Classification de documents métiers pour l'aide à l'extraction et la classification de relations lexico-sémantiques typées et pondérées

Camille Gosset, Mokhtar Boumedienne Billami, Mathieu Lafourcade, Christophe Bortolaso

Classification de documents métiers pour l'aide à l'extraction et la classification de relations lexico-sémantiques typées et pondérées

Camille GOSSET[1,2], Mokhtar Boumedyen BILLAMI[2], Mathieu LAFOURCADE[1], Christophe BORTOLASO[2]

[1] LIRMM, Montpellier
[2] Berger-Levrault, Labège

Camille.gosset@berger-levrault.com

Résumé

La langue française est riche et complexe dans sa nature, offrant un vaste choix de mots et de possibilités d'expression. Cependant, cette richesse peut également conduire à la polysémie des mots. L'extraction de relations lexico-sémantiques dans ce cadre peut être une tâche complexe et difficile, en raison de la nature de la langue et de la façon dont les phrases sont construites. Après une extraction terminologique, l'extraction de relations entre ces termes de mêmes domaines peut être effectuée à l'aide d'une classification de type de relations. Le classifieur nous permet de confirmer un type donné entre deux termes du corpus. Ensuite, un filtre sur les relations extraites peut être appliqué en utilisant la classification par domaine métier. Cependant, sur des textes de spécialité, les algorithmes de classification multi-classe sont-ils aussi performants comme leur application sur les textes génériques ? Dans cet article, nous évaluons plusieurs modèles de classifications ayant différents types de représentation de documents pour une application à des domaines métiers. Cette évaluation a permis de constater une extraction de relations plus fine avec certains modèles.

Mots-clés

Classification de textes/ documents, Classification de domaines, Représentation de documents, Classification de relation, Vocabulaire de spécialité.

1 Introduction

L'extraction de relations à partir de textes est un domaine actif dans le traitement automatique des langues. Les relations se forment entre différents termes clés, mais leur pertinence dépend de la façon dont ils sont utilisés dans le contexte. Pour filtrer les relations qui ne sont pas pertinentes, une méthode possible est de les classifier selon leur domaine d'application. Ainsi, si deux termes clefs d'une relation t_1, t_2 appartiennent à un domaine *d*, alors la relation *r* appartient également à *d*. De plus, en classifiant les documents, les termes clefs rattachés au document sont également classifiés, permettant ainsi de rattacher un document à l'un des domaines.

Il existe différentes méthodes de classification de domaines, notamment la classification automatique, la classification manuelle par des experts humains et la classification hybride combinant les deux approches. L'efficacité de ces méthodes dépend de plusieurs facteurs, tels que la qualité et la quantité des données d'entraînement, la pertinence des catégories de classification et la précision des algorithmes utilisés.

Dans ce papier, nous nous intéressons à la question de la classification de documents selon des domaines de spécialités. En effet, nous nous demandons si la classification de documents peut aider l'extraction de relations. L'objectif final est de pouvoir filtrer les relation d'une ontologie construite à partir de textes en fonction de huit domaines spécifiques : État civil et Cimetières, Élections, Commande publique, Urbanisme, Comptabilité et Finances locales, Ressources Humaines Territoriales, Justice et Santé. Nous avons à notre disposition des articles déjà classifiés dans un domaine précis, mais aussi des ouvrages dont il n'est pas possible de récupérer leurs domaines. Pour résoudre ce problème, nous avons élaboré trois stratégies possibles : la classification de documents, la classification directe des termes clés des experts et la classification directe des relations. Nous avons effectué une analyse des données et un développement expérimental comparatif des performances des différentes stratégies pour évaluer leur efficacité.

