pas observé de gains significatifs en utilisant une combinaison linéaire (cf. Table 10), quelle que soit la valeur de α_i (c.f. Équation 2). Cela suggère que la combinaison linéaire n'a pas vraiment tiré parti des critères spécifiques associés aux scores combinés, mais les a plutôt équilibrés.

$\alpha_i =$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Précision Moyenne	.56	.57	.57	.57	.56	.56	.55	.54	.54
R-Precision	.58	.59	.59	.58	.58	.58	.58	.58	.58

TABLE 10 – Précisions des instances de Packaging selon $Linear(CD^{node}_{target}, TF^{term}_{document})$

Combinaison séquentielle. La combinaison séquentielle a amélioré les mesures de pertinence des instances d'arguments symboliques. La Figure 16b montre les effets de la combinaison séquentielle pour les instances de Packaging : $Sequence(CD^{node}_{target}, TF^{term}_{document})$, où CD^{node}_{target} a été utilisé pour filtrer une proportion θ des résultats avant d'utiliser $TF^{term}_{document}$ pour l'ordonnancement. Étant donné que les instances d'arguments non spécifiques peuvent être très fréquentes dans les documents (par exemple le mot *'packaging'*), cette combinaison séquentielle a permis d'obtenir un meilleur ordonnancement final que l'un ou l'autre des deux scores seuls. La Table 11 montre l'impact de différentes proportions de réordonnancement, selon différentes valeurs de θ (c.f. Section 2.4.3), sur la combinaison séquentielle des scores sémantiques et lexicaux pour l'ordonnancement des instances de Packaging. Le filtrage d'une petite partie (environ 30 à 20 %) des instances en utilisant le score sémantique avant le score lexical a permis d'améliorer l'effet des mesures de pertinence. La R-Précision de Packaging avec cette combinaison séquentielle est de .63, alors que la R-Précision des scores sémantiques et lexicaux précédemment utilisés était respectivement de .50 et .59. Cela représente une amélioration de .04, ce qui est non négligeable.

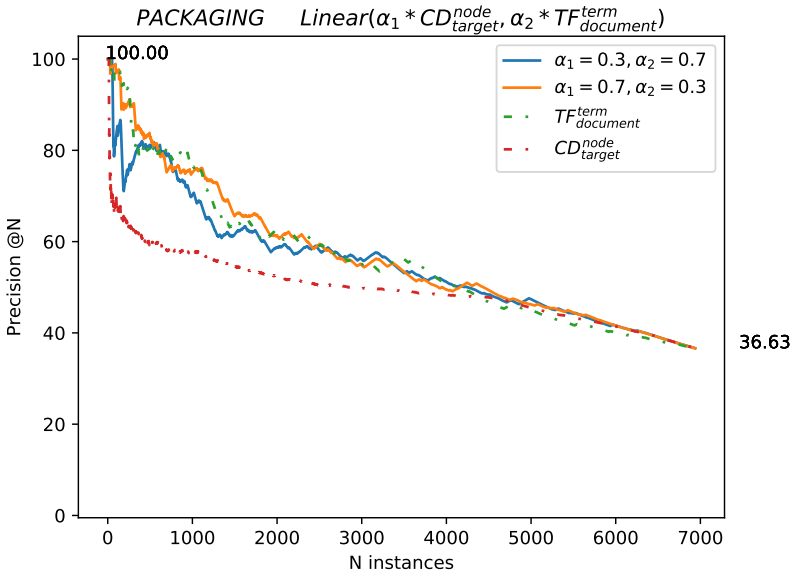
$\theta(\%) =$	10	20	30	40	50	60	70	80	90
Précision Moyenne	.51	.53	.53	.54	.55	.56	.57	.57	.57
R-Precision	.50	.50	.49	.51	.57	.61	**.63**	**.63**	.60

TABLE 11 – Précisions des instances de *Packaging* selon $Sequence(CD^{node}_{target}, TF^{term}_{document})$

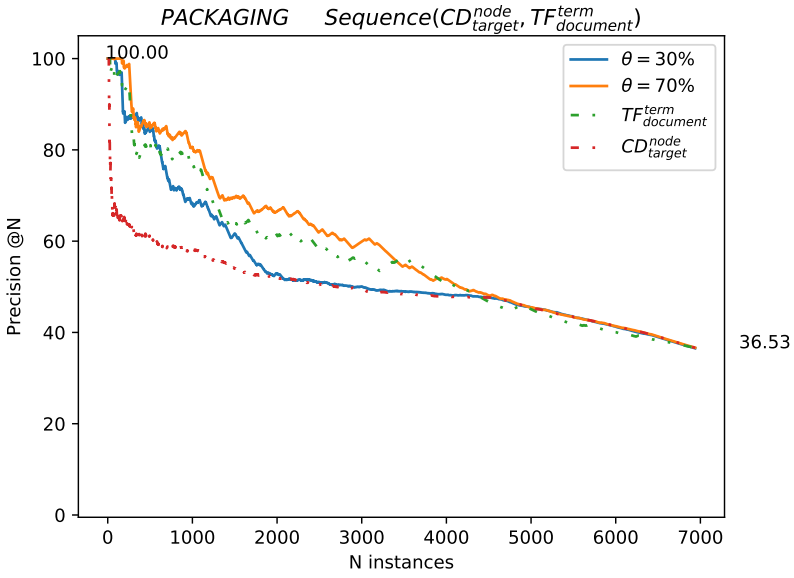
Un comportement similaire a été observé avec les instances de Component. La combinaison séquentielle $Sequence(CD^{node}_{target}, TF^{term}_{segment})$ a eu un comportement encore meilleur pour les instances de Method, confirmant que l'utilisation de segments des documents peut être plus efficace pour ces instances de concepts symboliques.

La combinaison séquentielle des scores s'est donc avérée plus adaptée que la combinaison linéaire pour les instances des arguments symboliques. Il est ainsi possible de combiner des critères de différents types (sémantique et lexical) pour mesurer la pertinence de ces instances.

Nos expérimentations n'ont pas révélé de combinaisons de scores adaptés à l'ordonnancement des instances d'arguments quantitatifs. De plus, les scores



(a)



(b)

FIGURE 16 – Combinaisons linéaires et séquentielles

sémantiques tels que CD^{node}_{target} ne sont pas adaptés à ces instances, car elles sont généralement décrites sur un petit nombre de niveaux dans l'ontologie. Nous recommandons donc de filtrer les instances des arguments quantitatifs en utilisant des scores de type icf selon les segments des documents identifiés dans les publications scientifiques. Ceci conforte l'intuition que les sections des articles ont un pouvoir discriminant important qui devrait être inclus dans le processus d'extraction des données expérimentales ciblées.

Ces scores de pertinence et leurs combinaisons pourraient présenter un comportement similaire lors d'une tâche d'extraction de données expérimentales dans un autre domaine. En effet, la méthode d'extraction basée sur la RTO présentée dans la section 2.3 est indépendante du domaine. Cette hypothèse nécessite une RTO d'un autre domaine ainsi que, si l'on désire évaluer la performance sur un autre domaine, un nouveau Gold Standard. Cela permettra d'obtenir une évaluation plus robuste de la méthode proposée.

2.6.4 Discussion

Les résultats des expérimentations présentées en Section 2.6.2 montrent que les scores de pertinence lexicaux et sémantiques peuvent être utilisés pour classer les informations expérimentales.

La méthode d'extraction appliquée aux instances d'arguments liées à la perméabilité des emballages alimentaires a conduit à une proportion élevée de faux positifs. La représentation SciPuRe des instances extraites peut être employée afin de calculer des scores de pertinence lexicale et sémantique qui trient les résultats. Les scores lexicaux de type tf en utilisant la fréquence dans les documents ou dans certains cas dans des segments textuels (c'est-à-dire des sections) des articles, ont obtenu les meilleurs résultats en termes de pertinence des instances d'arguments symboliques extraits. Le score lexical de type icf qui utilise les segments textuels des articles a obtenu les meilleurs résultats en termes de pertinence des instances d'arguments quantitatifs extraits. Des combinaisons de scores ont également été évaluées pour renforcer les effets des scores lexicaux en utilisant un score sémantique prenant en compte la spécificité du concept. La combinaison linéaire n'a donné lieu à aucune amélioration en raison de son incapacité à tirer parti des critères spécifiques qui soutiennent les différents scores. Cependant, la combinaison

séquentielle a eu des effets intéressants lorsque le score sémantique a été utilisé pour filtrer une petite proportion des résultats les moins spécifiques avant d'utiliser un score lexical pour classer les instances symboliques. Les scores de pertinence et leur combinaison peuvent donc être utilisés pour trier une partie des instances extraites, permettant ainsi de trouver des compromis entre l'exhaustivité et la validité des résultats. Les instances sélectionnées peuvent alors être proposées aux experts ou être intégrées dans des processus ultérieurs de traitement automatique de l'information.

Il existe des perspectives pour augmenter l'application du score de pertinence. Plusieurs niveaux de segments textuels (par exemple, sections, sous-sections, légendes de tableaux) pourraient être utilisés ou combinés pour étendre les scores de pertinence lexicaux fondés sur les segments des documents. Considérer, dans un score sémantique, la fréquence à laquelle les instances d'un concept apparaissent dans un texte est également un moyen courant pour mesurer la spécificité des concepts d'une ontologie [Harispe, 2014]. Des combinaisons plus complexes, comprenant plus de deux scores, et des techniques effectuant de l'apprentissage par renforcement sur les paramètres α et θ de combinaisons des scores, peuvent également constituer des pistes d'amélioration de cette approche.

Applicabilité à d'autres domaines. Notre méthode d'extraction d'instances d'arguments pourrait être appliquée à d'autres domaines expérimentaux. Elle est adaptée aux domaines dans lesquels les instances sont présentes sous des formes variées et comprennent à la fois des noms d'objets étudiés avec des variations terminologiques et des unités de mesure complexes. Elle permet également de traiter les corpus où seul un sous-ensemble des données expérimentales présentes dans les articles est jugé d'intérêt et doit être extrait.

Comme présenté en Section 2.3, le processus d'extraction d'instances d'arguments de relations n-Aires que nous avons développé s'appuie sur une RTO structurée autour de relations n-Aires entre concepts symboliques, quantitatifs et unités de mesure [Guillard et al., 2018] et formalisées dans la core-ontology. Ainsi celle-ci est facilement adaptable à d'autres domaines expérimentaux. Le changement de domain-ontology dans la RTO pour un autre domaine expérimental est la principale exigence pour appliquer notre méthode, ainsi que l'existence d'un corpus annoté pour l'évaluation. Cette RTO comprend ainsi une composante

terminologique pour chacun de ses concepts, qui est utilisée pour piloter le processus d'extraction des instances d'arguments. L'ensemble du processus d'extraction dépend fortement de la couverture de la description de la RTO, tant pour la détection de nouveaux termes ou unités de mesure que pour la désambiguïsation des concepts de quantité. L'utilisation d'une RTO de domaine est également nécessaire pour la construction de notre représentation SciPuRe, qui permet elle-même le calcul des scores de pertinence. Une discussion supplémentaire sur les critères de généricité de notre approche est présente en Section 4.2.

Une illustration d'application à un autre domaine est fournie ci-dessous. Le vocabulaire du domaine considéré est spécialisé et contient des instances d'arguments symboliques et quantitatifs qui peuvent présenter un intérêt. La bioraffinerie et l'altération des aliments par des microorganismes pathogènes sont des domaines d'application qui génèrent des données expérimentales relatives à des tâches spécifiques. Des études préliminaires ont déjà été menées sur la réutilisation de données expérimentales issues d'articles scientifiques, et annotées manuellement en utilisant une RTO, concernant l'altération des aliments [Guillard et al., 2017], la bioraffinerie [Lousteau-Cazalet et al., 2016b]. D'autres RTO existent et des jeux de données annotés manuellement suivant ces RTO existent [Fabre et al., 2020, Buche et al., 2021]. Voici un exemple d'extraction utilisant la RTO Valorcarn [15] [Roche et al., 2020] dans le domaine de l'altération des aliments carnés par des agents pathogènes. Cette RTO définie les relations n-Aires concernant les conditions de croissance des microorganismes et définit des instances d'arguments symboliques d'intérêt telles que Microorganisme ou Matrice et des instances d'arguments quantitatifs telles que Temperature ou Time. La relation de croissance microbienne décrite par cette RTO couvre 49 concepts symboliques, 10 concepts de quantité et 14 concepts d'unité.

Exemple 2 *Instances d'arguments reconnus :*
For macro-morphological observations, the isolates were three-point inoculated on (*MEA*) *medium and grown for* 7 [*days*] *at* 25 [*°C*] *in the dark. The isolates from the genera* (*Aspergillus*) *and* (*Penicillium*) *were additionally three-point inoculated ...*

Legend : (*label de concept symbolique*) [*unité de mesure*] *valeur numérique*

15. `https://ico.iate.inra.fr/atWeb/`

L'exemple 2 est un extrait de [Sonjak et al., 2011], disponible sur ScienceDirect. Deux Microorganisme peuvent être reconnus : *Aspergillus* et *Penicillium*. La Matrice est *MEA*. Comme il s'agit de l'acronyme d'un terme de trois mots, il faudrait probablement extraire une variation du label du concept pour le reconnaître. Les paramètres de contrôle, Time et Temperature sont extraits selon le processus détaillé dans la section 2.3.

Ces instances d'arguments peuvent ensuite être représentées à l'aide d'une représentation SciPuRe sur la base des descripteurs dépendants de l'ontologie, de la structure et du lexique, tel que décrit dans la Section 2.3.3. Tout comme dans le domaine des emballages alimentaires, de nombreuses informations sont présentes dans les articles. Les instances d'arguments extraites ne concernent pas systématiquement les conditions de croissance des microorganismes étudiées dans l'article. Il peut s'agir d'informations provenant d'une source externe, citées à titre de comparaison, ou d'instances d'arguments spécifiques à l'article, mais non liées aux conditions de croissance des microorganismes (par exemple, le temps et la température auxquels un organisme a été stocké avant l'expérimentation). Les scores de pertinence peuvent ensuite être calculés à l'aide des descripteurs de SciPuRe et utilisés pour trier les faux positifs comme dans la Section 2.4.

Pour conclure, notre méthode d'extraction des instances d'arguments de relations n-Aires présentée dans ce chapitre propose un pipeline fondé sur une RTO qui conceptualise les connaissances du domaine et guide l'ensemble des étapes d'extraction grâce à sa formalisation du domaine d'intérêt. Le processus d'extraction capable d'étendre le vocabulaire de la RTO et génère un ensemble de descripteurs dans une représentation SciPuRe pour chaque instance d'argument extraite. La représentation SciPuRe exploite les informations sémantiques, lexicales et structurelles pour décrire chaque instance d'argument de manière unique dans une représentation multi-descripteurs. Enfin, des combinaisons d'indicateurs de pertinences, calculés en exploitant les différents descripteurs de SciPuRe, permettent d'ordonner les instances d'arguments extraites en tenant compte de leurs contextes d'apparition dans le texte. Ces scores combinés de pertinence ont montrés des effets positifs grâce à un ordonnancement des instances d'arguments selon ces scores permettant de trier une partie des résultats faux-positifs extraits.

La Phase II, décrite dans le chapitre suivant, sera consacrée à l'utilisation des instances d'arguments issues de la Phase I pour reconstituer les instances de relations n-Aires présentent dans les documents. Cette phase débute par l'extraction des instances de relations partielles présentes dans les tableaux des articles et cherche à compléter les instances d'arguments manquantes en exploitant la structure des documents, les fréquences de cooccurrences entre instances d'arguments et leurs proximités sémantiques exprimées dans des modèles de word-embedding.

Chapitre 3

Reconstitution d'Instances de Relations n-Aires

Sommaire

Nous présentons ici la seconde étape de notre processus d'extraction de connaissances expérimentales depuis des articles scientifiques et leur structuration sous forme de relations n-Aires. Cette Phase II se consacre à la reconstitution des instances de relations n-Aires présentes dans les documents. Pour cela, notre méthode utilise les relations n-Aires partielles déjà présentes au sein des tableaux des articles scientifiques et cherche à les compléter avec les instances d'arguments issues de la Phase I. Nous avons conçu et évalué différentes approches. Une approche Structurelle, qui repose sur la proximité entre la relation partielle et les instances d'arguments pouvant la compléter ainsi que sur le ciblage de sections spécifiques pour certains arguments. Notre approche Fréquentiste recherche les cooccurrences fréquentes entre les instances d'arguments présentes dans une relation partielle et les instances d'arguments du texte qui peuvent venir compléter cette relation partielle. Enfin, l'approche par Plongements Lexicaux utilise des modèles de langages *word-embedding* afin d'obtenir une mesure d'association entre les instances d'arguments d'une relation partielle et les instances d'arguments candidates à sa complétion. Enfin, nous avons également conçu notre démarche de reconstitution des relations n-Aires dans une optique d'assistance aux experts. Ainsi, notre méthode de Phase II est en mesure de proposer plusieurs instances d'arguments pour chaque instances d'argument manquante dans une relation partielle, un expert pouvant ensuite sélectionner l'instance adéquate.

3.1 Etat de l'Art - extraction et reconstitution des relations n-Aires

3.1.1 Les relations n-Aires en extraction d'information

La notion de relation est couramment utilisée afin de représenter des connaissances dans le domaine du Web-Sémantique. Celle-ci est majoritairement basée sur le modèle *Resource Description Framework* (*RDF*) permettant d'énoncer des relations binaires (i.e. propriétés) soit entre deux ressources qui peuvent être

des instances de concepts (i.e. arguments de la relation binaire) A et B, soit entre une ressource et un littéral, au sein d'un triplet $A :: relation :: c$. La propriété de ce triplet est orientée, faisant de l'argument A le sujet de la relation (la ressource à décrire) et de B son objet, lui-même ressource (e.g. *luc* :: *aime* :: *jean*) ou littéral (e.g. *luc* :: *a_pour_taille* :: 185). Cependant lorsque l'on souhaite définir des relations entre un nombre plus important d'arguments, il est nécessaire d'utiliser des relations d'arité n nommées relations n-Aires. Le W3C définit cela comme *la manière naturelle de relier un individu à strictement plus d'un autre individu* [Giunti et al., 2019]. L'emploi du terme de relation n-Aire couvre donc toute relation impliquant au minimum trois arguments, matérialisée par un ensemble de relations binaires.

Le W3C définit différents cas d'usages des relations n-Aires :

1. L'ajout d'attributs supplémentaires à une relation binaire, comme illustré dans la Figure 17, dans laquelle on souhaite conserver un rôle prépondérant à l'un des arguments de la relation n-Aire jouant le rôle de sujet de la relation. Si l'on souhaite ajouter un argument exprimant une probabilité à une relation *christine* :: *has_diagnosis* :: *breast_tumor* alors il est nécessaire de créer deux instances de relations supplémentaires (e.g. *diagnosis_value* et *diagnosis_proba*) et un argument (e.g. *Diagnosis_Relation*) reliant les trois instances de relations. Le sujet de la relation est lié à un argument de relation (e.g. *christine* :: *has_diagnosis* :: *Diagnosis_Relation*). Cet argument est lui-même relié aux informations additionnelles (e.g. *Diagnosis_Relation* :: *diagnosis_value* :: *breast_tumor_christine* et *Diagnosis_Relation* :: *diagnosis_probability* :: *high*). L'instance de *Diagnosis_Relation* représente alors un unique objet contenant l'information des trois arguments *christine*, *breast_tumor* et *high*.

2. Une autre configuration de relation n-Aire relie des objets sans que l'un ne constitue le sujet de la relation, illustré dans la Figure 18. Dans ce cas l'instance de relation présente des liens avec chacun des arguments (e.g. *Purchase_Relation* :: *has_buyer* :: *john*, *Purchase_Relation* :: *has_seller* :: *books.example.com*, *Purchase_Relation* :: *has_object* :: *lenny_the_lion*). Chaque argument joue alors un rôle différent dans la relation sans que l'un ne ressorte comme le sujet de celle-ci. C'est ce schéma de relation qui est utilisée dans la RTO n-AryQ utilisée

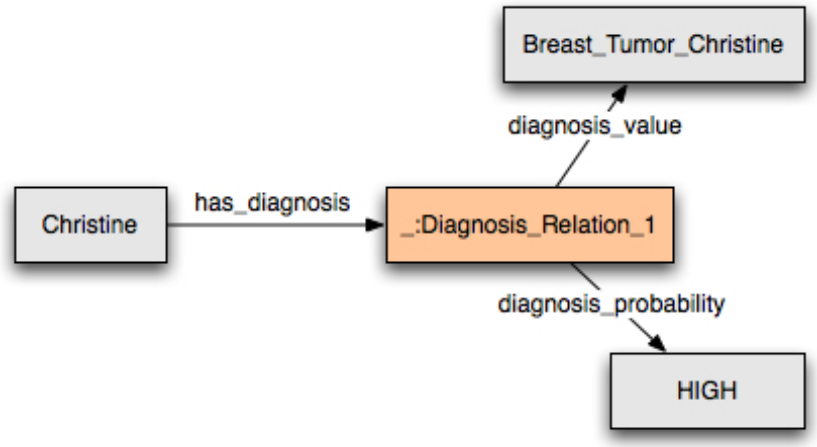


FIGURE 17 – Relation n-Aire pour : *'Christine has breast tumor with high probability'*

dans cette thèse.

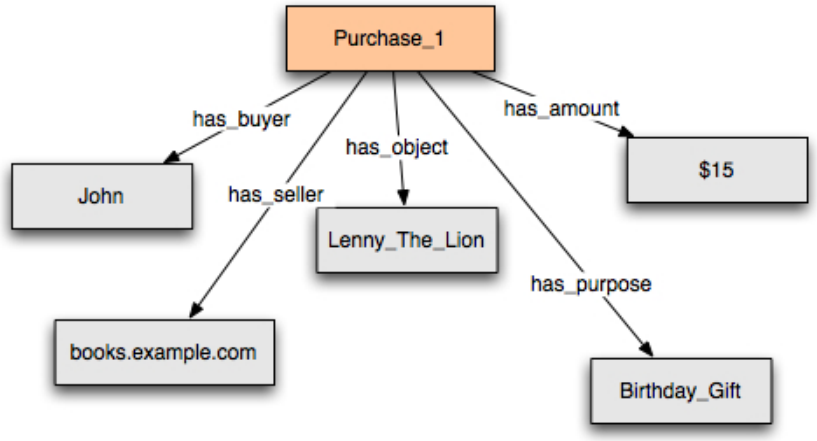


FIGURE 18 – Relation n-Aire pour : *'John buys a "Lenny the Lion" book from books.example.com for $15 as a birthday gift'*

3. Un dernier cas de relation n-Aire permet de prendre en compte une séquence d'arguments dans les relations n-Aires, lorsque les propriétés des arguments sont indistinctes, mais l'ordre de ceux-ci est important. Ce cas est illustré dans la Figure 19. L'instance de relation est alors entre un argument spécifique (e.g. $UA\_1337 :: flight\_sequence :: UA\_1337\_Relation$) et une liste ordonnée des autres arguments (e.g. $UA\_1337\_1\_Relation :: destination :: LAX$, $UA\_1337\_1\_Relation :: next\_segment :: UA\_1337\_2\_Relation$,

UA_1337_2_Relation :: *destination* :: *DFW*). Ce type de relation n-Aire est principalement employé afin de représenter une information ordonnée.

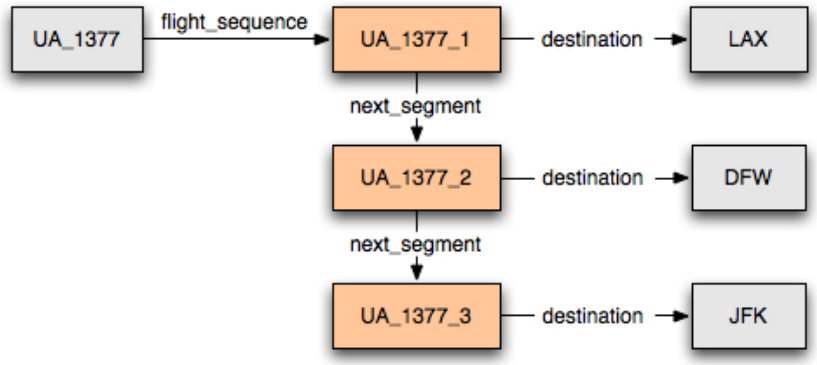


FIGURE 19 – Relation n-Aire pour : *'United Airlines flight 3177 visits the following airports : LAX, DFW, and JFK'*

Le format de relation n-Aire utilisé dans cette thèse correspond à celui illustré dans la Figure 18. En effet il n'existe pas toujours un sujet (i.e. tel que dans la Figure 17) unique lorsque l'on modélise une donnée expérimentale. Par exemple, pour la description de la croissance d'un micro-organisme dans un aliment, le-dit micro-organisme tout comme l'aliment sont éligibles à être sujets de la relation. Le cas de relation illustré dans la Figure 19 est lui très spécifique car le besoin de modéliser un ordonnancement des arguments est rare. Une instance d'un concept de relation n-Aire est liée, en RDF, à chacun des arguments grâce à un nombre n de propriétés. Celles-ci sont spécifiques à chaque argument, tel qu'illustré dans la Figure 20, et indiquent la propriété liant chaque argument à l'instance du concept de relation.

La description d'une connaissance sous forme d'une relation n-Aire peut être définie en utilisant le modèle de description RDFS, *'S'* signifiant *'Schema'*[1] [Brickley et al., 2014] et fournissant un ensemble de spécifications pour l'expression de ressources en RDF. Construit selon ces modèles de description, le langage *OWL* (*Web Ontology Language*) permet de représenter un champ de connaissances sous forme d'une ontologie structurée. Une fois l'ontologie définie, des instances de relations peuvent être ajoutées et constituer une base de connaissances afin de servir de source d'informations et/ou être utilisées par des processus avancés de raisonnement. L'alimentation en connaissances de ces bases peut être faite

1. `https://www.w3.org/TR/rdf-schema/`

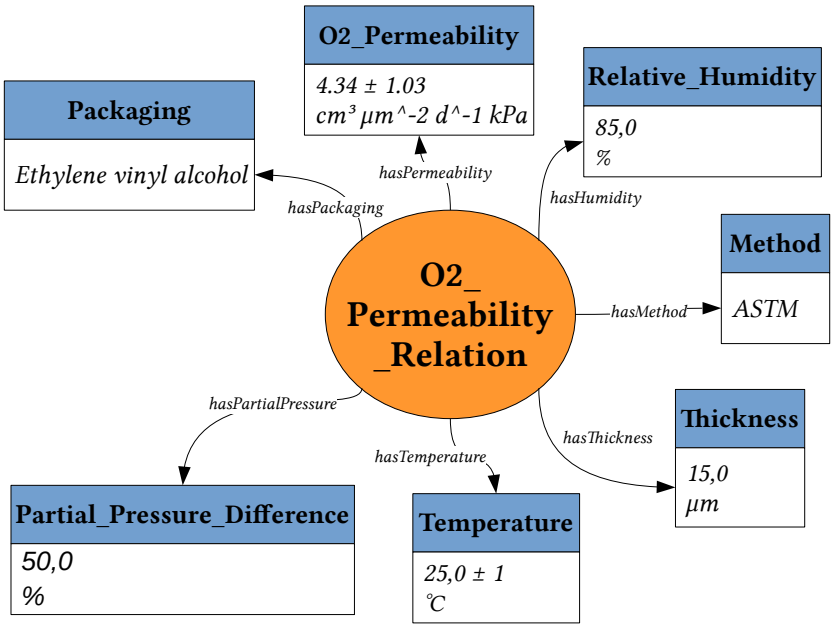


FIGURE 20 – Instance de relation n-Aire décrivant une mesure de perméabilité à l'oxygène

manuellement, le plus souvent par des experts du domaine considéré, ou assisté par des processus automatiques. L'extraction de relations depuis des textes afin de représenter de la connaissance est depuis longtemps un champ de recherche important en traitement automatique du langage [Grishman and Sundheim, 1996].

De manière générale, la notion de relation est utilisée afin de désigner les connexions qui existent entre différentes entités, notamment dans les textes. Sans nécessairement être formalisées selon les standards du web sémantique, les relations textuelles peuvent être de différentes natures : sémantiques, syntaxiques, coréférentielles etc. Leur extraction a tout d'abord consisté à reconnaître des relations binaires présentes dans une phrase grâce à un ensemble varié de techniques [Pawar et al., 2017], de l'utilisation de l'expertise linguistique humaine et de ressources externes à l'emploi de différents processus d'apprentissage. Le contexte d'extraction des relations s'est ensuite étendu, en considérant les relations inter-phrases ainsi que les coréférences [Grishman and Sundheim, 1996]. La complexité des relations à reconnaître a également augmenté en considérant un nombre d'arguments plus importants.

La reconnaissance d'événements est un exemple de relations à plusieurs arguments. La mention d'un certain événement dans un texte permet de déclencher

la détection de ses différents éléments : ses acteurs, son lieu, sa date ou toute autre information relative à celui-ci [Davidson, 1980]. La notion de relation n-Aire n'est pas nécessairement employée dans l'extraction d'événements. La notion d'événement correspond alors au cas où une relation n-Aire présente un argument principal, tel que dans la Figure 17. La nature d'un événement est généralement définie par son terme déclencheur [McGrath et al., 2011]. Par exemple, dans la phrase d'exemple de la Figure 18, la présence du verbe *'buys'* va être interprétée comme le déclencheur de la recherche d'un événement du type *Purchase*. Ce terme est utilisé afin de définir les informations à rechercher et caractériser l'événement. Les relations n-Aires ne possèdent pas de terme déclencheur, le concept de relation n'en définissant pas. Comme mentionné précédemment, les relations n-Aires ne présentent pas nécessairement un unique sujet d'intérêt mais potentiellement plusieurs.

Dans notre approche, les relations n-Aires sont définies par leur *argument résultat*, la nature de cet argument définissant celle de la relation. Par exemple, dans la Figure 20, la présence de l'argument *O2_Permeability* caractérise la relation n-Aire en *O2_Permeability_Relation*. Cependant l'argument résultat ne constitue pas le sujet de la relation, dans le cas de la relation illustrée ici, ce sujet pouvant être vu comme la valeur de perméabilité mais également l'emballage alimentaire. Malgré ces points communs entre la recherche d'événements et la recherche de relations n-Aires, le contexte de la reconnaissance d'événements est généralement limité au niveau de la phrase ou inter-phrases et à la détection de coréférences. Dans certaines recherches, la notion de relation n-Aire n'est pas toujours formulée explicitement. Il est alors question d'extraction de *'relations complexes'* [Minard et al., 2013], relations possédant de nombreux arguments, mais n'étant pas toujours formalisées selon les standards de la relation n-Aire en web sémantique.

