Étiquetage thématique automatisé de corpus par représentation sémantique

Lucie Martinet[**,***], Hussein T. Al-Natsheh[*,***,****], Fabien Rico[*,♯], Fabrice Muhlenbach[*,♯♯], Djamel A. Zighed[*,***]

[*]Université de Lyon, France
[**]CESI EXIA/LINEACT, 19 Avenue Guy de Collongue, F-69130 Écully, France
[***]Lyon 2, ERIC EA 3083, 5 Avenue Pierre Mendès France - F69676 Bron Cedex
[****]CNRS, ISH FRE 3768, 14 avenue Berthelot - 69363 Lyon Cedex 07
[♯]Lyon 1, ERIC EA 3083, 5 Avenue Pierre Mendès France, F69676 Bron Cedex
[♯♯]UJM-Saint-Etienne, CNRS, Lab. Hubert Curien UMR 5516, F-42023 Saint Etienne

Résumé. Dans les corpus de textes scientifiques, certains articles issus de communautés de chercheurs différentes peuvent ne pas être décrits par les mêmes mots-clés alors qu'ils partagent la même thématique. Ce phénomène cause des problèmes dans la recherche d'information, ces articles étant mal indexés, et limite les échanges potentiellement fructueux entre disciplines scientifiques.

Notre modèle permet d'attribuer automatiquement une étiquette thématique aux articles au moyen d'un apprentissage des représentations sémantiques d'articles du corpus déjà étiquetés. Passant bien à l'échelle, cette méthode a pu être testée sur une bibliothèque numérique d'articles scientifiques comportant des millions de documents. Nous utilisons un réseau sémantique de synonymes pour extraire davantage d'articles sémantiquement similaires et nous les fusionnons avec ceux obtenus par un modèle de classement thématique. Cette méthode combinée présente de meilleurs taux de rappel que les versions utilisant soit le réseau sémantique seul, soit la seule représentation sémantique des textes.

1 Introduction

L'activité des chercheurs a été bouleversée par un accès toujours plus important aux bibliothèques numériques en ligne. La recherche d'information dans ces bibliothèques numériques se fait le plus souvent au moyen de mots-clés entrés dans des moteurs de recherche. Néanmoins, l'appariement entre les mots-clés entrés et ceux utilisés pour décrire les documents scientifiques pertinents présents dans ces bibliothèques numériques peut s'avérer limité si la terminologie employée n'est pas la même dans les deux cas. Tout chercheur appartient à une communauté avec laquelle il partage des connaissances et un vocabulaire communs. Cependant, lorsque celui-ci souhaite étendre l'exploration bibliographique au-delà de sa communauté d'appartenance afin de recueillir des éléments d'information qui le conduisent à de nouvelles connaissances, il convient de lever plusieurs verrous scientifiques et techniques induits par la grande taille des bibliothèques numériques, l'hétérogénéité des données et la complexité du

langage naturel. Les chercheurs qui travaillent dans un contexte pluri- et trans-disciplinaire doivent pouvoir accéder aux documents qui les intéressent sans pour autant être bloqués par la barrière d'un cloisonnement disciplinaire induit par une méconnaissance du vocabulaire employé par d'autres disciplines scientifiques. Le plus souvent, les réseaux sémantiques sont une bonne réponse aux problèmes de variations linguistiques en retrouvant des synonymes ou des champs lexicaux communs. Dans le domaine scientifique, toutefois, cette approche n'est pas suffisante car elle se heurte à la terminologie propre au jargon scientifique et technique qui, par nature, est très spécifique, et qui a la particularité d'évoluer très rapidement. Une autre solution pourrait être apportée par le plongement lexical (ou "*word embedding*"). Cette technique permet de retrouver des termes liés par une proximité au sein d'un même document et, de là, de déduire une proximité sémantique. Cette approche présente malgré tout les problèmes de ne pas donner d'information sur le nombre de termes dont il faut tenir compte pour être encore considéré comme sémantiquement proche du terme initial et de ne pas trop bien fonctionner quand il s'agit d'un concept composé de plusieurs termes plutôt que d'un seul et unique terme.