2 Présentation des données

Notre étude analyse un corpus de documents spécialisés en français comprenant 172 ouvrages et 12 838 articles dans le domaine juridique et pratique, couvrant 8 secteurs du secteur public : État civil et Cimetières, Élections, Commande publique, Urbanisme, Comptabilité et Finances locales, Ressources Humaines Territoriales, Justice et Santé. Les articles sont classés dans un domaine spécifique, tandis que les ouvrages ne le sont pas. Pour déterminer le domaine auquel appartient chaque ouvrage ou partie d'ouvrage, il est nécessaire de les classifier. Le corpus a été annoté manuellement avec des termes-clés choisis par des experts spécialisés dans différents domaines du secteur public. Les annotations sont manuelles et les termes-clés ont été décrits avec plusieurs formes fléchies et parfois avec des informations supplémentaires non pertinentes. La liste unique de termes-clés représentatifs comprend 46 142 termes-clés répartis entre les ouvrages et les articles.

Les termes-clés utilisés ne sont pas toujours exclusifs à un seul domaine, ce qui peut conduire à une intersection entre différents domaines.

Les résultats de la matrice de corrélation des différents domaines ont montré que le domaine de la santé et celui des ressources humaines territoriales présentent la plus forte corrélation, atteignant 54%. Il y a également une corrélation

légèrement inférieure à 30% entre "commande publique" et "comptabilité et finances locales", ainsi qu'entre "justice" et "état civil et cimetières". Nous avons analysé en détail les données des articles pour déterminer les domaines des ouvrages. Les résultats montrent que 41 935 termes-clés ne sont présents soit dans les articles soit dans les ouvrages, et que 3 141 termes-clés sur un total de 5 167 sont exclusivement présents dans les articles, tandis que 38 794 termes-clés sur un total de 40 975 sont exclusivement présents dans les ouvrages.

3 Approche

Nous avons mis en place un système expérimental pour notre étude comparative d'extraction de relations. L'architecture de ce système est illustrée sur la Figure 1. À partir de documents textuels bruts, le but est de pouvoir extraire des relations entre les idées clés et de pouvoir les filtrer en fonction du domaine.

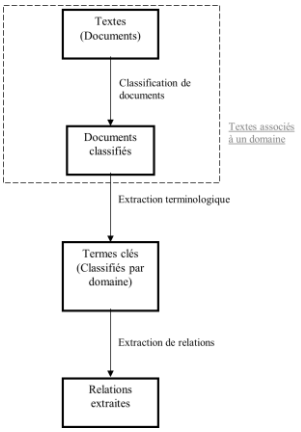

Figure 1 - Système général d'extraction de relations en utilisant la classification de relations

2.1 Classification de documents

Tout d'abord, nous avons développé un système de classification de documents. Celui-ci attribut chacun des paragraphes de chacun des textes un des 8 domaines après les avoir prétraités. Le système prend en entrée d'une part les articles qui sont déjà associés à un domaine spécifique et d'une autre part les ouvrages dont le domaine est à déterminer. Dans le cas des ouvrages, nous utilisons un système de classification qui prend en entrée une représentation du texte plutôt que le texte en lui-même qui peut être un sac de mots ou une représentation vectorielle.

2.1.1 Pré-traitement des textes

Les textes ont été prétraités selon les méthodes classiques. Pour ce faire, nous avons utilisé la bibliothèque Stanza [1]. Nous avons réalisé les étapes suivantes dans cet ordre : mise en minuscule, suppression de la ponctuation, tokenisation des phrases, suppression des *stopwords* en français, lemmatisation.

2.1.2 Classification de documents

Nous utilisons la classification automatique et supervisée pour attribuer une catégorie à chaque document, mais cela est difficile à faire directement avec les données textuelles brutes. Ainsi, nous utilisons différentes méthodes de représentation des documents. Nous utilisons des modèles de classification multi-classes pour attribuer l'une des 8 classes apprises. Pour ce faire, nous avons utilisé plusieurs modèles de classification multi-classes tels que *MultinomialNB*, *LinearSVC* et *SGDClassifier*.