L'extraction de relations n-Aire dans des données textuelles consiste à donc à regrouper un ensemble de n éléments (i.e. instances d'arguments) les uns avec les autres au sein d'une instance de concept relation. Pour cela il est nécessaire de définir des critères permettant de détecter l'appartenance d'instances d'arguments à une même instance de relation. Des méthodes existent pour l'extraction de relations binaires [Pawar et al., 2017]. Celles-ci se basent sur différentes approches comme l'utilisation d'analyses syntaxiques [Chan and Roth, 2011,

Bunescu and Mooney, 2005, Gamallo et al., 2012], la recherche d'associations fréquentes [Greenwood and Stevenson, 2006, Gamallo et al., 2012] ou encore de l'apprentissage automatique [Kumar, 2017, Grave, 2014]. Des ressources externes telles que des bases de connaissances ou des ontologies permettent également de définir les relations à extraire ainsi que les arguments les composants et/ou guider le processus d'extraction [Berrahou et al., 2017]. Ces ressources sont fréquemment utilisées afin de guider l'extraction des relations et, lorsque cela est un enjeu, classifier les instances d'arguments et les relations. Par exemple, [Groth et al., 2018] évalue un système pour l'extraction de toutes instances d'arguments présents conjointement aux instances d'un argument défini, afin de former des n-uplets (e.g. $Patient :: was\_treated\_with :: Emtricitabine, Etravirine, Darunavir$). L'alignement de ces n-uplets avec des schémas de relation existants est alors l'enjeu principal. La question que nous examinerons par la suite est alors celle-ci : peut-on utiliser, ou du moins adapter, les méthodes d'extraction des relations binaires pour l'extraction des relations n-Aires ?

L'hypothèse de Davidson [Davidson, 1980] présuppose qu'une relation n-Aire peut être réduite à une somme de relations binaires. C'est ainsi que fut abordée l'extraction de relations complexes dans le domaine biomédical [McDonald et al., 2005]. Une relation n-Aire à n arguments se décompose alors en $n-1$ relations binaires. En revanche, la revue de littérature de [Zhou et al., 2014] a montré que cette approche résulte en une explosion combinatoire, cela ayant un impact négatif sur les temps de calcul, mais également sur les résultats. [Zhou et al., 2014] estime expérimentalement qu'une méthode obtenant une précision de p dans l'extraction de relations binaires obtiendra alors une précision de p^{n-1}, avec n le nombre d'arguments de la relation, dans l'extraction de relations n-Aires en suivant l'hypothèse de Davidson.

Il semble donc nécessaire de dépasser les méthodes classiquement employées pour l'extraction de relations binaires et ainsi répondre aux spécificités des relations n-Aires. Cela peut être réalisé en utilisant simultanément les propriétés des relations n-Aires, et plus particulièrement des instances d'arguments présents dans les textes. Augmenter ainsi les sources d'informations employées afin de reconstituer les relations n-Aires, dans une approche multicritères et sans passer par une étape d'extraction de relations binaires à recomposer, est la piste que nous

explorerons par la suite.

3.1.2 Approches non supervisées

La premières catégorie d'approches afin d'extraire des relations entre entités dans des documents est l'approche non supervisée. Celle-ci n'emploie pas de corpus annotés pour l'apprentissage (ces corpus pouvant être employés pour la mesure des résultats de l'extraction). En l'absence d'instances d'arguments et de relations annotées, les approches non supervisées sont principalement basées sur différents systèmes de règles. Cela est spécialement nécessaire dans les domaines où les ressources annotées disponibles ne sont pas suffisantes pour déployer une méthode d'apprentissage automatique. De plus lorsque le besoin d'information n'est pas toujours précisé, c'est-à-dire que la forme et le type des relations n-Aires doivent être découverts, il peut être difficile d'employer de l'apprentissage. L'une des premières approches [Blaschke et al., 1999] se concentrait sur des relations simples (binaires) entre entités similaires dans les abstracts d'articles dans le domaine biomédical. Nommée *Protein-Protein-Interactions* (PPI), l'extraction de cette relation est une tâche standard dans le domaine biomédical [Zhou et al., 2014].

De manière générale, les travaux d'extraction des relations se basent sur deux étapes (1) la reconnaissance des entités d'intérêt, et (2) la reconstitution des relations. La possibilité d'une propagation d'erreur de (1) à (2) étant une préoccupation importante [Roth and Yih, 2004], l'identification des instances d'arguments doit donc posséder une couverture et une précision suffisante. Le point (1) est couvert dans le Chapitre 2 et le second dans ce Chapitre.

Lorsque les entités ont été reconnues dans les documents, plusieurs critères permettent de déterminer les candidats pertinents pour la mise en relation. Tout d'abord les connaissances linguistiques et l'analyse du texte des documents peuvent se révéler utiles pour identifier les liens entre entités et les règles associées. Ces règles sont constituées d'expressions régulières, ou motifs linguistiques, utilisant différents niveaux d'informations textuelles. [Proux et al., 2000] utilise par exemple des motifs construits manuellement afin de détecter la présence de deux noms de gênes proches d'un verbe d'interaction (e.g. $Gene1$ :: $interacts\_with$ :: $Gene2$). L'utilisation de règles peut également considérer l'éloignement des deux entités, un dictionnaire de verbes d'interactions ou la présence de négation

[Proux et al., 2000]. Cela se traduit alors par la construction manuelle de motifs lexico-syntaxiques, considérant l'arbre de dépendance syntaxique de la phrase afin d'y analyser le rôle de chaque terme. Cependant, la construction manuelle de règles linguistiques exhaustives est coûteuse et est très dépendante du domaine. La construction manuelle de motifs linguistiques semble techniquement adaptable à l'extraction de relations n-Aires, par exemple en les décomposant en sous-motifs pour l'extraction des relations binaires. Cependant elle nécessite de construire un nombre encore plus important de motifs linguistiques. Ces motifs requièrent généralement la présence d'un terme déclencheur de la relation (e.g. *'interract with'*, *'has a valu of'*). Cela nécessite alors la création d'un dictionnaire pour chacune des sous-relations binaires. Dans le cadre des relations n-Aires, cette approche manuelle semble bien trop coûteuse en termes de temps sans garantir d'être exhaustif.

D'autres critères que les informations linguistiques peuvent être employées afin de lier des instances d'arguments par une relation. Tout d'abord une hypothèse simple est que les instances qui cooccurrent fréquemment dans un même contexte ont une probabilité plus importante d'appartenir à une même relation [Shahab, 2017]. Cela peut être employé dans le contexte de la découverte de relations (i.e. quand les schémas de relation ne sont pas connus), ou de reconstitution de relations définies en amont. Dans ce dernier cas, une présélection des instances à associer est alors faite en suivant les modélisations des relations recherchées. Pour ce faire, plusieurs mesures sont couramment employées [Bach and Badaskar, 2007, Lenca et al., 2007]. Ces travaux considèrent alors les occurrences de deux entités (ou plus) comparées à leur nombre de cooccurrences dans un contexte défini. Ce contexte est généralement la phrase, mais il peut être d'une taille différente comme un ensemble de phrases ou encore une distance maximale entre les entités dans un texte. Les nombreuses mesures existantes ne reflétant pas les mêmes natures d'associations, les résultats obtenus diffèrent selon la mesure choisie [Lenca et al., 2007]. La mesure de la fréquence de cooccurrence [Agrawal et al., 1993] permet de déterminer les entités les plus probables d'êtres associées.

Prenons par exemple un ensemble de cinq phrases T, dont deux contiennent les instances conjointes des arguments Packaging : $LDPE$ et Thickness : $0.5\mu m$. La fréquence de cooccurrence des instances, $\{LDPE, 0.5\mu m\}$, est alors de

$\frac{|\{LDPE,0.5m\}|}{|T|} = \frac{2}{5} = 0.4$. La mesure *Lift* [Brin et al., 1997] étend cet usage en mesurant la dépendance entre les instances. Elle considère la fréquence de la cooccurrence des entités relativement à leurs scores de fréquences respectives. Ainsi, si l'instance Packaging : $LDPE$ apparaît trois fois et Thickness : $0.5\mu m$ deux fois, $lift(LDPE, 0.5\mu m) = \frac{supp(\{LDPE,0.5\mu m\})}{supp(LDPE)*supp(0.5\mu m)} = \frac{0.4}{0.6*0.4} = 1.66$. Un score supérieur à 1 indiquant la dépendance des deux instances. D'autres mesures existent afin de déterminer des règles d'associations (e.g. score de confiance d'une règle [Agrawal et al., 1993]) et calculer la corrélation entre entités (e.g. coefficient de Pearson [Pearson, 1896]). Ces mesures sont généralement employées dans le domaine de la fouille de données afin de déterminer les entités à associer ainsi que l'influence de la présence d'une entité sur une autre. Elles sont également applicables à un nombre plus important d'entités. D'autres mesures existent et sont plus spécifiquement employées en recherche d'information afin de mesurer la similarité ou l'association entre entités. La mesure de distance de Jaccard [Jaccard, 1901] est classiquement utilisée en recherche d'information. Elle est composée du rapport entre l'intersection de deux ensembles et leur union : $Jaccard(A, B) = \frac{|A \bigcap B|}{|A \bigcup B|}$. Ce score est généralement utilisé pour estimer la similarité entre deux classes d'objets, par exemple afin de déterminer la similarité de différentes instances candidates de relations [Ru et al., 2018]. L'indice de Dice [Dice, 1945] mesure également la similarité entre deux échantillons : $Dice(A, B) = \frac{2*|A \bigcap B|}{|A|+|B|}$. Celui-ci est également utilisé en recherche d'information, par exemple afin de mesurer l'association entre deux entités, ou entre une entité et un terme déclencheur d'événement [Sun, 2009]. La mesure de *Point-wise Mutual Information* (PMI) [Church and Hanks, 1990] permet de mesurer la coïncidence de deux instances d'arguments selon leurs distributions individuelles : $PMI(a, b) = log\frac{p(a,b)}{p(a)*p(b)}$. Celle-ci peut être utilisée afin de déterminer des scores d'associations entre instances d'arguments et ainsi former des relations [Cafarella et al., 2005, Yates et al., 2007]. Il a été démontré que la mesure de PMI a tendance à privilégier les associations rares, mais fortes [Role and Nadif, 2011]. Ces différentes approches, reposant sur la fréquence des cooccurrences d'instances d'arguments pour opérer une mise en relation, sont adaptables à l'extraction des relations n-Aires.

D'autres approches considèrent les informations sémantiques afin de déterminer les liens entre instances d'arguments. Des premiers travaux [Ghersedine et al., 2012] utilisent les propriétés de certaines relations n-Aires afin de guider l'identification de sous-relations binaires, suivant l'hypothèse de

Davidson, en contexte local (i.e. phrase) et de les recombiner. Il s'agit alors d'identifier une instance d'argument 'pivot' sur la base de sa cooccurrence fréquente avec une instance d'argument résultat (argument définissant la relation). [Ghersedine et al., 2012] a en effet constaté que la recombinaison de relations n-Aires basées sur des sous-relations binaires donne de meilleurs résultats, lorsque faite autour d'un seul argument, ici nommé argument pivot. Les instances d'autres arguments apparaissant alors avec cet argument pivot forment les relations binaires à recombiner pour obtenir la relation. Cependant, s'appuyer uniquement sur les associations entre instances d'arguments et instances de l'argument pivot limite les associations pouvant d'être détectées, d'autant plus lorsque le nombre d'arguments (i.e. l'arité de la relation) et d'instances dans le texte est important.

L'analyse des coréférences permet également de déterminer les instances d'arguments similaires présentes dans des contextes différents. Cette tâche consiste à détecter la référence à une même entité à travers des termes différents (e.g. *'Jean'* et *'garçon'* dans *'Jean va à l'école. Ce petit garçon est sage en classe.*). Les approches classiques pour résoudre ce problème [Elango, 2005] consistent par exemple à analyser l'arbre syntaxique d'une phrase selon l'algorithme de Hobbs [Hobbs, 1978], détectant l'usage de pronoms. Les travaux récents [Kobayashi and Ng, 2020] font appel à des ensembles variés d'approches. Si l'on omet les travaux nécessitant de l'apprentissage, l'utilisation de règles linguistiques jointes à l'utilisation de réseaux lexicaux tels que WordNet [Miller, 1995] et des connaissances qu'ils contiennent (e.g. liens de synonymie stricte) est une méthode classique pour la détection de coréférences. Une ressource ontologique semble pouvoir répondre à cette tâche en exploitant les différents labels existant dans un même contexte lexical, ainsi qu'en considérant les propriétés liant les concepts génériques et spécifiques [Ferré et al., 2020]. La détection de coréférences permet d'établir des liens entre des instances d'arguments et ainsi établir une équivalence entre instances apparaissant dans des contextes différents.

Toutes les approches non supervisées présentées ici sont actuellement rarement utilisées seules, mais plutôt complétées par l'utilisation de ressources externes, servant par exemple à définir les relations et leurs possibles arguments. Ceci concerne également la méthode que nous développons. Nous présentons ci-dessous les approches supervisées puis l'utilité de ressources externes dans les tâches d'extraction de relations.

3.1.3 Approches supervisées

Les approches supervisées utilisent des corpus de données annotées afin de procéder à l'apprentissage d'un modèle. Ces apprentissages peuvent ensuite être utilisés afin de prédire les instances d'arguments appartenant à des relations. Il existe plusieurs techniques permettant de faire de l'apprentissage sur différents critères.

Motifs. Pour commencer, les motifs linguistiques construits manuellement dans les approches non supervisées pour extraire les relations peuvent être appris automatiquement [Meng and Morioka, 2015]. La similarité des contextes d'apparition des arguments et relations est généralement captée grâce à l'utilisation de patterns considérant les informations lexicales séquentielles et syntaxiques [Lin and Pantel, 2001]. La recherche de ces patrons englobe également différentes approches en fusionnant les critères relatifs aux cooccurrences, au contexte lexical ou aux relations syntaxiques [Charnois et al., 2009]. De l'information sémantique peut également être incluse grâce à l'utilisation de ressources externes labélisant sémantiquement les termes importants dans les textes [Berrahou et al., 2017]. Ces motifs peuvent ensuite être utilisés afin de découvrir les associations entre instances d'arguments dans d'autres documents n'ayant pas étés annotés. La limitation principale de cette méthode d'apprentissage est le nombre important de motifs qu'elle génère. Il devient alors nécessaire de définir des critères permettant de sélectionner un sous-ensemble des motifs offrant le meilleur compromis entre exhaustivité et validité des relations qu'ils extraient.

Classifieurs. La reconstitution des relations peut également être envisagée comme un problème de classification. À partir de l'ensemble des instances d'arguments, les possibles relations n-Aires sont alors décomposées en relations binaires candidates. Il s'agit donc de classifier chaque relation selon un ensemble de descripteurs. Ces descripteurs peuvent être constitués d'informations lexicales (i.e. le texte des instances d'arguments), syntaxiques (étiquetage grammatical et arbre de dépendances syntaxiques) et sémantiques (rôle des instances d'arguments dans la relation). [Björne et al., 2009] note par exemple que les descripteurs reposant sur l'analyse de l'arbre de dépendance syntaxique sont essentiels afin de correctement capturer les instances de sous-relations binaires. Au contraire, [Móra et al., 2009] et

[McGrath et al., 2011] constatent que l'information lexicale et sémantique associée à la paire d'instances d'arguments et au terme déclencheur sont les descripteurs les plus à même de discriminer les sous-relations binaires candidates. Il est ainsi essentiel de posséder une description des instances d'arguments présentant des informations adaptées aux techniques employées pour leur mise en relation.

Deep Learning. L'apprentissage de l'extraction de relations dans les textes utilise également des réseaux de neurones artificiels pour l'apprentissage profond (i.e. deep learning). Différents modèles de réseaux de neurones existent et permettent deux apports dans l'extraction de relations. Tout d'abord il est possible d'utiliser le deep learning afin d'entraîner un modèle pouvant procéder à l'extraction ou à la classification de relations. Tout comme l'utilisation de classifieurs, cela nécessite de disposer d'une quantité suffisante de ressources textuelles annotées et de déterminer les descripteurs pertinents à considérer dans la tâche d'apprentissage. Il est également possible d'utiliser des modèles préexistants, ayant été entraînés sur de larges corpus de données, et permettant de réaliser différentes tâches de traitement automatique du langage utiles dans la détection de relations (i.e. transfer learning).

L'utilisation de deep learning s'est développée durant les dernières années pour de nombreux domaines, dont le traitement automatique du langage [Torregrossa et al., 2021]. Le deep learning donne de bons résultats dans les tâches de classification, dont la classification de relations binaires. Ces modèles utilisent de l'apprentissage supervisé sur des données présentant des relations entre entités : textes annotés avec des informations lexicales, syntaxiques et sémantiques et/ou des relations structurées au sein de bases de connaissances [Kumar, 2017]. Différents modèles de réseaux de neurones sont utilisés, les meilleurs résultats pour l'extraction de relations n-Aires étant actuellement obtenus par les réseaux de la famille des *long short-term memory* (LSTM) [Kumar, 2017]. Au contraire des réseaux de neurones récurrents (RNN), les LSTM possèdent des connexions de rétroaction entre neurones de couches différentes. Cela permet de réduire les phénomènes de dissipation ou de surcharge d'une information dans le réseau de neurones.

L'entraînement de réseaux de neurones pour extraire des relations n-Aires suivant ces approches permet d'extraire les sous-relations binaires avant de

recomposer les relations. Afin de recombiner les relations partielles trouvées dans un contexte donné, [Akimoto et al., 2019] propose de décomposer les schémas de relations n-Aires en sous-relations unaires et binaires pour apprendre à les recomposer. Des poids sont ainsi appris dans un réseau de type bi-LSTM pour chaque sous-relation et une fonction d'agrégation permet de valider ou non la relation finale recomposée. Mais différentes stratégies incluant des informations spécifiques ont été testées afin d'entraîner des réseaux de neurones à extraire des relations n-Aires permettant d'améliorer ou d'éviter cette tâche de recomposition des relations.

[Quirk and Poon, 2016] élabore une représentation du document sous forme de graphe. Chaque phrase est d'abord représentée comme un graphe, les termes (sous forme de vecteurs) étant les nœuds et les arêtes étant créés entre termes adjacents et selon les dépendances syntaxiques. Ce modèle est étendu au niveau du document en liant les graphes phrastiques par de nouvelles arêtes entre les racines syntaxiques de phrases successives. Cela permet d'entraîner un réseau LSTM à extraire simultanément l'ensemble des instances d'argument d'une relation. En effet, les arguments étant dispersés dans les documents, cette représentation permet de le considérer dans son intégralité. [Peng et al., 2017] divise cette représentation en deux sous-graphes directionnels, ne conservant dans l'un que les arêtes allant dans le sens de lecture du document et l'autre dans le sens inverse. Cependant cela entraîne une perte d'information, notamment en divisant l'arbre de dépendance syntaxique. Afin de remédier à cela, [Song et al., 2018] ne sépare pas le graphe-document en graphes directionnels, mais lui ajoute une fonction propageant l'information contextuelle portée par chaque vecteur de mot.

Une autre stratégie [Jia et al., 2019] est de considérer plusieurs niveaux de représentation des données textuelles comme données d'apprentissage pour l'extraction de relations n-Aires. Ces niveaux de représentations doivent pallier l'explosion combinatoire qui résulte de la combinaison de toutes les instances d'arguments ou relations partielles issues de contextes réduits. Dans chaque unité de discours définie (e.g. phrase, paragraphe, document) des n-uplets sont formés par l'ensemble des instances d'argument co-occurrentes. Un réseau LSTM est utilisé afin d'obtenir des vecteurs des sous-relations formées par ces n-uplets. Un opérateur d'agrégation regroupe ensuite chaque vecteur avec les instances d'arguments correspondant. L'ensemble des représentations agrégées

d'un document sont ensuite concaténées afin de prédire la relation finale. Ses expérimentations dans le domaine biomédical (i.e. oncologie) [Jia et al., 2019] ont montré que chacun des différents niveaux de représentation augmente significativement le rappel sans dégrader la précision. La considération d'un document à différents niveaux est donc une information utile pour la mise en relation des instances d'arguments.

Word Embeddings. Des avancés récentes dans le domaine du traitement automatique du langage ont vu le développement de modèles de langages utilisant largement les techniques de *word embedding*. De nombreux modèles existent et sont construits selon différentes approches : *continuous bag-of-words* (CBOW), visant à entraîner un réseau de neurones à prédire un terme selon son contexte [Mikolov et al., 2013a] et *Skip-Gram*, visant à prédire le contexte d'un mot en entrée [Mikolov et al., 2013b]. Entraînés sur de larges corpus de données, ils permettent d'obtenir un modèle de langage représentant un nombre important de mots dans des vecteurs à hautes dimensions.

Ces modèles peuvent ensuite être utilisés dans différentes tâches de TAL : traitement lexical et syntaxique (tokenisation, lemmatisation, étiquetage morphosyntaxique, analyse de dépendances) ou sémantique (détection de négation, analyse de polarité). Ces modèles peuvent donc être utilisés afin d'améliorer les tâches nécessaires à l'extraction de relations, en fournissant des informations lexicales et syntaxiques. De larges corpus existent pour l'entraînement et l'évaluation de ces modèles [Wang et al., 2018] et permettent d'obtenir des modèles pré-entraînés pouvant être appliqués sur d'autres documents. Leurs représentations vectorielles sont également plus denses, avec un nombre de dimensions moins important, que celles obtenues par une approche sac de mots et permettent alors un meilleur calcul de distances sémantiques entre instances d'arguments [Mikolov et al., 2013a, Mikolov et al., 2013b]. Les avancés les plus récentes dans le domaine du word embedding utilisent des modèles de type Transformers obtenus par des méthodes Seq2Seq (Sequence-to-Sequence). Ces modèles reposent sur des mécanismes d'attention afin de déterminer les termes les plus importants à considérer en entrée permettant de prédire le plus précisément le mot en sortie [Vaswani et al., 2017]. Les modèles de langage suivant cette méthode d'apprentissage sont actuellement les plus performants dans un ensemble

varié de tâches. Ils peuvent ainsi être utilisés afin de compléter les tâches de reconstitution de relations en apportant de l'information grammaticale, syntaxique ou sémantique.

Les tâches d'apprentissage sont généralement réalisées sur des corpus larges, mais génériques (e.g. blog, journaux en ligne, articles Wikipédia) permettant ainsi de modéliser le langage courant, ou au contraire spécialisé (e.g. articles scientifiques) afin d'obtenir des modèles répondant aux spécificités d'un domaine. Cela pose alors la question du transfert de l'apprentissage : un modèle appris sur un corpus peut-il être réutilisé tel quel sur des données concernant un domaine différent ?

Lors de l'utilisation de modèles de langage, la proximité des domaines est essentielle pour permettre ce transfert, comme démontré par [Peng et al., 2019]. Leur travail, *Biomedical Language Un-derstanding Evaluation* (BLUE), a consisté à évaluer les performances de trois modèles (ELMo et BioBERT, entrainés sur un corpus PubMed, et des versions de BERT ré-entraîné sur différents corpus) sur dix corpus différents et dans cinq tâches classiques de TAL (similarité de phrases, reconnaissance d'entités nommées, extraction de relations binaires, classification de documents et tâches d'inférences). Leurs résultats montrent l'importance de la robustesse du modèle pour la stabilité des résultats sur les différents corpus, ELMo ayant généralement des résultats plus faibles que BioBERT [Peng et al., 2019]. Les différentes versions de BERT mettent en évidence l'importance de la couverture du corpus d'apprentissage en termes de domaines. Les meilleurs résultats sont obtenus par un modèle réduit ré-entraîné sur des corpus différents. Le modèle plus large de BERT ré-entraîné obtient de meilleurs résultats uniquement dans les tâches de classification de documents et d'extraction de relations binaires [Peng et al., 2019]. Cela tend à démontrer qu'un transfert d'apprentissage considéré comme efficace est également dépendant de la tâche que l'on considère. En effet les techniques efficaces pour extraire les relations entre certains types d'entités peuvent se révéler inopérantes avec d'autres entités [Shahab, 2017]. La reconnaissance d'entités nommées en domaine spécialisé requiert des connaissances spécifiques au domaine, alors que l'extraction de relation repose également sur une connaissance générique du langage. Les modèles pré-entraînés sur des domaines scientifiques peuvent être employés en l'état sur notre corpus, mais avec certaines limitations dues aux différences de domaines, afin de fournir des critères de mise en relation.

Apprentissage semi-supervisés et supervision distante. Dans le cas où les données annotées disponibles sont rares, d'autres stratégies semi-supervisées peuvent être adoptées. Il s'agit alors d'utiliser un nombre d'exemples réduits, et de bonne qualité, afin d'amorcer la tâche d'apprentissage [Zhu and Goldberg, 2009]. Il est par exemple possible d'obtenir quelques motifs linguistiques à partir d'un ensemble d'exemples réduits de phrases [Chen et al., 2006]. Une fonction de similarité permet ensuite de déterminer les motifs à conserver parmi ceux obtenus depuis un ensemble de données plus larges, mais non annotées. Dans cette même catégorie, l'apprentissage actif consiste à sélectionner des cas devant être évalués. Son principe repose sur l'interaction avec un oracle, ensemble réduit de données étiquetées utilisé pour générer des exemples [Zhang et al., 2012] ou expert humain, et permet de maximiser le gain d'informations.

Une autre approche est d'automatiser le processus d'annotation des données nécessaires à l'apprentissage. Nommée supervision distante [Mintz et al., 2009], elle repose sur l'utilisation d'un outil permettant d'annoter automatiquement une large quantité de données. En revanche cela entraîne la création de bruit dans les données annotées, résultant en une détérioration des résultats. Ces différentes approches permettent d'utiliser de l'apprentissage sur un jeu de données annotées réduit ou inexistant et sont parfois utilisées dans l'extraction et la classification de relations binaires [Quirk and Poon, 2017, Zeng et al., 2015]. Cependant aucune d'elles n'a encore été largement évaluée dans le cadre des relations n-Aires. De plus, comme nous l'avons vu dans la Section 2.4, nombre d'entités détectées automatiquement ne constituent pas des instances d'arguments valides, mais des faux positifs a priori indistinguables des instances composant les relations n-Aires (e.g. une valeur de température n'est pas toujours la température de contrôle utilisée dans une relation de perméabilité).

Les processus supervisés et d'apprentissage nécessitent donc encore de larges corpus de données afin d'être performants. Ceux-ci sont particulièrement coûteux à obtenir. Cette limitation est encore plus présente dans les domaines spécialisés. Les domaines de spécialité requièrent également des experts pour réaliser l'annotation, ou a minima des annotateurs familiarisés avec le domaine. Dans le cadre des relations n-Aires, la plus grande arité de celles-ci et la dispersion des arguments dans les documents augmentent encore la complexité des tâches d'annotation.

3.1.4 Apport des ressources externes

L'utilisation de ressources externes peut être utile pour une grande variété de tâches dans l'extraction des relations n-Aires. La nécessité de disposer de corpus annotés, de grande taille et de bonne qualité est une évidence pour les tâches d'apprentissage. En dehors de cela, ces ressources permettent généralement d'améliorer l'extraction des arguments et leur mise en relation. Les modèles de langages issus de techniques de word embedding améliorent par exemple diverses tâches de TAL, comme les traitements lexico-syntaxiques ou l'identification d'entités nommées. Les ressources fournissant un vocabulaire permettent également de cibler les instances d'arguments à extraire. De plus amples détails sur ce premier point sont disponibles dans la Section 2.2.

Mais des ressources peuvent être employées spécifiquement pour les tâches d'extraction des relations. Le point essentiel dans l'extraction de relations est la formulation des relations d'intérêts : définir les arguments composant la relation ainsi que les liens de propriétés. Une ressource définissant cela peut prendre différentes formes : ontologie, thésaurus ou base de données.

L'utilisation d'une ontologie est courante dans le domaine de l'extraction de connaissance [Nédellec and Nazarenko, 2006] et un standard en Web sémantique afin de structurer l'information. Une ontologie permet par exemple de définir les connaissances d'intérêt sous forme de relations n-Aires. En plus de fournir un vocabulaire utile pour l'identification des instances d'arguments, les propriétés liant les différents concepts sont également utiles. Un cas typique est décrit dans [McDowell and Cafarella, 2006b], où les propriétés d'héritage permettent de considérer les instances textuelles des sous-concepts d'un concept d'intérêt et de les catégoriser sous un même argument. L'utilisation d'une ontologie permet aussi de guider la construction de motifs syntaxiques [Li et al., 2015] en contrôlant la formation de ceux-ci (e.g. un patron pour une relation $Person$ :: $wedding$ :: $Person$ doit contenir deux unités lexicales reconnues comme concepts de $Person$ et un terme déclencheur relatif à $wedding$). D'autres approches peuvent utiliser des ressources externes pour ajouter un niveau de représentation sémantique aux données textuelles [Ramadier and Lafourcade, 2016]. [Berrahou et al., 2017] s'appuie par exemple sur une RTO de domaine afin de représenter l'information sémantique dans les textes

et inclure ces informations sémantiques dans l'apprentissage de motifs. D'autres travaux n'adjoignent pas l'utilisation d'une ressource externe aux textes, mais utilisent des bases de données structurées afin d'en apprendre des 'Schémas Universels' de relations [Rosenfeld and Feldman, 2006].

Comme mentionné précédemment, d'autres ressources tels que les modèles de langage sont utilisés afin d'améliorer les tâches nécessaires à l'extraction de relations n-Aires. Il est par exemple possible d'utiliser le modèle BERT afin de labéliser les termes d'une phrase et ainsi améliorer l'entraînement de classifieurs pour les termes déclencheur, les instances d'arguments contenus par les phrases et les relations qu'ils entretiennent [Yang et al., 2019]. [Yu et al., 2020] procède similairement, mais greffe au modèle BERT une base de connaissance externe pour labéliser les arguments, tout en pondérant leur mise en relation selon leur distance.

L'utilisation de ressources externes reste donc généralement une nécessité dans l'extraction de relations, particulièrement en domaine de spécialité comme le nôtre. Le besoin de définir les schémas de relations augmente avec l'arité des relations visées et l'ambiguïté de leurs arguments.

3.1.5 Délimitation de la thèse

L'extraction de relations n-Aires en domaine de spécialité exploite un ensemble varié de techniques, listées dans la Table 12. La majorité de celles-ci emploient des ressources externes, a minima pour définir les schémas des relations à extraire, ou pour guider la reconnaissance des instances d'arguments et la formation des instances de relations. Les techniques récentes obtenant les meilleurs résultats [Peng et al., 2017, Song et al., 2018, Jia et al., 2019] emploient de l'apprentissage basé sur l'utilisation de réseaux de neurones de type LSTM. Les spécificités de notre travail, tel que l'absence de corpus d'apprentissage suffisant, limitent potentiellement leurs utilisations.

Notre travail consiste à extraire les connaissances relatives aux données expérimentales et leurs conditions opératoires. Ces connaissances sont formalisées sous forme de relations n-Aires dans une Ressource Termino-Ontologique. Le vocabulaire de cette ressource a été particulièrement utile lors de la reconnaissance des instances d'arguments dans les textes.

