Alignement de deux espaces sémantiques à des fins d'indexation automatique

Jean-François Chartier, Dominic Forest et Olivier Lacombe
Université de Montréal, École de bibliothéconomie et des sciences de l'information,
C.P. 6128, succursale Centre-ville, Montréal (Québec), H3C 3J7, Canada
Chartier.jf@gmail.com, dominic.forest@umontreal.ca,
lacombe.olivier.ol@gmail.com

RESUME

Cet article présente la méthode et les résultats de l'équipe de l'Université de Montréal à la 12e édition du Défi Fouille de Textes. La méthode développée repose sur une procédure d'apprentissage automatique supervisée. Elle est basée sur un espace sémantique des mots-clés d'indexation (ESMC) induit à partir de la base d'apprentissage et d'un espace sémantique de documents construit à partir de la base de test (ESD). La prédiction des mots-clés d'indexation pour un document de la base de test est réalisée en calculant la proximité entre les documents de l'ESD et les mots-clés de l'ESMC. Les k mots-clés les plus proches d'un document sont considérés être les mots-clés les plus pertinents pour son indexation.

ABSTRACT

This article presents the method and the results of the Université de Montréal team in the 12th edition of the Text Mining Challenge. The method developed is a supervised machine learning procedure. It is based on a keyword semantic space (ESMC) induced from the training set and a document semantic space (ESD) built from the test set (ESD). The keywords prediction for the test set is performed by calculating the proximity between the documents of the ESD and the keywords of the ESMC. The k nearest keywords of a document are considered to be the most relevant keywords for the document indexation.

MOTS-CLES : Indexation automatique ; Espace sémantique ; Corrélation ; Assignation de mots-clés ; apprentissage automatique

KEYWORDS: Automatic indexation ; Semantic Vector Space ; Correlation ; Keywords assignment ; Machine Learning

1 Introduction

La masse documentaire disponible en format numérique augmente rapidement. Cette réalité rend l'indexation manuelle des documents de plus en plus difficile à maintenir. Pour gérer cette masse documentaire des méthodes d'indexation automatique ont été proposées. Deux types de méthodes existent pour attribuer automatiquement des mots-clés d'indexation à un document : les approches par extraction et les approches par assignation.

L'indexation automatique par extraction consiste à attribuer à un document des mots-clés qui sont extraits directement du document en question. Une approche classique de type statistique consiste à extraire des mots-clés d'un document en se basant sur leur degré de représentativité pour le document et sur leur degré de discrimination par rapport à l'ensemble des documents d'un corpus. C'est par exemple ce que permet le coefficient de pondération TF-IDF (Jones 1972). Selon ce coefficient, un mot d'un document sera sélectionné comme mot-clé s'il est à la fois très fréquent dans celui-ci et rare dans l'ensemble du corpus.

Certaines méthodes d'extraction reposent sur des algorithmes d'apprentissage automatique. Un exemple classique de méthode à base d'apprentissage automatique non-supervisé est l'indexation sémantique latente (LSI) (Deerwester et al. 1990) et un exemple important de méthode à base d'apprentissage automatique supervisé est l'algorithme KEA (Witten et al. 1999). De nombreux autres types de méthodes d'extraction existent, tout particulièrement des méthodes à base de graphes, qui centrent l'analyse sur la connexité des mots d'un document (Mihalcea and Tarau 2004).

L'indexation automatique par assignation consiste à aligner un document avec des mots-clés issus d'un vocabulaire contrôlé. Le défi particulier de cette approche est qu'elle implique d'attribuer à un document des mots-clés d'indexation qui ne sont pas nécessairement présents dans le texte de ce dernier. Cette approche a été l'objet d'importants travaux dans les dernières années, parmi lesquels on retrouve majoritairement des approches à base de graphes, par exemple la méthode TopicCoRank (Bougouin 2015) et des méthodes par apprentissage automatique supervisé, par exemple la méthode KEA++ (Medelyan et Witten 2008). Contrairement aux méthodes d'extraction qui consistent à sélectionner directement dans un document les mots saillants pour son indexation, les méthodes d'assignation opèrent en amont un travail d'analyse des équivalences entre les mots d'un document et un vocabulaire contrôlé. Ceci permet d'établir des relations de substitution. À la place de sélectionner comme mot-clé d'indexation un mot extrait d'un document, on sélectionne son ou ses meilleurs substituts dans un thésaurus.