2.1.3 Représentation de documents

Nous utilisons trois méthodes courantes afin de représenter les documents : les méthodes classiques, les embeddings et les modèles de langage. La première façon de représenter les documents consiste à les représenter par sac de mots (ou *bag of words - BoW*, en anglais). Nous nous intéressons à deux façons de le décrire : *Classic BoW* et *TF-IDF*. Ensuite, en ce qui concerne les embeddings, deux méthodes sont possibles : utilisation des plongements lexicaux (*word embeddings*) avec le modèle word2vec ou utilisation de doc2vec qui consiste à créer un vecteur du document directement. Finalement, concernant les modèles de langues, nous utilisons la méthode de représentation phrase par phrase grâce à Sentence-CamemBERT et Sentence-Flaubert (récupérés depuis *HuggingFace)* pour créer une représentation de chacune des phrases, puis nous appliquons la moyenne.

2.2 Extraction terminologique

Dans la deuxième étape du système, un extracteur terminologique est utilisé pour extraire des expressions. Cependant, le système automatique utilisé par notre système effectue une unification des expressions extraites afin de regrouper les différentes formes fléchies d'un texte sous une même forme représentative. Les annotations des experts sont utilisées comme extraction terminologique, mais d'autres méthodes telles que pke [2], TopicRank [3], keyBERT [4] ou YAKE [5] peuvent également être utilisées.
Pour l'étape d'unification, les données sont prétraitées pour éliminer les premiers mots outils tels que les déterminants, par exemple. Nous utilisons la lemmatisation pour unifier les annotations d'un même identifiant sous un même terme-clé représentatif dit référent. Nous avons divisé notre problème en deux cas de figure : les termes simples et complexes. Pour les termes complexes, la forme fléchie ayant le plus grand nombre d'occurrences dans le corpus est choisie comme substitut pour représenter un terme-clé donné. Pour les termes simples, une forme canonique dite standard est privilégiée pour privilégier la forme singulière et générique à la forme plurielle. Ainsi, "de restauration immobilière", "restauration immobilière" et "Restauration immobilière" font tous les trois références au même terme-clé.

2.3 Extraction de relations

La dernière étape du système général (voir Figure 1) consiste à extraire les relations entre deux termes clés. Pour ce faire, plusieurs techniques d'extraction de relations sont disponibles et peuvent être utilisées, mais notre système utilise une classification de type de relations pour extraire des relations potentielles à partir des termes clés extraits précédemment. Nous utilisons une base de connaissance lexicale pour identifier les types de relations appropriées entre les termes clés d'un même domaine. Ensuite, les relations extraites sont vectorisées pour apprendre le type de relation associé à l'aide d'une classification. Après l'apprentissage, un vecteur de relations est construit en sélectionnant des paires de termes clés, qui peut être associé ou non à un type de relation donné.

2.3.1 Instances de relations lexico-sémantiques à partir d'une base de connaissance

En utilisant une base de connaissance lexico-sémantique, nous récupérons des relations potentielles entre les termes-clés unifiés de même domaine. Nous avons choisi d'utiliser le réseau lexico-sémantique JeuxDeMots [6] pour le français car il contenait 14 millions de nœuds et 320 millions de relations à l'époque de notre choix, et était librement

disponible. Cette ressource nous permet de récupérer différents types de relations tels que l'hyperonymie, l'hyponymie, les constituants (*has_part*), la synonymie et l'antonymie. Nous récupérons l'instance de relation sous la forme Terme1−Type_relation−Terme2−Poids, où le poids représente le niveau de confiance attribué à la relation. Ce score est obtenu grâce aux joueurs qui ont attribué chacun un score de confiance à une relation donnée, et nous filtrons tout poids inférieur ou égal à 0 car cela signifie que peu de joueurs ont donné leur confiance à cette relation. D'autres ressources lexico-sémantiques existent, telles BabelNet [7] ou WordNet [8], mais nous avons choisi d'utiliser JeuxDeMots pour nos besoins.

2.3.2 Création de représentations vectorielles des paires de termes

Dans cette section nous apprenons des plongements lexicaux et en déduisons part une opération arithmétique simple une représentation vectorielle des instances de relations extraites depuis notre base de connaissances.