Approches Non Supervisée	Exemples de travaux
Motifs lexicaux manuels	[Proux et al., 2000]
+ syntaxe	[Elango, 2005, Kobayashi and Ng, 2020]
+ modèle sémantique	[Miller, 1995, Ferré et al., 2020]
Règles sémantiques	[Ghersedine et al., 2012, Ghersedine et al., 2012]
Calcul de cooccurrence	[Ru et al., 2018, Cafarella et al., 2005]
	[Yates et al., 2007]
Calcul de similarités	[Shahab, 2017, Sun, 2009]
Domain Transfert	[Shahab, 2017]

Approches Supervisée	Exemples de travaux
Extractions de motifs lexicaux fréquents	[Charnois et al., 2009]
+ syntaxe	[Meng and Morioka, 2015]
	[Lin and Pantel, 2001]
+ sémantique	[Berrahou et al., 2017]
	[Ramadier and Lafourcade, 2016]
	[Li et al., 2015]
Classifieurs de relations candidates	[McGrath et al., 2011]
	[Móra et al., 2009]
	[Björne et al., 2009]
+ modèle Word Embedding	[Yang et al., 2019]
Recomposition par Deep Learning	[Akimoto et al., 2019]
	[Quirk and Poon, 2016]
	[Song et al., 2018, Jia et al., 2019]
	[Peng et al., 2017]
+ modèle Word Embedding	[Yu et al., 2020]
Extraction de motifs Semi-supervisée	[Chen et al., 2006]
Apprentissage actif	[Zhang et al., 2012]
Apprentissage sur base de données	[Rosenfeld and Feldman, 2006]

TABLE 12 – Familles de méthodes existantes pour l'extraction de relations n-Aires

Les travaux de la littérature abordent la question complexe de l'extraction des relations n-Aires selon plusieurs angles. Tout d'abord une partie se consacre à la question de la classification de relations. Cette question n'est pas abordée dans nos travaux, car les propriétés liant les arguments sont clairement définies dans notre RTO. De plus il n'y a pas d'ambiguïté possible sur les propriétés de ces arguments (i.e. un argument donné ne possède qu'une de propriété le liant à la relation). Cette question de la classification des propriétés est abordée dans la littérature principalement dans le cadre de la reconnaissance d'événements.

Généralement, les travaux sur l'extraction de relations n-Aire se consacrent à l'extraction de relations possédant une arité de trois ou quatre arguments et dans un contexte réduit. Cela est particulièrement visible dans les nombreuses études visant à l'extraction de connaissances dans le domaine biomédical [Bravo et al., 2015, Zhou et al., 2014, Shahab, 2017]. Le corpus de ces études est généralement composé uniquement d'abstracts et les travaux visent à extraire les interactions relatives à une procédure médicale (e.g. $Gene - Disease - Drug$). Les relations n-Aire représentant des connaissances expérimentales possèdent généralement une arité plus grande que les relations classiquement abordées dans la littérature, telle que les mesures de perméabilité considérées dans notre travail (7 arguments). De plus, nous considérons l'extraction à l'échelle d'un article scientifique en entier. Cela nous rapproche d'une taille de document déjà considérée dans quelques travaux sur l'extraction des relations n-Aires [Quirk and Poon, 2016, Peng et al., 2017, Song et al., 2018, Jia et al., 2019], mais l'extraction d'instances de relations à l'échelle d'un document reste moins abordée dans la littérature.

La taille des documents influence l'espace de recherche dans lequel il sera nécessaire d'effectuer l'extraction. Un nombre plus important d'instances d'arguments est alors disponible, ce qui augmente les combinaisons possibles lors de la reconstitution des instances de relations [Zhou et al., 2014]. Il ne s'agit alors plus seulement de trouver la bonne instance d'argument à ajouter dans une relation, mais de détecter l'instance de relation spécifique devant accueillir cette instance d'argument. La possibilité qu'une même information soit présente à différents endroits nécessite également un travail de détection des coréférences. De plus les informations d'intérêts sont présentes dans des contextes hétérogènes : tableaux, légendes, sections et sous-sections. Cela constitue une information structurelle pouvant être mise à contribution. Dans un grand nombre de publications scientifiques, une part importante des informations est structurée et présentée au sein de tableaux. Ainsi le Gold Standard de relations utilisé dans cette thèse, et présenté en Section 3.3.2, est constitué de 1096 instances d'arguments regroupés dans 204 instances de relations. Parmi les arguments de ces 204 relations, 412 instances sont présents dans les tableaux des articles ($\approx 38\%$). Il est également attendu que les différentes sections des articles scientifiques contiennent des informations différentes.

Si la taille des documents est admise comme ayant une influence sur les résultats à travers les facteurs mentionnés précédemment [Zhou et al., 2014], l'utilisation de la structure des documents n'est généralement pas un facteur déterminant utilisé pour la reconstitution des relations. Quelques travaux mettent cette information structurelle à contribution : [Jia et al., 2019] considère par exemple une représentation du texte différente pour chaque niveau (i.e. phrase, paragraphe, document) et les utilise pour la mise en relation des arguments. Habituellement, les travaux considérant la structure des documents analysent alors les différences dans l'information portée par les sections des documents [Cohen et al., 2010, Kando, 1997] et l'impact de la prise en compte des différentes sections des articles scientifiques sur des résultats d'extraction [Shah et al., 2003, Hofmann et al., 2009].

Le travail de reconstitution des relations n-Aires présenté dans cette thèse en Section 3.2 s'appuie alors sur les hypothèses suivantes :

— l'identification des instances d'argument fondé sur l'utilisation d'une RTO, présenté en Section 2.1, possède une couverture suffisante.

— les informations contenues au sein des tableaux des articles scientifiques sont pertinentes en tant qu'instances d'arguments. De plus, la structuration des tableaux permet de déduire automatiquement les liens entre ces arguments [Hignette et al., 2009, Buche et al., 2013b]. Ces instances d'arguments constituent ainsi des relations n-Aires partielles, mais de bonne qualité car fondées sur la structuration par les tableaux des informations. Les identifier peut permettre de les utiliser comme base pour la reconstitution de relations complètes. De plus amples détails sur cette étape sont présentés en Section 3.3.

— les instances d'arguments présents dans les tableaux possèdent également des manifestations dans le texte de documents. Celles-ci peuvent être directes (i.e. présence de la valeur de l'instance) ou indirectes (i.e. présence d'un terme se référant à l'argument). Davantage de détails sur ce point sont présentés en Section 2.1.

— la détection de liens entre les manifestations des instances d'arguments des tableaux et les instances d'arguments des textes permet d'associer les instances d'arguments du texte aux relations n-Aires partielles.

Les sections suivantes présentent plus en détail les étapes de prétraitements

essentiels à la reconstitution des relations n-Aires, ainsi que les trois approches choisies pour leur applicabilité à notre approche. Ces approches se fondent sur l'utilisation de la structure propre aux articles scientifiques, l'analyse des cooccurrences indiquant l'association entre instances d'arguments et leurs proximités sémantiques mesurées par des modèles de word embedding entraînés sur des corpus d'articles scientifiques.

3.1.6 Approches des relations n-Aires par interconnexion

La majorité des travaux sur l'extraction de relations n-Aires se situent dans l'hypothèse de Davidson [Davidson, 1980] stipulant qu'une relation n-Aire peut être vue comme une chaîne de relations binaires entre ses arguments. Une relation entre n arguments résultant alors en une chaîne de $n-1$ relations binaires. Les travaux à l'état de l'art ont montré que cette approche résultait en une diminution de la précision des résultats [Zhou et al., 2014]. De plus, les relations n-Aires articulant un ensemble d'arguments autour d'un concept de relation ne possédant pas d'instance indépendamment des instances d'arguments, déterminer la chaîne de relations binaires correspondant à une relation n-Aire peut être difficile, car il existe un nombre important de possibilités. Des travaux présentant des caractéristiques similaires à l'extraction de relation n-Aires, comme l'extraction d'événements, recherche un argument pivot [Ghersedine et al., 2012] afin de le substituer au concept de relation et détecter les relations binaires entre une instance de cet argument pivot et les autres instances d'arguments. Nous faisons ici l'hypothèse que considérer l'ensemble de toutes les relations possibles peut aider à améliorer l'extraction d'une instance de relation n-Aire.

Notre méthode d'extraction des relations n-Aires exploite les informations présentes dans les tableaux des articles. Celles-ci constituent des relations n-Aires partielles et nous cherchons alors à les compléter avec des instances d'arguments issues du texte (cf. Section 3.2). Notre approche étend ainsi l'hypothèse de Davidson [Davidson, 1980], qui stipule que les relations n-Aires peuvent être décomposées en une suite de relations binaires, leur chaîne formant la relation n-Aire. Nous recherchons l'ensemble maximum de relations binaires pouvant être détectées entre les instances d'arguments de la relation n-Aire partielle et les instances d'arguments issues des textes. Cela forme alors une interconnexion

de relations binaires, tel que illustré dans la La Figure 21. Cette approche est originale au vu de l'état de l'art, aucuns travaux ne reposant explicitement sur cette hypothèse (cf. Hypothèse 3) à notre connaissance. Nous ne nous reposons alors pas sur des relations spécifiques entre arguments (classiquement les relations binaires formant la chaîne de relations entre arguments doivent être choisies ou définies par l'utilisateur ou par une méthode d'apprentissage). Cette méthode est également adaptée à notre travail de complétion des relations, plusieurs instances d'argument pouvant être candidates à la complétion d'une relation n-Aire partielle, la mesure de l'interconnexion entre chaque candidat permet de les discriminer. En revanche cette approche augmente la combinatoire et donc le nombre de calculs nécessaires. Nous mettons en œuvre notre approche dans une démarche itérative et globale et qui intègre des critères sémantiques, à travers notamment l'utilisation de la représentation SciPuRe, et statistiques.

Définition 3 *L'interconnexion des relations entre l'ensemble des arguments d'une relation n-Aire partielle et l'ensemble des arguments manquant est composé de toutes les relations binaires pouvant être détectées entre les arguments de chacun de ces deux ensembles. Pour n arguments, le nombre de relations binaires formant cette l'interconnexion de relation est donc dans :* $[m; n * m]$*, avec n le nombre d'arguments présents dans la relation partielle et m le nombre d'arguments manquants à celle-ci.*

Il est important de préciser que l'ordre dans lequel les instances candidates sont ajoutées aux relations partielles n'est pas pris en compte ici. En effet le calcul des scores discriminant les candidats n'est fait qu'en considérant uniquement les instances d'arguments présentes dans la version d'origine de la relation partielle, et non pas d'éventuelles instances d'arguments ayant été ajoutés lors d'une itération précédente. Ce choix a été fait afin de minimiser les risques de propagation d'erreur. Il serait néanmoins possible d'élaborer une approche utilisant les instances déjà ajoutées à une relation afin d'aider à la découverte de nouveaux candidats valides. Ce point est discuté plus en détail dans la Section 4.3.

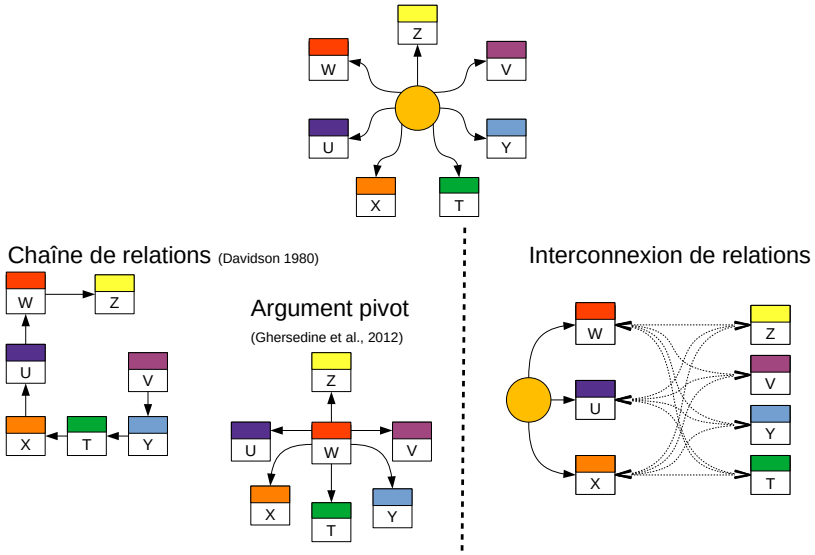


FIGURE 21 – Décomposition d'une relation n-Aire en chaîne et ensemble de relations binaires

3.2 Méthodologie - extraction et reconstitution des relations n-Aires

L'objectif de cette seconde Phase est de reconstituer les instances de relations n-Aires, concernant les connaissances expérimentales présentes dans des articles scientifiques. Ces instances de relations sont constituées d'un ensemble d'arguments regroupés autour d'un concept de relation. Les instances des arguments composant les instances de relations sont présentes dans les documents. Cependant l'appartenance d'une instance d'argument à une instance de relation n'est pas explicitée dans les textes. De plus, si certaines instances d'arguments n'appartiennent qu'à une instance de relation (e.g. la valeur de perméabilité), d'autres instances d'arguments peuvent appartenir à des relations différentes (e.g. un même emballage dans des relations de perméabilité à l'oxygène, au dioxyde de carbone et au monoxyde de dihydrogène) ou à des instances différentes d'une même relation (e.g. une même température comme paramètre de contrôle de différentes instances de différentes relations).

3.2.1 Approche de reconstitution des relations n-Aires

Dans chaque document, les instances d'arguments sont dispersées au sein des différentes sections et sous-sections. Les articles scientifiques contiennent également des figures et des tableaux contenant de l'information. Les instances d'arguments présentes dans les tableaux peuvent être reliées grâce à la structure de ces dernières. Les relations n-Aires formées par les informations des tableaux constituent des instances de relations partielles, certaines des instances des arguments définis dans le schéma de la relation étant manquant (e.g. la relation en haut à gauche dans la Figure 22). Pour reconstituer des instances de relations n-Aires complètes, nous nous appuyons sur ces instances de relations partielles issues des tableaux. En effet, les auteurs font apparaître dans les tableaux l'essentiel de l'information recherchée, essentielle dans le sens de l'information la plus importante et pertinente. L'objectif de notre travail est donc de rechercher les instances d'arguments issues des textes pouvant venir compléter celles-ci (e.g. les instances en bas à gauche dans la Figure 22). Ces instances d'arguments sont issues de la Phase I et possèdent une représentation SciPuRe.

Notre travail de reconstitution d'une instance de relations n-Aire comporte différentes étapes illustrées dans la Figure 22 :

(A) **Construction de l'ensemble des candidats.** Un ensemble d'instances candidates est construit pour chaque argument manquant dans la relation partielle. La RTO indiquant la structure des relations n-Aires, des ensembles sont formés à partir des instances d'arguments issues du document. Chacun de ces ensembles forme ainsi un ensemble de candidats pouvant compléter un argument de la relation partielle. Par exemple, si une relation partielle issue d'un tableau ne possède pas d'instance de l'argument Température, alors l'ensemble de candidats est constitué des instances d'arguments Température issues du même document. Comme illustré dans la Figure 22, nous souhaitons compléter l'instance d'argument manquante présente en vert dans la relation partielle. L'ensemble des instances sélectionnées est donc $\{D, E, H, E'\}$, ce sont les candidats à la complétion de la relation. Ce processus est détaillé dans la Section 3.2.3.

(B) **Fusion des doublons.** Au sein d'un ensemble des candidats, les instances considérées comme similaires sont fusionnées. Le nombre d'occurrences d'un candidat peut alors constituer une information utilisable par le travail

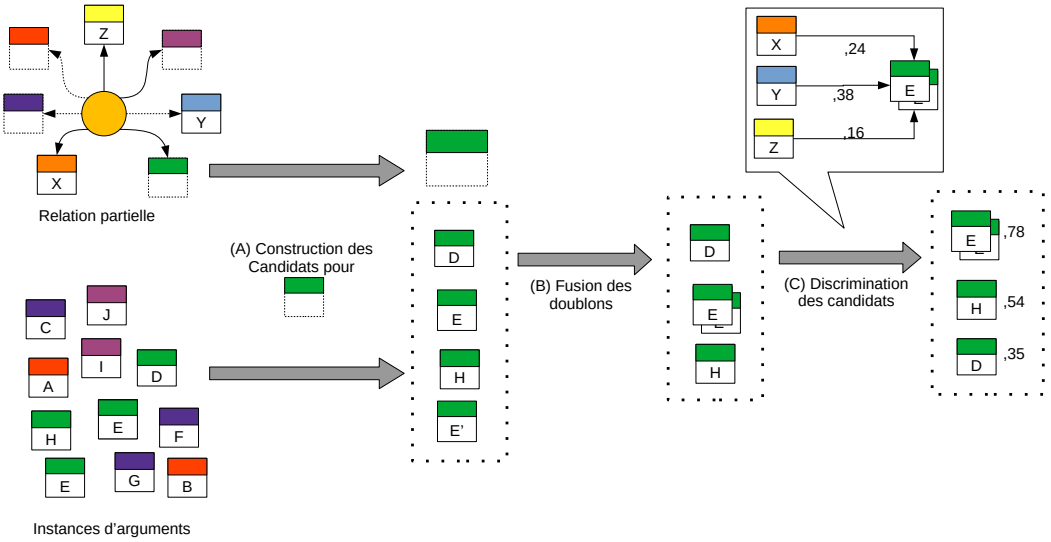


FIGURE 22 – Processus de reconstitution des relations n-Aires

de discrimination. Par exemple, les instances symboliques possédant des descripteurs lexicaux *'LDPE'* et *'low density polyethylene'* se réfèrent au même concept. Cela est indiqué par les descripteurs Node dans leurs représentations SciPuRe. Ces instances sont donc fusionnées. Cette étape est décrite plus en détail dans la Section 3.2.4.

(C) **Discrimination des candidats.** Afin de discriminer les candidats, un score est attribué à chacun et permet de décider de celui à joindre à la relation partielle. Ce score dépend de l'association mesurée entre le candidat et les arguments existants dans la relation partielle. Nous avons adapté et étendu différentes mesures de qualité à notre problématique, dans des approches Structurelles, Fréquentistes et par Plongements Lexicaux. Ces approches sont décrites en Section 3.2.5.

Les étapes (A), (B) et (C) sont répétées pour chaque argument manquant de la relation partielle.

3.2.2 Représentation des relations partielles issues des tableaux

Les tableaux des articles scientifiques possédant des informations structurées, et celles-ci pouvant être extraites sous forme de relations n-Aires partielles,

nous cherchons à les compléter avec des instances d'arguments issues du texte. La méthode d'extraction des relations partielles [Hignette et al., 2009, Buche et al., 2013b] à partir des tableaux est détaillée dans la Section 3.3. Les instances de relations partielles issues des tableaux présentent différentes caractéristiques. Celles-ci sont regroupées au sein d'une représentation des relations décrite dans la Table 13, conçue sur le modèle de SciPuRe. Cette représentation, nommée STaRe (Scientific Table Representation), regroupe les descripteurs décrivant les instances d'arguments faisant partie de la relation n-Aires. STaRe comprend les descripteurs ontologiques et lexicaux des instances d'arguments ainsi que les descripteurs ontologiques et structurels de l'instance de relation n-Aire partielle. STaRe factorise ainsi les descripteurs structurels des instances d'arguments tout en indiquant le contexte d'apparition de la relation partielle et sa formalisation dans l'ontologie.

Les descripteurs de la représentation STaRe de la relation partielle décrivent cette relation à travers deux types de descripteurs :

— **Descripteurs Ontologiques** : Ces descripteurs classifient la relation selon le formalisme de la RTO. Le descripteur *Relation* indique la relation n-Aire ciblée dans la RTO (e.g. H2O_Permeability_Relation, Impact_Factor_Component). Les arguments composants la relation n-Aire sont décrits par deux descripteurs ontologiques. Result_Argument indique l'argument résultat de la relation dans la RTO, l'argument caractérisant la relation. Les valeurs du descripteur Arguments donne le reste des arguments composant la relation n-Aire. Ces Arguments possèdent eux-mêmes des descripteurs obtenus durant le processus d'extraction :

  (i) la première valeur indique l'argument concerné (e.g. H2O_Permeability, Temperature), ce qui correspond au descripteur ontologique Target de la représentation SciPuRe des instances d'arguments.

  (ii) Une liste de valeur est ensuite indiquée, comprenant les valeurs de descripteurs ontologiques, Node (le concept spécifique de l'argument), et lexicaux, Original_Value (le contenu de la cellule) et Attached_Value (le contenu de l'en-tête du tableau).

— **Descripteurs Structurels** : Cette représentation détaille également des informations relatives au contexte du tableau et à sa place dans le document. Table donne le titre du tableau et Caption indique sa légende. Segment

décrit la section et la sous-section desquelles le tableau a été extrait. Enfin, Document donne le titre de l'article d'origine et DOI le référence de manière unique.

	Descripteur	Valeur
ONTOLOGIQUE	Relation	`H2O_Permeability_Relation`
	Result_Argument	***SciPure***
	Arguments	Target: `H2O_Perm.` / Node: `H2O_Perm.` / Original_Value: $1.27 * 10^2$ $cm^3mm^{-2}s^{-1}bar$ / Attached_Value: *Water Perm.* Target: `Packaging` / Node: `Chitosan` / Original_Value: *Chitosan films* / Attached_Value: *Chitosan films* Target: `Method` / Node: ∅ / Original_Value: ∅ / Attached_Value: ∅ Target: `R_H` / Node: ∅ / Original_Value: ∅ / Attached_Value: ∅ Target: `Temperature` / Node: `Temperature` / Original_Value: $25°C$ / Attached_Value: *Temp. (°C)* Target: `Thickness` / Node: ∅ / Original_Value: ∅ / Attached_Value: ∅
STRUCT.	Table	*Table 3*
	Caption	*Water permeability of tested packaging at* $25°C$
	Segment	*Results and Discussion*
	Document	*Barrier properties of chitosan coated polyethylene*
	DOI	*10.1016/j.memsci.2012.02.037*

TABLE 13 – Exemple d'instance de représentation STaRe d'instance de relation n-Aire partielle

Cette représentation indique les arguments manquants dans l'instance de la relation (e.g. Temperature = ∅). Cela permet de définir les ensembles d'instances d'arguments candidates parmi les instances d'arguments textuelles disponibles dans le document. La représentation des relations partielles est également exploitée afin de détecter des manifestations de ses arguments dans les textes. La valeur de perméabilité donnée dans un tableau est par exemple également citée dans le texte. Ces manifestations constituent un critère utile à l'établissement de liens entre une instance d'argument candidate et les arguments existants dans les relations partielles. Ces deux points sont détaillés dans les sections suivantes.

Certains des descripteurs de la représentation STaRe sont partagés avec la représentation SciPuRe des instances d'arguments extraits lors de la Phase I. Par exemple, le texte des légendes des tableaux étant fouillé durant la phase I à la recherche d'instances d'arguments, une instance reconnue ainsi partagerait alors ses descripteurs Segment et Document avec la représentation d'une relation partielle et son descripteur Sentence correspondrait à Caption. Le corpus des instances de relations n-Aires partielles extraites des tableaux et utilisé comme Gold Standard est disponible en ligne [Lentschat et al., 2021d].

3.2.3 Constitution des ensembles de candidats (A)

La première étape pour compléter les relations partielles issues des tableaux consiste à construire les ensembles d'instances d'arguments candidates parmi les instances d'arguments issues de la Phase I. Pour cela nous utilisons les correspondances entre les descripteurs de la représentation SciPuRe des instances d'arguments issus de la Phase I et les descripteurs de la représentation STaRe des relations partielles issues des tableaux. Pour chaque instance d'argument non renseignée dans une relation partielle, un ensemble de candidats est formé en sélectionnant les instances textuelles de cet argument issues du même document. Cela est fait sur la base des descripteurs Argument et Document de SciPuRe et de la représentation des relations partielles. Les codes de l'ensemble de la Phase II sont disponibles en ligne [2].

Exemple 3 *Une relation partielle a été extraite à partir d'un tableau de l'article "Packaging design for potato chips". Celle-ci est partielle, car les valeurs des arguments Temperature et Relative_Humidity sont absentes. Deux ensembles de candidats sont donc formés : le premier en sélectionnant les instances de la Phase I ayant comme descripteurs Argument == "Temperature" et Document == "Packaging design for potato chips", et le second les instances avec les descripteurs Argument == "Relative_Humidity" et Document == "Packaging design for potato chips".*

3.2.4 Fusion des doublons (B)

Chaque ensemble de candidats peut contenir des doublons. Une même valeur numérique constituant une instance d'argument quantitatif peut par exemple être présente à plusieurs endroits dans un même document. Un argument symbolique peut également être présent sous différentes formes à travers des variations terminologiques. Il est utile d'identifier ces cas. En effet il est inutile de différentier des instances portant la même information. De plus des méthodes d'associations fréquentistes peuvent utiliser l'information constituée par la mesure du nombre d'occurrences d'un même candidat.

2. https ://github.com/Eskode/ARTEXT4LOD

Pour chaque ensemble de candidats, nous cherchons donc à les regrouper selon les valeurs de leurs descripteurs. Les codes de la Phase II sont disponibles en ligne [3]). Ce dédoublonnage est réalisé sur des critères différents selon les arguments considérés. En effet l'équivalence de deux instances ne peut pas être établie sur les mêmes critères selon les types d'arguments considérés. Les informations issues d'un même document et constituant des instances d'un même argument (i.e. descripteur Target) sont considérées comme des doublons selon des critères différents :

— Les instances d'arguments symboliques sont fusionnées sur un critère sémantique en utilisant un descripteur ontologique de SciPuRe. Le descripteur Node permet de déterminer les instances d'arguments symboliques appartenant au même concept spécifique dans la RTo et sont donc considérés comme des doublons.

— Les instances quantitatives sont fusionnées selon des critères lexicaux, en utilisant les descripteurs lexicaux de SciPuRe. Nous considérons alors que les instances présentant le même descripteur Original_Value, i.e. la même valeur numérique et unité de mesure, portent la même information et constituent donc des doublons.

Exemple 4 *Dans le premier ensemble de candidats formé précédemment, des instances de "Temperature" possèdent différentes valeurs numériques : { 25°C; 20°C; 15°C; 25°C .... }. Certaines de ces valeurs sont identiques. Nous rassemblons les candidats équivalents en un seul : { [25°C, 25°C]; 20°C; 15°C; .... }. Ceux-ci partageront alors le même score d'association dans l'étape suivante, (C), de discrimination des candidats.*

3.2.5 Discrimination des candidats (C)

Ici nous présentons deux méthodes endogènes (dépendant de ressources internes) et une méthode hybride, exogène (dépendant de ressources externes) et endogène, permettant de discriminer les candidats. Ces méthodes constituent le cœur de la Phase 2, notre contribution à la reconstitution des relations n-Aires. Les trois méthodes élaborées est expérimentées sont introduites ci-dessous puis décrites dans la suite de cette section. Ces méthodes sont évaluées en Section 3.4.

3. https ://github.com/Eskode/ARTEXT4LOD

La première méthode est la méthode Structurelle (endogène). Elle constitue une approche volontairement simple, employée comme *baseline* afin d'observer l'apport des deux autres approches plus élaborées. Elle se fonde sur l'hypothèse que, dans un article scientifique, les informations complétant les données présentes dans un tableau sont localisées à proximité de celle-ci. Dans chaque lot de candidats, elle sélectionne donc simplement le plus proche du tableau. Les méthodes Fréquentistes et par Plongements Lexicaux que nous proposons reposent sur des mesures d'associations entre chacun des candidats et les arguments présents dans la relation partielle à compléter. La méthode Fréquentiste se fonde sur l'hypothèse que les instances d'arguments des relations partielles possèdent des manifestations présentes dans le texte des documents. La mesure des cooccurrences entre ces manifestations et les candidats permet alors de calculer des valeurs d'associations. La méthode par Plongements Lexicaux utilise des modèles de langages obtenus selon des techniques de word-embedding. Ces modèles représentent les termes sous forme de vecteurs à hautes dimensions. La mesure d'une distance cosinus entre deux vecteurs de mots permet ainsi d'estimer la similarité entre les termes [Sitikhu et al., 2019]. Les modèles de word-embedding permettent alors, en utilisant les descripteurs lexicaux des instances d'arguments candidates et des instances d'arguments présentent dans la relation partielle, de mesurer des valeurs de similarité entre leurs termes qui soient utilisables comme critères de discriminations.

Méthode Structurelle. La méthode Structurelle pour la mesure de l'association d'une instance d'argument candidate et des instances d'arguments de la relation partielle repose sur la structure des articles scientifiques. Il a été largement montré que différentes sections des articles scientifiques contiennent différentes informations [Hofmann et al., 2009, Shah et al., 2003, Kando, 1997, Cohen et al., 2010]. Plus de détails sur l'impact de la structure des articles scientifiques dans le domaine de l'extraction d'information sont présentés dans la Section 2.2. Certains scores de pertinence employés dans la Phase I (cf. Chapitre 2) utilisent cette structure. L'intuition de cette approche est donc que les meilleurs candidats pour compléter les relations partielles se situent dans les mêmes sections que les tableaux dont ces dernières sont issues. De plus les instances candidates de certains types d'arguments ont une plus grande probabilité d'être pertinentes dans certaines sections que dans d'autres (e.g. les meilleurs candidats à venir compléter

un argument Method vont se situer dans la section *'Materials and Methods'*).

Des traces de la structure des articles sont conservées dans la représentation SciPuRe des instances d'arguments ainsi que dans la représentation des instances des relations partielles via des descripteurs structurels. Ces descripteurs localisent les instances dans les documents. La mesure d'association structurelle recherche, dans chaque ensemble de candidats, le plus proche du tableau dont on souhaite compléter les informations. Pour cela elle s'appuie les descripteurs des relations partielles et de SciPuRe.

Il s'agit alors dans chaque ensemble de candidats de rechercher le sous-ensemble minimal, possédant au moins un candidat, dans le contexte le plus réduit possible. Cette recherche débute par la légende contenue dans le tableau en comparant le descripteur Caption et le descripteur Sentence de SciPuRe. Elle passe ensuite au niveau des segments, d'abord les sous-sections puis la section. Cela est fait en comparant les descripteurs Segment des instances candidates et de la représentation de la relation partielle à compléter. Les sections proches sont ensuite considérées, toujours en exploitant les descripteurs Segment, en recherchant la section la plus proche possible de celle où est située le tableau. Cela est répété jusqu'à ce que le document entier ait été considéré. Ce processus est stoppé au moment où au minimum un candidat est présent dans le contexte sélectionné. Dans la situation où plusieurs candidats sont présents dans le sous-ensemble minimal ainsi constitué, c'est la distance minimale en nombre de tokens avec le tableau contenant la relation (i.e. entre le descripteur Attached_Value de chacun des candidats et le descripteur Table de la relation partielle) qui permet de discriminer les candidats restants. Les codes de la Phase II sont disponibles en ligne [4]).

Exemple 5 *Parmi les candidats de l'ensemble Temperature, l'instance ayant comme valeur 20°C est présente dans la légende du tableau. Cela nous est indiqué, car le descripteur de la représentation SciPuRe Sentence de ce candidat correspond au descripteur Caption de la représentation de la relation partielle. Considérée comme la plus proche, cette instance constitue alors le meilleur candidat.*

Une variante de cette méthode repose sur les associations constatées entre arguments et sections spécifiques. En effet dans les articles des sections différentes

4. https ://github.com/Eskode/ARTEXT4LOD

portent des informations de natures différentes. Cela apparaît par exemple lors de l'extraction automatique de mots-clés [Shah et al., 2003], où l'exploration de différentes sections donne des résultats différents, mais également dans les tâches d'extraction d'entités nommées dans le domaine médical, les résultats variant significativement selon les sections [Cohen et al., 2010]. Ainsi dans notre domaine, les données de contrôle vont être plus présentes dans la section *Materials and Methods* alors que les valeurs de perméabilités mesurées seront dans *Results and Discussion*. Afin de cibler la recherche de certains arguments dans des sections, il est possible de manuellement définir des priorités en se référant à une table d'association argument/section. Cette table, dont un exemple est présenté dans la Table 14, indique quelles sections des articles scientifiques sont à examiner en premier lieu lorsque l'on recherche des instances d'arguments spécifiques. La méthode Structurelle Guidée recherche alors parmi les candidats ceux dont le descripteur SciPuRe Segment possède la plus forte priorité selon l'ordre indiqué dans la Table 14, priorité établie en collaboration par les trois des trois annotateurs de la Section 2.5. Les critères précédemment indiqués de la méthode structurelle s'appliquent ensuite si plusieurs candidats se situent dans la section prioritaire de l'argument.