La littérature dans le domaine tend à démontrer que l'indexation automatique, par extraction ou assignation, ne permet pas d'émuler de manière satisfaisante l'indexation manuelle des professionnels. Le développement de méthodes automatiques pour accomplir cette tâche représente toujours un important défi informatique.

La 12e édition du défi Fouille de Textes est une invitation à la communauté scientifique à apporter des éléments de réponses à la problématique de l'indexation automatique. Le reste de l'article présente les modalités du défi de la compétition, la méthode développée par notre équipe et l'analyse de ses performances.

2 Présentation du défi et des corpus

La tâche demandée dans le cadre de cette édition du défi est une tâche d'indexation automatique de notices bibliographiques de langue française. Ces notices sont regroupées en quatre corpus appartenant à quatre domaines de spécialité différente, soit la chimie (CHIMIE), les sciences de l'information (INFO), la linguistique (LING) et l'archéologie (ARCHEO). Le corpus CHIMIE contient 782 notices, le corpus INFO 706 notices, le corpus LING 715 notices et le corpus ARCHEO contient 718 notices. Chaque notice contient un document textuel (titre et résumé) et des mots-clés attribués manuellement par des indexeurs professionnels. Ces mots-clés peuvent être de type contrôlé (appartenant à un thésaurus) ou non-contrôlé.

Chaque corpus a été séparé de manière aléatoire en une base d'apprentissage et en base de test selon les proportions 2/3 et 1/3. Pour la compétition, les mots-clés des notices de la base de test de chaque corpus ont été supprimés. Le défi de la compétition consiste à développer une méthode qui permet de prédire automatiquement ces mots-clés. En d'autres mots, à partir de l'analyse des bases d'apprentissage, il faut prédire pour chaque document des notices des bases de test, quels sont les mots-clés attribués par les indexeurs professionnels.

3 Description de la méthode

La méthode développée par notre équipe est une méthode avec apprentissage automatique supervisée. La méthode est basée sur la construction de deux types d'espace sémantique. Le premier espace, dénoté ESMC, est un espace sémantique des mots-clés d'indexation. Cet espace est le modèle induit à partir de la base d'apprentissage. Le deuxième espace, dénoté ESD, est un espace sémantique des documents de la base de test. La prédiction des mots-clés pour un document de la base de test est réalisée en alignant les deux espaces ESMC et ESD. Les sous-sections suivantes présentent les différents paramètres de notre méthode.

3.1 L'espace sémantique des mots-clés d'indexation (ESMC)

L'ESMC modélise les relations de substitution ou d'équivalence entre d'une part les différentes chaînes de caractères qui composent les documents des notices d'une base d'apprentissage et d'autre part les mots-clés qui leur ont été attribués par des indexeurs professionnels. Dénotons par $\mathbb{T} = \{t_1 \dots t_n\}$ un ensemble de n mots-clés, par $\mathbb{C} = \{c_1 \dots c_m\}$ un ensemble de m chaînes de caractères et par θ_{ij} la valeur d'un coefficient d'association entre un mot-clé t_i et une chaîne c_j. Un ESMC peut être représenté par une matrice $[\theta_{ij}] \in \mathbb{R}^{n \times m}$ dans laquelle chaque ligne forme un vecteur $\vec{t}_i = (\theta_{i1} \dots \theta_{im})$ modélisant le degré d'équivalence entre le mot-clé t_i et l'ensemble des chaînes de caractères $c_j \in \mathbb{C}$.

Les mots-clés et les chaînes de caractères formant l'ESMC sont sélectionnés de la manière suivante. Les mots-clés $t_i \in \mathbb{T}$ sélectionnés correspondent à tous les mots-clés attribués par les indexeurs professionnels à un moins deux notices dans la base d'apprentissage d'un corpus. Ils incluent à la fois des mots-clés contrôlés membres d'un thésaurus et des mots-clés non-contrôlés. La sélection des chaînes de caractères $c_j \in \mathbb{C}$ est effectuée dans les titres et résumés des notices de la base d'apprentissage d'un corpus. Toutes les graphies sont premièrement récupérées. Ensuite, sont supprimées les graphies qui correspondent soit à des nombres, des singletons (un caractère) ou des mots fonctionnels. Les graphies retenues sont ensuite normalisées à l'aide d'un algorithme de racinisation. Finalement, les graphies racinées sont décomposées en chaînes de 5grams de caractères (Mcnamee and Mayfield 2004).