Entrainement des plongements lexicaux. Nous utilisons des termes-clés, qu'ils soient des expressions comme "Projet et construction soumis à enquête publique" ou des mots singuliers tels que "caravanage". Cependant, les modèles entraînés sur des corpus standards ne sont pas adaptés pour traiter des plongements lexicaux pour des termes et des expressions spécifiques identifiés à l'avance par des experts. Nous souhaitons donc apprendre des plongements lexicaux pour les termes et les expressions. Nous extrayons des vecteurs moyens à partir des différents composants d'une expression pour obtenir des représentations vectorielles continues de termes-clés. Pour ce faire, nous entraînons des plongements lexicaux sur le corpus MÉTIER en faisant une substitution lexicale dans l'ensemble du corpus MÉTIER de chaque annotation experte par son terme-clé référent. Nous utilisons ensuite l'architecture CBOW (*Continuous Bag Of Words*) de Word2Vec avec un paramétrage par défaut, en considérant les termes-clés comme des entités plutôt que la moyenne des mots les composant. Nous cherchons ainsi à forcer la compréhension des suites de mots comme un élément indivisible (par exemple, "dépenses_publiques" pour le terme-clé "dépenses publiques").

Déduction des représentations vectorielles de relations typées. Après avoir entraîné nos plongements lexicaux à l'aide de *Word2vec*, nous pouvons en déduire des représentations vectorielles de relations dites typées. Nous disposons d'un ensemble de paires de termes-clés reliées par un type de relation, et les représentations vectorielles varient en fonction du type de relation. Nous extrayons un ensemble d'apprentissage de paires de termes-clés pour un type de relation donné à partir de la liste de termes-clés. Nous récupérons ensuite les vecteurs associés aux termes-clés que nous avons précédemment construits avec *Word2Vec*. Nous formons une paire de vecteurs de termes-clés à partir des vecteurs de termes-clés, puis nous calculons un nouveau vecteur en effectuant une opération entre les deux vecteurs de chaque paire en entrée. Mathématiquement, si V_1 représente un terme-clé source et V_2 un terme-clé cible, tous deux liés par une relation R, alors operationRelation (V_1, V_2) est le vecteur de relation (source, R, cible). Les relations peuvent être symétriques ou non symétriques, ce qui détermine l'opération à effectuer sur les vecteurs de mots. Si la relation est non symétrique, nous utilisons la différence, tandis qu'en cas de symétrie, nous appliquons une valeur absolue à cette différence. L'hyperonymie *(r_isa*) et l'hyponymie (r_*hypo*) sont des exemples de relations non symétriques, tandis que la synonymie est une relation symétrique. Nous utilisons l'exemple concret de la relation d'hyperonymie pour illustrer le calcul du vecteur de relation dans le cas d'une relation non symétrique, et l'exemple de la relation de synonymie pour illustrer le calcul du vecteur de relation dans le cas d'une relation symétrique.

2.3.3 Classification : apprentissage du type de relations

Nous avons développé des classifieurs binaires pour prédire le type de relations entre deux termes, en utilisant un ensemble équilibré de relations sémantiques tirées de JeuxDeMots pour l'apprentissage et la validation. Chaque classificateur s'applique sur deux relations qui vont de pair et associe un type de relation, regroupant les hyponymes et hyperonymes, les synonymes et antonymes, et les constituants "fait partie de" et "est une partie de". Plusieurs modèles de classification binaire ont été créés allant des arbres de décision et des méthodes ensemblistes jusqu'aux machines à vecteur de support (*Support Vector Machine* - SVM), pour déterminer le modèle le plus pertinent pour classifier les relations à partir de nos vecteurs de relations et pour la classification de domaine. Les vecteurs embeddings de relations sont fournis aux classificateurs pour prédire si une paire de termes fait référence à une association soit entre une paire de relations du premier type dans un paquet, soit de l'autre type de relation du même paquet. Les classifieurs sont utilisés de manière imbriquée pour prédire de nouvelles relations sémantiques non connues en exécutant les classifieurs successivement et en sélectionnant ceux qui ont obtenu un score supérieur à 95% de confiance. Le but est de déterminer de nouvelles relations pour un domaine métier spécifique, plutôt que de filtrer un réseau lexical et d'obtenir uniquement des relations présentes dans une base de connaissances. Le développement de classificateurs pour la prédiction de nouvelles relations peut être vu comme un axe d'enrichissement de telles ressources.