`Argument`	`Segment`				
	abstract	Intro.	Mat. & Meth.	Res. & Dis.	Conc.
Permeability	2	4		1	3
Packaging	3	1	2		4
Method	3	2	1	4	
Partial_Pressure			1		
Relative_Humidity	3	2	1		4
Temperature	3	2	1		4
Thickness	3	2	1		4

TABLE 14 – Priorités de recherche des Argument selon les Segments pour la méthode Structurelle Guidée

Exemple 6 *Parmi les candidats de l'ensemble Temperature, nous recherchons donc les instances se situant en premier lieu dans la section "Materials and Methods", en suivant les valeurs des descripteurs Segment de SciPuRe. Si aucun candidat n'est trouvé nous regarderons ensuite dans la section "Introduction", puis dans "Abstract", etc. Dès qu'au moins un candidat est retenu, ou que tous les*

Segment associés à l'Argument ont été parcourus, nous appliquons les critères classiques de l'approche structurelle afin de discriminer les candidats restants.

La deuxième méthode de reconstitution des relations n-Aires que nous proposons utilise les fréquences de cooccurrences entre les instances d'arguments candidates à compléter la relation et les manifestations des arguments de la relation partielle dans le texte.

Méthode Fréquentiste. La méthode Fréquentiste est fondée sur des mesures d'associations entre un candidat, et ses doublons, avec les manifestations d'arguments d'une relation partielle. Les manifestations d'arguments d'une relation partielle correspondent aux instances présentent dans le texte d'un document et se rapportant aux instances présentent dans les tableaux, selon la Définition 4 et comme illustré dans l'Exemple 7.

Définition 4 *Une manifestation d'arguments est la coréférence entre une instance d'argument présente dans un tableau et une instance d'argument présente dans le texte d'un document. La manifestation d'une instance d'argument d'une relation partielle est établie en comparant la valeur des descripteurs Argument et Document de sa représentation aux descripteurs SciPuRe des instances extraites du texte.*

De plus nous définissons une manifestation comme directe ou indirecte :

- *une manifestation est considérée comme directe lorsque les valeurs des descripteurs Original_Value correspondent.*
- *une manifestation est considérée comme indirecte lorsque les valeurs des descripteurs Node, pour les arguments symboliques, ou Attached_Value, pour les arguments quantitatifs, correspondent.*

Exemple 7 *Une relation partielle dans un tableau présente une instance de l'argument Temperature. Son descripteur Original_Value, la valeur de la cellule, est de [25, °C]. Son descripteur Attached_Value, l'en-tête de la colonne, est ['temperature']. Ainsi, toute présence de ces termes dans le texte est considérée comme des manifestations directes ou indirectes de cette instance Temperature dans le document.*

Les cooccurrences entre les instances candidates et les manifestations des arguments de la relation partielle sont employées pour mesurer les associations dans l'approche Fréquentiste. Dans la recherche des cooccurrences $candidat - manifestation$, nous pouvons considérer sa nature directe ou indirecte ainsi que différentes tailles du contexte (par exemple, phrase ou section) établissant cette cooccurrence. Ainsi, dans une configuration établissant une cooccurrence, nous nommons w_m le descripteur lexical SciPuRe de la manifestation, avec $w_m \in \{Original\_Value; Attached\_Value\}$. Le contexte est défini par w_c, le descripteur SciPuRe structurel de la manifestation, avec $w_c \in \{Sentence, Window, Segment, Document\}$

De nombreuses mesures existent afin de mesurer les cooccurrences entre entités. Nous avons choisi d'en évaluer trois : Dice [Dice, 1945], Jaccard [Jaccard, 1901] et Point-wise Mutual Information (PMI) [Church and Hanks, 1990]. Ces mesures ont été choisies pour leurs simplicités et leur usage courant en extraction d'information. Dice et Jaccard sont des mesurent de l'association d'entités lexicales via leur cooccurrence [Ru et al., 2018, Sun, 2009, Cafarella et al., 2005, Yates et al., 2007]. Ces mesures sont également employées dans d'autres domaines, comme l'écologie, afin de mesurer la similarité de deux échantillons [Jaccard, 1901, Dice, 1945]. PMI privilégie les cooccurrences rares, mais fortes (i.e. nombreuses cooccurrences relativement aux probabilités d'occurrences seules) [Role and Nadif, 2011]. Afin de mesurer l'association entre un candidat et une manifestation d'un argument existant dans la relation, les ensembles à comparer sont composés pour l'un des doublons de ce candidat et l'autre des instances de la manifestation de l'argument existant dans la relation. Les formules de ces mesures sont détaillées en Section 3.1.2.

Nous avons évalué l'effet de la taille du contexte de cooccurrence w_c ainsi que celui des différentes manifestations w_m dans les mesures utilisées dans la Section 3.4.2. Une illustration du calcul d'un score d'association de Dice entre un candidat et différentes manifestations d'une instance d'argument d'une relation partielle, dans différents contextes, est donnée dans la Figure 23.

Le processus calculant l'association entre les instances d'un candidat et les manifestations des arguments de la relation partielle nécessite de dénombrer leurs cooccurrences. Le processus de dénombrement utilise pour cela différents descripteurs de SciPuRe. Les descripteurs structurels de SciPuRe indiquent les

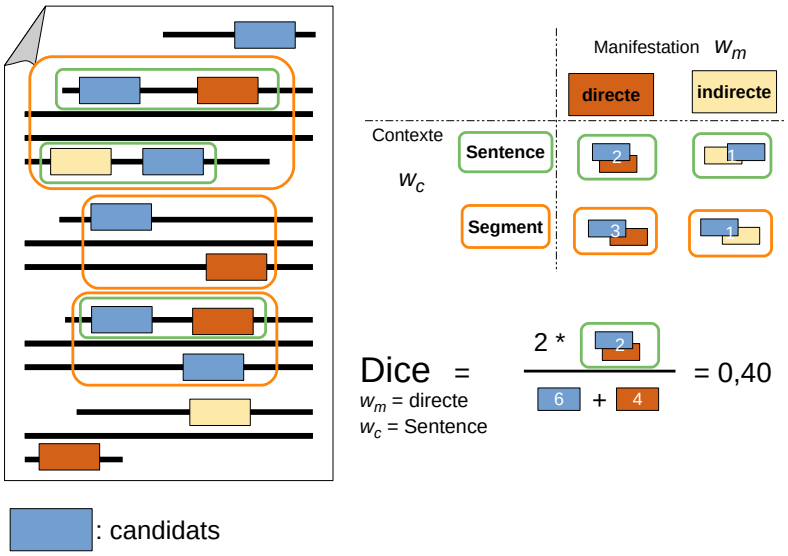


FIGURE 23 – Exemple de mesure de cooccurrences (Dice)

différents contextes w_c dans lesquels deux instances peuvent être cooccurrentes. Nous comparons la valeur du descripteur Sentence, Window, Segment ou Document des candidats et des manifestations d'arguments. Les descripteurs lexicaux d'une manifestation définissent son type, w_m, directe ou indirecte. Le dénombrement final de cooccurrences entre les instances d'un candidat et les manifestations des arguments de la relation partielle correspond au nombre de paires candidats-manifestations répondant aux critères w_c et w_m. Ce processus est illustré dans la Figure 23 (les codes de la Phase II sont disponibles en ligne [5])).

Les cooccurrences candidat-manifestation dans le contexte w_c et du type de manifestation w_m sont employées dans le calcul des scores Jaccard, Dice ou PMI. Ce nombre de cooccurrences constitue alors le numérateur des scores de Jaccard et Dice : $Jaccard = \frac{cooccurrences}{union(candidats, manifestations.w_m)}$ et $Dice = 2 * \frac{cooccurrences}{candidats + manifestations.w_m}$. Le score de PMI estime la probabilité d'une cooccurrence comparée aux probabilités individuelles d'instance du candidat et des manifestations : $PMI = \log \frac{p(cooccurrence)}{p(candidat) * p(manifestation.w_m)}$. Comparer les scores des différents candidats permet de discriminer les candidats pouvant venir compléter une instance de relation partielle. Les résultats de la complétion des relations partielles en associant les candidats selon la méthode Fréquentiste sont explorés dans la Section 3.4.

5. https ://github.com/Eskode/ARTEXT4LOD

Enfin, notre troisième méthode de complétion des relations n-Aires exploite des modèles pré-entraînés de word-embedding afin de mesurer la similarité entre termes des instances d'arguments candidates et les instances d'arguments présentent dans la relation n-Aire partielle à compléter.

Méthode par Plongements Lexicaux Le word embedding désigne les méthodes d'apprentissage ayant comme objectif de représenter les termes d'un corpus dans des vecteurs de nombres réels. Dans le domaine du traitement automatique, ces techniques sont couramment employées afin de créer des modèles de langage. Ces modèles de word-embedding sont appris par des réseaux de neurones profonds sur de larges corpus suivant les méthodes CBOW (continuous bag of words) [Mikolov et al., 2013a], Skip-Gram [Mikolov et al., 2013b]. Les modèles de word-embedding représentent ainsi les termes d'un domaine à travers des vecteurs à hautes dimensions (cf. Section 3.1.3). Les modèles de spaCy utilisent par exemple des vecteurs à 300 dimensions et le modèle BERT des vecteurs à plus de 750 dimensions. Ces modèles permettent de calculer des scores sémantiques de similarité entre deux termes. Cela est réalisé en calculant la similarité entre les vecteurs de ces deux termes, couramment via un calcul du cosinus de leurs angles mesurant la similarité entre ceux-ci. Ces valeurs de similarité peuvent être employées comme scores d'association entre les termes d'un candidat et des arguments d'une relation partielle.

Des modèles pré-entraînés existent et peuvent être directement employés [Jiao and Zhang, 2021]. La majorité de ceux-ci sont entraînés sur des corpus constitués des textes obtenus sur des sites de médias, blogs ou œuvres littéraires [Li and Yang, 2018]. Certains sont obtenus à partir de documents scientifiques (articles et rapport médicaux, publications en chimie et biologie) [Khattak et al., 2019]. Les sources choisies pour entraîner un modèle de word-embedding ont un impact important sur les performances de ce modèle par la suite [Peng et al., 2019]. L'adaptation de domaine est une question largement discutée en apprentissage automatique : le fait qu'un corpus présente un vocabulaire ou des formes syntaxiques spécifiques a en effet une influence sur le modèle, et le réutiliser dans un domaine différent peut alors se révéler difficile [Peng et al., 2019, Shahab, 2017].

Il n'existe pas de modèles issus de notre domaine d'application (i.e. les

emballages alimentaires). Cela est dû à la rareté des sources, ce domaine étant largement moins fourni en article que par exemple le domaine biomédical, ainsi que de l'intérêt sociétal et économique secondaire que ce domaine représente comparé par exemple à nouveau au domaine biomédical. Les modèles se rapprochant le plus du domaine que nous étudions sont ceux issus de corpus entraînés sur des articles scientifiques, majoritairement dans des domaines tels que la médecine, la biologie et la chimie [Neumann et al., 2019b].

Utiliser un modèle existant est une solution rapide à déployer, mais ce modèle peut ne pas être parfaitement adapté à notre domaine. Afin de tester l'effet du transfert d'un modèle à notre domaine, nous avons testé huit modèles différents. Ils sont tous entraînés sur différents corpus en anglais, issus du domaine scientifique ou non (cf. Table 15). Nous utilisons la librairie spaCy [6] (v3.0) choisie pour ses performances et sa simplicité d'utilisation. Les modèles utilisés sont issus des modèles de spaCy ou adaptés pour être utilisés par cette librairie.

Pour tous les modèles utilisés, nous appliquons la même méthode afin de déterminer le score de similarité sémantique entre un candidat et les instances d'arguments d'une relation partielle. Ce score a pour but d'identifier les meilleurs candidats pour compléter les relations partielles. Pour chaque candidat, nous considérons l'ensemble des termes, c'est-à-dire ses descripteurs lexicaux SciPuRe Original_Value et Attached_Value, de chacune des instances de son ensemble. Ces mêmes descripteurs lexicaux sont également récupérés dans les instances des arguments de la relation partielle. Les valeurs de similarité entre chacun des termes du candidat et chacun des termes des manifestations des arguments de la relation partielle sont calculées (en utilisant la fonction *similarity()* de spaCy). Cette fonction calcule le score de similarité correspondant au cosinus entre deux vecteurs de termes (mot ou syntagme). Une valeur est ainsi obtenue pour chaque combinaison possible entre les termes des instances du candidat et les manifestations. Les termes ne possédant pas de vecteur dans le modèle de langage utilisé, tel que les valeurs numériques, les unités de mesure complexes ou les termes spécialisés sont ignorés. Afin de ne conserver qu'une valeur associée à chaque candidat, et ainsi être capables de les discriminer, nous avons évalué deux variantes de la méthode par Plongements Lexicaux. La première consiste à conserver le score de similarité le plus élevé, maximum, entre un candidat et une

6. `https://spaCy.io/`

Modèle	Corpus sources
en_core_sci_lg	GENIA [a] (biomédical), Pubmed Central Open Access Subset [b] (médical), The MedMentions Entity Linking datasest [c] (médical), Ontonote [d] (blogs, news, commentaires)
en_core_sci_scibert	modèle SciBERT [e] (articles scientifiques[f])
en_ner_craft_md	CRAFT [g] (biomédical, chimie)
en_ner_jnlpba_md	JNLPBA [h] (biomédical)
en_ner_bc5cdr_md	BC5CDR [i] (biomédical)
en_ner_bionlp13cg_md	BIONLP13CG [j] (biologie)
en_core_web_lg	Ontonote [d] (blogs, news, commentaires)
en_core_web_trf	modèle RoBERTa [k] (Wikipedia, littérature)

[a] : `https://nlp.stanford.edu/~mcclosky/biomedical.html`
[b] : `https://evexdb.org/pmresources/vec-space-models/`
[c] : `https://github.com/chanzuckerberg/MedMentions`
[d] : `https://catalog.ldc.upenn.edu/LDC2013T19`
[e] : `https://github.com/allenai/scibert`
[f] : `https://semanticscholar.org/`
[g] : `https://bionlp-corpora.sourceforge.net/CRAFT/`
[h] : `https://github.com/spyysalo/jnlpba`
[i] : `https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/`
[j] : `https://2013.bionlp-st.org/`
[k] : `https://github.com/pytorch/fairseq/tree/master/examples/roberta`

TABLE 15 – Modèles de word-embeddings utilisés

manifestation d'argument. La deuxième consiste à utiliser la moyenne arithmétique des scores de similarités. Ces deux variantes sont évaluées en Section 3.4.3. Le calcul du score de similarité sémantique d'un candidat est illustré dans la Figure 24 et les codes de la Phase II sont disponibles en ligne [7]).

Nous avons également ajouté à ces trois méthodes le filtrage des instances d'arguments selon leurs scores de pertinence obtenus à l'issue de la Phase I. Nous avons également élaboré une approche s'inscrivant dans une démarche d'assistance aux experts. Cela est fait en sélectionnant, selon leurs valeurs d'association, plusieurs instances d'arguments candidates pour chaque argument d'une relation

7. https ://github.com/Eskode/ARTEXT4LOD

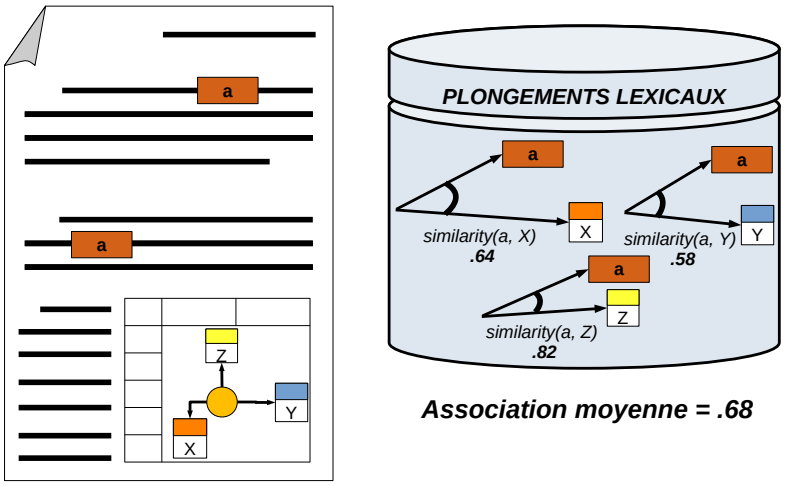


FIGURE 24 – Calcul d'une similarité sémantique entre un candidat et les arguments d'une relation partielle

partielle à compléter. Celles-ci sont ensuite proposées à un expert qui peut ensuite sélectionner l'instance d'argument pertinente.

3.2.6 Filtrage et sélection des candidats

En complément des méthodes de complétion des instances de relations n-Aires présentées en Section 3.2.5, nous avons identifié deux points pouvant être manipulés :

— le **filtrage** des instances d'arguments candidates en préalable au processus de complétion des relations. Nous avons en effet vu dans la Section 2.6 qu'un nombre important de faux-positifs était présent à la suite du processus d'extraction des instances d'arguments, mais que l'attribution de scores de pertinences à ces instances fondées sur les descripteurs de SciPuRe (c.f. Section 2.4) permettait d'écarter une partie de ces faux-positifs. Le filtrage des instances d'arguments candidates possédant les plus faibles scores de pertinence, issus de la Phase I, devrait ainsi permettre d'éliminer une partie des candidats ne constituant pas des instances d'arguments des relations n-Aires d'intérêts. Ce filtrage est fait avant de débuter la recherche des instances d'arguments à ajouter aux relations par les méthodes Structurelle, Fréquentiste ou par Plongements Lexicaux. Nous évaluons dans la Section 3.4 les effets d'un filtrage de 0%, 20%, 40% et 60% des instances d'arguments

candidates possédant les plus faibles scores de pertinence. Ce filtrage est fait selon les scores de $Sequence(CD^{node}_{target}, TF^{term}_{document})$ pour les instances d'arguments symboliques et de $ICF^{term}_{segment}$ pour les instances d'arguments quantitatifs. En effet ce sont ces scores qui ont montré, dans la Section 2.6, le meilleur comportement pour écarter les faux-positifs à la suite de l'extraction de la Phase I.

— la **sélection** de plusieurs instances candidates à une instance de relation. En effet les méthodes de complétion des instances de relation, Structurelle, Fréquentiste et Plongements Lexicaux, fournissent des mesures qui ordonnent les instances pouvant compléter les relations partielles. Le candidat présentant le meilleur score est alors sélectionné, mais il est aussi possible de retenir un ensemble de plusieurs candidats. Cela s'inscrit dans une approche d'assistance et d'accompagnement des experts, qui peuvent ensuite choisir l'information à retenir pour chaque relation. Lors de l'évaluation des instances de relations n-Aires complétées, en Section 3.4, nous considérons alors que l'ensemble d'instances d'argument ajoutées à la relation est valide si l'une des instances constitue une instance d'argument valide pour compléter la relation. Les effets de la sélection de un, trois, cinq ou dix candidats pour chaque instance d'argument à compléter dans une instance de relation partielle sont évalués dans la Section 3.4.

3.3 Instances de relation n-Aires - présentation du corpus

Les méthodes décrites en Section 3.2 se fondent sur des instances de relations n-Aires partielles extraites des tableaux de données présentes dans les articles scientifiques. Ces relations partielles sont ensuite complétées par les instances d'arguments présents dans les textes. Nous présentons ici des méthodes d'annotation manuelle et automatique des tableaux afin de former un Gold Standard de relations partielles. En fin de section, nous présentons également le Gold Standard des relations n-Aires complètes permettant l'évaluation des méthodes de reconstitution des relations présentées en Section 3.2.

3.3.1 Gold Standard des relations partielles dans les tableaux

Quatre relations n-Aires d'intérêt ont été sélectionnées dans notre Ressource Termino-Ontologique (RTO) pour être annotées : les relations de perméabilité (perméabilité à l'oxygène, au dioxyde de carbone et à l'eau) et la composition des emballages alimentaires (composante du facteur d'impact). Le corpus d'article contenant ces relations n-Aires a été créé à partir de 10 articles scientifiques en anglais, enregistrés au format HTML de plusieurs revues internationales sur le site ScienceDirect. Ces articles font également partie du Gold Standard utilisé pour la phase I (cf. Section 2.5) et permettent de constituer un Gold Standard composé de 31 tableaux, contenant 779 instances d'arguments, formant 332 instances de relations.

Annotation manuelle de relations partielles dans les tableaux. Les approches pour la Phase II décrite dans la Section 3.2 se basent sur les relations partielles reconnues dans les tableaux des articles scientifiques. Nous avons donc manuellement composé un Gold Standard constitué d'instances de relations n-Aires partielles contenues dans les tableaux des articles et concernant la composition des emballages alimentaires et la perméabilité au gaz (i.e. relations Impact_Factor_Component_Relation et CO2_Permeability_Relation, H2O_Permeability_Relation, O2_Permeability_Relation). Ce Gold Standard est également comparé aux résultats d'une extraction automatique de données selon la méthode décrite précédemment et dans [Buche et al., 2011, Hignette et al., 2009].

Les tableaux ont d'abord tous été annotés par un unique annotateur. Cette annotation a ensuite été re-croisée par trois autres annotateurs afin de corriger, compléter et valider l'annotation. L'objectif de la tâche d'annotation était de reconnaître toutes les données présentes dans les tableaux relatives aux mesures de perméabilité des emballages et à leur composition. Cette tâche d'annotation est pilotée par notre RTO Transmat [Guillard et al., 2018]. Un guide d'annotation a ainsi été conçu dans une approche collective et itérative impliquant l'ensemble des annotateurs. Un exemple de tableau sélectionné pour l'annotation est présenté dans la Figure 25. La figure 26 est un exemple du résultat de l'annotation manuelle de la première ligne de ce tableau. Ce Gold Standard, ainsi que le guide d'annotation et

Table 1 . Water vapour permeability (WVP) at 25 °C of PE films coated with chitosan compared to chitosan self standing films prepared with different casting solvents and plasticizers.

Sample	WVP × 10 $^{-13}$ (g/m s Pa) Δ RH 70%	WVP × 10 $^{-13}$ (g/m s Pa) Δ RH 45%	WVP × 10 $^{-13}$ (g/m s Pa) Δ RH 33%
PE	4.62 ± 0.73f	5.55 ± 0.23f	7.72 ± 2.58f
CS coated PE	12.37 ± 1.14f	6.67 ± 0.23f	7.88 ± 2.39f
PECSEinv	9.14 ± 1.09f	6.41 ± 3.28f	2.85 ± 0.34f
CSA	4161.31 ± 656.17a,b	2199.80 ± 1048.33d,e	25.71 ± 2.20f
CSE	4100.77 ± 588.88a,b	2884.37 ± 346.43b,c,d,e	38.71 ± 2.61f
CSAGLY	5410.08 ± 1543.67a	1905.39 ± 149.64e	26.14 ± 1.24f
CSEGLY	3481.46 ± 343.88b,c,d	2635.38 ± 414.28c,d,e	105.17 ± 6.57f

PE, polyethylene; CS coated PE, chitosan (CSE) coated polyethylene; PESCEinv, coating exposed to dry compartment; CSA, chitosan film prepared with aqueous acid solvent; CSE, chitosan film prepared with hydroalcoholic acid solvent; CSAGLY and CSEGLY, glycerol plasticized samples.

Different letters (a–f) indicate significant differences between formulations ($p < 0.05$).

FIGURE 25 – Exemple de tableau à annoter

les codes de prétraitements, pour l'extraction de données depuis des tableaux dans des articles scientifiques est disponible sur le Dataverse [Lentschat et al., 2021d].

L'annotation manuelle a été réalisée directement dans le code HTML des tableaux. Différentes balises et attributs ont été définis afin d'annoter les concepts de l'ontologie, instances d'argument et de relations :

— Les balises *qc* et *sc*, signifiant respectivement *'quantity concept'* et *'symbolic concept'*, annotent les concepts génériques de l'ontologie reconnus dans l'entête des tableaux. La reconnaissance de ces concepts et des arguments qu'ils représentent permet ensuite de définir les relations n-Aires présentes dans le tableau.

— Les relations n-Aires présentent dans le tableau sont indiquées dans la légende du tableau par des balises *rc* (i.e. *relation concept*).

— Les lignes du tableau sont annotées pour décrire les instances de relations n-Aires présentes, indiquées par une balise *ri*.

— Les instances d'arguments présentes dans la ligne sont indiquées par des balises *ai* au sein de chaque cellule du tableau.

— Les informations relatives à des instances d'arguments présentes en dehors des cellules des tableaux sont signalées par des balises *aii* (*'argument instance implicite'*). Cela concerne par exemple les valeurs d'humidité relative

```
<ri type="H2O Permeability_Relation" id="2">
    <ai type="Polyethylene" id="0"> </ai>
    <ai type="H2O Permeability" id="2"> </ai>
    <aii type="Temperature"> </aii>
    <aii type="Relative_Humidity"> </aii>
</ri>
<td class="align-left">
    <ai type="Polyethylene" id="0">
    PE
    </ai>
</td>
<td class="align-char">
    <ai type="H2O Permeability" id="0">
    4.62 ± 0.73</ai>f
</td>
<td class="align-char">
    <ai type="H2O Permeability" id="1">
    5.55 ± 0.23</ai>f
</td>
<td class="align-char">
    <ai type="H2O Permeability" id="2">
    7.72 ± 2.58</ai>f
</td>
```

FIGURE 26 – Exemple de l'annotation de la première ligne du tableau dans la Figure 25

contenues dans l'entête du tableau de la Figure 25. Ces instances d'arguments peuvent être annotées manuellement, mais ne sont pas considérées par l'extraction automatique des informations des tableaux présentée dans la Section 3.3.1.

Ces annotations peuvent ensuite être extraites des tableaux sous forme de relation n-Aires partielles suivant la représentation des relations n-Aires présentée en Section 3.2.2 dans la Table 13. Le nombre de relations n-Aires partielles annotées dans le corpus et la répartition de leurs instances d'arguments est illustrés dans la Table 16.

	Relation		
Arguments	CO2_Perm.	H2O_Perm.	O2_Perm.
Permeability	25	75	104
Packaging	11	51	36
Method	0	0	0
Thickness	14	19	43
Temperature	0	6	6
Relative_Humidity	0	6	16
Partial_Pressure	0	0	0

	Relation
Arguments	Impact_Factor_Component
Impact_Factor_Component	128
Packaging	44
Packnumber	67
Component_Qty_Value	128

TABLE 16 – Composition des 332 relations n-Aires partielles

Annotation automatique de relations partielles dans les tableaux. Les relations n-Aires contenues dans les tableaux des articles scientifiques peuvent être annotées automatiquement. Dans cette optique, un travail a été réalisé lors d'un stage d'assistant-ingénieur [Guari, 2021] dans le cadre du projet ARTEXT4LOD [8]. Celui-ci a consisté à implémenter une méthode d'annotation automatique fondée sur des travaux de l'état de l'art [Buche et al., 2011, Hignette et al., 2009]. Le résultat de ce travail a été comparé au Gold Standard des relations partielles annoté manuellement dans les tableaux et est disponible sur le Dataverse

8. https://www.theses.fr/s213955

[Lentschat et al., 2021d]. Disposer d'une méthode d'extraction automatique des relations n-Aires partielles présentes dans les tableaux est nécessaire à notre processus d'extraction. Cette solution vient s'insérer dans notre méthode et nous permet de proposer un pipeline complet d'extraction d'instances de relations n-Aires.

Dans cette annotation automatique, les tableaux ont d'abord été prétraités automatiquement afin de normaliser leurs mises en forme. Cette étape a été soumise à une vérification manuelle. Le but était d'obtenir des tableaux dont la structure est conforme à celle attendue pour réaliser une annotation automatique avec les approches sélectionnées [Buche et al., 2011, Hignette et al., 2009]. La méthode choisie [Buche et al., 2011, Hignette et al., 2009] pour l'extraction des relations entre les données expérimentales contenues dans les tableaux est guidée par la même RTO que celle utilisée dans les Phases I et II de cette thèse. Cette méthode utilise l'ontologie afin de définir les données à extraire et ainsi reconnaître les arguments présents dans un tableau, et regrouper leurs instances dans des relations n-Aires, en utilisant la signature de la relation ainsi que les labels des concepts portés par ses arguments. Cette méthode donne de bons résultats et est par exemple employée comme outil pour assister des méta-analyses de résultats expérimentaux en extrayant les données des tableaux des articles scientifiques [Wolf et al., 2018]. Plus de détails sur l'extraction automatique d'information depuis les tableaux sont disponibles dans l'Annexe A.2.

3.3.2 Gold Standard des relations n-Aires des documents

Le Gold Standard des relations n-Aires complétées, utilisé pour l'évaluation des méthodes de complétion des relations n-Aires partielles décrites en Section 3.2, s'appuie sur les relations partielles reconnues précédemment. Les relations n-Aires partielles présentées en Section 3.3.1 sont complétées avec les instances d'arguments provenant des textes. Ces instances textuelles d'arguments proviennent du Gold Standard utilisé afin d'évaluer la Phase I [Lentschat et al., 2021a], présentées en Section 2.5.

Le Gold Standard de relations n-Aires pour l'évaluation des méthodes de la Phase II a été composé manuellement. La méthode choisie se base sur les relations n-Aires partielles et procède à la recherche des instances d'arguments dans les

textes pouvant les compléter. Ce Gold Standard des relations complétées a été réalisé par un unique annotateur. Un dépôt en ligne Dataverse [Lentschat, 2021] met ce Gold Standard à disposition.

Pour chaque instance de relation contenue dans le Gold Standard manuel des relations partielles, l'annotateur a débuté par l'identification des instances d'arguments manquants. Il a ensuite recherché au sein des documents l'information correspondant à chacune de ces instances d'arguments. Une vérification est ensuite faite parmi les instances d'arguments du Gold Standard utilisé en Phase I, afin de trouver l'instance correspondant à l'information manquante. Par exemple, si une valeur de Temperature manque à une relation partielle, l'annotateur va tout d'abord vérifier que cette donnée existe dans le document. Si c'est le cas, il va ensuite rechercher l'instance d'argument correspondante ayant été annotée dans le Gold Standard utilisé dans la phase I. Aucun cas d'instance d'argument devant compléter une instance de relation partielle, mais qui serait manquante dans le Gold Standard n'a été relevé. Cela a permis à la fois de re valider le Gold Standard des instances d'arguments, en s'assurant que l'information pertinente a bien été annotée, et de déterminer spécifiquement l'instance d'argument devant être ajoutée à la relation partielle en se basant sur la lecture du texte des documents. La Table 17 récapitule la composition des instances de relations n-Aires en indiquant les instances d'argument présentes dans les relations.