Dans cet article, nous proposons une solution combinant deux sources d'information sémantique : la première est issue de l'ensemble de synonymes déduits d'un réseau sémantique, la seconde provient de la représentation sémantique d'une projection vectorielle des articles.

2 État de l'art

La recherche de documents sémantiquement similaires n'est pas un problème nouveau en fouille de textes. Dans les bibliothèques numériques, les documents peuvent être enrichis par des méta-données qui permettent de les qualifier, les étiqueter et les classer. Ces enrichissements (*tags*, mots-clés ou catégories de sujets) manquent cependant d'une taxonomie standardisée et sont pénalisés par la subjectivité du jugement des personnes impliquées dans le processus d'annotation manuel (Abrizah et al., 2013).

Dans ces bibliothèques numériques, pour parvenir à atteindre des documents sémantiquement liés à des documents ou des mots-clés fournis en entrée, l'emploi de sources d'extension sémantique nous semble être une piste incontournable. Une première solution consiste à utiliser des bases de données lexicales comme *WordNet* (Miller, 1995) ou des bases de connaissances telles que *BabelNet* (Navigli et Ponzetto, 2012), *DBpedia* (Lehmann et al., 2015) ou *YAGO* (Mahdisoltani et al., 2015).

Une autre solution consiste à utiliser des techniques de plongement lexical (Bojanowski et al., 2017) pour trouver des terminologies sémantiquement similaires. Malgré l'avantage de ces techniques, celles-ci ne fournissent pas de critère permettant de définir précisément une proximité et ainsi de concevoir qu'un terme proche dans la projection puisse être encore considéré comme étant sémantiquement proche du terme initial. Les modèles thématiques, tels que l'*allocation de Dirichlet latente*, ou LDA (Blei et al., 2003), ainsi qu'une version supervisée de LDA (Ramage et al., 2009), sont parmi ceux qui semblent être les plus appropriés pour résoudre le problème qui nous intéresse. Cependant, dans le cas d'une application réelle à des millions de documents, telle qu'une bibliothèque numérique comportant des collections d'articles scientifiques touchant de nombreuses disciplines, et cela sur un grand nombre d'années, même les approches évolutives récentes demandent l'utilisation de puissances de calcul vraiment conséquentes, comme l'emploi d'une ferme de calcul (*computer cluster*) (Liang et al., 2015).

3 Modèle *Æ2TS* et expérimentations

Notre méthode de classification de documents, destinée à l'attribution d'étiquettes aux articles d'une bibliothèque numérique, est une combinaison de deux autres méthodes connues.

Corpus scientifique et catégories de sujets. La première méthode repose sur la construction d'un modèle d'apprentissage supervisé effectué à partir d'une représentation vectorielle sémantique des articles de la bibliothèque numérique *ISTEX* [1], qui offre les avantages d'être pluridisciplinaire et issue de collections provenant de différents éditeurs scientifiques. Elle propose plus de 4 millions d'articles, dont les méta-données sont publiques, correspondant à nos critères de sélection (publication durant ces vingts dernières années, en anglais, ayant les méta-données titres, résumés, mots-clés et sujets). Les étiquettes à attribuer aux articles sont issues de la collection "Web of Science" [2] contenant plus de 250 sujets faisant consensus dans le monde de la recherche. Ces sujets sont dits *aplanis*, étant présentés avec leur domaine père, sous forme de liste comme par exemple [informatique, intelligence artificielle]. Nous travaillons ici uniquement sur des sujets simples, c.-à-d. sans combinaison de sujets reliés par des mots de liaison, composés de termes uniques ou d'un nom et d'un adjectif (p. ex. « intelligence artificielle »), afin d'éviter toute confusion entre sujet combiné et liste aplanie. Seuls les sujets connus du réseau sémantique sont utilisés, afin de bénéficier d'une liste de synonymes conséquente. Au total, nous avons recensé 33 sujets d'étiquettes, en anglais, constitués de sujets simples, qui permettent de construire des ensembles de tests positifs significatifs de plus de cent articles.