4 Evaluation

3.1 Corpus de test

Afin d'évaluer différents modèles de classification de documents, nous utilisons les articles déjà catégorisés entre les différents domaines. Cela permet d'avoir un corpus déjà labélisé. Nous avons constaté que nous avions une composition très déséquilibrée. Pour cela nous avons équilibré notre corpus afin de créer un corpus de test équilibré. Pour cela, nous avons réduit le nombre d'échantillons au nombre d'articles de la plus petite classe à savoir "élections". Nous nous retrouvons donc avec 250 éléments dans chacune des classes après le rééquilibrage.

3.2 Résultats

Nous allons maintenant présenter les résultats de notre expérimentation pour la classification de documents. Nous allons tout d'abord exposer les scores obtenus pour l'attribution des textes aux 8 domaines. Ensuite, nous

sélectionnons les résultats les plus performants pour la classification afin de vérifier leur pertinence pour l'extraction de relations. Enfin, nous comparons les résultats obtenus avec et sans classification de documents. Tous les résultats sont évalués selon une mesure unique : la F-mesure.

Catégories	Représentation de documents	Prédiction	Commande publique	Comptabilité et finances locales	Justice	RH territoriales	Santé	Urbanisme	Elections	Etat civil & cimetières	Moyenne
Plongements lexicaux	Doc2Vec	*most_similar*	0.93	0.89	0.91	0.77	0.78	0.92	0.98	0.92	0,89
	Word2Vec	Linear SVC	0.48	0.32	0.02	0.22	0.38	0.07	0.40	0.57	0,31
Méthodes classiques	Classic BoW	Linear SVC	**0.99**	**0.95**	0.88	**0.96**	0.94	**0.96**	**0.99**	0.91	**0.95**
	TF-IDF	Linear SVC	0.94	0.91	0.93	0.95	**0.97**	0.92	**0.99**	0.94	0,94
		SGDClassif	0.94	0.89	**0.94**	0.95	0.95	0.93	**0.99**	0.95	0,94
Modèles de langues	Sentence CamemBERT	Linear SVC	0.94	0.89	0.90	0.90	0.92	0.92	0.94	**0.97**	0,92

Table 1 - Résultats de classification de documents

3.2.1 Résultats de classification de documents

Les résultats de la classification de documents sont présentés dans la Table 1 qui contient uniquement les meilleurs résultats pour chaque catégorie de représentation de documents. Les meilleurs résultats sont mis en évidence pour chacun des 8 domaines. Bien que d'autres modèles aient également donné de bons résultats, nous avons choisi d'utiliser les méthodes classiques, telles que *Classic BoW* et *TF-IDF*, car elles ne nécessitent pas de modèle spécifique pour une langue en particulier. Cela signifie que notre modèle peut être utilisé pour d'autres langues sans avoir à entraîner un nouveau modèle pour chaque langue. Bien que les méthodes classiques semblent plus simples, les résultats obtenus peuvent sembler surprenants. Cependant, une analyse plus fine sera effectuée pour confirmer ces résultats.

3.2.2 Résultats dans le cadre de l'extraction de relations

Les résultats pour l'extraction de relations, avec et sans classification, sont présentés dans la Table 2. Les résultats pour chaque domaine classé sont spécifiés, mais les résultats moyens sont présentés dans les deux dernières colonnes, car le modèle d'extraction de relations est moyenné par défaut, quand le domaine n'est pas connu. Nous constatons que la classification de domaine conduit à de meilleurs résultats pour les relations asymétriques (hyperonymie VS hyponymie et fait partie de VS est une partie de) et pour les relations symétriques (synonymie et antonymie) pour deux des quatre classifieurs (*SVC* et *Decision Tree*). Nous pouvons donc conclure qu'il est utile d'utiliser une classification de domaine avant d'extraire les relations. Nous supposons que cela améliore également la qualité des relations extraites, mais une analyse qualitative est nécessaire pour confirmer cette hypothèse, que nous envisageons de réaliser dans le futur.

