Indexation et appariement de documents cliniques avec le modèle vectoriel

Khadim Dramé[1, 2] Ibrahima Diop[1, 2] Lamine Faty[1, 2] Birame Ndoye[1]
(1) Université Assane Seck de Ziguinchor, Diabir, Ziguinchor, Sénégal
(2) Laboratoire d'Informatique et d'Ingénierie pour l'Innovation, Ziguinchor, Sénégal
khadim.drame@univ-zig.sn, ibrahima.diop@univ-zig.sn,
lamine.faty@univ-zig.sn, b.ndoye5360@zig.univ.sn

RÉSUMÉ

Dans ce papier, nous présentons les méthodes que nous avons développées pour participer aux tâches 1 et 2 de l'édition 2019 du défi fouille de textes (DEFT 2019). Pour la première tâche, qui s'intéresse à l'indexation de cas cliniques, une méthode utilisant la pondération TF-IDF (term frequency – inverse document frequency) a été proposée. Quant à la seconde tâche, la méthode proposée repose sur le modèle vectoriel pour apparier des discussions aux cas cliniques correspondants ; pour cela, le cosinus est utilisé comme mesure de similarité. L'indexation sémantique latente (latent semantic indexing – LSI) est également expérimentée pour étendre cette méthode. Pour chaque méthode, différentes configurations ont été testées et évaluées sur les données de test du DEFT 2019.

ABSTRACT

Indexing and matching clinical documents using the vector space model.

In this paper, we present the methods that we developed to participate in tasks 1 and 2 of the 2019 edition of the french text mining challenge (DEFT 2019). For the first task, which focuses on the indexing of clinical cases, a method using TF-IDF weighting (term frequency - inverse document frequency) has been proposed. For the second one, the proposed method is based on the vector space model to match discussions with corresponding clinical cases; for this, the cosine is used as similarity measure. The latent semantic indexing (LSI) is also used to extend this method. For each method, different configurations were tested and evaluated on the test data of DEFT 2019.

MOTS-CLÉS : indexation, modèle vectoriel, TF-IDF, indexation sémantique latente, similarité sémantique, cas cliniques.

KEYWORDS: indexing, vector space model, TF-IDF, latent semantic indexing, semantic similarity, clinical cases.

1 Introduction

Le défi fouille de textes (DEFT) est une campagne d'évaluation visant à promouvoir le développement de méthodes et d'applications dans le domaine du traitement automatique de langues naturelles (TALN). Dans son édition de 2019, il s'intéresse à l'analyse de cas cliniques ; il comporte trois tâches traitant essentiellement l'indexation, la recherche et l'extraction d'informations à partir de textes biomédicaux (Grabar *et al.*, 2019).

La tâche 1 consiste à identifier, à partir d'une liste de mots clés, ceux qui sont pertinents pour représenter un couple cas clinique/discussion donné. Cette question d'indexation où chaque document est associé à un ou plusieurs mots clés, peut être considérée comme un problème de classification multi-label. Dans la littérature, ce problème a suscité un grand engouement et différentes approches sont proposées. Certaines approches prônent la décomposition du problème en sous-problèmes de classification binaire (Read *et al.*, 2011), et d'autres l'adaptation des méthodes existantes et notamment l'algorithme des k plus proches voisins (Huang *et al.*, 2011; Dramé *et al.*, 2016). La méthode que nous proposons s'inscrit dans la deuxième approche et utilise la méthode de pondération TF-IDF pour déterminer les mots clés pertinents pour indexer un document.

La tâche 2, quant à elle, s'intéresse à l'appariement des cas cliniques et des discussions. L'idée est de déterminer, pour chaque cas clinique, la discussion correspondante à partir d'un ensemble de discussions. Pour traiter ce type de problème, la similarité (sémantique) est communément utilisée. Dans la littérature, deux approches différentes sont développées : une exploitant des ressources externes (Schuhmacher & Ponzetto, 2014) et une autre basée sur la représentation vectorielle. La deuxième approche est largement explorée; différents modèles sont utilisés : le modèle vectoriel (Vector Space Model - VSM)(Salton *et al.*, 1975), l'indexation sémantique latente (Latent Semantic Indexing - LSI) (Deerwester *et al.*, 1990), l'allocation de Dirichlet latente (Latent Dirichlet Allocation - LDA) (Blei *et al.*, 2003) et récemment les plongements lexicaux (word embeddings) (Mikolov *et al.*, 2013; Le & Mikolov, 2014) . Nous proposons une méthode d'appariement inspirée de cette approche et fondée principalement sur la représentation vectorielle des documents. Le VSM et la LSI sont expérimentés avec la mesure de cosinus.