	Relation		
Arguments	CO2_Perm.	H2O_Perm.	O2_Perm.
Permeability	25	75	104
Packaging	25	75	104
Method	11	62	77
Thickness	17	47	88
Temperature	25	75	104
Relative_Humidity	11	75	90
Partial_Pressure	0	4	2

	Relation
Arguments	Impact_Factor_Component
Impact_Factor_Component	128
Packaging	128
Packnumber	67
Component_Qty_Value	128

TABLE 17 – Composition des 332 relations n-Aires

On observe, en comparant les Tables 16 et 17, que des instances d'arguments spécifiques sont rarement présentes dans les relations partielles. Par exemple la méthode et certains paramètres de contrôle, température et humidité relative, ne sont presque jamais indiqués dans les tableaux des articles. Les valeurs indiquant le nom et l'épaisseur de l'emballage le sont partiellement, mais une part importante des instances de relations nécessitent toujours d'être complétées. En ce qui concerne les relations de composition des emballages, la quantité du composant employé est systématiquement présente au côté du nom de ce composant. En revanche l'emballage associé est souvent manquant. Même après complétion manuelle, certaines instances de relations restent incomplètes. Cela provient de l'absence occasionnelle d'instances informant certains arguments des relations. Cela peut être par exemple le cas de la méthode de mesure utilisée (i.e. argument Method), parfois décrite, mais pas nommée, ou de la différence de pression partielle (i.e. argument Partial_Pressure), paramètre pouvant être inclus dans l'unité de mesure de la perméabilité.

3.4 Résultats et discussions

Nous évaluons ici les méthodes de reconstitution des instances de relations n-Aires décrites en Section 3.2. Ces méthodes utilisent comme base les instances de relations extraites des tableaux (cf. Section 3.3.1) et recherche les instances d'arguments pouvant les compléter en utilisant des critères structurels, fréquentistes et sémantiques. Nous évaluons les effets de ces méthodes de complétion des relations en comparant les instances de relations obtenues avec le Gold Standard des relations décrit en Section 3.3.2.

Les scores d'évaluation choisis s'appuient sur les valeurs de rappel, précision et f-score (micro). Afin de mesurer ces valeurs, nous comparons les instances d'arguments d'une relation obtenues à l'issue de la Phase II aux instances d'arguments de la relation correspondante dans le Gold Standard présenté en Section 3.3.2.

Pour chacune de ces instances d'arguments, plusieurs cas de figure se présentent :

- l'instance d'argument était déjà présente dans la relation partielle. N'ayant

pas eu besoin d'être complétée, nous ne prenons pas en compte cette instance dans les calculs de rappel et précision.

- l'instance d'argument a été complétée dans la relation et :
  - cette même instance est présente dans l'instance de relation du Gold Standard. La méthode de complétion a donc été efficace en détectant une instance valide à ajouter à la relation. Il s'agit d'un cas d'instance **valide** (i.e. vrai positif).
  - l'instance est erronée, car elle ne correspond pas à l'instance d'argument de l'instance de la relation complétée présente dans le Gold Standard. La méthode n'a donc pas ajouté la bonne instance d'argument à la relation. Il s'agit d'une **erreur** d'instance (faux positif).
  - le Gold Standard n'indiquait pas d'instance d'argument. La méthode a donc ajouté une information qui n'était pas requise, et donc invalide. Il s'agit d'un cas d'instance en **excès** (faux positif).
- l'instance d'argument est manquante dans la relation complétée et :
  - cette instance est également manquante dans la relation du Gold Standard. L'information n'étant pas présente dans le document, il n'y avait donc pas d'instance à ajouter à cet argument de la relation. Il s'agit d'un cas d'instance **vide** (vrai négatif).
  - cette instance est présente dans le Gold Standard et aurait dû être complétée. La méthode de complétion des relations partielles n'a donc pas réussi à détecter une instance à ajouter à la relation, celle-ci étant potentiellement manquante à l'issue de la phase I. Il s'agit d'un cas d'instance **manquante** (faux négatif).

Ces critères permettent de calculer les scores de rappel (i.e. proportion des instances d'arguments valides retrouvées), précision (i.e. proportion des instances d'arguments ajoutées valides) et f-score (i.e. la moyenne harmonique du rappel et de la précision). Les Tables 18 et 19 détaillent les différents cas retrouvés dans les instances d'arguments complétant les instances de relations n-Aires partielles dans l'approche Structurelle.

Dans l'évaluation que nous faisons des approches de complétion des instances de relations partielles, nous considérons également les deux points détaillés dans la Section 3.2.6 : (i) le filtrage des instances d'arguments candidates avant la

complétion des instances de relations et (ii) la sélection de plusieurs candidats pour un argument de l'instance de la relation partielle à compléter.

(i) le **filtrage** des instances d'arguments issues de la Phase I est fait selon les scores de pertinence calculés en Phase I. Lors d'un filtrage des instances d'argument, cela est fait avant de procéder aux différentes étapes de reconstitution des instances de relation n-Aires.

(ii) la **sélection** de plusieurs instances candidates pour compléter un argument de relation n-Aire partielle de plus d'une des instances candidates pour compléter un argument manquant dans une instance de relation partielle est faite afin de proposer un choix aux experts pouvant alors sélectionner le résultat valide. Pour l'évaluation de cette stratégie, nous considérons que l'instance d'argument ajoutée à la relation est valide si l'une des instances du lot de candidats sélectionnés correspond à l'instance attendue dans la relation à compléter.

Ces deux critères permettent de filtrer une partie des faux-positifs présents dans les résultats de la Phase I et de positionner les approches dans une démarche d'assistance aux experts.

3.4.1 Méthode Structurelle

La méthode Structurelle utilise la proximité au sein des documents entre les instances d'arguments candidates et le tableau contenant la relation partielle à compléter, et recherche les plus proches. Une variante, la méthode Structurelle Guidée, utilise des associations déterminées manuellement entre les sections des documents et certains types d'arguments afin de privilégier les candidats contenus par certaines sections des documents en utilisant l'information portée par les descripteurs Segment et Argument des représentations SciPuRe des instances d'arguments. Celles-ci sont décrites en détail dans la Section 3.2.5

Les résultats présentés dans les Tables 18 et 19 présentent le détail des instances d'arguments qui composent les relations complétées par la méthode Structurelle, respectivement sans et avec guide. Les résultats donnés dans ces Tables sont mesurés en ne sélectionnant qu'un unique candidat par instance d'argument à compléter et sans filtrage préalable des instances d'arguments selon leurs scores

de pertinence. Nous observons une légère différence dans les valeurs de précision entre les deux versions de la méthode, en faveur de la méthode Structurelle Guidée (.20 vs .25). Cela provient du nombre plus important des instances de Packaging valides qui sont sélectionnées par le guidage de la méthode Structurelle (70 vs 148). Il y a également une légère amélioration dans la récupération des instances de Température valides (0 vs 17). En revanche ce guidage dégrade notablement la qualité des instances des arguments Thickness (48 vs 0). Cela indique qu'il serait nécessaire de réviser l'ordre des sections ayant été identifiées comme les plus susceptibles de contenir des instances valides de Thickness. Les scores de rappel et de précisions des méthodes Structurelle et Structurelle Guidée correspondent aux valeurs des premières colonnes dans les Figures 27 et 28.

Arguments	Instances présentes	Instances dans les Relations Complétées				
		Valide	Erreur	Excès	Vide	Manquant
Impact_Factor_ Component	8		115	5		
H2o_Permeability	75					
Co2_Permeability	25					
O2_Permeability	104					
Partial_Pressure				87	240	6
Thickness	76	48	28	52		
Packnumber	67				61	
Relative_Humidity	22	20	134		28	
Component_Qty_ Value	128					
Packaging	142	70	120			
Method		74	76	28	26	
Temperature	12		192			

pas de filtrage des instances d'arguments, sélection d'un unique candidat par instance d'argument manquant

rappel = .97 **précision** = .20 **f-score** = .33

TABLE 18 – Résultats de la méthode Structurelle

La Figure 27 présente l'évolution des scores de rappel des méthodes Structurelles, non guidées et Guidées, en faisant varier le pourcentage d'instances filtrées en fin de Phase I (différentes couleurs de colonnes) et le nombre de candidats sélectionnés pour compléter les relations partielles (différents lots d'histogrammes). Il apparaît que filtrer une partie des candidats entraîne une diminution du

Argument	Instances présentes	Instances dans les Relations Complétées				
		Valide	Erreur	Excès	Vide	Manquant
Impact_Factor_ Component	8	1	114	5		
H2o_Permeability	75					
Co2_Permeability	25					
O2_Permeability	104					
Partial_Pressure				87	240	6
Thickness	76		76	52		
Packnumber	67				61	
Relative_Humidity	22	21	133		28	
Component_Qty_ Value	128					
Packaging	142	148	42			
Method		74	76	28	26	
Temperature	12	17	175			

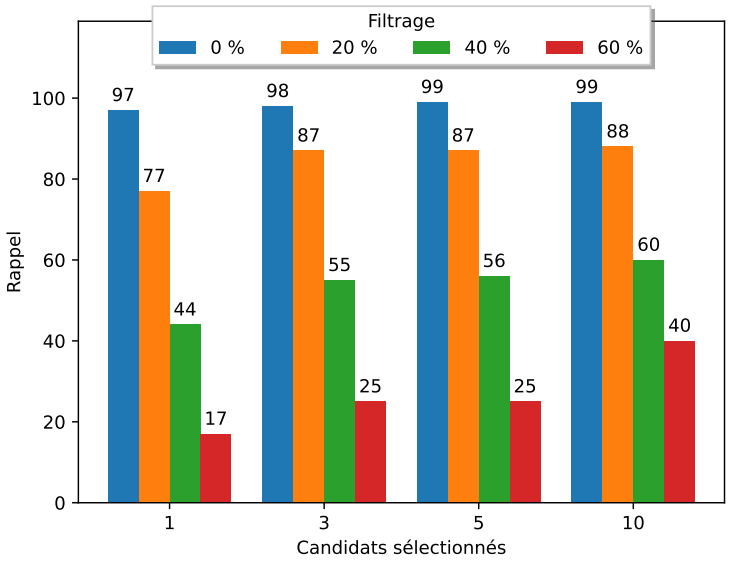
pas de filtrage des instances d'arguments, sélection d'un unique candidat par instance d'argument manquant

rappel = .98 **précision** = .25 **f-score** = .40

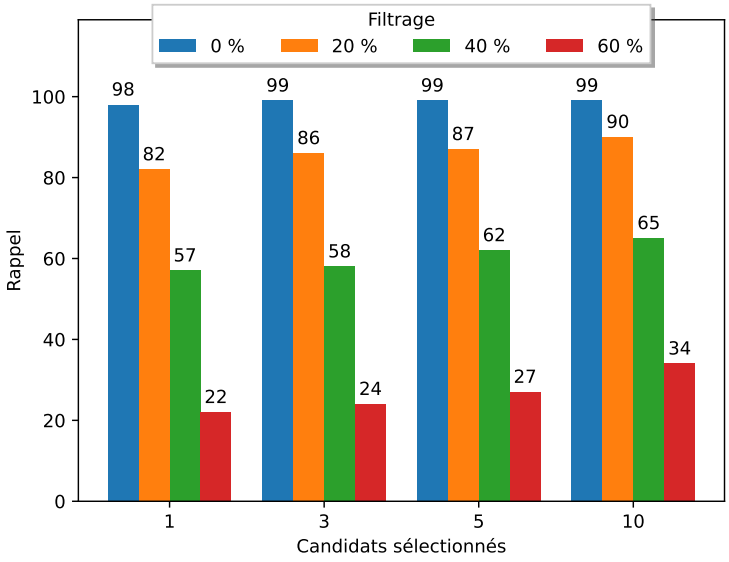
TABLE 19 – Résultats de la méthode Structurelle Guidée

rappel proportionnelle au filtrage introduit. Cela signifie que certains candidats valides sont toujours éliminés par ce filtrage, leurs scores de pertinence calculés en Section 2.4 étant trop faibles. En revanche l'augmentation du nombre de candidats sélectionnés permet de réduire cet impact, en particulier lors d'un filtrage important. Le rappel lors de filtrages à 60% double en passant d'une sélection d'un candidat à une sélection de dix candidats par instance d'argument manquant. Ces comportements se retrouvent pour les deux approches de la méthode Structurelle.

La Figure 28 présente l'évolution des scores de précision des méthodes Structurelles, non guidées et guidées, en faisant varier le pourcentage d'instances filtrées en fin de Phase I et le nombre de candidats sélectionnés pour compléter les relations partielles. Il apparaît naturellement une augmentation de la précision selon le nombre de candidats sélectionnés pour chaque argument manquant. En sélectionnant trois candidats au lieu d'un seul, cette précision augmente en moyenne de 66% (e.g. .20− > .36 et .25− > .38) et augmente encore lors de la sélection de cinq ou dix candidats. Cette augmentation est davantage marquée pour la méthode Structurelle Guidée. Cela tend donc à signifier que le guidage de



(a) méthode Structurelle



(b) méthode Structurelle Guidée

FIGURE 27 – Effet du filtrage et du nombre de candidats sélectionnés sur le Rappel

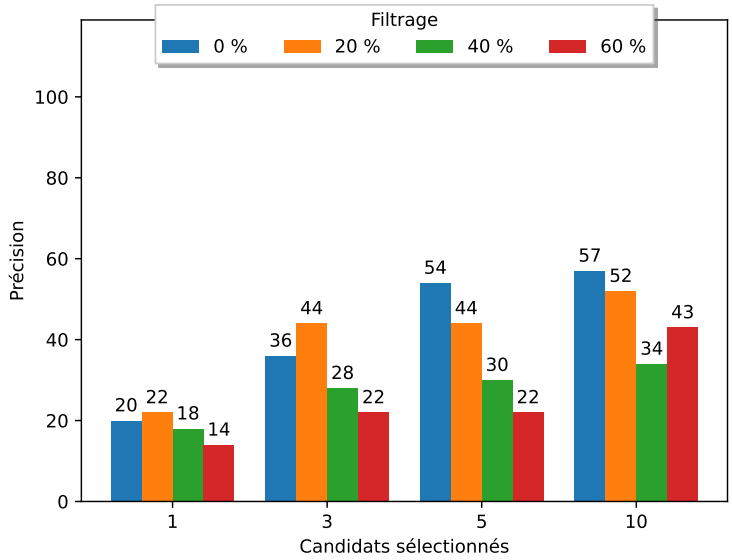
la recherche des candidats dans des sections spécifiques est pertinent, mais qu'il est nécessaire de sélectionner plusieurs candidats dans cette section. L'exemple 8 détaille un cas où la sélection de plusieurs candidats est adaptée. Le filtrage préalable des instances candidates non pertinentes présente un effet positif sur la précision uniquement lorsque celle-ci est faite sur une petite partie des candidats (i.e. 20%). Au delà la précision diminue, indiquant que des candidats pertinents sont écartés par le filtrage.

Exemple 8 ***Confusion de relations dans la sélection des candidats***

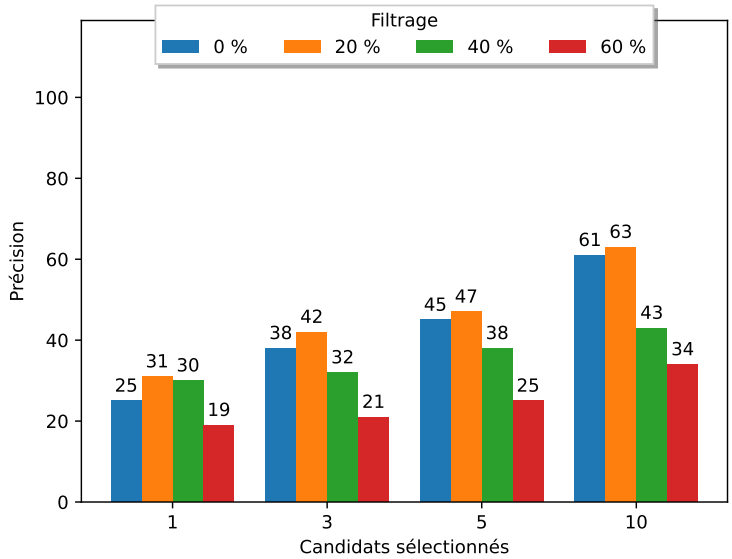
Prenons deux relations partielles, A (H2O_Permeability_Relation) et B (CO2_Permeability_Relation), ayant chacune des instances de l'argument Relative_Humidity manquants. Les valeurs valides des instances d'argument à ajouter à ces relations sont respectivement '50%' pour A et '0%' pour B. Des instances d'arguments avec ces valeurs ont été obtenues à l'issue de la Phase I et doivent être attribuées aux bonnes relations (e.g. A →' 50%' et B →' 0%'), effectuer l'inverse résulterait en deux erreurs (e.g. A →' 0%' et B →' 50%').

Il y a un potentiel de confusion entre une instance d'argument invalide (i.e. ne devant pas être attribuée à aucune relation) et une instance valide, mais mal attribuée (i.e. ayant été ajoutée à la mauvaise relation). Dans cet exemple, attribuer les deux instances candidates à chacune des instances de relations résulte alors en deux validations des instances ajoutées aux arguments (A →' 50%'/'0%' et B →' 0%'/'50%'). Ainsi, plus le nombre d'instances d'une même relation est important dans un même document, et plus ce potentiel de confusion augmente. Cela augmente également avec le nombre d'instances d'arguments candidates pour une relation. En revanche, sélectionner plusieurs candidats pour une instance d'argument à compléter permet de diminuer l'influence de cette confusion sur les résultats.

L'approche Structurelle présente dans l'ensemble d'excellents scores de rappel, autour de .98, celui-ci ne diminuant qu'avec un fort filtrage préalable des instances de candidats. En revanche, obtenir une précision dépassant .50 nécessite d'augmenter le nombre de candidats sélectionnés pour chaque argument manquant à 5 ou 10. Appliquer un léger filtrage des résultats à l'issue de la Phase I (i.e. 20%, en orange dans les Figures 27 et 28) a également un léger effet favorable à la précision sans trop dégrader le rappel.



(a) méthode Structurelle



(b) méthode Structurelle Guidée

FIGURE 28 – Effet du filtrage et du nombre de candidats sélectionnés sur la Précision

3.4.2 Méthode Fréquentiste

La méthode Fréquentiste mesure les associations entre un candidat et les manifestations textuelles des instances d'arguments d'une relation partielle. Pour cela trois mesures ont été testées : Jaccard, Dice et Point-Wise Mutual Information. Nous avons également considéré plusieurs contextes, w_c de cooccurrences grâce aux descripteurs structurels de SciPuRe : Sentence, Window, Segment et Document. Différentes manifestations, w_m, des arguments des relations partielles sont également considérées via les descripteurs lexicaux : Original_Value et Attached_Value. Les configurations possibles étant ainsi nombreuses, nous présentons dans la Table 20 les plus élevés et plus faibles f-scores de chacune des mesures.

Il apparaît dans la Table 20 peu de différences entre les trois mesures testées. Bien que Dice présente systématiquement des valeurs supérieures à Jaccard et PMI, cette différence n'est pas significative (e.g. valeurs de rappel respectivement de .71, .69 et .68). En revanche, il apparaît clairement que le contexte le plus favorable à la mesure de l'association entre un candidat et les manifestations des arguments d'une relation partielle est le Document (e.g. le meilleur f-score est $Dice(Document, AttachedValue) = .40$). Le Segment est le contexte le moins favorable. Considérer Sentence ou Window comme contexte amène des résultats intermédiaires. Cela signifie que la fréquence globale (i.e. au niveau du Document) des instances d'arguments (i.e. candidats ou manifestations) est le facteur permettant le plus de détecter les associations entre candidats et relations. Nous supposons que le fait que le Segment soit le contexte le moins favorable est probablement dû à la confusion entre relations et candidats illustrés dans l'Exemple 8. En effet les instances d'arguments similaires devant être attribuées chacun à des instances de relations différentes sont généralement présentes dans les mêmes sections. L'évaluation montre que considérer les manifestations indirectes des instances des arguments des relations partielles donne les meilleurs résultats. Les descripteurs Attached_Value manifestent en effet généralement le concept derrière l'argument (e.g. *'température'*, *'thickness'*) et sont donc pertinents à considérer dans des contextes de cooccurrences larges tels que Document. Les manifestations directes, Original_Value, sont plus généralement associées aux contextes réduits, Sentence ou Window, car elles permettent d'associer des instances plus spécifiques. Les meilleurs scores des approches Fréquentistes restent faibles. Un rappel autour

de .70 peut être considéré comme acceptable lorsque l'on souhaite adopter une démarche d'assistance aux experts, mais une précision maximale de .28 est insuffisante. L'ensemble complet des résultats pour les différentes configurations de l'approche Fréquentiste est présenté en Annexe A.3.

Mesure	w_c	w_m	Rappel	Précision	f-score
Dice	Document	Attached_Value	**.71**	**.28**	**.40**
Jaccard	Document	Attached_Value	.69	.27	.39
PMI	Document	Original_Value	.68	.25	.36
...					
Dice	Segment	Attached_Value	.52	.13	.20
Jaccard	Segment	Attached_Value	.50	.12	.19
PMI	Segment	Attached_Value	.50	.12	.19

TABLE 20 – Scores de rappel, précision et f-score des mesures fréquentistes selon w_c et w_m

La Figure 29 présente l'évolution du rappel d'une méthode Fréquentiste, utilisant une mesure de Dice avec $w_c = Document$ et $w_m = Attached\_Value$, en faisant varier le pourcentage d'instances filtrées en fin de Phase I et le nombre de candidats sélectionnés pour compléter les relations partielles. Il apparaît qu'augmenter le nombre de candidats sélectionnés produit une légère augmentation du rappel. Filtrer une partie des candidats permet d'augmenter le rappel lorsque ce filtrage est réglé à 20%, au delà ce rappel diminue. Il semble donc qu'un filtrage léger des candidats, 20%, a un meilleur effet sur le rappel des méthodes Fréquentistes que d'augmenter le nombre de candidats sélectionnés.

La Figure 30 présente l'évolution des scores de précisions d'une méthode Fréquentiste, utilisant une mesure de Dice avec $w_c = Document$ et $w_m = Attached\_Value$, en faisant varier le pourcentage d'instances filtrées en fin de Phase I et le nombre de candidats sélectionnés pour compléter les relations partielles. Il apparaît une légère augmentation de la précision selon le nombre de candidats sélectionnés pour chaque argument manquant. Le filtrage préalable des instances candidates non pertinentes présente un effet positif sur la précision majoritairement lorsque celle-ci est faite sur une petite partie des candidats (i.e. 20% et 40%). Il apparaît également dans la Figure 30 que filtrage et sélection de plusieurs candidats peuvent rentrer en interférence. En effet l'augmentation du filtrage lorsque l'on sélectionne dix candidats ne fait que dégrader la précision. Cela

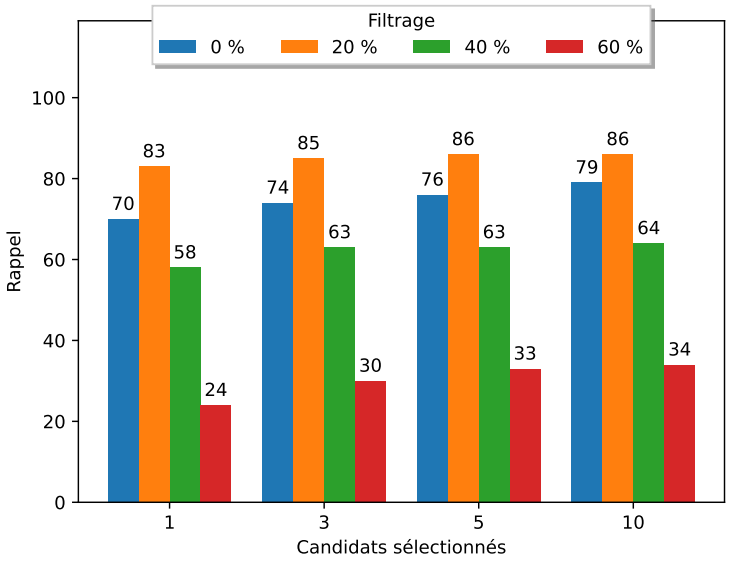


FIGURE 29 – Effet du filtrage et du nombre de candidats sélectionnés sur le Rappel - Dice, $w_c = Document$, $w_m = Attached\_Value$

montre clairement qu'un filtrage trop important écarte des candidats pertinents du processus de mise en relation.

Dans l'ensemble, il est possible d'obtenir un rappel acceptable (i.e. $>$ .80) en appliquant un léger filtrage des candidats sélectionnés à l'issue de la Phase I. La sélection de plusieurs candidats n'a que peu d'effets pour les méthodes fréquentistes. En revanche la précision reste un problème avec ces méthodes, aucun des paramètres testés n'ayant permis de dépasser .50.

3.4.3 Méthode par Plongements Lexicaux

La méthode par Plongements Lexicaux utilise les scores de similarité, calculés à partir de modèles de langage de word-embedding, entre les termes des instances des arguments d'une relation partielle et les termes d'une instance d'argument candidate. L'ensemble des scores de similarité entre ces termes est réduit à une valeur unique, par moyenne arithmétique ou en sélectionnant la valeur maximale.

Les résultats détaillés présentés dans les Tables 21 et 22 présentent les scores de rappel, précision et f-score pour chacun des modèles de langage testés, respectivement avec moyenne arithmétique ou par valeur maximale. Les résultats

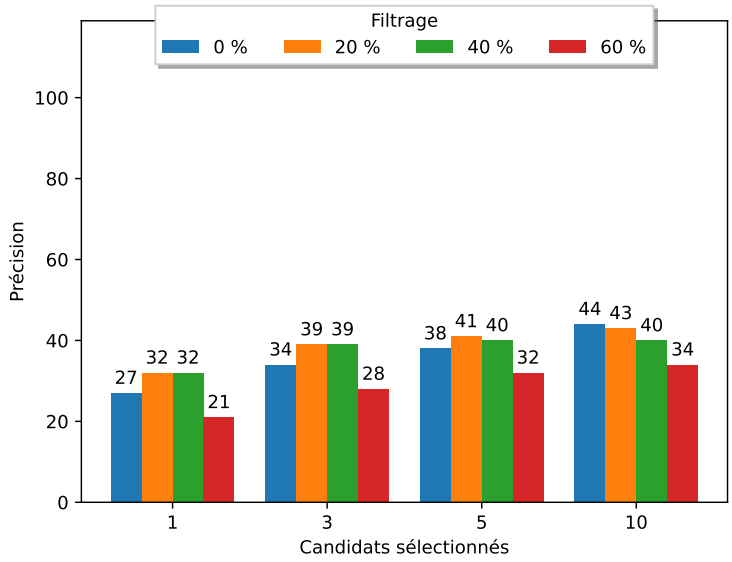


FIGURE 30 – Effet du filtrage et du nombre de candidats sélectionnés sur la Précision - Dice, $w_c = Document$, $w_m = Attached\_Value$

des Tables 21 et 22 sont donnés en ne sélectionnant qu'un unique candidat par instance d'argument à compléter et sans filtrage préalable des instances d'arguments selon leurs scores de pertinence.

Nous observons tout d'abord que la majorité des modèles évalués présentent des scores de rappel et de précision équivalents, autour de .50 pour le rappel et .12 pour la précision. Utiliser la moyenne arithmétique des scores de similarités montre des résultats supérieurs de quelques points à l'utilisation de la valeur maximale, mais cette différence n'est pas significative. Le point principal à noter est le fait que les modèles core_web_trf et core_sci_scibert présentent un f-score supérieur de près de .15 aux autres modèles. Ces deux modèles sont basés sur les Transfomers de BERT, core_web_trf est entrainé sur des sources web et core_sci_scibert sur des articles scientifiques de diverses sources (cf. Table 15 en Section 3.2.5). Cette différence montre la qualité de l'approche de BERT comparée aux modèles basés sur des approches Skip-Gram et CBOW. En revanche il n'y a pas de différences significatives entre ces deux modèles, bien que core_sci_scibert soit entraîné sur des sources scientifiques. Cela peut provenir de l'absence d'articles appartenant spécifiquement à notre domaine dans le corpus d'entraînement de core_sci_scibert. En effet, il a été montré qu'un transfert d'apprentissage efficace

ne peut se faire que sur des domaines réellement proches [Peng et al., 2019]. Ceci pouvant aussi expliquer l'absence de variations significatives dans les scores d'autres modèles. Ces modèles sont entraînés sur des domaines variés, mais aucun d'eux n'est suffisamment proche de notre domaine d'application. Les performances supérieures des modèles suivant la méthode de BERT, conjuguées à l'absence de différences selon les domaines d'apprentissage, peuvent également suggérer que la connaissance générale du langage acquise par le modèle de word-embedding est plus importante que les spécificités du domaine. Cette hypothèse a également été faite dans des études comparant les performances des modèles de word-embedding sur différents domaines [Peng et al., 2019, Wang et al., 2018, Shahab, 2017].