Données. Nous construisons un ensemble d'entraînement, pour chaque sujet recherché, grâce à une requête sélectionnant les articles de la bibliothèque scientifique numérique comportant les étiquettes associées (issues de *Web of Science*) dans leur titre ou leur résumé. Cette requête s'effectue par le moteur de recherche *Elasticsearch* (Dixit, 2017) d'*ISTEX*. L'ensemble de tests est construit par des articles contenant les mots recherchés dans leur liste de mots-clés ou de sujets, mais absents de leur titre et résumé.

Méthode de Représentation Sémantique par Projection Vectorielle : « RSPV ». Tous les articles de la bibliothèque sémantique numérique sont préalablement transformés dans leur représentation dans un espace sémantique vectoriel (utilisation de LSA (Halko et al., 2011), sur la matrice de sac de bi-grammes et uni-grammes de mots). Seuls les mots, non vides, ayant une fréquences d'au moins 20 apparitions sont considérés. Nous construisons ensuite un modèle de classement pour chaque sujet avec les forêts aléatoires. Cette méthode s'appuie sur un ensemble d'entraînement sur des ensembles d'exemples positifs et négatifs de même taille. Les exemples positifs sont extraits du corpus *ISTEX* avec *Elasticsearch* et les exemples négatifs sont retournés de façon aléatoire. Tous les articles du corpus sont ainsi ordonnés suivant leur probabilité d'appartenir au sujet recherché en une liste qui est ensuite tronquée (100 000 premiers) pour donner la sortie de RSPV.

Méthode d'ensemble de synonymes : « Synset ». La méthode d'attribution d'étiquette *Synset* s'appuie sur une banque de synonymes, telle que *BabelNet*, pouvant être utilisée à la fois comme un dictionnaire encyclopédique, un réseau sémantique ou une base de connaissances. À partir d'une étiquette issue de termes de référence de la base *Web of Science*, nous composons un groupe de mots synonymes « synset », ou « *synonym set* », ayant une équivalence

1. http://www.istex.fr/
2. https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html

sémantique. Une requête recherchant l'étiquette et tous ses synonymes est ensuite lancée dans le moteur de recherche d'*ISTEX* sur les méta-données des articles. Cette requête exécutée, nous obtenons une liste d'articles ordonnés par pertinence que nous appellons « liste *synset* ».

Méthode de Auto-Etiquetage Thématique de Texte basé sur la Sémantique : « *Æ2TS* ». Notre méthode *Æ2TS* est une combinaison des deux méthodes décrites précédemment, soit une fusion des résultats des listes *synset* et *RSPV*. Nous appliquons la moyenne des rangs attribués à un article donné dans *RSPV* et *synset*, puis nous ré-ordonnons les articles avec cette nouvelle valeur de rang moyen, ce qui peut s'exprimer comme suit : soit s_A le rang attribué à l'article A par la méthode *synset* et r_A le rang attribué à A par la méthode *RSPV*, la valeur t_A utilisée pour réaliser l'ordre des listes fusionnées sera $t_A = \frac{s_A + r_A}{2}$. Lorsque la méthode *synset* n'attribue aucun rang à un article A, nous appliquons la formule suivante : $t_A = r_A \times |S|$, où S est l'ensemble des résultats donnés par la méthode *synset*, et $|S|$ le nombre de résultats de cette liste. Notons que nous restreignons volontairement la liste des résultats de la méthode *Æ2TS* à au plus le double de la taille des résultats obtenus avec la méthode *synset*, la liste *synset* ayant un nombre de résultats plus petit que celui de la liste *RSPV*.

