*Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles*
*Lille, France, 28 juin au 2 juillet 2021*

# Extraction automatique de relations sémantiques d'hyperonymie et d'hyponymie dans un corpus métier

Camille Gosset[1], Mokhtar Boumedyen Billami[1], Mathieu Lafourcade[2], Christophe Bortolaso[1], Mustapha Derras[1]
(1) Berger-Levrault, Labège, France
(2) LIRMM, Université Montpellier, Montpellier, France
(1){camille.gosset, mb.billami, christophe.bortolaso, mustapha.derras}@berger-levrault.com
(2) mathieu.lafourcade@lirmm.fr

## RESUME

Nous nous intéressons dans cet article à l'extraction automatique de relations sémantiques d'hyperonymie et d'hyponymie à partir d'un corpus de spécialités métier. Le corpus regroupe des ouvrages et articles en français d'expertise juridique et a été partiellement annoté en termes-clés par des experts. Nous prétraitons ces annotations afin de pouvoir les retrouver dans ce corpus et obtenir un concept général pour extraire les relations entre ces termes. Nous décrivons une étude expérimentale qui compare plusieurs méthodes de classification appliquées sur des vecteurs de relations construits à partir d'un modèle Word2Vec. Nous comparons les résultats obtenus grâce à un jeu de données construit à partir de relations d'hyperonymie tirées d'un réseau lexico-sémantique français que nous inversons pour obtenir les relations d'hyponymie. Nos résultats montrent que nous obtenons une classification pouvant atteindre un taux d'exactitude de 92 %.

## ABSTRACT

**Automatic extraction of hypernym and hyponym relations in a professional corpus.**
In this paper, we are interested in the automatic extraction of hypernymy and hyponymy semantic relations from a corpus of business specialties. The corpus includes books and articles in French of legal expertise and has been partially annotated with keywords by experts. We pre-process these annotations to be able to find them in this corpus and to obtain a general concept to extract the relationships between these terms. We describe an experimental study which compares several classification methods applied on vectors of relations constructed using Word2Vec. We compare the results obtained using a dataset constructed from hypernymy relations drawn from a French lexico-semantic network that we invert to obtain the hyponymy relations. Our results show that we obtain a classification that can reach an accuracy rate of 92%.

MOTS-CLES : Extraction de relations d'hyperonymie et d'hyponymie, Word2Vec, réseau lexico-sémantique, apprentissage automatique, classification.

KEYWORDS: Extraction of hypernymy and hyponymy relations, Word2Vec, lexico-semantic network, machine learning, classification.

Volume 1 : conférence principale, pages 162–170.

# 1 Introduction

L'identification de concepts et de relations dans des documents est une phase clé dans la construction d'une base de connaissances ontologiques (Asim et al., 2018 ; Buitelaar et al., 2005). La construction manuelle d'un tel type de ressources est une tâche à forte intensité de main-d'œuvre. Bien que les données non structurées puissent être transformées en données structurées, cette construction englobe un processus très long et coûteux (Cullen & Bryman, 1988). Plusieurs travaux de recherche tendent actuellement vers l'apprentissage automatique des ontologies afin de conjurer le goulot d'étranglement d'acquisition des connaissances (Konys, 2019). L'acquisition automatique des ontologies à partir de textes est devenue ainsi un domaine de recherche prépondérant du fait de la grande masse de données présentes sur le web dont ces dernières peuvent être décrites sous-forme structurées, semi-structurées ou non structurées (Deb et al., 2018). Buitelaar et al. (2005) ont proposé une méthodologie permettant de construire des ontologies à partir de textes et cela en plusieurs étapes. L'ensemble du processus de leur méthodologie est connu sous le nom de « mille-feuille d'apprentissage de l'ontologie » (*Ontology Learning Layer Cake*). L'extraction de relations est l'une des étapes de ce processus et constitue un niveau essentiel pour la structuration des connaissances.