Ce papier décrit ces méthodes, développées pour participer à ces deux tâches du DEFT 2019. Le reste du papier est structuré comme suit : nos méthodes d'indexation et d'appariement sont présentées respectivement dans les sections 2 et 3 ; les résultats obtenus sont décrits et discutés respectivement dans les sections 4 et 5.

2 Indexation de cas cliniques

Dans cette section, nous décrivons notre méthode d'indexation et ses différentes extensions, visant à améliorer ses résultats.

L'approche proposée est basée sur la représentation vectorielle des documents. Dans la phase de prétraitement, chaque document est d'abord segmenté en phrases et les phrases en tokens. Ensuite, l'ensemble des n-grams (séquences de 1 à n tokens, n fixé empiriquement à 6) sont extraits et appariés aux mots clés ; pour l'appariement des n-grams aux mots clés, ces derniers sont préalablement normalisés (suppression de mots vides, racinisation). A l'issue de cette étape, on a une liste de mots clés extraits avec leurs poids respectifs dans le document. Pour identifier les mots clés les plus pertinents pour représenter un document (couple cas clinique/discussion), nous avons expérimenté les mesures comme la fréquence du mot clé dans le document (TF), la TF-IDF, la première occurrence du mot et la combinaison des ces deux mesures dans un modèle supervisé. Les trois configurations de cette méthode sont décrites ci-dessous :

- *uasz-indexer1* : les fréquences des mots clés (TF) dans le document sont utilisées pour les classer ;
- *uasz-indexer2* : les scores TF-IDF des mots clés sont utilisées pour les classer ;

– *uasz-indexer3* : les mesures utilisées dans les configurations précédentes (TF et TF-IDF) sont combinées avec la première occurrence du mot clé dans une méthode supervisée. Ainsi, pour chaque mots clé extrait, sa pertinence pour le document cible est prédite par un modèle entrainé (en utilisant un algorithme d'apprentissage automatique) sur le corpus d'entrainement. Ensuite, les mots clés sont classés en fonction de leur pertinence et les top K les plus pertinents pour le document sont retournés, K étant fourni. Nous avons choisi le classifieur *Naive Bayes* qui, dans les tests que nous avons réalisés sur le corpus d'entrainement, a donné de meilleurs résultats.

Nous avons aussi expérimenté la méthode proposée dans (Dramé *et al.*, 2016), basée sur l'algorithme des k plus proches voisins, mais les résultats obtenus sont mitigés.

3 Appariement de cas cliniques et discussions

Dans cette section, nous présentons notre méthode d'appariement et son extension avec l'indexation sémantique latente (LSI).

La méthode développée repose sur le modèle vectoriel et utilise la mesure cosinus pour calculer la similarité entre cas cliniques et discussions. Dans la première étape, les documents (cas cliniques et discussions) sont d'abord segmentés en phrases et les phrases en tokens. Les mots vides sont ensuite élagués. Enfin, l'ensemble des mots (concepts dans le cas du modèle LSI) du corpus constitue les dimensions du modèle. Chaque document est ainsi représenté dans cet espace (de grande dimension) par un vecteur de mots (ou concepts) dont les composants sont les poids de ces derniers. La similarité entre deux documents, représentés dans ce modèle, est ainsi assimilée au cosinus de l'angle formé par les vecteurs correspondants.

Plusieurs modèles de représentation vectorielle sont explorés (VSM, LSI, LDA, doc2vec embeddings) mais le modèle vectoriel (VSM) et l'indexation sémantique latente (LSI) ont donné de meilleurs résultats sur nos tests. Nous avons ainsi soumis les trois configurations suivantes :

– *uasz-app1* : elle utilise le modèle vectoriel sur le corpus de test ;

– *uasz-app2* : elle utilise le modèle vectoriel sur l'ensemble du corpus (corpus d'entrainement + corpus de test) ;

– *uasz-app3* : elle utilise l'indexation sémantique latente sur le corpus de test.

4 Evaluation

Dans cette section, nous allons d'abord présenter les jeux de données et les métriques utilisées pour évaluer les systèmes participants au DEFT 2019. Ensuite, les résultats de nos méthodes seront analysés et discutés.

4.1 Jeux de données

Le DEFT 2019 a porté sur l'analyse de cas cliniques rédigés en français. Il est constitué de trois tâches : la première s'intéresse à l'indexation de cas cliniques (tâche 1), la deuxième à l'appariement de cas cliniques aux discussions (tâche 2) et la troisième se focalise sur l'extraction d'informations (tâche 3) (Grabar *et al.*, 2019).

Pour chaque tâche, les organisateurs ont fourni des corpus d'entrainement et de test (Grabar *et al.*, 2018). Pour la première tâche, un corpus d'entrainement constitué de 290 couples de cas cliniques/discussions a été fourni avec, pour chaque couple, les mots clés associés dans l'ordre décroissant de leur pertinence. Le corpus de test est quant à lui constitué de 213 couples de cas cliniques/discussions avec le nombre de mots clés attendu. En ce qui concerne la tâche 2, un corpus d'entrainement constitué de 290 cas cliniques et 290 discussions est fourni avec un appariement de chaque cas à la discussion correspondante. Le corpus de test comporte 214 cas cliniques ainsi que le même nombre de discussions.