4 Résultats et discussion

Synset	**RSPV**	**Æ2TS**	**Sujets**	$\|test\|$	$\|S\|$	**Synset**	**RSPV**	**Æ2TS**	**Sujets**	$\|test\|$	$\|S\|$
6.54%	12.18%	**19.18%**	Artificial Intelligence	657	7903	14.16%	8.58%	**28.33%**	Substance Abuse	466	7893
22.70%	5.41%	**24.98%**	Information Systems	1313	8440	14.37%	5.39%	**18.12%**	Thermodynamics	2727	8375
0.00%	5.69%	**10.35%**	Rehabilitation	773	7449	**7.16%**	3.69%	5.83%	Psychology	1871	7187
16.25%	13.29%	**20.68%**	Philosophy	677	7116	**5.45%**	3.05%	**5.45%**	Ophthalmology	459	2096
3.71%	6.39%	**11.17%**	Microscopy	6819	8547	0.00%	7.92%	**12.46%**	Ceramics	1276	8249
3.64%	0.22%	**9.89%**	Infectious Diseases	1375	8343	9.41%	7.93%	**12.69%**	Toxicology	1552	5563
9.52%	2.12%	**16.40%**	Respiratory System	189	8968	9.96%	**32.95%**	28.54%	Neuroimaging	522	3679
12.44%	5.12%	**12.44%**	Literature	860	7357	**7.59%**	4.44%	7.31%	Sociology	698	4718
32.40%	14.46%	**35.48%**	Robotics	747	3705	14.71%	5.53%	**18.76%**	Psychiatry	1700	7448
29.85%	7.10%	22.76%	Pediatrics	747	8233	3.64%	**10.62%**	9.74%	Oncology	2937	5705
0.02%	4.40%	**5.63%**	Mechanics	4640	8222	5.88%	**8.98%**	8.05%	biophysics	323	3674
0.07%	**5.81%**	1.19%	Condensed Matter	1514	1523	4.91%	4.21%	**7.72%**	Emergency Medicine	285	1379
18.21%	14.07%	**35.98%**	Transplantation	4997	8975	8.81%	10.73%	**15.47%**	Surgery	6412	8271
16.70%	18.06%	**19.93%**	Religion	587	6956	**4.45%**	0.11%	3.40%	Physiology	2761	8494
6.38%	2.82%	**7.74%**	Pathology	2726	8544	0.57%	**1.89%**	0.57%	Mycology	530	542
4.26%	2.85%	**8.45%**	Immunology	4769	8787	9.02%	12.65%	**16.37%**	Biomaterials	1020	3649
8.41%	28.31%	**37.45%**	Nursing	3282	8252						
						9.75%	8.81%	**15.82%**	**moyenne**	1885	6492
						6 / 33	5 / 33	**25 / 33**	**meilleurs résultats**		

TAB. 1 – *Valeurs de rappel des trois méthodes : liste issue d'un réseau sémantique (Synset), liste issue d'une représentation sémantique par projection vectorielle (RSPV) et méthode globale (Æ2TS). Les résultats sont donnés pour 33 sujets scientifiques issus de "Web of Science". Le nombre d'articles utilisés pour le test est noté $|test|$ et le nombre d'articles retournés par la méthode synset est noté $|S|$. Le nombre d'articles retournés par la méthode RSPV est toujours plus grand que $|S|$, et le nombre d'articles retournés par la méthode globale (Æ2TS), fusion des deux précédentes méthodes, est nécessairement plus grand encore. Pour chaque sujet, le meilleur résultat des trois méthodes est indiqué en caractères gras.*

Les résultats obtenus par chacune des trois méthodes selon le protocole décrit précédemment sont présentés dans le Tableau 1. La qualité des résultats de chaque méthode est évaluée

au moyen du rappel. Notons que pour savoir si une réponse est correcte ou non pour un article donné, il faudrait avoir une évaluation humaine experte dans tous les domaines, ce qui n'est pas envisageable. Nous avons ainsi utilisé pour nos expérimentations un petit jeu de test déjà étiqueté (au minimum 100 articles par sujet). En raison du petit nombre d'articles présents dans cet ensemble de test, les valeurs de rappel sont globalement faibles pour les trois méthodes.

La combinaison des deux approches dans la méthode *Æ2TS* est celle qui fournit les meilleurs taux de rappel (15,82%) pour le plus grand nombre de sujets testés (24/33). Les quelques cas où la seule méthode d'utilisation du réseau sémantique (*Synset*) dépasse les deux autres ne concerne que des sujets pour lesquels le concept est constitué d'un seul terme (comme « psychologie »). Les concepts issus de plusieurs termes (comme « intelligence artificielle ») semblent mieux retrouvés pour les deux autres méthodes que pour la méthode classique *Synset*, et tout particulièrement pour la méthode *Æ2TS*. Les résultats présentés ici, bien que devant être confortés par d'autres expériences, sont déjà encourageants et confirment l'intérêt de l'apport d'une représentation sémantique issue d'une projection vectorielle pour pouvoir auto-étiqueter des documents scientifiques avec des étiquettes composées d'un ou de plusieurs termes.