Plusieurs travaux de recherche se recentrent sur l'extraction de relations d'hyperonymie et d'hyponymie (par symétrie) car elles permettent d'avoir la hiérarchie de concepts dans une ontologie (Panchenko et al., 2013 ; Mintz et al., 2009 ; Bunescu & Mooney, 2005). Nous nous intéressons, dans cet article, à cette particularité d'extraction de relations de type '*Hyperonyme*' et '*Hyponyme*' au sein de textes provenant de 8 domaines du secteur public, à savoir : (1) '*État civil et Cimetières'*, (2) '*Élections'*, (3) '*Commande publique'*, (4) '*Urbanisme'*, (5) '*Comptabilité et finances locales'*, (6) '*Ressources humaines territoriales'*, (7) '*Justice'* et (8) '*Santé'*. Par définition, l'hyperonymie permet de représenter les termes pouvant être associés aux génériques d'une cible donnée, par exemple la *grippe aviaire* a pour termes génériques *grippe*, *maladie*, *pathologie* voire un *processus pathologique*. L'hyponymie, quant à elle, permet de représenter les termes pouvant être associés aux spécifiques d'une cible donnée, par exemple *engagement* a pour termes spécifiques *mariage*, *engagement à long terme*, *engagement social*, *engagement de servir* voire *engagement d'achat ferme*. Chaque relation peut faire référence à un domaine bien défini. Dans les exemples que nous venons de citer, c'est le domaine de la « santé » qui peut être identifié lorsque nous évoquons la *grippe aviaire* alors qu'il s'agit du domaine de « l'état civil » lorsque nous évoquons l'*engagement en mariage*.

Nous présentons tout d'abord dans la section 2 un état de l'art sur les méthodes d'extraction de relations. Ensuite, nous décrirons dans la section 3 les données exploitées ainsi que les prétraitements effectués sur ces données. Par la suite, dans la section 4, nous détaillons notre méthodologie pour l'extraction de relations. Enfin, nous discutons dans la section 5 de nos résultats d'expérimentation obtenus avant de conclure et présenter des perspectives de notre travail (cf. section 6).

# 2 Travaux antérieurs d'extraction de relations

Plusieurs travaux de recherche sur l'extraction de relations ont été proposés, deux grandes catégories d'approches existent (Granada et al., 2018 ; Wang et al., 2017), à savoir : (1) les approches à base de patrons lexico-syntaxiques (Panchenko et al., 2013 ; Hearst, 1992) ; et (2) les approches à base d'apprentissage automatique/apprentissage profond (Xue et al., 2018 ; Bunescu & Mooney, 2005).

L'approche d'extraction de relations la plus ancienne revient à Hearst (1992). Elle repose sur l'utilisation de patrons lexico-syntaxiques pour l'extraction des hyperonymes dont la langue est l'anglais. Pour la langue française, Panchenko et al. (2013) se sont intéressés aux patrons de Hearst : ils ont itéré sur l'étude et l'ont étendue afin de limiter un certain nombre de bruits. Par la suite, Panchenko et al. (2016) et Bordea et al. (2015) ont montré que ce type d'approches produit des résultats intéressants, notamment du point de vue de la précision, mais ne couvre qu'une partie de l'information. L'utilisation d'un nombre prédéfini de patrons lexico-syntaxiques pose les difficultés, d'une part, d'expliciter certaines relations et, d'autre part, de gérer l'ambiguïté de certains patrons.

D'autres approches à base d'apprentissage automatique ont été proposées afin de limiter l'utilisation des patrons lexico-syntaxiques, que ce soit par supervision (Pantel & Pennacchiotti, 2006 ; Bunescu & Mooney, 2005 ; Snow et al., 2004) ou non supervision (Mintz et al., 2009 ; Morin & Jacquemin, 2004). L'une des premières approches de cette catégorie a été proposée par Brin (1998). Elle s'appuie sur une technique de *bootstrapping* qui consiste à sélectionner des patrons par apprentissage semi-supervisé afin de construire une base d'apprentissage. Sur la même idée, des techniques de sélection de mots par analyse distributionnelle et de sélection de traits sémantiques ont été utilisées pour identifier des relations entre des entités nommées (Etzioni et al., 2004).