Modèle	Rappel	Précision	f-score
ner_jnlpba_md	.53	.13	.21
ner_craft_md	.52	.13	.21
ner_bionlp13cg_md	.52	.13	.21
ner_bc5cdr_md	.54	.14	.22
core_web_trf	**.67**	**.23**	**.35**
core_web_lg	.52	.13	.20
core_sci_scibert	**.65**	**.22**	**.33**
core_sci_lg	.52	.13	.20

TABLE 21 – Scores de rappel, précision et f-score selon les modèles de word-embedding : moyenne arithmétique des similarités

Modèle	Rappel	Précision	f-score
ner_jnlpba_md	.50	.12	.19
ner_craft_md	.49	.11	.18
ner_bionlp13cg_md	.51	.12	.20
ner_bc5cdr_md	.49	.11	.19
core_web_trf	**.67**	**.24**	**.35**
core_web_lg	.51	.12	.20
core_sci_scibert	**.65**	**.22**	**.33**
core_sci_lg	.55	.15	.23

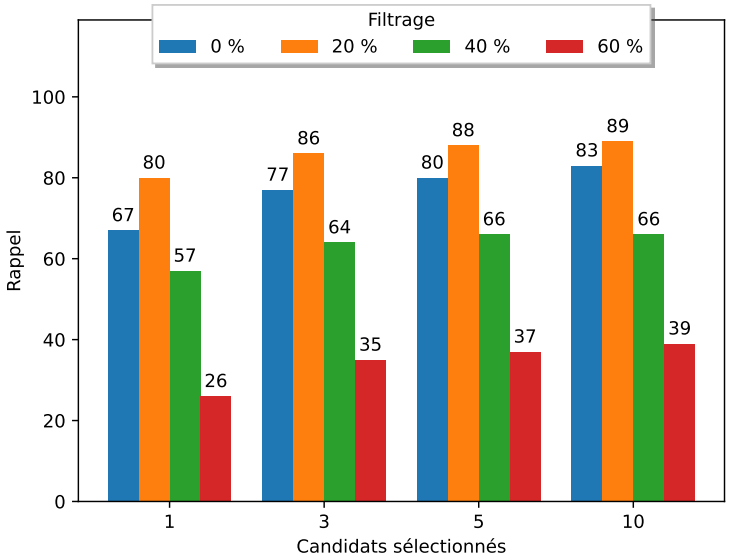
TABLE 22 – Scores de rappel, précision et f-score selon les modèles de word-embedding : valeur maximale des similarités

Nous évaluons l'effet du filtrage et du nombre de candidats sélectionnés sur les modèles core_web_trf et core_sci_scibert, car ceux-ci présentent les résultats les plus prometteurs. La Figure 31 présente l'évolution de leurs scores de rappel

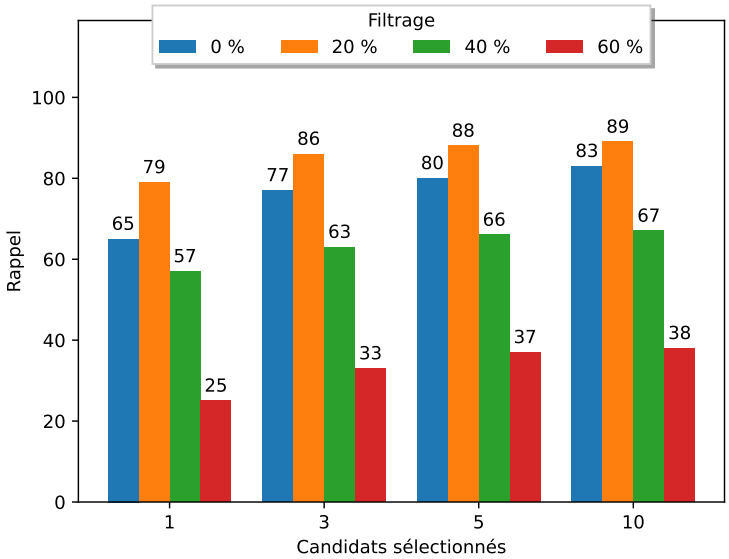
en faisant varier le pourcentage d'instances filtrées en fin de Phase I et le nombre de candidats sélectionnés pour compléter les relations partielles. Il apparaît qu'augmenter le nombre de candidats sélectionnés produit une augmentation du rappel. Cependant cette augmentation est principalement présente lors de la sélection de trois candidats, en sélectionner cinq ou dix n'offrant pas d'amélioration notable. Filtrer une partie des candidats permet d'augmenter le rappel lorsque ce filtrage est réglé à 20%, au delà ce rappel diminue. Cela signifie que certains candidats valides sont éliminés par ce filtrage, leurs scores de pertinences calculés en Section 2.4 étant trop faibles. Ces comportements sont présents pour chacun des deux modèles. Il semble donc qu'une augmentation légère, à trois, du nombre de candidats sélectionnés allié à un léger filtrage, de 20%, des candidats présentant les plus faibles scores de pertinence permet d'obtenir des scores de rappel dépassant .80.

La Figure 32 présente l'évolution des scores de précision des modèles core_web_trf et core_sci_scibert, en faisant varier le pourcentage d'instances filtrées en fin de Phase I et le nombre de candidats sélectionnés pour compléter les relations partielles. Il apparaît naturellement une augmentation de la précision selon le nombre de candidats sélectionnés pour chaque argument manquant. Ce gain est de presque .20 points en passant d'un candidat à trois puis augmente à nouveau d'environ .10 points à cinq et dix. Cette augmentation est similaire pour les modèles core_web_trf et core_sci_scibert. Le filtrage préalable des instances candidates non pertinentes présente un effet positif sur la précision majoritairement lorsque celui-ci est réalisé sur une petite partie des candidats (i.e. 20% ou 40%). Cet effet est visible, mais ne représente jamais un gain de plus de .05 points de précision. Avec un filtrage plus important, la précision diminue, indiquant que des candidats pertinents sont écartés par le filtrage.

L'approche par Plongements Lexicaux présente dans l'ensemble des scores de rappel et de précision faibles. Les modèles core_web_trf et core_sci_scibert, fondés sur le modèle BERT, présentent les meilleurs scores parmi l'ensemble des modèles évalués. Il est possible d'augmenter le rappel à un niveau acceptable (i.e. $> .80$) en appliquant un léger filtrage des instances d'arguments selon leurs scores de pertinence et en sélectionnant trois candidats par argument à compléter. En revanche, obtenir une précision dépassant .50 nécessite d'augmenter le nombre de candidats sélectionnés pour chaque argument manquant à 5 ou 10.

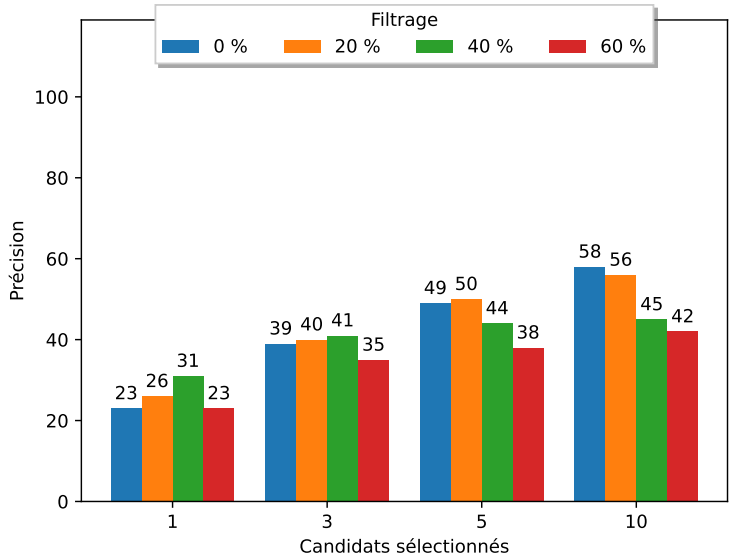


(a) core_web_trf

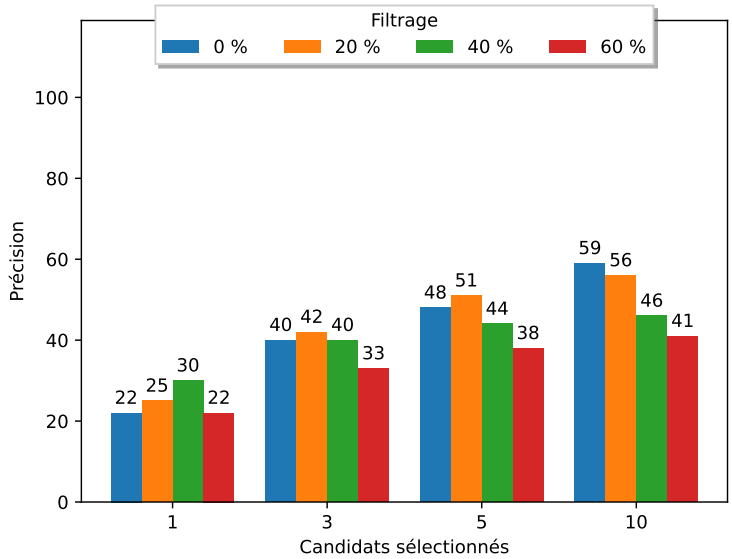


(b) core_sci_scibert

FIGURE 31 – Effet du filtrage et du nombre de candidats sélectionnés sur le Rappel de la méthode par Plongements Lexicaux



(a) core_web_trf



(b) core_sci_scibert

FIGURE 32 – Effet du filtrage et du nombre de candidats sélectionnés sur la Précision de la méthode par Plongements Lexicaux

3.4.4 Discussion

En observant les trois méthodes évaluées, Structurelle, Fréquentiste et par Plongements Lexicaux, ainsi que les variations des paramètres testés, il est possible de tirer des conclusions générales. Tout d'abord, utiliser les scores de pertinence des instances d'arguments mesurés lors de la Phase I pour appliquer un filtrage léger, de 20%, améliore les scores de rappel et de précision. Cela résulte en une amélioration générale du f-score de .05 points pour l'ensemble des approches testées. Un filtrage plus important des instances d'arguments candidates a au contraire des effets négatifs. Cela indique que les scores de pertinence des instances calculés ne permettent de filtrer efficacement que les faux-positifs les plus 'flagrants' (i.e. présentant les scores de pertinence les plus faibles), mais que des instances valides et devant être mises en relation peuvent toujours présenter des scores de pertinence moyens et être écartées dans un filtrage plus important. La Table 23 présente les meilleurs f-scores de chacune des approches, selon leurs différents critères et le nombre de candidats sélectionnés. Ces résultats sont présentés avec un filtrage de 20% des instances candidates. La sélection de plusieurs candidats pour les instances d'arguments à compléter permet d'augmenter le rappel et la précision des méthodes. Ce comportement était prévisible, en effet sélectionner un nombre plus important de candidats augmente les probabilités d'ajouter une instance d'argument valide à la relation partielle.

Approche	Critère	f-score candidats sélectionnés			
		1	3	5	10
Structurelle	simple	.35	.58	.58	.65
Structurelle	guidée	.45	.56	.61	**.74**
Fréquentiste*	Jaccard	$\mathbf{.48}^{d}_{a}$	$.54^{d}_{a}$	$.61^{p}_{o}$	$.66^{p}_{o}$
Fréquentiste*	Dice	$.46^{d}_{a}$	$.55^{d}_{o}$	$.60^{p}_{o}$	$.66^{p}_{o}$
Fréquentiste*	PMI	$.44^{d}_{o}$	$.53^{d}_{a}$	$.60^{p}_{o}$	$.68^{p}_{o}$
Plongements Lexicaux	core_web_trf	.40	**.59**	.64	.70
Plongements Lexicaux	core_sci_scibert	.39	.57	**.65**	.70

*contextes : $p = Sentence$, $w = Window$, $s = Segment$, $d = Document$
manifestations : $o = Original\_Value$, $a = Attached\_Value$

TABLE 23 – Meilleures valeurs de f-score des approches (filtrage = 20%)

La Table 23 permet de choisir l'approche à adopter en fonction du nombre de candidats sélectionnés. Cela permet ainsi de se positionner par rapport à une tâche

d'extraction visant à assister des experts dans la formation de relations n-Aires.

Si l'on désire fournir un résultat ne nécessitant pas d'intervention humaine, les approches Fréquentistes basées sur les cooccurrences des instances candidates avec les manifestations d'arguments des relations partielles au niveau du Document montrent les meilleurs résultats ($Jaccard = .48, Dice = .46$). L'approche Structurelle Guidée permet également d'obtenir un f-score proche (.45). Ce cas reste cependant limité, car ces f-score restent faibles.

Lorsque le travail peut être assisté par un expert qui sélectionnera dans un ensemble de trois candidats l'instance d'argument pouvant compléter une instance de relation n-Aire partielle, les approches ne présentent pas de nettes différences dans leurs résultats. L'ensemble des f-scores variant de .53 à .59.

À cinq candidats en revanche l'approche par Plongements Lexicaux apparaît comme la meilleure avec des f-scores de .65 et .64 pour les deux modèles basés sur BERT. Cela indiquerait que la similarité mesurée par ces modèles de *word-embedding* permettent bien de détecter le candidat devant venir compléter une instance de relation n-Aires partielle, mais que d'autres instances présentent des similarités plus importantes.

Enfin, dans le cadre demandant une grande intervention de l'expert pour trier les dix candidats attribués aux instances d'arguments manquants dans les instances des relations n-Aires partielles, l'approche Structurelle Guidée se distingue avec un f-score de .74. Cette approche ajoutant les candidats sur la base des sections dans lesquelles ils apparaissent, cette sélection permet de passer outre le fait que plusieurs instances différentes d'un même argument se retrouvent au sein d'une même section. Si l'on recherche par exemple une instance de Packaging pour compléter une relation partielle, il est probable de la trouver dans la section *Materials and Methods*, cependant des instances différentes de Packaging appartenant à des instances différentes de relations s'y trouvent également.

Il est également intéressant de noter que le type de manifestation et le contexte de cooccurrence de l'approche Fréquentiste présentant les meilleurs résultats changent avec le nombre de candidats sélectionnés. Il apparaît dans la Table 23 que les cooccurrences entre les candidats et les manifestations indirectes des arguments de la relation partielle (i.e. Attached_Value) au niveau du Document donnent les meilleurs résultats lorsqu'un ou trois candidats sont sélectionnés. En revanche

lorsqu'on élargit cette sélection à cinq ou dix candidats, ce sont les cooccurrences des candidats avec les manifestations directes (i.e. Original_Value) au niveau des Sentence qui sont privilégiées.

La combinaison de différentes méthodes pourrait permettre de pallier les défauts inhérents à chacune des approches de complétion des relations partielles. Plusieurs pistent peuvent être explorées. Tout d'abord des méthodes de combinaisons similaires à celles employées avec les scores de pertinence de la Phase I (cf. Section 2.4) pourraient être envisagées : combinaison linéaire ou séquentielle des scores d'un candidat selon les différentes approches. Des méthodes de *machine learning* pourraient également être employées, par exemple en apprenant des poids sur les différents contextes et manifestations impliquées dans l'approche Fréquentiste, ou en attribuant des poids à des approches différentes dans une combinaison de celles-ci. L'utilisation de règles additionnelles dépendant des connaissances du domaine, c'est-à-dire de l'ontologie utilisée, devrait également permettre de facilement écarter certains candidats invalides. Par exemple en définissant des valeurs minimales et maximales d'un argument Temperature ou en restreignant un certain sous-ensemble des concepts spécifiques d'un argument Method à un certain type de relation de Permeability. Enfin une vision globale des connaissances d'un document serait une piste intéressante à explorer. En effet certaines instances d'arguments spécifiques (e.g. les paramètres de contrôles) sont souvent communes à différentes instances de relations dans un même document. Au contraire, d'autres le sont rarement (e.g. les valeurs de perméabilité). Des heuristiques basées sur la propagation, ou restriction, d'une information ajoutée à une instance de relation n-Aire aux autres instances de cette même relation pourraient permettre de guider la reconstitution des relations n-Aires. Ces points sont présentés plus en détail dans la Section 4.3.

Chapitre 4

Conclusion et Perspectives

Sommaire

4.1 Conclusion

Dans cette thèse, nous avons mené des travaux sur l'extraction automatique d'instances de relations n-Aires dans des articles scientifiques en utilisant une Ressource Termino-Ontologique (RTO). En structurant ces connaissances, celles-ci peuvent être réutilisées dans des outils d'aide à la décision. La RTO que nous utilisons permet de représenter les informations sous forme de relations n-Aires, modélisant les observations d'une mesure, en incluant l'objet de l'expérimentation, son résultat, sa méthode et les valeurs des paramètres de contrôle. Dans le domaine d'application traité, elles permettent d'associer un emballage à ses propriétés de perméabilité et d'indiquer les valeurs des paramètres de contrôle employées pour la mesure. Ces informations sont présentes dans les textes ainsi que dans les tableaux des documents et doivent être reconnues, extraites et structurées sous forme d'instances de relations n-Aires afin de peupler la RTO du domaine. Cette thèse s'inscrit dans les domaines de recherche sur les smart data et la capitalisation des connaissances sur des données expérimentales d'un domaine scientifique particulier. La méthode proposée dans cette thèse a été expérimentée sur le domaine des emballages alimentaires mais est conçue pour être générique et

pourrait être appliquée à d'autres domaines scientifiques.

Nous avons conçu une méthode en deux Phases suivant l'approche classiquement utilisée pour l'extraction des relations n-Aires dans la littérature [Zhou et al., 2014, Berrahou et al., 2017]. La Phase I se consacre à l'extraction des instances d'arguments dans les textes des articles, guidée par les composantes sémantique et terminologique de la RTO. La Phase II reconstitue ensuite des instances de relations n-Aires s'appuyant sur les instances de relations partielles détectées dans les tableaux des articles et en recherchant les instances d'arguments issues de la Phase I pertinentes pour venir les compléter.

Dans le cadre de cette thèse, nous proposons trois contributions pour l'extraction de relations n-Aires. La première concerne la représentation des instances d'arguments de relations n-Aires extraites des articles scientifiques. En nous appuyant sur des descripteurs sémantiques, lexicaux et structurels, nous proposons la représentation *Scientific Publication Representation* (SciPuRe), une représentation originale des instances d'arguments. Cette représentation propose une approche multi-descripteurs permettant d'une part, la sélection des instances d'arguments pertinentes, et d'autre part la reconstitution des instances de relations n-Aires.

La représentation SciPuRe nous permet de concevoir des mesures multi-descripteurs de la pertinence des instances d'arguments. En effet notre méthode d'extraction des instances d'arguments obtient un bon score de rappel (i.e. .85) mais produit également un nombre important de faux positifs (i.e. précision de .41). La représentation SciPuRe permet de concevoir des mesures de la pertinence de chaque instance d'argument en considérant ses descripteurs ontologiques, lexicaux et structurels. Il est ainsi possible d'ordonner les instances extraites durant la Phase I selon leur pertinence et ainsi filtrer les faux positifs. Nous avons montré que l'utilisation des scores de pertinence, ainsi que leur usage en combinaison, permet d'obtenir de meilleurs résultats en filtrant une partie des faux-positifs. L'amélioration de la précision de l'extraction des instances d'arguments quantitatifs est obtenue grâce à des scores lexicaux (e.g. $ICF^{term}_{segment}$, $ICF^{argument}_{segment}$) utilisant le caractère discriminant du terme représentant le concept de quantité au regard de la section dans laquelle il apparaît. Dans l'application aux emballages alimentaires, la précision de valeurs de perméabilité identifiée passe ainsi de .16 à une R-Précision de .30. De la même manière, la précision des instances

de Relative_Humidity passe de .28 à une R-Précision de .49. L'amélioration de la précision de l'extraction des instances d'arguments symboliques est obtenue grâce à une combinaison séquentielle de scores exploitant la notion de spécificité des instances d'arguments selon la structure de la RTO (i.e. CD^{node}_{target}) ainsi que la fréquence d'apparition (i.e.$TF^{term}_{document}$). En opérant par exemple ainsi un premier filtrage de 20% des instances de Packaging les moins spécifiques selon la RTO et en re-ordonnant ensuite les instances restantes selon leur fréquence, nous obtenons une R-Précision des instances de Packaging selon le score $Sequential(CD^{node}_{target}, TF^{term}_{document})$ de .63, contre .37 à l'origine.

Notre deuxième contribution consiste à proposer différentes méthodes, appliquées de manière originale, pour la reconstitution d'instances de relations n-Aires. Cette Phase II utilise les informations structurées dans les tableaux des articles pouvant être extraites en tant qu'instances de relations n-Aires partielles puis complétées par les instances d'arguments pertinentes issues de la Phase I. La représentation SciPuRe des instances d'arguments est étendue en Phase II pour représenter les instances de relations partielles extraites des tableaux présents dans les articles. Cette représentation originale, nommée STaRe (Scientific Table Representation), spécifie la nature et la structure de la relation n-Aire et situe l'instance de relation partielle dans son document. STaRe comprend également les descripteurs ontologiques et lexicaux des instances d'argument présents dans la relation. Cette représentation des instances de relations n-Aires factorise ainsi les représentations SciPuRe des instances d'arguments présents dans les tableaux. Nous avons développé trois approches pour compléter les instances partielles de relations n-Aires issues des tableaux avec les instances d'arguments extraits du texte. La première approche repose sur l'utilisation de la structure des articles scientifiques, la seconde approche utilise des mesures de cooccurrences entre instances d'arguments et manifestations des arguments des relations partielles issues des tableaux dans différents contextes, et enfin la troisième approche emploie des modèles de langage de type word-embedding pour rechercher les associations sémantiques entre instances d'arguments des textes et instances d'argument des relations partielles.

Lors de la Phase II, la représentation SciPuRe est à nouveau employée conjointement à la représentation STaRe. Ses descripteurs offrent en effet toutes les informations nécessaires pour mettre en œuvre les différentes méthodes

de reconstitution des relations que nous avons proposées. Nous avons testé nos différentes méthodes en faisant varier le nombre d'instances d'arguments candidates sélectionnées pour les instances d'arguments manquantes dans les instances de relations n-Aires partielles. Cette sélection s'inscrit dans une démarche d'annotation semi-automatique au cours de laquelle les annotateurs ont la possibilité de sélectionner le candidat adéquat pour chaque instance d'argument de la relation. Nos résultats ont montré que la méthode de reconstitution des relations n-Aires la plus adaptée dépend du nombre d'instances d'arguments candidates sélectionnées pour chaque instance d'argument manquant dans les instances de relations partielles. La méthode Fréquentiste est la mieux adaptée lorsque seule une instance d'argument est sélectionnée, les associations de cooccurrence entre cette instance d'argument et les manifestations des arguments de la relation partielle permettant d'obtenir un f-score de .48. Lors de la sélection de 3 ou 5 instances candidates, l'approche par Plongements Lexicaux utilisant les modèles de word-embedding basés sur le modèle BERT a montré des f-scores respectivement de .59 et .65. Cela indique l'importance de la proximité sémantique entre les descripteurs lexicaux des instances d'arguments extraits des textes et ceux des instances de relations partielles. Enfin, lorsque 10 instances candidates sont sélectionnées pour chaque instance d'argument manquant dans une instance de relation partielle, la méthode Structurelle Guidée atteint un f-score de .74.

Notre troisième contribution consiste en trois corpus annotés, utilisés pour l'évaluation des résultats de cette thèse et réutilisables par les communautés de recherche en TAL et en IC. Notre corpus *TRANSMAT Gold Standard* [Lentschat et al., 2021a, Lentschat et al., 2021c] (cf. Section 2.5) est composé d'instances d'arguments symboliques et quantitatifs concernant la composition des emballages alimentaires et leurs perméabilités au gaz. Celui-ci est composé de 1772 instances d'arguments (1050 symboliques et 722 quantitatives) annotées manuellement dans 50 documents. Le corpus *TRANSMAT tables data* [Lentschat et al., 2021d] (cf. Section 3.3.1) contient des tableaux extraits des articles scientifiques, annotés manuellement et automatiquement [Buche et al., 2011, Hignette et al., 2009], par des relations n-Aires partielles. Ce corpus présente 31 tableaux issus de 10 articles et contenant 779 instances d'arguments liées dans 332 instances de relations n-Aires. Le corpus *TRANSMAT n-Ary relations* [Lentschat, 2021] (cf. Section 3.3.2) contient des instances de relations n-Aires issues du corpus *TRANSMAT tables data* et complétées

manuellement par les instances d'arguments de *TRANSMAT Gold Standard*. Ces 332 instances de relations couvrent alors 1547 instances d'arguments. Les corpus annotés avec des relations n-Aires sont rares, rareté encore davantage présente pour les corpus de publications scientifiques annotées au niveau du document entier [Zhou et al., 2014]. Des corpus considérés comme des standards permettant l'évaluation et la comparaison des approches proposées pour l'extraction de relations n-Aires en domaine de spécialité restent à construire. De plus, le domaine des emballages alimentaires n'est pas un domaine déjà couvert par d'autres Gold Standards. Enfin, ces corpus peuvent également être utiles à d'autres questions de recherche (e.g. extraction d'unités de mesure, extraction automatique de données dans les tableaux, identification d'entités nommées en domaine de spécialité).

Pour finir, nos représentations multi-descripteurs SciPuRe pour la représentation des instances d'arguments, ainsi que la représentation STaRe qui factorise les descripteurs dans une représentation d'instance de relation n-Aire, peuvent être exploités dans d'autres travaux. Les descripteurs proposés sont exploitables dans les domaines du traitement automatique du langage, notamment l'extraction d'information et l'extraction de relations, et le domaine de l'ingénierie des connaissances.

4.2 Discussion sur la généricité de l'approche

La méthode décrite dans cette thèse est conçue dans un cadre générique d'extraction de connaissances depuis des articles scientifiques en domaine expérimental. Les connaissances produites au sein de ces domaines partagent des caractéristiques communes importantes à identifier afin de déterminer les cas à notre méthode est applicable. L'évaluation et l'identification des caractéristiques de notre pipeline applicables à d'autres domaines se situe dans le cadre du transfert de l'apprentissage [Torrey and Shavlik, 2010], où les méthodes apprises sur un domaine spécifique sont appliquées dans un domaine différent. L'évaluation de notre pipeline sur d'autres domaines expérimentaux permettra également d'évaluer l'étendue de sa généricité. Comme montré en fin de la Section 2.6.4 la Phase I de notre méthode est directement applicable à d'autres domaines expérimentaux disposant d'un corpus de document et d'une RTO de domaine. L'Exemple 2 illustre ainsi une extraction d'instances d'arguments guidée par la

RTO Valorcarn [Roche et al., 2020] dans le domaine de l'altération des aliments carnés par des agents pathogènes. Le nombre et la variété des domaines dans lesquels notre méthode d'extraction de relations n-Aires est appliquée permettra ainsi de confirmer, ou infirmer, les intuitions sur lesquelles notre méthode se fonde tout en augmentant sa robustesse.

Une application de notre méthode doit alors respecter un ensemble de critères pour être adaptée à un autre domaine :

1. **la quantité et l'intérêt des sources**. Notre méthode extrait les connaissances depuis des articles scientifiques. Il est donc entendu que des articles décrivant les connaissances d'intérêt sont disponibles en quantité suffisante (e.g plusieurs dizaines d'articles). Notre méthode se plaçant dans une approche *smart data* [Marcia, 2017, Duong et al., 2017], le nombre d'articles n'a pas à être important. En effet nous visons à faire ressortir la valeur individuelle de chaque donnée, et non à détecter de grandes tendances au sein d'un jeu de données. En revanche, l'extraction automatique de connaissances depuis un corpus composé uniquement de quelques articles n'est pas nécessairement adaptée à notre méthode, car un expert humain peut réaliser une extraction manuelle en un temps acceptable.
2. **la présence d'une RTO de domaine**. Notre méthode exploite une RTO afin de guider l'extraction de relations n-Aires dans les documents. Celle-ci est composée d'une ontologie noyau formalisant les relations n-Aires ainsi que les principaux concepts des domaines expérimentaux (i.e. unités de mesure et différentiation entre concepts symboliques et quantitatifs). Une ontologie de domaine contient ensuite les concepts spécifiques au domaine, concepts auxquels est associée à chacun une composante terminologique sous forme de labels. Le fait que les relations n-Aires soient définies dans l'ontologie noyau appuie la généricité de notre approche pour les domaines expérimentaux. L'applicabilité à un autre domaine nécessite alors une nouvelle ontologie de domaine. La plateforme @Web [1] contient aujourd'hui 7 ontologies de domaine toutes structurées autour de la même ontologie noyau. Nous avons utilisé l'ontologie TRANSMAT [Guillard et al., 2018], dans le domaine des emballages alimentaire, et les autres domaines couverts concernant les caractéristiques des aliments (texture, composition, altération

1. `https://ico.iate.inra.fr/atWeb/`

par des micro-organismes), la microfiltration, la bioraffinerie, et le blé dur.

3. **les stratégies différenciées pour les données symboliques et quantitatives**. Notre méthode extrait des relations n-Aires composées d'instances d'arguments symboliques et quantitatifs. Des stratégies différenciées ont été mises en place afin de considérer les différences entre ces deux types d'arguments. Cela concerne par exemple leur extraction (e.g. besoin de désambiguïsation des instances quantitatives), les scores adaptés à la mesure de leur pertinence (combinaison séquentielle pour les instances symboliques et score lexical sur le modèle icf pour les instances quantitatives). Être ainsi capable d'extraire différents types d'arguments permet à notre méthode d'être adaptées aux différentes situations de présentation et d'apparition des données dans les documents.
4. **l'exploitation des tableaux**. Notre méthode repose sur l'extraction de relations n-Aires partielles depuis les tableaux des articles scientifiques devant ensuite être complétées en Phase II par des instances d'arguments issues du texte. Il s'agit d'un choix permettant d'exploiter les données importantes présentes dans ces tableaux, et parfois absentes du texte, tout en exploitant les liens détectables via la structure des tableaux. Les tableaux sont souvent présents dans les articles scientifiques, afin de présenter, résumer ou comparer les résultats d'une étude. De plus, des méthodes existent afin de transformer le graphique d'une figure en tableau (e.g. *graph2tab* [Brandizi et al., 2012]) et peuvent être ajoutées à notre pipeline avec un coût en travail humain réduit. Être capable de gérer ces cas permet à notre méthode d'être appliquée dans un grand nombre de domaines expérimentaux.

Des applications de notre pipeline sur des domaines différents permettront également d'observer les effets de certains critères pouvant influer les résultats de l'extraction :

1. **la composition des documents**. Les documents composant les corpus de différents domaines peuvent présenter des caractéristiques différentes dans leur composition. Dans ces caractéristiques, nous comptons par exemple la présence de tableaux. Leur nombre et l'exhaustivité des informations qu'ils contiennent peuvent fortement influencer notre méthode en fournissant en début de Phase II des instances de relations n-Aires de quantité et de partialité variable. Notre méthode repose également sur l'utilisation des

sections des articles. Bien que l'usage soit de présenter des résultats de recherche selon une norme *abstract-Introduction-Method-Results-Conclusion*, des variations à cette norme propres à certains domaines peuvent influencer les résultats. Une évaluation sur différents domaines sera ainsi l'occasion d'observer plus largement l'influence des sections sur l'extraction des instances d'argument, leur pertinence et la recomposition des relations n-Aires afin de tirer des conclusions plus larges et adapter le pipeline aux besoins.

2. **la composition de la RTO**. L'ontologie noyau définit les relations n-Aires et l'ontologie de domaine guide leur extraction, notamment grâce à sa composante terminologique. Cependant, les différences entre ontologies de domaines sont également à discuter. L'exhaustivité des concepts et de leurs vocabulaires peut par exemple être remise en cause par des résultats d'extraction insuffisants. La structure de l'ontologie de domaine est également importante. La profondeur des sous-graphs associés à chaque argument est par exemple exploitée par notre score de pertinence sémantique. Ainsi, l'efficacité de ce score est dépendante de la finesse de description des concepts dans la RTO.
3. **la composition des relations n-Aires**. Notre méthode est appliquée à l'extraction de relations n-Aires concernant les relations de perméabilité. Celles-ci possèdent une arité de 7 arguments et sont composées de 2 arguments symboliques et 5 arguments quantitatifs. Les arguments symboliques ne sont ici pas sujet à l'ambiguïté, tout comme deux arguments quantitatifs (Temperatur et Thickness). Les instances des autres arguments quantitatifs nécessitent une désambiguïsation. Ces caractéristiques de relations n-Aires vont varier selon la définition des connaissances dans différents domaines. Cela peut alors influer sur les résultats de l'extraction ainsi que les meilleures méthodes à employer (e.g. [Zhou et al., 2014] a par exemple démontré expérimentalement qu'une forte arité influence négativement les résultats d'extraction). Les arguments composant la relation à extraire sont également importants. Nous pressentons par exemple que le nombre d'arguments à désambiguïser dans la relation va influencer l'efficacité des méthodes de reconstitution des relations n-Aires reposants sur les cooccurrences, car celles-ci devront alors exploiter davantage le descripteur Attached_Value ne permettant pas nécessairement

de différentier les doublons des instances d'arguments quantitatifs.

Cette thèse contribue donc aux travaux en extraction semi-automatique de connaissances à partir de publications scientifiques en domaine expérimental. Ces connaissances doivent être décrites dans une RTO et peuvent être présentent dans le texte des articles ainsi que dans les tableaux, sous forme de termes ou de valeurs numériques avec leurs unités de mesure. Une étude pluridisciplinaire constituerait une contribution à la méthode présentée ici, en permettant une évaluation large des approches mises en œuvre et en étudiant les différences spécifiques entre domaines, permettant ainsi d'enrichir notre méthode.