5 Conclusion et perspectives

Dans ce travail, nous avons étudié trois méthodes permettant d'attribuer sémantiquement des étiquettes de sujets scientifiques aux articles d'un corpus. Ces étiquettes sont issues d'une taxonomie de la collection *Web of Science*. Or les bibliothèques numériques multidisciplinaires combinent des corpus provenant de nombreux éditeurs scientifiques utilisant chacun leur propre taxonomie. Ce phénomène freine l'accès de certains articles à des chercheurs d'autres disciplines par leur emploi d'une terminologie et d'une taxonomie différentes. En enrichissant la bibliothèque numérique avec plus de balises obtenues à travers la méthode d'auto-étiquetage thématique de textes scientifiques *Æ2TS* que nous proposons, la taxonomie et les balises étendront l'exploration de la recherche à davantage d'articles sémantiquement pertinents.

L'approche *Æ2TS* combine deux sources d'information sémantique (synonymes issus d'un réseau sémantique et résultats de la représentation sémantique d'une projection vectorielle). Notre étude expérimentale montre une amélioration significative en terme de rappel par rapport aux résultats obtenus en utilisant seulement les synonymes de sujets extraits des réseaux sémantiques. Ajoutons que lorsqu'une requête est menée sur un mode exploratoire dans une bibliothèque numérique scientifique, il est difficile de connaître directement les termes exacts de la thématique des documents recherchés. La requête sera donc le plus souvent une périphrase composée de plusieurs termes, situation où la méthode *Æ2TS* retourne les meilleurs résultats.

Remerciements

Les auteurs souhaitent remercier le projet ISTEX ainsi que la Région Auvergne-Rhône-Alpes (ARC6) pour leurs soutiens ayant permis la réalisation de ce travail.

Références

Abrizah, A., A. N. Zainab, K. Kiran, et R. G. Raj (2013). LIS journals scientific impact and subject categorization : a comparison between Web of Science and Scopus. *Scientometrics 94*(2), 721–740.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993–1022.

Bojanowski, P., E. Grave, A. Joulin, et T. Mikolov (2017). Enriching word vectors with subword information. *TACL 5*, 135–146.

Dixit, B. (2017). Chapter 2. The Improved Query DSL. In *Mastering Elasticsearch 5.x*, pp. 74–141. Birmingham, UK : Packt Publishing, Limited.

Halko, N., P.-G. Martinsson, et J. A. Tropp (2011). Finding structure with randomness : Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review 53*(2), 217–288.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, et C. Bizer (2015). DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web 6*(2), 167–195.

Liang, F., Y. Yang, et J. Bradley (2015). Large scale topic modeling : Improvements to LDA on Apache Spark. `https://tinyurl.com/y7xfqnze`.

Mahdisoltani, F., J. Biega, et F. M. Suchanek (2015). YAGO3 : A knowledge base from multilingual wikipedias. In *CIDR 2015, Asilomar, CA, USA, January 4-7, 2015*. www.cidrdb.org.

Miller, G. A. (1995). WordNet : A lexical database for English. *Communications of the ACM (CACM) 38*(11), 39–41.

Navigli, R. et S. P. Ponzetto (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence 193*, 217–250.

Ramage, D., D. Hall, R. Nallapati, et C. D. Manning (2009). Labeled lda : A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 248–256. ACL.

Summary

In scientific text corpus, some articles from different research communities are not tagged by the same keywords even if they share the same topic. This causes issues in information retrieval systems using limited number of tag variations and thus, lower chances of interdisciplinary exploration. Our approach automatically assigns a topic tag to articles by learning a classifier for each topic based on the semantics representation of the title and the abstract of already tagged articles. The approach requires much less computation power than using topic modeling on millions of documents. In our proposed model, we use topic sysnomyns to retrieve more semantically similar articles and merge them to the articles obtained by the topic classifier. The experiments show higher recall against two variations of the model, one only uses the synonyms set, and another one only uses the semantic representation of the text.