Par ailleurs, Cartier (2015) a proposé une approche hybride d'extraction de relations à partir d'un corpus d'un million de définitions provenant de deux ressources, à savoir : le TLFi, *Trésor de la Langue Française informatisé* (Dendien & Pierrel, 2003) et Wikipédia. Cette approche combine la précision des patrons lexico-syntaxiques et le rappel des méthodes statistiques par analyse distributionnelle. Plusieurs relations sémantiques ont été étudiées dont l'hyperonymie et l'hyponymie font partie. Le but du travail mené par Cartier (2015) est d'obtenir automatiquement une ressource sémantique pour le français contemporain à partir d'un corpus de textes.

Hashimoto et al. (2015) se sont intéressés à classifier les relations sémantiques en utilisant des plongements lexicaux (*Word Embeddings*). Dans le même principe, Mallart et al. (2020) ont proposé une approche d'identification de relations dans un corpus métier du journalisme par une modélisation LSTM (*Long Short-Term Memory*) et une utilisation de plongements lexicaux pré-entraînés avec un Word2Vec et une architecture Skip-Gram (Mikolov et al., 2013). Un modèle de classification binaire a été aussi proposé et a permis d'améliorer significativement les résultats obtenus.

Dans cet article, nous nous positionnons dans un travail d'extraction de relations par apprentissage automatique et à partir d'un corpus de données regroupant un ensemble de domaines métier. Cet apprentissage est guidé par un grand réseau lexico-sémantique où les sens communs/métiers sont définis de base. Notre objectif est un peu similaire à celui de Cartier (2015) puisque nous nous intéressons à produire des bases ontologiques métiers. De même, les travaux de Mallart et al. (2020) et Hashimoto et al. (2015) sont proches du nôtre puisque des plongements lexicaux et des modèles de classification sont utilisés. Cependant, l'originalité du travail que nous proposons dans cet article est double : (1) une extraction de relations sémantiques sans avoir besoin d'un corpus annoté sémantiquement (sens/relations) et (2) une prise en compte de tous les termes à classe ouverte (noms communs, entités nommés, adjectifs, adverbes et verbes).

# 3 Données de travail

Nous utilisons un corpus français contenant 172 ouvrages et 12 838 articles en ligne d'une expertise juridique et pratique. L'ensemble des documents de ce corpus est édité par la société Berger-Levrault.

Ce corpus traite 8 domaines de spécialité, à savoir : (1) '*État civil et Cimetières'*, (2) '*Élections'*, (3) '*Commande publique'*, (4) '*Urbanisme'*, (5) '*Comptabilité et finances locales'*, (6) '*Ressources humaines territoriales'*, (7) '*Justice'* et (8) '*Santé'*. Dans ce qui suit, nous faisons référence au corpus par le nom MÉTIER. Ce dernier a été partiellement annoté par des experts. Chaque paragraphe de chaque document (ouvrage ou article) est annoté avec des termes-clés. Cela constitue une représentation semi-structurée par le biais de balises HTML. Concrètement, plus de 45 000 annotations manuelles en termes-clés ont été effectuées. Par exemple, dans la phrase *« Démocratie de proximité, l'expression suscite immédiatement un doute, une inquiétude, un trouble […] »,* le terme *démocratie de proximité* est considéré comme terme-clé. Autre exemple, *« Une réforme à la recherche d'une démocratie participative à l'échelon local »*, ici *démocratie participative* est aussi un terme-clé. Dans le corpus MÉTIER, le nombre d'occurrences de chaque terme-clé est variable. Certains termes sont très fréquents comme *formation* ou *Association foncière urbaine* avec plus de 500 occurrences ; d'autres termes sont fréquents comme *prix*, *publicité* ou encore *accord-cadre* avec au moins 200 occurrences. Enfin, des termes peu fréquents existent aussi comme *survie*, *contamination*, *réception* ou encore *équipement* avec moins de 50 occurrences. Si nous prenons en considération seulement les articles, la TABLE *1* décrit le nombre d'annotations des experts pour un regroupement d'articles par domaine. Pour ces articles, il est à noter que l'ensemble des termes-clés d'un domaine donné fait référence à un thésaurus de ce domaine.