4.3 Perspectives

Plusieurs questions et thématiques ont émergé lors de nos travaux, pouvant venir les compléter ou constituer des débouchés de nos travaux de recherche. Quelques pistes existent pour enrichir les approches proposées dans ce manuscrit, comme la combinaison des scores Structurels, Fréquentistes et par Plongements Lexicaux utilisées durant la Phase II. Une méthode combinant les approches Structurelle et Fréquentiste permettrait par exemple d'accorder un poids plus important aux cooccurrences détectées dans des sections spécifiques des documents devant contenir les instances d'arguments pertinentes pour compléter la relation partielle.

D'autres perspectives existent pour le développement de nos travaux de thèse. Tout d'abord des extensions de notre méthode peuvent être développées : intégration de notre pipeline et des représentation SciPuRe et STaRe avec la plateforme @Web, ce qui permet ensuite de déployer une méthode d'apprentissage par renforcement. Parmi les perspectives méthodologiques, deux approches semblent prometteuses. La première consiste à une extraction simultanées des instances de relations n-Aires partielles et des instances d'arguments les complétant en utilisant des critères de raisonnement guidant la reconstitution. La seconde perspective méthodologique s'inspire de [Quirk and Poon, 2016, Song et al., 2018], qui représente les documents en graphes afin de réduire l'impacte de la dispersion des instances d'arguments dans les documents, et l'augmente avec les descripteurs des représentations SciPuRe et de STaRe.

4.3.1 Perspectives opérationnelles

L'intégration du pipeline d'extraction des relations n-Aires à @Web est un point à développer afin d'inclure cette solution d'extraction (semi-)automatique au processus d'ajout de connaissances à cette plateforme. Notre méthode reposant sur l'utilisation d'une RTO de domaine, et étant considéré que cette RTO couvre l'ensemble du domaine représenté (i.e. que l'ensemble des concepts du domaine et de leurs relations sont représentés), la mise en place d'une méthode semi-automatique d'élaboration de RTO de domaine serait un apport important afin d'améliorer les résultats de notre méthode. Dans cette optique, les pré-traitements que nous utilisons peuvent être employés afin de maintenir à jour la composante terminologique d'une RTO de domaine, en ajoutant aux concepts les nouveaux labels rencontrés durant la détection de variations terminologiques et d'unités de mesure.

Afin de contextualiser les instances d'arguments et de relations n-Aires dans la RTO, il est nécessaire d'intégrer les représentations SciPuRe et STaRe extraites dans une extension de la RTO. Cela passe par une extension de la core-ontology permettant d'aligner les descripteurs des représentations avec la structure de la RTO. Les descripteurs ontologiques font déjà référence à la RTO, leur alignement est donc trivial. Les descripteurs lexicaux font référence à la valeur de l'instance d'argument ainsi qu'aux labels employés, labels sont présents dans la RTO au format *skos*. Les descripteurs structurels peuvent eux profiter d'une formalisation en *skosxl* [2], une extension de *skos*, permettant de définir de nouveaux espaces de noms. Par exemple le descripteur Segment d'une instance d'argument pourrait être représenté sous forme de : $skosxl : SegmentLabel = "Matrials\_and\_Methods$.

Cette contextualisation des instances d'arguments dans la base de connaissance offre des critères interrogeables avec des langages de requête (e.g. SPARQL). Cela permet de créer un système d'interrogation d'une base de documents textuels avec un langage riche qui exploite à la fois les descripteurs de SciPure et de STaRe ainsi que la structure des relations n-Aires de la RTO. L'intérêt d'utiliser un tel langage d'interrogation pourrait par exemple être la possibilité de remonter des résultats contextualisés dans le texte et répondant à un ensemble de critères choisis par l'utilisateur (e.g. récupérer des extraits d'articles dans lesquels ont été trouvées

2. `https://www.w3.org/TR/skos-reference/skos-xl.html`

les instances de relations n-Aires de perméabilité à l'oxygène pour des emballages composites à base de PHBV dans une gamme de température entre 18 et 24 °C, valeur de température spécifiée dans la section *Materials and Methods*).

Les techniques d'apprentissage par renforcement [Sutton and Barto, 2018] pourraient également être appliquées sur les approches que nous déployons durant les Phases I et II. Durant la Phase I, différents scores sont utilisés afin d'ordonner les instances d'arguments selon leur pertinence et les filtrer. D'autres scores sont ensuite employés en Phase II afin de déterminer les instances d'arguments les plus adéquates à compléter les instances de relations partielles issues des tableaux. Nous pourrions employer des méthodes d'apprentissage par renforcement afin de déterminer progressivement les scores, et leurs paramètres, les mieux adaptés pour chaque tâche. Cette méthode serait déployée sur un corpus d'apprentissage, ou plus probablement dans une démarche de validation par les experts des résultats proposés, fournissant ainsi un retour au programme d'apprentissage actif. Une telle approche serait également intéressante à évaluer simultanément sur des domaines multiples. Cela permettrait de mesurer la portée du transfert de cet apprentissage, de l'applicabilité des scores déterminés pour un domaine et utilisés dans un autre domaine expérimental, et de déterminer les critères généraux les plus importants à considérer pour l'extraction de connaissances expérimentales dans des articles scientifiques.

4.3.2 Perspectives méthodologiques

Extraction simultanée des instances de relations et d'arguments. Dans cette thèse nous avons adopté l'approche majoritaire dans la littérature pour l'extraction d'instances de relations n-Aires [Zhou et al., 2014] en identifiant tout d'abord les instances d'arguments dans les textes puis en reconstituant les instances de relations n-Aires. Nous avons ajouté à cela l'extraction de relations n-Aires partielles depuis les tableaux des articles, étapes rendue possible par leur présence dans les publications scientifiques. Adopter une approche extrayant durant la même phase les instances de relations n-Aires et leurs instances d'arguments permettrait de raisonner sur d'autres critères tout au long de l'extraction. Dans ce cadre, une extraction simultanée des relations n-Aires et de leurs arguments se fonderait également sur l'utilisation des relations partielles présentes dans les tableaux.

Cette approche utiliserait également les instances de relations n-Aires partielles contenues dans les tableaux d'un document, car ces informations sont structurées et considérées comme pertinentes . Pour chacune de ces relations, nous connaissons ainsi les instances d'arguments qui les composent ainsi que les instances d'arguments manquantes. La recherche de ces instances d'arguments peut alors se faire selon différents critères, comme ceux déjà utilisés dans cette thèse ou de nouveaux à élaborer. L'utilisation d'un langage de requête permettrait dans un premier temps de contrôler l'ajout d'instances d'arguments aux relations n-Aires avec un ensemble de règles. Ces règles peuvent être fixées par des experts ou inférées en accord avec les connaissances déjà incluses dans une base de connaissance (e.g. la valeur du paramètre de contrôle de la température ne dépasse jamais $50°C$ dans une mesure de perméabilité). Les descriptions SciPuRe et STaRe sont pleinement employables pour ce travail de raisonnement sur les instances de relations n-Aires partielles et sur les instances d'arguments.

Le principal intérêt d'une extraction simultanée des instances de relations n-Aires est de procéder à des raisonnements durant l'ajout des instances d'arguments aux relations. Deux axes peuvent ainsi être explorés : l'ordre de complétion des arguments et un raisonnement sur le partage ou l'exclusivité des instances d'arguments entre les relations n-Aires d'un même document. L'ordre d'ajout des instances d'arguments dans la relation n-Aire partielle peut en effet avoir une influence sur les résultats (e.g. détecter en premier lieu la méthode permettrait de trouver plus aisément la température). Au contraire, l'ajout d'une instance invalide d'un argument pourrait se propager en entraînant l'ajout d'autres instances invalides pour les arguments. Une extraction simultanée des instances des relations n-Aires permettrait également de raisonner sur les instances d'arguments au niveau du document. En effet, certaines instances d'arguments sont spécifiques à une instance de relation en particulier (e.g. une valeur de perméabilité n'appartient jamais à plusieurs relations différentes). Au contraire, des instances d'arguments peuvent être propagées à toutes les instances d'une même relation n-Aire (e.g. la valeur du paramètre de contrôle relative à la température est la même pour toutes les relations de perméabilité). Pour finir, des déductions plus avancées peuvent être réalisées en comparant les instances de relations différentes. Les objets d'étude présentés dans un document (e.g. les emballages) ont été évalués afin d'en déduire plusieurs de leurs caractéristiques (e.g. perméabilités et compositions). Ainsi, si l'on a trouvé le Packaging d'une relation 02_Permeability_Relation, alors il existe

probablement une relation Impact_Factor_Component avec ce même Packaging.

Ces différents critères sont à considérer dans une approche simultanée de l'extraction des instances de relations n-Aires et de leurs arguments. Les règles qui découlent de leur prise en compte devraient tendre vers un maximum de généricité afin d'être applicables à d'autres domaines. Cela demande alors une étude de la signature des relations n-Aires afin d'en déduire des heuristiques contrôlant l'extraction (e.g. une instance d'un argument résultat ne peut être partagé entre plusieurs instances de relations). Cependant, il est probable que la création de certaines règles nécessite l'intervention d'experts du domaine, ou un apprentissage effectué sur un corpus annoté ou une base de connaissances contenant des instances de relations n-Aires existantes.

Méthode de représentation de document en graphe pour l'extraction d'instances de relations n-Aires. Les travaux de [Quirk and Poon, 2016, Song et al., 2018] se consacrent à l'extraction de relations n-Aires en utilisant des réseaux de neurones de type LSTM. La représentation que les auteurs font des documents pour extraire les relations n-Aires pourrait être adaptée à notre approche. Ceux-ci transforment les documents en graphes directionnels, dont les termes sont les sommets et les arêtes entre ceux-ci sont créées par leur séquence (i.e. l'ordre des mots dans la phrase), les dépendances syntaxiques à l'intérieur de chaque phrase et entre racines syntaxiques de phrases successives. L'intérêt principal de cette approche, qui effectue une représentation au niveau du document, est de réduire l'impact de la dispersion des instances d'arguments dans la tâche de reconstitution des relations n-Aires en connectant différentes parties des documents.

Cette représentation est d'intérêt, car elle pourrait être enrichie en exploitant les descripteurs de la représentation SciPuRe. L'utilisation de différents critères, les descripteurs de la représentation SciPuRe, permettrait non seulement d'obtenir cette représentation du document en graphe mais également de l'enrichir avec de l'information sémantique et ainsi augmenter les connexions existantes dans ce graphe. Les descripteurs structurels permettent de créer les sommets et arêtes permettant d'adapter l'approche de [Quirk and Poon, 2016] et de l'augmenter avec la considération de nouvelles informations (i.e. paragraphes, sections, sous-sections). Les descripteurs ontologiques enrichiraient cette approche en

identifiant les noeuds sémantiquement similaires, des arêtes pouvant ensuite être créées entre eux afin de lier les occurrences d'un même concept et ainsi réduire l'influence de la dispersion des instances d'arguments. En effet, [Song et al., 2018] a montré que l'augmentation de la connectivité du graphe améliore les résultats de la reconnaissance des relations n-Aires. Les descripteurs lexicaux permettent eux d'identifier les noeuds d'intérêt dans les graphes, les instances d'arguments entrant dans la composition des relations n-Aires. Si cette approche représentant un document en graphe est adoptée pour l'extraction de relations n-Aires, il sera alors important de considérer les critères sur lesquelles créer les connexions dans ce graphe et rechercher ceux offrant le gain de performance le plus important.

Bibliographie

[Agrawal et al., 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216.

[Akbik et al., 2019] Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). Flair : An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

[Akimoto et al., 2019] Akimoto, K., Hiraoka, T., Sadamasa, K., and Niepert, M. (2019). Cross-sentence n-ary relation extraction using lower-arity universal schemas. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6226–6232.

[Amami et al., 2012] Amami, M., Elkhlifi, A., and Faiz, R. (2012). Bioev : a system for learning biological event extraction. In *2012 International Conference on Information Technology and e-Services*, pages 1–5. IEEE.

[Andrade and Bork, 2000] Andrade, M. A. and Bork, P. (2000). Automated extraction of information in molecular biology. *FEBS letters*, 476(1-2) :12–17.

[Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia : A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

[Aussenac-Gilles et al., 2006] Aussenac-Gilles, N., Aussenac-Gilles, N., Condamines, A., and Sèdes, F. (2006). *Évaluation et maintenance des ressources termino-ontologiques : une question à approfondir*. Cépaduès.

[Avram et al., 2021] Avram, A.-M., Zaharia, G.-E., Cercel, D.-C., and Dascalu, M. (2021). Upb at semeval-2021 task 8 : Extracting semantic information on measurements as multi-turn question answering.

[Bach and Badaskar, 2007] Bach, N. and Badaskar, S. (2007). A review of relation extraction. *Literature review for Language and Statistics II*, 2 :1–15.

[Bachimont, 2000] Bachimont, B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, pages 305–323.

[Bellandi et al., 2007] Bellandi, A., Furletti, B., Grossi, V., and Romei, A. (2007). Ontology-driven association rule extraction : A case study. *Contexts and Ontologies Representation and Reasoning*, 10.

[Bellinger et al., 2004] Bellinger, G., Castro, D., and Mills, A. (2004). Data, information, knowledge, and wisdom. `http://dx.doi.org/10.18167/DVN1/FC2YXC`. Accessed : 2021-08-31.

[Bellot et al., 2019] Bellot, P., Cappellato, L., Ferro, N., Mothe, J., Murtagh, F., Nie, J.-Y., SanJuan, E., Soulier, L., and Trabelsi, C. (2019). Report on clef 2018 : Experimental ir meets multilinguality, multimodality, and interaction. *SIGIR Forum*, 52(2) :72–82.

[Beltagy et al., 2019] Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert : A pretrained language model for scientific text. *arXiv preprint arXiv :1903.10676*.

[Berrahou, 2015] Berrahou, S. L. (2015). *Extraction d'arguments de relations n-aires dans les textes guidée par une RTO de domaine*. PhD thesis, Université de Montpellier.

[Berrahou et al., 2017] Berrahou, S. L., Buche, P., Dibie, J., and Roche, M. (2017). Xart : Discovery of correlated arguments of n-ary relations in text. *Expert Systems with Applications*, 73 :115–124.

[Berrahou et al., 2015] Berrahou, S. L., Buche, P., Dibie-Barthelemy, J., and Roche, M. (2015). Identification des unités de mesure dans les textes scientifiques. In *Actes de la 22e conf\'erence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 88–94.

[Bhagavatula et al., 2015] Bhagavatula, C. S., Noraset, T., and Downey, D. (2015). Tabel : Entity linking in web tables. In Arenas, M., Corcho, O., Simperl, E., Strohmaier, M., d'Aquin, M., Srinivas, K., Groth, P., Dumontier, M., Heflin, J., Thirunarayan, K., Thirunarayan, K., and Staab, S., editors, *The Semantic Web - ISWC 2015*, pages 425–441, Cham. Springer International Publishing.

[Björne et al., 2009] Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., and Salakoski, T. (2009). Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 10–18, Boulder, Colorado. Association for Computational Linguistics.

[Blaschke et al., 1999] Blaschke, C., Andrade, M. A., Ouzounis, C. A., and Valencia, A. (1999). Automatic extraction of biological information from scientific text : protein-protein interactions. In *Ismb*, volume 7, pages 60–67.

[Bollacker et al., 2008] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase : a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

[Bourigault and Jacquemin, 1999] Bourigault, D. and Jacquemin, C. (1999). Term extraction-i-term clustering : An integrated platform for computer-aided terminology. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

[Boyce et al., 2017] Boyce, B. R., Boyce, B. R., Meadow, C. T., Kraft, D. H., Kraft, D. H., and Meadow, C. T. (2017). *Text information retrieval systems*. Elsevier.

[Brack et al., 2020] Brack, A., D'Souza, J., Hoppe, A., Auer, S., and Ewerth, R. (2020). Domain-independent extraction of scientific concepts from research articles. In Jose, J. M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M. J., and Martins, F., editors, *Advances in Information Retrieval*, pages 251–266, Cham. Springer International Publishing.

[Brandizi et al., 2012] Brandizi, M., Kurbatova, N., Sarkans, U., and Rocca-Serra, P. (2012). graph2tab, a library to convert experimental workflow graphs into tabular formats. *Bioinformatics*, 28(12) :1665–1667.

[Bravo et al., 2015] Bravo, À., Piñero, J., Queralt-Rosinach, N., Rautschka, M., and Furlong, L. I. (2015). Extraction of relations between genes and diseases from text and large-scale data analysis : implications for translational research. *BMC bioinformatics*, 16(1) :1–17.

[Brickley et al., 2014] Brickley, D., Guha, R. V., and McBride, B. (2014). Rdf schema 1.1. *W3C recommendation*, 25 :2004–2014.

[Brill, 1995] Brill, E. (1995). Transformation-based error-driven learning and natural language processing : A case study in part-of-speech tagging. *Computational linguistics*, 21(4) :543–565.

[Brin et al., 1997] Brin, S., Motwani, R., and Silverstein, C. (1997). Beyond market baskets : Generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 265–276.

[Buche et al., 2013a] Buche, P., Dervaux, S., Dibie-Barthelemy, J., Ibanescu, L., Soler, L., and Touhami, R. (2013a). Intégration de données hétérogènes et imprécises guidée par une ressource termino-ontologique. *Revue des Sciences et Technologies de l'Information-Série RIA : Revue d'Intelligence Artificielle*, 27(4-5) :539–568.

[Buche et al., 2012] Buche, P., Dervaux, S., Dibie-Barthelemy, J., Ibanescu, L., and Touhami, R. (2012). Vers la publication d'une rto dédiée à l'annotation de relations n-aires. In *IC'2012, 23. Journées francophones d'Ingénierie des Connaissances, Atelier «Ontologies et Jeux de Données pour évaluer le web sémantique»*, pages 3–p.

[Buche et al., 2021] Buche, P., Dervaux, S., Leconte, N., Belna, M., Granger-Delacroix, M., Garnier-Lambrouin, F., Gregory, G., Barrois, L., and Gesan-Guiziou, G. (2021). Milk microfiltration process dataset annotated from a collection of scientific papers. *Data in Brief*, 36 :107063.

[Buche et al., 2011] Buche, P., Dibie-Barthelemy, J., Ibanescu, L., and Soler, L. (2011). Fuzzy web data tables integration guided by an ontological and terminological resource. *IEEE Transactions on Knowledge and Data Engineering*, 25(4) :805–819.

[Buche et al., 2013b] Buche, P., Dibie-Barthélemy, J., Ibanescu, L., and Soler, L. (2013b). Fuzzy web data tables integration guided by an ontological and terminological resource. *IEEE Transactions on Knowledge and Data Engineering*, 25(4) :805–819.

[Bunescu and Mooney, 2005] Bunescu, R. and Mooney, R. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 724–731.

[Cafarella et al., 2005] Cafarella, M. J., Downey, D., Soderland, S., and Etzioni, O. (2005). Knowitnow : Fast, scalable information extraction from the web. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 563–570.

[Chan and Roth, 2011] Chan, Y. S. and Roth, D. (2011). Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 551–560.

[Charnois et al., 2009] Charnois, T., Plantevit, M., Rigotti, C., and Crémilleux, B. (2009). Fouille de données séquentielles pour l'extraction d'information dans les textes. *Traitement Automatique des Langues*, pages pp59–87.

[Chen et al., 2006] Chen, J., Ji, D., Tan, C. L., and Niu, Z.-Y. (2006). Semi-supervised relation extraction with label propagation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume : Short Papers*, pages 25–28.

[Church and Hanks, 1990] Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1) :22–29.

[Cohen et al., 2010] Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C., and Hunter, L. E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC bioinformatics*, 11(1) :492.

[Cohen et al., 2003] Cohen, W. W., Ravikumar, P., Fienberg, S. E., et al. (2003). A comparison of string distance metrics for name-matching tasks. In *IIWeb*, volume 3, pages 73–78.

[Cooper, 1971] Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information storage and retrieval*, 7(1) :19–37.

[Cram and Daille, 2016] Cram, D. and Daille, B. (2016). Terminology extraction with term variant detection. pages 13–18.

[Craswell, 2009] Craswell, N. (2009). *Precision at n*, pages 2127–2128. Springer US, Boston, MA.

[Dannélls, 2006] Dannélls, D. (2006). Automatic acronym recognition. In *Demonstrations*.

[Davidson, 1980] Davidson, D. (1980). *The logical form of action sentences. Essays on actions and events*. Oxford : Clarendon Press.[orginally published in 1967].

[Davletov et al., 2021] Davletov, A., Gordeev, D., Arefyev, N., and Davletov, E. (2021). LIORI at SemEval-2021 task 8 : Ask transformer for measurements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1249–1254, Online. Association for Computational Linguistics.

[De Mauro et al., 2016] De Mauro, A., Greco, M., and Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*.

[Dice, 1945] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3) :297–302.

[Doddington et al., 2004] Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., and Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

[Dou et al., 2015] Dou, D., Wang, H., and Liu, H. (2015). Semantic data mining : A survey of ontology-based approaches. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, pages 244–251.

[Duong et al., 2017] Duong, T. H., Nguyen, H. Q., and Jo, G. S. (2017). Smart data : Where the big data meets the semantics. *Computational Intelligence and Neuroscience*, 2017 :6925138.

[Eckart de Castilho and Gurevych, 2014] Eckart de Castilho, R. and Gurevych, I. (2014). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

[Eckart de Castilho et al., 2016] Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan. The COLING 2016 Organizing Committee.

[Ehrmann et al., 2013a] Ehrmann, M., della Rocca, L., Steinberger, R., and Tanev, H. (2013a). Acronym recognition and processing in 22 languages.

[Ehrmann et al., 2013b] Ehrmann, M., Della Rocca, L., Steinberger, R., and Tannev, H. (2013b). Acronym recognition and processing in 22 languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 237–244, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

[Elango, 2005] Elango, P. (2005). Coreference resolution : A survey. *University of Wisconsin, Madison, WI.*

[Etzioni et al., 2008] Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, 51(12) :68–74.

[Fabre et al., 2020] Fabre, C., Buche, P., Rouau, X., and Mayer-Laigle, C. (2020). Milling itineraries dataset for a collection of crop and wood by-products and granulometric properties of the resulting powders. *Data in brief*, 33 :106430.

[Ferré et al., 2020] Ferré, A., Bossy, R., Ba, M., Deleger, L., Lavergne, T., Zweigenbaum, P., and Nédellec, C. (2020). Handling entity normalization with no annotated corpus : weakly supervised methods based on distributional representation and ontological information. In *12. Conference on Language Resources and Evaluation*, pages 11–16.

[Foppiano et al., 2019] Foppiano, L., Romary, L., Ishii, M., and Tanifuji, M. (2019). Automatic identification and normalisation of physical measurements in scientific literature. In *Proceedings of the ACM Symposium on Document Engineering 2019*, pages 1–4.

[Franco-Salvador et al., 2014] Franco-Salvador, M., Rosso, P., and Navigli, R. (2014). A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 414–423.

[Gamallo et al., 2012] Gamallo, P., Garcia, M., and Fernández-Lanza, S. (2012). Dependency-based open information extraction. In *Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP*, pages 10–18.

[Gashteovski et al., 2017] Gashteovski, K., Gemulla, R., and Corro, L. d. (2017). Minie : minimizing facts in open information extraction. Association for Computational Linguistics.

[Gerard Salton, 1983] Gerard Salton, M. M. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

[Ghersedine et al., 2012] Ghersedine, A., Buche, P., Dibie-Barthélemy, J., Hernandez, N., and Kamel, M. (2012). Extraction de relations n-aires interphrastiques guidée par une rto. In *CORIA : Conférence en Recherche d'Information et Applications*, pages 179–190.

[Giunti et al., 2019] Giunti, M., Sergioli, G., Vivanet, G., and Pinna, S. (2019). Representing n-ary relations in the semantic web. *Logic Journal of the IGPL*.

[Grau et al., 2009] Grau, B., Ligozat, A.-L., and Minard, A.-L. (2009). Corpus study of kidney-related experimental data in scientific papers. In *International Workshop on Biomedical Information Extraction in conjunction with Recent Advances in Natural Language Processing*.

[Grave, 2014] Grave, E. (2014). A convex relaxation for weakly supervised relation extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1580–1590.

[Greenwood and Stevenson, 2006] Greenwood, M. A. and Stevenson, M. (2006). Improving semi-supervised acquisition of relation extraction patterns. In *Proceedings of the Workshop on Information Extraction beyond the Document*, pages 29–35.

[Grishman, 2015] Grishman, R. (2015). Information extraction. *IEEE Intelligent Systems*, 30(5) :8–15.

[Grishman, 2019] Grishman, R. (2019). Twenty-five years of information extraction. *Natural Language Engineering*, 25(6) :677–692.

[Grishman and Sundheim, 1996] Grishman, R. and Sundheim, B. (1996). Message Understanding Conference- 6 : A brief history. In *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*.

[Groth et al., 2018] Groth, P., Lauruhn, M., Scerri, A., and Daniel Jr, R. (2018). Open information extraction on scientific text : An evaluation. *arXiv preprint arXiv :1802.05574*.

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2) :199–220.

[Guari, 2021] Guari, R. (2021). Conception et implémentation d'une méthode semi-automatique d'annotation de tableaux de données issus d'articles

scientifiques guidée par ontologie. *Rapport interne - Université de Montpellier*. stage assistant ingénieur.

[Guarino et al., 2009] Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology ? In *Handbook on ontologies*, pages 1–17. Springer.

[Guillard et al., 2015] Guillard, V., Buche, P., Destercke, S., Tamani, N., Croitoru, M., Menut, L., Guillaume, C., and Gontard, N. (2015). A Decision Support System to design modified atmosphere packaging for fresh produce based on a bipolar flexible querying approach. *Computers and Electronics in Agriculture*, 111 :131–139.

[Guillard et al., 2018] Guillard, V., Buche, P., Menut, L., and Dervaux, S. (2018). Matter transfer ontology. `https://doi.org/10.15454/NK24ID`. Accessed : 2021-07-19.

[Guillard et al., 2017] Guillard, V., Couvert, O., Stahl, V., Buche, P., Hanin, A., Denis, C., Dibie-Barthelemy, J., Dervaux, S., Loriot, C., Vincelot, T., et al. (2017). Map-opt : a software for supporting decision-making in the field of modified atmosphere packaging of fresh non respiring foods. *Packaging Research*, 2(1) :28.

[Gwet, 2014] Gwet, K. L. (2014). *Handbook of inter-rater reliability : The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

[Hahn and Oleynik, 2020] Hahn, U. and Oleynik, M. (2020). Medical information extraction in the age of deep learning. *Yearbook of Medical Informatics*, 29(01) :208–220.

[Hao et al., 2017] Hao, T., We, Y., Qiang, J., Wang, H., and Lee, K. (2017). The representation and extraction of qunatitative information. In *Proceedings of the 13th joint ISO-ACL workshop on interoperable semantic annotation (ISA-13)*.

[Harispe, 2014] Harispe, S. (2014). *Knowledge-based semantic measures : From theory to applications*. PhD thesis.

[Harman, 1996] Harman, D. (1996). The text retrieval conferences (trecs). Technical report, NATIONAL INST OF STANDARDS AND TECHNOLOGY GAITHERSBURG MD.

[Harper et al., 2021] Harper, C., Cox, J., Kohler, C., Scerri, A., Daniel Jr, R., and Groth, P. (2021). Semeval 2021 task 8 : Measeval–extracting

counts and measurements and their related contexts. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021), Bangkok, Thailand (online). Association for Computational Linguistics*.

[Hignette et al., 2009] Hignette, G., Buche, P., Dibie-Barthélemy, J., and Haemmerlé, O. (2009). Fuzzy annotation of web data tables driven by a domain ontology. In *European Semantic Web Conference*, pages 638–653. Springer.

[Hobbs, 1978] Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, 44(4) :311–338.

[Hofmann et al., 2009] Hofmann, K., Tsagkias, M., Meij, E., and De Rijke, M. (2009). The impact of document structure on keyphrase extraction. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1725–1728. ACM.

[Jaccard, 1901] Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat*, 37 :241–272.

[Jain et al., 2007] Jain, A., Cucerzan, S., and Azzam, S. (2007). Acronym-expansion recognition and ranking on the web. In *2007 IEEE International Conference on Information Reuse and Integration*, pages 209–214.

[Jessop et al., 2011a] Jessop, D. M., Adams, S. E., and Murray-Rust, P. (2011a). Mining chemical information from open patents. *Journal of cheminformatics*, 3(1) :1–17.

[Jessop et al., 2011b] Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L., and Murray-Rust, P. (2011b). Oscar4 : a flexible architecture for chemical text-mining. *Journal of cheminformatics*, 3(1) :1–12.

[Ji and Grishman, 2011] Ji, H. and Grishman, R. (2011). Knowledge base population : Successful approaches and challenges. In *Proceedings of the 49th annual meeting of the association for computational linguistics : Human language technologies*, pages 1148–1158.

[Jia et al., 2019] Jia, R., Wong, C., and Poon, H. (2019). Document-level n-ary relation extraction with multiscale representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.

[Jiao and Zhang, 2021] Jiao, Q. and Zhang, S. (2021). A brief survey of word embedding and its recent development. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 5, pages 1697–1701.

[Jonnalagadda et al., 2015] Jonnalagadda, S. R., Goyal, P., and Huffman, M. D. (2015). Automating data extraction in systematic reviews : a systematic review. *Systematic reviews*, 4(1) :78.

[Kando, 1997] Kando, N. (1997). Text-level structure of research papers : Implications for text-based information processing systems. In *BCS-IRSG Annual Colloquium on IR Research*.

[Kapetanios et al., 2013] Kapetanios, E., Tatar, D., and Sacarea, C. (2013). *Natural language processing : semantic aspects*. CRC Press.

[Khattak et al., 2019] Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., and Rudzicz, F. (2019). A survey of word embeddings for clinical text. *Journal of Biomedical Informatics : X*, 4 :100057.

[Khetan et al., 2021] Khetan, V., M, A. K., Wetherley, E., Eneva, E., Sengupta, S., and Fano, A. E. (2021). Knowledge graph anchored information-extraction for domain-specific insights.

[Kim et al., 2017] Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G., and Olivetti, E. (2017). Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials*, 29(21) :9436–9444.

[Klink et al., 2000] Klink, S., Dengel, A., and Kieninger, T. (2000). Document structure analysis based on layout and textual features. In *Proc. of International Workshop on Document Analysis Systems, DAS2000*, pages 99–111. Citeseer.

[Kobayashi and Ng, 2020] Kobayashi, H. and Ng, V. (2020). Bridging resolution : A survey of the state of the art. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3708–3721, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[Konys, 2018] Konys, A. (2018). Towards knowledge handling in ontology-based information extraction systems. *Procedia Computer Science*, 126 :2208–2218. Knowledge-Based and Intelligent Information Engineering Systems : Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia.

[Krallinger et al., 2013] Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., and Valencia, A. (2013). Overview of the chemical compound and drug name recognition (chemdner) task. In *BioCreative challenge evaluation workshop*, volume 2, page 2. Citeseer.

[Kumar, 2017] Kumar, S. (2017). A survey of deep learning methods for relation extraction. *arXiv preprint arXiv :1705.03645*.

[Lee et al., 2020] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4) :1234–1240.