5 Conclusion

Pour conclure, dans cet article, nous avons étudié si la classification automatique de textes peut aider à l'extraction de relations. Nous avons évalué différentes représentations de documents et constaté que la représentation sous sac de mots est efficace pour la classification de textes de spécialité, qui a permis de classifier nos documents dans huit domaines

Type de relation	Classifier	Commande publique	Comptabilité et finances locales	Justice	RH territoriales	Santé	Urbanisme	Elections	Etat civil & cimetières	Moyenne x domaine	Moyenne sans la classification de domaine
Hyperonymie VS Hyponymie	SVC	0.78	0.80	0.79	0.86	0.83	0.84	0.89	0.83	**0.83**	0.75
	DT	0.57	0.68	0.71	0.77	0.79	0.72	0.78	0.71	**0.72**	0.71
	RF	0.75	0.74	0.76	0.81	0.83	0.78	0.85	0.80	**0.79**	0.78
	k-NN (k=5)	0.65	0.81	0.80	0.84	0.78	0.82	0.87	0.77	**0.79**	0.78
Synonymie VS Antonymie	SVC	0.85	0.74	0.74	0.79	0.85	0.78	0.84	0.80	**0.80**	0.73
	DT	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	**0.75**	0.72
	RF	0.79	0.61	0.64	0.69	0.75	0.65	0.76	0.73	0.70	**0.78**
	k-NN (k=5)	0.65	0.61	0.64	0.66	0.66	0.63	0.65	0.6	0.64	**0.74**
Fait partie de VS est une partie de	SVC	0.75	0.84	0.83	0.84	0.91	0.80	0.66	0.84	**0.82**	0.79
	DT	0.76	0.72	0.75	0.72	0.81	0.61	0.54	0.61	**0.68**	0.65
	RF	0.79	0.81	0.76	0.86	0.89	0.79	0.63	0.86	**0.80**	0.73
	k-NN (k=5)	0.28	0.80	0.78	0.81	0.81	0.72	0.55	0.79	**0.75**	0.75

Table 2 - Résultats de l'extraction de relations à la fois lors de la classification de domaine et sans

différents. Nous avons ensuite utilisé cette classification pour l'extraction de relations, en passant par l'extraction terminologique, et avons observé que la classification de documents améliore la qualité des résultats pour la plupart des relations. En résumé, cette étude montre que la classification automatique de documents peut aider à l'extraction de relations. Nous prévoyons toutefois de poursuivre nos recherches en effectuant une étude approfondie sur des corpus de référence en anglais, ainsi qu'en modifiant nos systèmes d'extraction terminologique et d'extraction de relations pour trouver une combinaison performante pour des textes de spécialité et de référence.

6 Références

[1] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, et C. D. Manning, « Stanza: A Python Natural Language Processing Toolkit for Many Human Languages », dans *System Demonstrations, ACL*, 2020.

[2] F. Boudin, « pke: an open source python-based keyphrase extraction toolkit », dans *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, Osaka, Japan, déc. 2016, p. 69-73.

[3] A. Bougouin, F. Boudin, et B. Daille, « TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction », 2013.

[4] M. Grootendorst, « KeyBERT: Minimal keyword extraction with BERT. » Zenodo, 2020.

[5] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, et A. Jatowt, « YAKE! Keyword extraction from single documents using multiple local features », *Information Sciences*, vol. 509, p. 257-289, janv. 2020.

[6] M. Lafourcade, « JeuxDeMots : Un réseau lexico-sémantique pour le français, issu de jeux et d'inférences », p. 40, 2020.

[7] R. Navigli et S. P. Ponzetto, « BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network », *Artificial Intelligence*, vol. 193, p. 217-250, déc. 2012.

[8] G. A. Miller, « WordNet: A Lexical Database for English », *Communications of the ACM*, vol. 38, nº 11, p. 39-41, 1995, doi: 10.1145/219717.219748.