| **Domaine** | **Nombre de termes-clés** | **Nombre d'articles** | **Nombre d'annotations** |
|---|---|---|---|
| *État civil et Cimetières* | 642 | 2 767 | 2 169 |
| *Élections* | 108 | 152 | 150 |
| *Commande publique* | 876 | 1 354 | 1 201 |
| *Urbanisme* | 327 | 1 357 | 554 |
| *Comptabilité et finances locales* | 981 | 1 971 | 1 957 |
| *Ressources humaines territoriales* | 293 | 361 | 122 |
| *Justice* | 1 447 | 3 980 | 870 |
| *Santé* | 491 | 896 | 830 |

TABLE 1 : Taille du vocabulaire et nombre d'annotations des experts

Nous avons récupéré l'ensemble des termes-clés représentant les annotations des experts (à la fois les thésaurus mais aussi des tags utilisés dans les ouvrages). Ces termes-clés ont été décrits avec plusieurs formes fléchies et parfois avec des informations additionnelles non pertinentes comme les déterminants. Par exemple, *des frais* ou *associations syndicales autorisées*. Cela est tout à fait normal puisqu'il n'y avait pas une ressource lexicale de référence au moment de l'annotation, raison pour laquelle tous les experts n'annotent pas de la même façon. Un prétraitement est donc nécessaire. Nous utilisons le parseur Stanza (Qi et al., 2020) pour éliminer les premiers mots outils. Par exemple, le terme *des frais* est transformé en *frais*. Toutefois, *frais* est ambiguë syntaxiquement (nom/adjectif). Afin de lever cette ambiguïté, nous utilisons Stanza pour fournir la classe grammaticale en contexte. Ainsi, chaque terme-clé est associé avec (1) sa classe grammaticale et (2) sa forme lemmatisée prétraitée. Cette combinaison représente l'identifiant d'un terme-clé. Pour les multi-mots, la classe du gouverneur est celle qui est attribuée au terme-clé.

# 4 Méthodologie

Cette section décrit l'architecture de notre approche et le modèle d'apprentissage que nous avons développé. Nous répondons au problème d'extraction de relations en plusieurs étapes : (1) Quelle forme fléchie est la plus adéquate pour un terme-clé donné ? (2) Quel est le réseau lexico-sémantique français à utiliser pour la validation des relations pouvant être extraites ? (3) Comment les vecteurs de relations sont construits à partir d'un modèle Word2Vec ? et (4) Quel modèle de classification est développé pour juger la pertinence des relations d'hyperonymie et d'hyponymie ? Dans ce qui suit, nous faisons référence à l'hyperonymie par la relation *r_isa* et à l'hyponymie par la relation *r_hypo.*

Dans la première étape, les termes-clés d'un même identifiant vont être unifiés. Nous souhaitons donner une même forme aux termes dits équivalents mais de forme fléchie différente. Par exemple, *personne âgée* et *personnes âgées* font référence au même terme-clé. Afin de choisir le bon représentant, nous avons fait le choix de prendre la forme fléchie ayant le plus grand nombre d'occurrences dans le corpus MÉTIER. Pour cela, une analyse statistique est effectuée sur tout le corpus afin de calculer le nombre d'occurrences de chaque forme fléchie pour chaque terme-clé annoté par les experts. Ainsi, la forme la plus fréquente représente le substitut pouvant faire référence à un terme-clé donné. Pour les termes-clés dits simples (par exemple, *réception*, *équipement*, *marché*, etc.), nous privilégions la forme singulière à la forme plurielle. Cela se justifie par le fait d'avoir une forme standard pouvant être retrouvée facilement dans des dictionnaires. Afin de satisfaire ce besoin, nous utilisons en plus de Stanza la ressource Lexique3 (New et al., 2007) pouvant avoir une priorité plus élevée dans notre processus.