Différents coefficients d'association peuvent être utilisés pour calculer le degré d'équivalence θ_{ij} entre un mot-clé t_i et une chaîne c_j (e.g. information mutuelle, corrélation de Matthews, PMI). Ces coefficients doivent permettre de calculer la dépendance statistique entre t_i et c_j dans un corpus donné. Plus t_i est spécifique à c_j, plus la valeur de θ_{ij} doit être élevée. Dans nos expérimentations, le coefficient χ^2 (chi2) est celui qui a optimisé les performances de notre méthode (Manning, Raghavan, and Schütze 2008, 275).

3.2 L'espace sémantique des documents (ESD)

L'ESD modélise le contenu des documents en termes de distribution de chaînes de caractères. Dénotons par $\mathbb{D} = \{d_1 \dots d_n\}$ un ensemble de n document et par ω_{zj} la valeur d'un coefficient de pondération d'une chaîne de caractères $c_j \in \mathbb{C}$ dans le document $d_z \in \mathbb{D}$. Un ESD peut être représenté par une matrice $[\omega_{zj}] \in \mathbb{R}^{n \times m}$ dans lequel chaque ligne forme un vecteur $\vec{d}_z = (\omega_{z1} \dots \omega_{zm})$ modélisant la saillance des différentes chaînes de caractère $c_j \in \mathbb{C}$ dans le document d_z. L'ESD est forme essentiellement une variante basée sur des 5grams de caractères de l'espace vectoriel classique de Salton (Salton, Wong, and Yang 1975).

Cet espace est construit de la manière suivante. L'ensemble $\mathbb{D}$ correspond aux documents (titre et résumé) des notices de la base de test d'un corpus, alors que $\mathbb{C}$ représente l'ensemble des chaînes de caractères sélectionnés précédemment pour construire l'ESMC de telle sorte que l'ESMC et l'ESD d'un corpus partagent la même base vectorielle et sont alignés. Plusieurs coefficients de pondération peuvent être utilisés, la plupart étant des variantes du TF-IDF. Dans nos expérimentations, les performances de notre méthode ont été optimisées avec le coefficient Okapi BM25 (Robertson et al. 1999).

3.3 Prédiction des *k* mots-clés d'indexation

Notre méthode est basée sur l'hypothèse que la prédiction des mots-clés les plus pertinents pour indexer un document sont les mots-clés qui se substituent le mieux au contenu de ce document. La recherche de ces mots-clés est réalisée en comparant à l'aide d'une métrique les vecteurs $\vec{t}_i \in ESMC$ avec les vecteurs $\vec{d}_z \in ESD$. En d'autres mots, notre méthode est basée sur l'hypothèse que les k mots-clés les plus pertinents pour indexer le document d_z sont les k mots-clés dont les vecteurs $\vec{t}_i$ sont les plus similaires à $\vec{d}_z$. La prédiction consiste alors simplement à maximiser une métrique de similarité $\forall_{i=1}^{k} \underset{\vec{t}_i \in ESMC}{\operatorname{argmax}} S(\vec{d}_z, \vec{t}_i)$. Plusieurs métriques peuvent être utilisées pour comparer les vecteurs d'un espace sémantique (Kiela and Clark 2014, 23). Dans nos expérimentations, la métrique angulaire du cosinus est celle qui a optimisé les performances de notre méthode. Le nombre de k mots-clés prédit pour chaque document test a été fixé en utilisant le nombre moyen de mots-clés attribués par les indexeurs professionnels aux documents de la base d'apprentissage.

4 Résultats

La méthode présentée dans la section précédente est simple et construite à partir de techniques classiques qui ont fait leurs preuves dans de nombreux domaines de la recherche d'informations. Les résultats obtenus avec cette méthode sont toutefois parmi les meilleurs de la compétition. Sur les quatre tâches d'indexation automatique de la compétition, notre équipe a terminé première pour les tâches d'indexation des corpus INFO et LING, deuxième pour la tâche d'indexation du corpus CHIMIE et quatrième pour la tâche d'indexation du corpus ARCHEO. Le tableau 1 présente l'évaluation des performances de notre méthode ainsi que celle des méthodes des quatre autres équipes participantes. Les performances ont été évaluées selon la précision, le rappel et la F-mesure entre les mots-clés prédits par notre méthode et les mots-clés de référence attribués par les indexeurs professionnels. Les scores associés à une étoile sont ceux obtenus avec notre méthode.