[Lenca et al., 2007] Lenca, P., Vaillant, B., Meyer, P., and Lallich, S. (2007). Association rule interestingness measures : Experimental and theoretical studies. In *Quality Measures in Data Mining*, pages 51–76. Springer.

[Lentschat, 2021] Lentschat, M. (2021). Transmat n-ary relations. `https://doi.org/10.18167/DVN1/1BBJBQ`.

[Lentschat et al., 2021a] Lentschat, M., Buche, P., Dibie-Barthelemy, J., Menut, L., and Roche, M. (2021a). Food packaging permeability and composition dataset dedicated to text-mining. *Data in Brief*, 36 :107135.

[Lentschat et al., 2020] Lentschat, M., Buche, P., Dibie-Barthelemy, J., and Roche, M. (2020). Scipure : a new representation of textual data for entity identification from scientific publications. In *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics*, pages 220–226.

[Lentschat et al., 2021b] Lentschat, M., Buche, P., Dibie-Barthelemy, J., and Roche, M. (2021b). Representation and relevance scores of experimental data extracted with an ontological and terminological resource. *International Journal of Intelligent Information and Database Systems (IJIIDS)*, a paraître :–.

[Lentschat et al., 2021c] Lentschat, M., Buche, P., and Menut, L. (2021c). Transmat gold standard. `http://doi.org/10.18167/DVN1/U7HK8J`.

[Lentschat et al., 2021d] Lentschat, M., Buche, P., Menut, L., and Guari, R. (2021d). TRANSMAT tables data. `https://doi.org/10.18167/DVN1/GCZBC9`.

[Levenshtein, 1966] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

[Li et al., 2015] Li, H., Krause, S., Xu, F., Moro, A., Uszkoreit, H., and Navigli, R. (2015). Improvement of n-ary relation extraction by adding lexical semantics to distant-supervision rule learning. In *ICAART (2)*, pages 317–324.

[Li and Yang, 2018] Li, Y. and Yang, T. (2018). Word embedding for understanding natural language : a survey. In *Guide to big data applications*, pages 83–104. Springer.

[Lin and Pantel, 2001] Lin, D. and Pantel, P. (2001). Dirt@ sbt@ discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328.

[Lin et al., 2020] Lin, Y., Ji, H., Huang, F., and Wu, L. (2020). A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

[Liu and El-Gohary, 2017] Liu, K. and El-Gohary, N. (2017). Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. *Automation in Construction*, 81 :313–327.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

[Lousteau-Cazalet et al., 2016a] Lousteau-Cazalet, C., Barakat, A., Belaud, J.-P., Buche, P., Busset, G., Charnomordic, B., Dervaux, S., Destercke, S., Dibie, J., Sablayrolles, C., et al. (2016a). A decision support system for eco-efficient biorefinery process comparison using a semantic approach. *Computers and Electronics in Agriculture*, 127 :351–367.

[Lousteau-Cazalet et al., 2016b] Lousteau-Cazalet, C., Barakat, A., Belaud, J.-P., Buche, P., Busset, G., Charnomordic, B., Dervaux, S., Destercke, S., Dibie, J., Sablayrolles, C., et al. (2016b). A decision support system for eco-efficient biorefinery process comparison using a semantic approach. *Computers and Electronics in Agriculture*, 127 :351–367.

[Luan et al., 2018] Luan, Y., He, L., Ostendorf, M., and Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv :1808.09602*.

[Magerman, 1995] Magerman, D. M. (1995). Statistical decision-tree models for parsing. *arXiv preprint cmp-lg/9504030*.

[Makarov, 2018] Makarov, P. (2018). Automated acquisition of patterns for coding political event data : Two case studies. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 103–112, Santa Fe, New Mexico. Association for Computational Linguistics.

[Malmasi et al., 2015] Malmasi, S., Hassanzadeh, H., and Dras, M. (2015). Clinical information extraction using word representations. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 66–74.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). Evaluation in information retrieval. *Introduction to information retrieval*, 1 :188–210.

[Mao et al., 2003] Mao, S., Rosenfeld, A., and Kanungo, T. (2003). Document structure analysis algorithms : a literature survey. In *Document Recognition and Retrieval X*, volume 5010, pages 197–207. International Society for Optics and Photonics.

[Marcia, 2017] Marcia, L. Z. (2017). Smart data for digital humanities. *Journal of data and information science*, 2(1) :1.

[Marsi and Öztürk, 2015] Marsi, E. and Öztürk, P. (2015). Extraction and generalisation of variables from scientific publications. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 505–511.

[McDonald et al., 2005] McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., and White, P. (2005). Simple algorithms for complex relation extraction with applications to biomedical ie. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 491–498.

[McDowell and Cafarella, 2006a] McDowell, L. K. and Cafarella, M. (2006a). Ontology-driven information extraction with ontosyphon. In *The Semantic Web - ISWC 2006*, pages 428–444, Berlin, Heidelberg. Springer Berlin Heidelberg.

[McDowell and Cafarella, 2006b] McDowell, L. K. and Cafarella, M. (2006b). Ontology-driven information extraction with ontosyphon. In *International Semantic Web Conference*, pages 428–444. Springer.

[McGrath et al., 2011] McGrath, L. R., Domico, K., Corley, C. D., and Webb-Robertson, B.-J. (2011). Complex biological event extraction from full text using signatures of linguistic and semantic features. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 130–137, Portland, Oregon, USA. Association for Computational Linguistics.

[Meng and Morioka, 2015] Meng, F. and Morioka, C. (2015). Automating the generation of lexical patterns for processing free text in clinical documents. *Journal of the American Medical Informatics Association*, 22(5) :980–986.

[Mesquita et al., 2013] Mesquita, F., Schmidek, J., and Barbosa, D. (2013). Effectiveness and efficiency of open relation extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 447–457.

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

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Miller, 1995] Miller, G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, 38(11) :39–41.

[Minard et al., 2013] Minard, A.-L., Grau, B., Ligozat, A.-L., and Thomas, S. R. (2013). Extraction de relations complexes. application à des résultats expérimentaux en physiologie rénale. *Revue des Sciences et Technologies de l'Information-Série TSI : Technique et Science Informatiques*, 32(1) :75–108.

[Minard et al., 2010] Minard, A.-L., Ligozat, A.-L., and Grau, B. (2010). Extraction de résultats expérimentaux d'articles scientifiques pour le peuplement d'une base de données. In *10th International Conference on statistical analysis of textual data (JADT)*, volume 73, Roma, Italy.

[Mintz et al., 2009] Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

[Mizzaro, 1998] Mizzaro, S. (1998). How many relevances in information retrieval ? *Interacting with computers*, 10(3) :303–320.

[Mohit and Hwa, 2005] Mohit, B. and Hwa, R. (2005). Syntax-based semi-supervised named entity tagging. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 57–60.

[Mondal, 2020] Mondal, I. (2020). Bertchem-ddi : Improved drug-drug interaction prediction from text using chemical structure information. *arXiv preprint arXiv :2012.11599*.

[Móra et al., 2009] Móra, G., Farkas, R., Szarvas, G., and Molnár, Z. (2009). Exploring ways beyond the simple supervised learning approach for biological event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 137–140, Boulder, Colorado. Association for Computational Linguistics.

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1) :3–26.

[Nasar et al., 2018] Nasar, Z., Jaffry, S. W., and Malik, M. K. (2018). Information extraction from scientific articles : a survey. *Scientometrics*, 117(3) :1931–1990.

[Nédellec, 2004] Nédellec, C. (2004). Machine learning for information extraction in genomics—state of the art and perspectives. In *Text Mining and its Applications*, pages 99–118. Springer.

[Nédellec and Nazarenko, 2006] Nédellec, C. and Nazarenko, A. (2006). Ontologies and information extraction. *arXiv preprint cs/0609137*.

[Neumann et al., 2019a] Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019a). Scispacy : Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv :1902.07669*.

[Neumann et al., 2019b] Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019b). Scispacy : Fast and robust models for biomedical natural language processing. *ArXiv*, abs/1902.07669.

[Ng, 2017] Ng, V. (2017). Machine learning for entity coreference resolution : A retrospective look at two decades of research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

[Nguyen et al., 2019] Nguyen, P., Kertkeidkachorn, N., Ichise, R., and Takeda, H. (2019). Mtab : Matching tabular data to knowledge graph using probability models. *CoRR*, abs/1910.00246.

[Niklaus et al., 2018a] Niklaus, C., Cetto, M., Freitas, A., and Handschuh, S. (2018a). A survey on open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[Niklaus et al., 2018b] Niklaus, C., Cetto, M., Freitas, A., and Handschuh, S. (2018b). A survey on open information extraction.

[Norman et al., 1965] Norman, R. Z. et al. (1965). Structural models : An introduction to the theory of directed graphs.

[Okazaki and Ananiadou, 2006] Okazaki, N. and Ananiadou, S. (2006). A term recognition approach to acronym recognition. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 643–650. Association for Computational Linguistics.

[Pawar et al., 2017] Pawar, S., Palshikar, G. K., and Bhattacharyya, P. (2017). Relation extraction : A survey. *arXiv preprint arXiv :1712.05191*.

[Pazienza et al., 2005] Pazienza, M. T., Pennacchiotti, M., and Zanzotto, F. M. (2005). Terminology extraction : an analysis of linguistic and statistical approaches. In *Knowledge mining*, pages 255–279. Springer.

[Pearson, 1896] Pearson, K. (1896). Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187) :253–318.

[Peng et al., 2017] Peng, N., Poon, H., Quirk, C., Toutanova, K., and Yih, W.-t. (2017). Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5 :101–115.

[Peng et al., 2019] Peng, Y., Yan, S., and Lu, Z. (2019). Transfer learning in biomedical natural language processing : An evaluation of BERT and elmo on ten benchmarking datasets. *CoRR*, abs/1906.05474.

[Perera et al., 2020] Perera, N., Dehmer, M., and Emmert-Streib, F. (2020). Named entity recognition and relation detection for biomedical information extraction. *Frontiers in Cell and Developmental Biology*, 8 :673.

[Pershina et al., 2014] Pershina, M., Min, B., Xu, W., and Grishman, R. (2014). Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 732–738.

[Poerner et al., 2020] Poerner, N., Waltinger, U., and Schütze, H. (2020). Inexpensive domain adaptation of pretrained language models : Case studies on biomedical NER and covid-19 QA. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, pages 1482–1490, Online. Association for Computational Linguistics.

[Proux et al., 2000] Proux, D., Rechenmann, F., Julliard, L., et al. (2000). A pragmatic information extraction strategy for gathering data on genetic interactions. In *Ismb*, volume 8, pages 279–285.

[Pustejovsky et al., 2001] Pustejovsky, J., Castano, J., Cochran, B., Kotecki, M., and Morrell, M. (2001). Automatic extraction of acronym-meaning pairs from medline databases. In *MEDINFO 2001*, pages 371–375. IOS Press.

[Qi et al., 2020] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza : A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*.

[Quirk and Poon, 2016] Quirk, C. and Poon, H. (2016). Distant supervision for relation extraction beyond the sentence boundary. *arXiv preprint arXiv :1609.04873*.

[Quirk and Poon, 2017] Quirk, C. and Poon, H. (2017). Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, pages 1171–1182, Valencia, Spain. Association for Computational Linguistics.

[Ramadier and Lafourcade, 2016] Ramadier, L. and Lafourcade, M. (2016). Patrons sémantiques pour l'extraction de relations entre termes-application aux comptes rendus radiologiques. In *TALN : Traitement Automatique des Langues Naturelles*.

[Reymonet et al., 2007] Reymonet, A., Thomas, J., and Aussenac-Gilles, N. (2007). Modelling ontological and terminological resources in owl dl. In *Proceedings of ISWC*, volume 7.

[Rijgersberg et al., 2013] Rijgersberg, H., Van Assem, M., and Top, J. (2013). Ontology of units of measure and related concepts. *Semantic Web*, 4(1) :3–13.

[Roche et al., 2015] Roche, M., Fortuno, S., Lossio-Ventura, J. A., Akli, A., Belkebir, S., Lounis, T., and Toure, S. (2015). Extraction automatique des

mots-clés à partir de publications scientifiques pour l'indexation et l'ouverture des données en agronomie. *Cahiers Agricultures*, 24(5) :313–320.

[Roche et al., 2020] Roche, M., Teisseire, M., and Shrivastava, G. (2020). Valorcarn-tetis : Terms extracted with fastr (free extraction). `http://dx.doi.org/10.18167/DVN1/FC2YXC`. Accessed : 2021-06-10.

[Role and Nadif, 2011] Role, F. and Nadif, M. (2011). Handling the impact of low frequency events on co-occurrence based measures of word similarity. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR-2011). Scitepress*, pages 218–223.

[Rosenfeld and Feldman, 2006] Rosenfeld, B. and Feldman, R. (2006). Ures : an unsupervised web relation extraction system. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 667–674.

[Roth et al., 2013] Roth, B., Barth, T., Wiegand, M., and Klakow, D. (2013). A survey of noise reduction methods for distant supervision. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 73–78.

[Roth and Yih, 2004] Roth, D. and Yih, W.-t. (2004). A linear programming formulation for global inference in natural language tasks. Technical report, ILLINOIS UNIV AT URBANA-CHAMPAIGN DEPT OF COMPUTER SCIENCE.

[Rowley, 2007] Rowley, J. (2007). The wisdom hierarchy : representations of the dikw hierarchy. *Journal of information science*, 33(2) :163–180.

[Ru et al., 2018] Ru, C., Tang, J., Li, S., Xie, S., and Wang, T. (2018). Using semantic similarity to reduce wrong labels in distant supervision for relation extraction. *Information Processing & Management*, 54(4) :593–608.

[Sang and De Meulder, 2003] Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *arXiv preprint cs/0306050*.

[Shah et al., 2003] Shah, P. K., Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2003). Information extraction from full text scientific articles : where are the keywords ? *BMC bioinformatics*, 4(1) :20.

[Shah and Jain, 2014] Shah, R. and Jain, S. (2014). Ontology-based information extraction : An overview and a study of different approaches. *International journal of computer Applications*, 87(4).

[Shahab, 2017] Shahab, E. (2017). A short survey of biomedical relation extraction techniques. *arXiv preprint arXiv :1707.05850*.

[Shin et al., 2020] Shin, H. J., Park, J. Y., Yuk, D. B., and Lee, J. S. (2020). BERT-based spatial information extraction. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 10–17, Online. Association for Computational Linguistics.

[Singh, 2018] Singh, S. (2018). Natural language processing for information extraction. *arXiv preprint arXiv :1807.02383*.

[Sitikhu et al., 2019] Sitikhu, P., Pahi, K., Thapa, P., and Shakya, S. (2019). A comparison of semantic similarity methods for maximum human interpretability. In *2019 artificial intelligence for transforming business and society (AITB)*, volume 1, pages 1–4. IEEE.

[Song et al., 2018] Song, L., Zhang, Y., Wang, Z., and Gildea, D. (2018). N-ary relation extraction using graph-state LSTM. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2226–2235, Brussels, Belgium. Association for Computational Linguistics.

[Sonjak et al., 2011] Sonjak, S., Ličen, M., Frisvad, J. C., and Gunde-Cimerman, N. (2011). The mycobiota of three dry-cured meat products from slovenia. *Food Microbiology*, 28(3) :373–376.

[Steiner et al., 2012] Steiner, T., Verborgh, R., Troncy, R., Gabarro, J., and Van de Walle, R. (2012). Adding realtime coverage to the google knowledge graph. In *11th International Semantic Web Conference (ISWC 2012)*, volume 914, pages 65–68. Citeseer.

[Sun, 2009] Sun, A. (2009). A two-stage bootstrapping algorithm for relation extraction. In *Proceedings of the Student Research Workshop*, pages 76–82.

[Sutton and Barto, 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning : An introduction*. MIT press.

[Thompson and Taylor, 2008] Thompson, A. and Taylor, B. N. (2008). Use of the international system of units (si).

[Tissaoui et al., 2013] Tissaoui, A., Aussenac-Gilles, N., Laublet, P., and Hernandez, N. (2013). Evonto : un outil d'évolution de ressource termino-ontologique pour l'annotation sémantique. *Tech. Sci. Informatiques*, 32(7-8) :817–840.

[Torregrossa et al., 2021] Torregrossa, F., Allesiardo, R., Claveau, V., Kooli, N., and Gravier, G. (2021). A survey on training and evaluation of word embeddings. *International Journal of Data Science and Analytics*, pages 1–19.

[Torrey and Shavlik, 2010] Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends : algorithms, methods, and techniques*, pages 242–264. IGI global.

[Touhami et al., 2011] Touhami, R., Buche, P., Dibie-Barthélemy, J., and Ibănescu, L. (2011). An ontological and terminological resource for n-ary relation annotation in web data tables. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 662–679. Springer.

[Uban et al., 2021] Uban, A. S., Chulvi, B., and Rosso, P. (2021). Understanding patterns of anorexia manifestations in social media data with deep learning. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology : Improving Access*, pages 224–236, Online. Association for Computational Linguistics.

[Van Assem et al., 2010] Van Assem, M., Rijgersberg, H., Wigham, M., and Top, J. (2010). Converting and annotating quantitative data tables. In *International Semantic Web Conference*, pages 16–31. Springer.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

[Veyseh et al., 2021] Veyseh, A. P. B., Dernoncourt, F., Chang, W., and Nguyen, T. H. (2021). Maddog : A web-based system for acronym identification and disambiguation. *arXiv preprint arXiv :2101.09893*.

[Voorhees et al., 2005] Voorhees, E. M., Harman, D. K., et al. (2005). *TREC : Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, MA.

[Wang et al., 2018] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue : A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv :1804.07461*.

[Wang and Zhang, 2010] Wang, D. and Zhang, H. (2010). Inverse-category-frequency based supervised term weighting scheme for text categorization. *arXiv preprint arXiv :1012.2609*.

[Wolf et al., 2018] Wolf, C., Angellier-Coussy, H., Gontard, N., Doghieri, F., and Guillard, V. (2018). How the shape of fillers affects the barrier properties of polymer/non-porous particles nanocomposites : A review. *Journal of Membrane Science*, 556 :393–418.

[Wood et al., 2021] Wood, I., Johnson, M., and Wan, S. (2021). Integrating lexical information into entity neighbourhood representations for relation prediction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 3429–3436.

[Wren and Garner, 2002] Wren, J. D. and Garner, H. R. (2002). Heuristics for identification of acronym-definition patterns within text : towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of information in medicine*, 41(05) :426–434.

[Xu and Huang, 2005] Xu, J. and Huang, Y.-L. (2005). A machine learning approach to recognizing acronyms and their expansion. In *2005 International Conference on Machine Learning and Cybernetics*, volume 4, pages 2313–2319. IEEE.

[Yang et al., 2019] Yang, S., Feng, D., Qiao, L., Kan, Z., and Li, D. (2019). Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294.

[Yates et al., 2007] Yates, A., Banko, M., Broadhead, M., Cafarella, M. J., Etzioni, O., and Soderland, S. (2007). Textrunner : open information extraction on the web. In *Proceedings of Human Language Technologies : The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 25–26.

[Yu et al., 2020] Yu, H., Cao, Y., Cheng, G., Xie, P., Yang, Y., and Yu, P. (2020). Relation extraction with bert-based pre-trained model. In *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pages 1382–1387. IEEE.

[Zaenen et al., 2010] Zaenen, A., Condoravdi, C., Bobrow, D., and Hoffmann, R. (2010). Supporting rule-based representations with corpus-derived lexical information. In *Proceedings of the NAACL HLT 2010 First International*

Workshop on Formalisms and Methodology for Learning by Reading, pages 114–121.

[Zeng et al., 2015] Zeng, D., Liu, K., Chen, Y., and Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.

[Zhang et al., 2012] Zhang, C., Hoffmann, R., and Weld, D. (2012). Ontological smoothing for relation extraction with minimal supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26.

[Zhang et al., 2020] Zhang, Y., Zhang, Y., Qi, P., Manning, C. D., and Langlotz, C. P. (2020). Biomedical and clinical english model packages in the stanza python nlp library.

[Zhang, 2017] Zhang, Z. (2017). Effective and efficient semantic table interpretation using tableminer+. *Semantic Web*, 8(6) :921–957.

[Zhou et al., 2014] Zhou, D., Zhong, D., and He, Y. (2014). Biomedical relation extraction : from binary to complex. *Computational and mathematical methods in medicine*, 2014.

[Zhu and Goldberg, 2009] Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1) :1–130.

Annexe A

Annexes

A.1 Liste des publications dans le cadre de cette thèse

Revues Internationales

— Lentschat, M. ; Buche, P. ; Dibie-Barthelemy, J. ; Roche, M. (2021) *Towards combined semantic and lexical scores based on a new representation of textual data to extract experimental data from scientific publications*. Int. J. Intelligent Information and Database Systems Processing, *IN PRESS*.

— Lentschat, M. ; Buche, P. ; Dibie-Barthelemy, J. ; Menut, L. ; Roche, M. (2021) *Food packaging permeability and composition dataset dedicated to text-mining*. Data in Brief, Elsevier, 2021, 36, pp.107135.

Actes de conférences internationales

— Lentschat, M. ; Buche, P. ; Dibie-Barthelemy, J. ; Roche, M. (2020) *SciPuRe : a new Representation of textual data for entity identification from scientific publications*. In Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics, ACM, pp. 220-226.

Actes de conférences nationales

— Lentschat, M. ; Buche, P. ; Dibie-Barthelemy, J. ; Roche, M. (2021) *Combinaison de mesures lexicales et sémantiques pour l'extraction de données expérimentales dans des articles scientifiques*. Revue des Nouvelles Technologies de l'Information, RNTKI-E-37. Actes de la Conférence Extraction et gestion des connaissances (EGC 2021), Montpellier, France,

25 Janvier 2021/29 Janvier 2021, pp. 501-502.

— Lentschat, M. ; Buche, P. ; Dibie, J. ; and Roche, M. (2021). *Extraction et reconstitution de relation n-Aires issues d'articles scientifiques en domaine expérimental guidées par une Ressource Termino-Ontologique*. (EGC2022), Blois, France, 24 Janvier 2022/28 Janvier 2022. *ACCEPTED*.

Conférences sans actes

— Lentschat, M. ; Buche, P. ; Dibie-Barthelemy J. ; Roche, M. (2021) *Extraction d'instances de relations n-Aires issues d'articles scientifiques guidée par une ontologie*. Séminaire INRAE Semantic Linked Data. 11-14 octobre 2021.

— Lentschat, M. ; Buche, P. ; Dibie-Barthelemy J. ; Roche, M. (2021) *Extraction d'instances de relations n-Aires guidée par une ontologie : comparaison de méthodes structurelles, fréquentistes et sémantiques*. Séminaire IN'OVIVE. 21 septembre 2021.

— Lentschat, M. (2019) *Fouille de données textuelles guidée par ontologie*. Workshop département INRAE Transform. 25 novembre 2019. Paris.

— Lentschat, M. ; Buche, P. ; Dibie-Barthelemy J. ; Roche, M. (2019) *Towards the extraction of partial instances of N-Ary relations in textual data*. Workshop IN'OVIVE associé à IC'2019. 1 Juillet 2019. Toulouse.

Jeux de données

— Lentschat, M. (2021) *TRANSMAT n-Ary relations*, `https://doi.org/10.18167/DVN1/1BBJBQ`, CIRAD Dataverse, DRAFT VERSION.

— Lentschat, M. ; Buche, P. ; Menut, L. ; Guari, R. (2021) *TRANSMAT tables data*, `https://doi.org/10.18167/DVN1/GCZBC9`, CIRAD Dataverse, V1.

— Lentschat, M. ; Buche, P. ; Menut, L. (2020) *TRANSMAT Gold Standard*, `https://doi.org/10.18167/DVN1/U7HK8J`, CIRAD Dataverse, V3.

A.2 Annotation automatique de relations n-Aires dans les tableaux

Les travaux récents dans le domaine de l'extraction de données depuis des tableaux se concentrent sur la reconnaissance des concepts représentés et la mise en relation des données. La plupart des approches utilisent des ressources externes afin de guider la catégorisation des données et relations identifiées. Par exemple *TabEL* [Bhagavatula et al., 2015] développe une méthode probabiliste basée sur les cooccurrences de concepts observées sur Wikipédia afin de catégoriser les données et déterminer celles à lier dans des relations. Un autre système, *TableMiner+* [Zhang, 2017], exploite l'information contextuelle (e.g. la légende du tableau) afin d'améliorer la désambiguïsation des concepts présents dans les tableaux. Cela permet à *TableMiner+* d'utiliser plusieurs bases de connaissances (e.g. WordNet, Wikipedia, Freebase, DBpedia) pour catégoriser les données et relations trouvées. Il utilise également une annotation itérative, chaque étape étant guidée par les annotations faites lors de la précédente. Un travail récent, MTab [Nguyen et al., 2019], s'attaque au challenge SemTab 2019, consistant à aligner les données de tableaux avec une base de connaissance (e.g. DBpedia), et procède en deux temps. MTab débute par estimer les données présentes dans les cellules, les catégories des colonnes définies par l'entête du tableau et les relations possibles en accord avec la base de connaissance. Il effectue ensuite un second passage sur les cellules en agrégeant les estimations faites précédemment, ce qui permet ensuite de résoudre les ambiguïtés restantes dans la catégorisation des colonnes ainsi que dans les relations.

La méthode que nous avons choisie [Buche et al., 2011, Hignette et al., 2009] est fondée sur l'utilisation d'une RTO. Celle-ci a été sélectionnée pour la qualité de ses résultats à l'état de l'art, sa robustesse face aux différents formats de données présents dans les tableaux et sa compatibilité avec la RTO que nous utilisons. De plus, cette méthode crée une annotation des tableaux directement exploitable par notre méthode. Cette méthode a également déjà été utilisée dans des travaux s'appuyant sur le modèle de RTO que nous utilisons [Buche et al., 2011, Hignette et al., 2009].

La méthode d'extraction de relations partielles depuis les tableaux suit un processus en plusieurs étapes similaire à celles de l'annotation manuelle décrite

précédemment. Les balises générées par l'annotation automatique correspondent également à celles de l'annotation manuelle. Tout d'abord les entêtes des tableaux sont annotés automatiquement afin de distinguer les colonnes numériques et symboliques. Les relations présentes dans le tableau sont ensuite identifiées de manière automatique. Chaque ligne est traitée afin de découvrir les instances de relation qu'elle contient et les instances d'arguments, contenues dans les cellules, qui la composent. Des scores de similarités sont employés afin de remplacer l'expertise humaine dans l'alignement des concepts de l'ontologie avec les entités reconnues, tels que les entêtes des colonnes et les nouvelles unités de mesure. Ce travail a fait l'objet d'un rapport de stage d'assistant-ingénieur [Guari, 2021].

Le processus d'annotation automatique a été évalué sur le corpus de 31 tableaux présentés précédemment. Les balises attribuées automatiquement aux colonnes (i.e. *sc* et *qc*), aux concepts de relation (i.e. *rc*), aux lignes (i.e. *ri*) et aux cellules (i.e. *ai*) ont été comparées à celles définies lors du processus d'annotation manuelle. La Table 24 compare en détail les annotations faites automatiquement et manuellement.

tag	Manuelle	Automatique
sc	48	50
qc	172	170
rc	73	86
ri	828	925
ai	1273	1454
aii	717	∅
TOTAL	3111	2685

TABLE 24 – Distribution des annotations faites manuellement et automatiquement

La Table 25 donne l'évaluation de l'annotation automatique en termes de rappel, de précision et de f-score. Les résultats sont cohérents et comparables aux études à l'état de l'art [Buche et al., 2011, Hignette et al., 2009]. La précision sur les balises *sc* est faible (i.e. .38). Ceci est dû au nombre important de colonnes contenant un terme lié à un concept symbolique, et donc identifié comme un concept symbolique. Comme ces colonnes contiennent ensuite rarement des instances d'arguments appartenant aux relations n-Aires, le risque de propagation des erreurs est faible et ne constitue pas un problème majeur. Les erreurs dans l'annotation des balises *qc* proviennent principalement de la reconnaissance des

unités de mesure. L'annotation des balises *ri* est satisfaisante, d'autant plus qu'elle repose principalement sur l'identification des balises *sc* et *qc*. De même, l'annotation des tags *ri* et *ai* repose sur l'annotation correcte des tags précédents, leurs résultats sont donc proches.

Balise	recall	precision	f-score
sc	.81	.38	.52
qc	.75	.62	.68
rc	.79	.67	.73
ri	.77	.69	.73
ai	.75	.66	.70

TABLE 25 – Évaluation de la méthode d'annotation automatique des tableaux

A.3 Résultats de l'approche Fréquentistes

L'approche Fréquentiste de reconstitution des instances relations n-Aires mesure l'association entre une instance d'argument candidate et les manifestations textuelles des instances d'arguments présentent dans la relation n-Aire partielle par un score de cooccurrence. Trois mesures ont été évaluées : Dice, Jaccard et *Point-wise Mutual Information* (PMI). Les cooccurrences sont considérées dans différents contextes w_c : la phrase (i.e. Sentence), la fenêtre textuelle (i.e. Window, ± 1 phrase autour de l'instance d'argument) ou l'intégralité du document (i.e. Document). Deux types de manifestations textuelles des instances d'arguments des relations n-Aires partielles sont également pris en compte : les manifestations directes (i.e. Original_Value, la manifestation a la même valeur que l'instance d'argument de la relation) ou indirectes (i.e. Attached_Value, la manifestation est un terme dénotant le concept spécifique de l'instance d'argument de la relation).

La Table 26 présente l'ensemble des scores de rappel, précision et f-score pour les différentes configurations de l'approche Fréquentiste. Ce tableau est ordonné selon des f-scores décroissants. Ces résultats sont mesurés sans filtrage des instances d'arguments selon leur pertinence et en ne sélectionnant qu'une instance d'argument candidate pour chaque argument à compléter dans les relations partielles.

Mesure	w_c	w_m	Rappel	Précision	f-score
Dice	Document	Attached_value	.71	.28	.40
Jaccard	Document	Attached_value	.69	.27	.39
Dice	Document	Original_value	.69	.27	.38
Jaccard	Document	Original_value	.68	.25	.37
PMI	Document	Original_value	.68	.25	.36
PMI	Document	Attached_value	.67	.24	.35
PMI	Sentence	Original_value	.62	.19	.30
Jaccard	Sentence	Attached_value	.60	.17	.27
Dice	Sentence	Original_value	.59	.17	.27
PMI	Sentence	Attached_value	.58	.16	.26
Dice	Window	Original_value	.58	.16	.25
Jaccard	Window	Original_value	.58	.16	.25
Jaccard	Sentence	Original_value	.58	.16	.25
PMI	Window	Attached_value	.57	.16	.25
PMI	Segment	Original_value	.57	.15	.25
PMI	Window	Original_value	.57	.15	.25
Dice	Window	Attached_value	.56	.15	.24
Jaccard	Window	Attached_value	.56	.15	.23
Dice	Segment	Original_value	.55	.14	.23
Dice	Segment	Attached_value	.55	.14	.23
Jaccard	Segment	Original_value	.54	.14	.22
Dice	Sentence	Attached_value	.52	.13	.20
Jaccard	Segment	Attached_value	.50	.12	.19
PMI	Segment	Attached_value	.50	.12	.19

TABLE 26 – Scores de rappel, précision et f-score des mesures fréquentistes selon w_c et w_m