Pour la seconde étape, notre choix s'est porté sur le réseau lexico-sémantique JeuxDeMots[1] (Lafourcade, 2007). En effet, à ce jour, JeuxDeMots est le plus grand réseau français librement disponible avec 14 millions de nœuds et environ 320 millions de relations. Il permet d'avoir le sens commun et les sens métiers des termes polysémiques français. Son utilisation nous permet d'extraire un ensemble de relations de type *r_isa* et *r_hypo* et cela à partir de la liste des termes-clés substituts du corpus MÉTIER. Dans cet article, nous nous intéressons seulement à ces deux types de relations. L'utilisation de JeuxDeMots nous a permis de récupérer un ensemble de 110 118 relations d'hyperonymes. En inversant ces relations, cela nous permet d'avoir un même nombre de relations d'hyponymes. Les paires extraites de JeuxDeMots constituent un jeu d'apprentissage et d'évaluation.

Pour la troisième étape, nous entraînons des plongements lexicaux sur le corpus MÉTIER afin d'obtenir des représentations vectorielles continues de termes-clés. Pour cela, nous faisons tout d'abord une substitution lexicale dans le corpus de tout terme-clé par son référant (forme fléchie la plus fréquente ou représentation dans Lexique3 pour les mots simples). Ensuite, le modèle Word2Vec (Mikolov et al., 2013) avec une architecture CBOW (*Continuous Bag-of-Words*) est utilisé pour l'entraînement. Nous avons utilisé Gensim[2] pour satisfaire cet entraînement (avec un paramétrage par défaut). Par ailleurs, nous avons privilégié l'utilisation de Word2Vec à la place des modèles d'entraînement de plongements lexicaux contextualisés comme BERT (Devlin et al., 2019). Cela se justifie par le fait de ne pas se limiter seulement à des contextes où le terme-source et le terme-cible d'une relation donnée doivent co-occurrer dans une même phrase ou un même paragraphe. Le fonctionnement de BERT est limité à 512 mots en contexte et les termes reliés par hyperonymie ne sont pas forcément en simultané dans un même paragraphe. Il est donc intéressant d'englober une grande quantité de textes avec Word2Vec. Par ailleurs, et pour un principe un peu différent, nous

[1] http://www.jeuxdemots.org/rezo.php

[2] https://radimrehurek.com/gensim/models/word2vec.html

n'avons pas utilisé FastText (Bojanowski et al., 2017). Le concept de *n*-grammes tel qu'il est décrit par cette bibliothèque d'apprentissage pose un problème de compositionnalité des mots. Par exemple, le vecteur associé au terme *fiche de paie* par FastText est le vecteur moyen des mots qui le composent, à savoir : *fiche* et *paie*. Avec Word2Vec, nous pouvons forcer la compréhension des suites de mots comme un élément à part entière (i.e. *fiche_de_paie*). Ainsi, dans notre travail, les plongements lexicaux représentant nos termes-clés sont appris en les considérant comme des entités et non comme une moyenne des mots qui les composent. Pour cela, nous utilisons les phrases du module *gensim*[3].

La suite de notre processus d'entraînement avec Word2Vec consiste à construire des vecteurs de relations à partir des vecteurs de termes-clés. Pour cela, nous proposons de calculer un nouveau vecteur par soustraction des vecteurs en entrée. Concrètement, si $V_1$ est le vecteur du terme-clé *marché public* et $V_2$ est le vecteur du terme-clé *marché* alors ($V_1$–$V_2$) est le vecteur de la relation (*marché public*, *r_isa*, *marché*). Cette représentation vectorielle répond à la contrainte de non-symétrie des relations *r_isa* et *r_hypo*. Ces plongements lexicaux sont le point d'entrée pour la classification.