Corpus	Évaluations	1er	2e	3e	4e	5e	moy.
INFO	Précision	***31.03**	21.26	21.92	11.72	13.83	19.952
	Rappel	***28.23**	30.32	21.83	23.54	12.01	23.186
	F-mesure	***28.98**	23.86	21.45	15.34	12.49	20.424
LING	Précision	***30.26**	23.28	23.16	13.98	15.67	21.27
	Rappel	***34.16**	32.73	25.85	30.81	16.1	27.93
	F-mesure	***31.75**	26.3	24.19	19.07	15.63	23.388
CHIMIE	Précision	24.92	***19.67**	21.15	10.88	18.19	18.962
	Rappel	21.73	***25.07**	17.54	30.25	14.9	21.898
	F-mesure	21.46	***21.07**	18.28	15.31	15.29	18.282
ARCHEO	Précision	43.48	55.26	53.77	***30.77**	33.93	43.442
	Rappel	52.71	38.03	33.46	***43.24**	31.25	39.738
	F-mesure	45.59	43.46	40.11	***34.96**	30.75	38.974

Tableau 1: Évaluation des performances des méthodes des participants de la compétition. Les scores associés à une étoile sont ceux obtenus avec notre méthode.

La performance de notre méthode sur les corpus INFO et LING dominent significativement celles des méthodes des autres participants. Sur le corpus CHIMIE, la performance de notre méthode est comparable à la méthode développée par l'équipe gagnante. Finalement, pour la tâche d'indexation automatique du corpus ARCHEO, la performance de notre méthode nous positionne en queue de peloton, bien que, en termes de performance absolue, elle soit la meilleure que nous ayons obtenue.

5 Discussion

Les résultats précédents sont caractérisés par d'importantes variations selon les corpus, ce qui suggère que notre méthode serait plus adaptée à certains types de corpus que d'autres. Les meilleures performances de notre méthode ont été obtenues sur le corpus ARCHEO, mais c'est sur les corpus INFO et LING que notre méthode se démarque le plus des autres. De plus, bien que notre méthode se positionne au deuxième rang pour la tâche d'indexation du corpus CHIMIE, les performances sont étonnamment faibles comparativement à celles obtenues sur les trois autres corpus.

Le tableau 2 présente quelques statistiques comparatives qui nous permettent d'expliquer ces résultats. La colonne « $\mathbb{C}_a \cap \mathbb{C}_t$ » correspond à la proportion de chaînes de caractères communes aux documents des notices d'apprentissage et de test. La colonne « $\mathbb{T}_a \cap \mathbb{T}_t$ » correspond à la proportion des mots-clés communes aux notices d'apprentissage et aux notices de test. La colonne « $\mathbb{T} \notin \mathbb{D}$ » correspond à la proportion de mots-clés attribué à une notice par les indexeurs professionnels qui ne sont pas occurrents dans le document (ni dans le titre ou le résumé). Finalement, la colonne « $\boldsymbol{I}(\mathbb{C}, \mathbb{T})$ » correspond à la quantité d'information mutuelle (Bouma 2009) entre les chaînes de caractères d'un corpus et ses mots-clés de références. Ce coefficient permet de mesurer, en termes de dépendance statistique, avec quelle régularité les indexeurs professionnels ont associé mots-clés et chaînes de caractères dans leur travail d'indexation. Plus la valeur de $\boldsymbol{I}(\mathbb{C}, \mathbb{T})$ est élevée, plus ce travail a été systématique : la présence de certaines chaînes de caractères dans un document est systématiquement associée à l'attribution de mots-clés spécifiques.

Corpus	F-mesure	Rang	$\mathbb{C}_a \cap \mathbb{C}_t$	$\mathbb{T}_a \cap \mathbb{T}_t$	$\mathbb{T} \notin \mathbb{D}$	$I(\mathbb{C}, \mathbb{T})$
INFO	28.98	1er	0.67	0.55	0.68	7.15
LING	31.75	1er	0.71	0.58	0.61	8.33
CHIMIE	21.07	2e	0.70	0.44	0.76	3.82
ARCHEO	34.96	4e	0.73	0.60	0.37	10.10

Tableau 2 : Statistiques comparatives des corpus.