En dernière étape, nous avons créé plusieurs classifieurs binaires : de l'utilisation des arbres de décision et des méthodes ensemblistes vers l'utilisation des machines à vecteur de support (*Support Vector Machine – SVM*). Pour cela, nous avons utilisé la bibliothèque Scikit-learn (Buitinck et al., 2013). Chaque classifieur permet de prédire si une paire de termes fait référence à une association soit entre un terme-cible et un terme générique, soit entre un terme-cible et un terme spécifique. Les vecteurs *embeddings* de relations sont fournis aux classifieurs. Nous avons à disposition un ensemble de 220 236 paires de relations (moitié *r_isa*/moitié *r_hypo*). La base d'exemples est équilibrée puisqu'elle a été construite à partir de la relation *r_isa*.

Le principe de notre approche est d'apprendre des relations sémantiques tirées de JeuxDeMots et de les évaluer. L'objectif, par la suite, étant d'utiliser ces classifieurs binaires pour prédire de potentielles relations sémantiques non reconnues à l'heure actuelle par JeuxDeMots. En effet, même si JeuxDeMots représente l'un des plus grands réseaux lexico-sémantiques du français, la couverture des relations sémantiques contenant des termes provenant de domaines métier reste à améliorer. Ainsi, le développement de classifieurs pour la prédiction de nouvelles relations peut être vu comme un axe d'enrichissement de telles ressources. Dans la section qui suit, nous discutons des résultats obtenus.

# 5 Résultats et discussion

Nous avons appliqué nos classifieurs binaires sur un jeu de données équilibré de 220 236 paires de relations (110 118 relations *r_isa* / 110 118 relations *r_hypo)*. Les classifieurs utilisés ont été instanciés avec les paramètres par défaut et une identification des meilleurs paramètres à l'aide de *GridSearchCV*[4] n'a pas eu lieu. Afin d'obtenir des résultats fiables, nous utilisons le principe de la validation croisée (*Cross-validation*). Une fonction est disponible dans Scikit-learn[5]. Afin d'évaluer la qualité de nos classifieurs, nous utilisons deux mesures d'évaluations, à savoir : (1) le taux d'exactitude (*accuracy rate*) et (2) F-mesure. Ces deux mesures ont été indiquées directement dans un paramètre d'entrée à la fonction de validation croisée de Scikit-learn. Les résultats obtenus sont présentés dans le tableau 2.

[3] https://radimrehurek.com/gensim/models/phrases.html
[4] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[5] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html

| Modèle | F-mesure | Taux d'exactitude |
|---|---|---|
| LinearSVC | 0,73 | 0,85 |
| RidgeClassifier | 0,72 | 0,79 |
| KNeighborsClassifier | 0,82 | 0,91 |
| RandomForestClassifier | 0,69 | 0,76 |
| DecisionTreeClassifier | 0,67 | 0,73 |
| LogisticRegression | 0,71 | 0,83 |
| SupportVectorClassification | **0,83** | **0,92** |

TABLE 2 : Résultats d'évaluation par application de différents classifieurs

Nous constatons des résultats entre 67 % et 83 % pour la F-mesure, ce qui nous montre que cette voie est prometteuse. Nous constatons aussi que les moins bons résultats sont obtenus sur les classifieurs de type arbres de décision. En effet, les *embeddings* qui sont entraînés présentent en sortie un certain nombre de dimensions. Les arbres de décisions, eux, extraient les dimensions qu'ils considèrent comme importantes. Toutes les dimensions ne sont donc pas traitées. Par ailleurs, nous avons de bons résultats avec un autre type de classifieurs, à savoir : les $k$ plus proches voisins (avec $k = 3$). La famille des machines à vecteurs de support (SVM) se servent de vecteurs afin de discriminer une classe. Elles se sont avérées plus adaptées à l'utilisation des plongements lexicaux. De plus, nous pouvons obtenir encore de meilleurs résultats en utilisant *GridSearchCV* pour fixer les meilleurs hyperparamètres.

# 6 Conclusion et perspectives

Dans cet article, nous avons proposé une approche d'extraction de relations sémantiques basée sur des modèles de classification. Deux relations ont été traitées, à savoir : l'hyperonymie et l'hyponymie. Des plongements lexicaux ont été entraînés sur un corpus de spécialité métier décrivant 8 domaines. À partir de ces plongements lexicaux, des vecteurs de relations ont été créés et fournis à plusieurs classifieurs binaires pour prédire la relation pouvant lier deux termes donnés. L'originalité de ce travail réside dans l'apprentissage des vecteurs de relations guidé par une ressource lexico-sémantique, à savoir JeuxDeMots. Les résultats obtenus sont prometteurs : une F-mesure de 83 % et un taux d'exactitude de 92 % par utilisation de l'algorithme de classification SVC (*Support Vector Classification*) faisant référence à la famille des machines à vecteur de support (SVM).

Trois perspectives s'ouvrent à nous à la suite de ce travail : (1) nous envisageons d'intégrer une couche de déduction des relations par transitivité. Par exemple, si ($terme_A$ *r_isa* $terme_B$) et ($terme_B$ *r_isa* $terme_C$) alors ($terme_A$ *r_isa* $terme_C$). Cela permettrait, d'une part, d'améliorer le taux de performance de nos classifieurs en augmentant la base d'apprentissage, et d'autre part, de structurer les données par une extraction de la hiérarchie de termes pouvant être obtenue ; (2) nous envisageons d'extraire des relations comme précédemment énoncé mais pour un seul domaine de spécialité à la fois. Cela permettrait, par exemple, d'orienter la construction d'une ontologie à un seul domaine prédéfini. Pour cela, JeuxDeMots peut être utilisé en sélectionnant les termes liés à la relation *Domaine* ; et (3) notre étude peut être étendue à d'autres relations, par exemple la synonymie ou l'antonymie qui peuvent fonctionner aussi par paires de termes. Pour de telles relations symétriques, la modification du vecteur de relation ($V_{termeA} - V_{termeB}$) vers ($|V_{termeA} - V_{termeB}|$) est essentielle. Toutefois, le non-respect de la symétrie peut être un indicateur de problèmes dans l'apprentissage et/ou la polysémie (sens commun/métier).

# Références


ASIM M. N., WASIM M., KHAN M. U. G., MAHMOUD W. & ABBASI H. M. (2018). A survey of ontology learning techniques and applications. *Database*, 2018: article ID bay101. DOI: 10.1093/database /bay101.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, **5**, p. 135–146.

BORDEA G., BUITELAAR P., FARALLI S. & NAVIGLI R. (2015). SemEval-2015 Task 17: Taxonomy Extraction Evaluation (TExEval). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015@NAACL-HLT)*. **452**(465), p. 902–910.

BRIN S. (1998). Extracting Patterns and Relations from the World Wide Web. In *International Workshop on The World Wide Web and Databases*, Springer, p. 172–183.

BUITELAAR P., CIMIANO P. & MAGNINI B. (2005). Ontology learning from text: an overview. In *Ontology Learning from Text: Methods, Evaluation and Applications,* IOS Press, **123**, p. 3–12.

BUITINCK L., LOUPPE G., BLONDEL M., PEDREGOSA F., MUELLER A., GRISEL O., NICULAE V., PRETTENHOFER P., GRAMFORT A., GROBLER J., LAYTON R., VANDERPLAS J., JOLY A., HOLT B. & VAROQUAUX G. (2013). API design for machine learning software: experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, p. 108–122.

BUNESCU R. & MOONEY R. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, p. 724–731.