Ces statistiques comparatives nous permettent d'avancer quelques conjectures. Premièrement, les performances de notre méthode sont liées au degré de représentativité des bases d'apprentissage des corpus. C'est ce que montre la colonne « $\mathbb{T}_a \cap \mathbb{T}_t$ ». Plus les mots-clés de référence des notices de test sont absents des notices d'apprentissage, moins bonnes sont les prédictions réalisées par notre méthode. Ceci n'est pas étonnant puisqu'il s'agit d'une méthode avec apprentissage automatique de type supervisé, par conséquent la méthode ne peut prédire un mot-clé absent de la base d'apprentissage. Deuxièmement, la colonne « $\mathbb{T} \notin \mathbb{D}$ » montre que lorsqu'une proportion importante des mots-clés d'une notice sont également des mots qui figurent directement dans le document de la notice, comme c'est le cas pour le corpus ARCHEO, notre méthode performe moins bien que celles des autres équipes participantes. Troisièmement, les performances de notre méthode sont liées à la dépendance statistique entre mots-clés et chaînes de caractères. C'est ce que suggère la colonne « $I(\mathbb{C}, \mathbb{T})$ ». Plus le travail des indexeurs professionnels sur un corpus a été systématique, c'est-à-dire plus l'association entre le contenu d'un document et ses mots-clés de référence est spécifique, plus les prédictions de notre méthode sont bonnes.

6 Conclusion

Les analyses précédentes suggèrent que notre méthode est particulièrement adaptée à des tâches d'indexation automatique qui nécessite une part importante d'assignation de mots-clés contrôlés. En effet, les performances de notre méthode se démarquent des celles des autres lorsque les mots-clés à prédire ne sont pas occurrents directement dans le document de la notice et par conséquent qu'une approche par extraction est peu pertinente. C'est la raison pour laquelle notre méthode domine la compétition pour les tâches d'indexation des corpus INFO et LING, mais qu'en contrepartie elle performe moins bien sur le corpus ARCHEO. Nous conjecturons que les méthodes qui ont bien performé sur ce dernier corpus sont des approches qui exploitent davantage des techniques d'extraction de mots-clés. Les analyses précédentes suggèrent également que cet avantage de notre méthode sur les autres est toutefois conditionnel à la systématicité des indexations de référence. C'est pour cette raison que les performances de notre méthode sont moins dominantes pour la tâche d'indexation du corpus CHIMIE.

Références

Bougouin, Adrien. 2015. “Indexation Automatique Par Termes-Clés En Domaines de Spécialité.” Université de Nantes.

Bouma, Gerlof. 2009. “Normalized (pointwise) Mutual Information in Collocation Extraction.” *Proceedings of GSCL*, 31–40.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. “Indexing by Latent Semantic Analysis.” *Journal of the American Society for Information Science* 41 (6): 391–407.

Jones, Karen Spark. 1972. “A Statistical Interpretation of Term Specificity and Its Application in Retrieval.” *Journal of Documentation* 28 (1): 11–21.

Kiela, Douwe, and Stephen Clark. 2014. “A Systematic Study of Semantic Vector Space Model Parameters.” In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and Their Compositionality (CVSC) at EACL*, 21–30.

Manning, C., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Mcnamee, Paul, and James Mayfield. 2004. “Character N-Gram Tokenization for European Language Text Retrieval.” *Information Retrieval* 7 (1-2): 73–97.

Mihalcea, Rada, and Paul Tarau. 2004. “TextRank: Bringing Order into Texts.” In . Association for Computational Linguistics.

Robertson, Stephen E., Steve Walker, Micheline Beaulieu, and Peter Willett. 1999. “Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive Track.” *Nist Special Publication SP*, 253–64.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. 1975. “A Vector Space Model for Automatic Indexing.” *Communications of the ACM* 18 (11): 613–20.

Witten, Ian H., Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. “KEA: Practical Automatic Keyphrase Extraction.” In *Proceedings of the Fourth ACM Conference on Digital Libraries*, 254–55. ACM.