CARTIER E. (2015). Extraction automatique de relations sémantiques dans les définitions : approche hybride, construction d'un corpus de relations sémantiques pour le français. *Traitement Automatique des Langues Naturelles (TALN),* Caen, France. HAL : halshs-01412736.

CULLEN J. & BRYMAN A. (1988). The Knowledge Acquisition Bottleneck: Time for Reassessment? *Expert Systems*, **5**, p. 216–225.

DEB C. K., MARWAHA S., ARORA A. & DAS M. (2018). A Framework for Ontology Learning from Taxonomic Data. In *Big Data Analytics, Springer*, p. 29–37.

DENDIEN J. & PIERREL J.-M. (2003). Le Trésor de la Langue Française Informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues, HERMÈS.*

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-Human Language Technology*.

ETZIONI O., CAFARELLA M. J., DOWNEY D., POPESCU A., SHAKED T., SODERLAND S., WELD D. S. & YATES A. (2004). Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison. *American Association for Artificial Intelligence (AAAI 2004),* p. 391–398.

GRANADA R., VIEIRA R., TROJAHN C. & AUSSENAC-GILLES N. (2018). Evaluating the Complementarity of Taxonomic Relation Extraction Methods Across Different Languages. arXiv: 1811.03245v1.

HASHIMOTO K., STENETORP P., MIWA M. & TSURUOKA Y. (2015). Task-Oriented Learning of Word Embeddings for Semantic Relation Classification. *Computational Natural Language Learning (CoNLL)*.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics*, **2**, *COLING '92*, p. 539–545, Stroudsburg, PA, USA.

KONYS A. (2019). Knowledge Repository of Ontology Learning Tools from Text. In *the 23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, p. 1614–1628.

LAFOURCADE M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th International Symposium on Natural Language Processing*, Dec 2007, Pattaya, Chonburi, Thailand. HAL : lirmm-00200883.

MALLART C., NOUY M. L., GRAVIER G. & SEBILLOT P. (2020). Relation, es-tu là ? Détection de relations par LSTM pour améliorer l'extraction de relations. *JEP-TALN-RECITAL*.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations*, p. 1–12.

MINTZ M., BILLS S., SNOW R., & JURAKSKY D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, p. 1003–1011.

MORIN E. & JACQUEMIN C. (2004). Automatic Acquisition and Expansion of Hypernym Links. *Computers and the Humanities*, 38, p. 363–396.

NEW B., BRYSBAERT M., VÉRONIS J. & PALLIER C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, p. 661–677.

PANCHENKO A., FARALLI S., RUPPERT E., REMUS S., NAETS H., FAIRON C., PONZETTO S. P. & BIEMANN C. (2016). TAXI at SemEval-2016 Task 13: a Taxonomy Induction Method based on Lexico-Syntactic Patterns, Substrings and Focused Crawling. In *Proceedings of SemEval-2016 (SemEval@NAACL-HLT)*, p. 1320–1327.

PANCHENKO A., NAETS H., BROUWERS L. & FAIRON C. (2013). Recherche et visualisation de mots sémantiquement liés. *TALN-RÉCITAL 2013*, p. 747–754.

PANTEL P. & PENNACCHIOTTI M. (2006). Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, p. 113–120.

QI P., ZHANG Y., ZHANG Y., BOLTON J. & MANNING C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *System Demonstrations, Association for Computational Linguistics (ACL)*.

SNOW R., JURAFSKY D. & NG A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems (NIPS 2004)*.

WANG C., HE X. & ZHOU A. (2017). A Short Survey on Taxonomy Learning from Text Corpora: Issues, Resources and Recent Advances. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP),* p. 1201–1214.

XUE L., QING S. & PENGZHOU Z. (2018). Relation Extraction Based on Deep Learning. *The 17th International Conference on Computer and Information Science (ICIS)*, p. 687–691, DOI: 10.1109/ICIS.2018.8466437.