4.2 Mesures d'évaluation

La précision moyenne (Mean Average Precision – MAP) et la précision au rang N (P@N) sont utilisées pour mesurer les performances des systèmes participants à la tâche 1. Pour la tâche 2, les mesures classiques (précision et rappel) sont utilisées pour évaluer les appariements cas cliniques/discussions.

4.3 Résultats

Les résultats de nos différents systèmes participants à la tâche 1 sont présentés dans TABLE 1. Nous remarquons que le système *uasz-indexer2*, utilisant la pondération TF-IDF, a obtenu des résultats largement meilleurs selon les deux mesures d'évaluation utilisées (MAP et P@N). Notons également que le système *uasz-indexer3*, qui utilise une méthode supervisée, est plus performante que *uasz-indexer1*, qui lui se sert de la fréquence des mots clés pour les classer. Nous avons également expérimenté une approche supervisée combinant les attributs TF.IDF, TF et la première occurrence du mot clé dans le document avec différents classifieurs (Naive Bayes, Neural Network et Random Forest) mais les résultats obtenus sont moins intéressants. Enfin, l'approche développée dans (Dramé *et al*., 2016), utilisant la méthode des k plus proches voisins, a été explorée mais elle n'a pas permis d'améliorer les résultats.

Systèmes	MAP	P@N
uasz-indexer1	0,2761	0,3433
uasz-indexer2	**0,3957**	**0,4547**
uasz-indexer3	0,3174	0,3783

TABLE 1 : Résultats de nos systèmes d'indexation à la tâche 1 du DEFT 2019

Comparé aux systèmes participants à la tâche 1, *uasz-indexer2*, notre meilleur système, a obtenu des résultats moyens (MAP de 0,395 contre 0,478 pour le système le plus performant) dépassant légèrement la moyenne (0,385) ; il a obtenu une précision moyenne comparable à la médiane (0,401) sans utiliser aucune ressource externe supplémentaire.

Les résultats des systèmes développés pour l'appariement des cas cliniques et des discussions sont présentés dans TABLE 2. Nous constatons que les système *uasz-app1* et *uasz-app2*, basés tous sur le modèle vectoriel, ont obtenu des résultats meilleurs que ceux de *uasz-app3*, qui implémente l'indexation sémantique latente. Les trois systèmes ont obtenu dans l'ensemble des résultats satisfaisants.

Quand la fréquence documentaire (IDF) est calculée sur tout le corpus (données d'entrainement et de test) (*uasz-app2*) plutôt que sur le corpus de test seulement (*uasz-app1*), les résultats sont légèrement améliorés avec le modèle vectoriel.

Systèmes	Précision	Rappel
uasz-app1	0,8738	0,8738
uasz-app2	**0,8832**	**0,8832**
uasz-app3	0,8318	0,8318

TABLE 2 : Résultats de nos systèmes d'appariement à la tâche 2 du DEFT 2019

D'autres approches telles que l'allocation de Dirichlet latente (Blei *et al.*, 2003) et les plongements lexicaux (Le & Mikolov, 2014) sont aussi explorées mais elles ont donné des résultats mitigés. En plus, elles nécessitent plus de ressources et un temps d'exécution plus important.

Comparés aux résultats globaux de la tâche 2, nos systèmes ont obtenu des performances satisfaisantes ; tous les trois ont dépassé la moyenne (0,803) sans l'utilisation d'aucune ressource externe supplémentaire. En plus, *uasz-app2*, bien qu'étant simple, a donné des résultats prometteurs dépassant la précision médiane (0,862). Toutefois, comparé au système le plus performant (0,953), nos résultats restent à améliorer.

5 Discussion

L'évaluation de notre méthode d'indexation sur les données de test du DEFT 2019 montre que cette dernière, bien qu'étant simple, reste performante. L'utilisation de la pondération TF.IDF s'est montrée pertinente. Le score TF.IDF d'un mot clé reste ainsi un bon indicateur pour mesurer son poids dans un document. La fréquence aussi reste un indicateur intéressant. Toutefois, la combinaison de ces deux mesures et la première occurrence dans une méthode supervisée a, à notre surprise, donné des résultats mitigés. Cette faible performance peut s'expliquer par la taille moins conséquente de notre corpus d'entrainement (290 documents).

Les résultats de la TABLE 1 montre que cette méthode permet de prédire correctement 40% des mots clés permettant d'indexer des couples de cas clinique/discussion. Une analyse détaillée des résultats a permis de constater que notre méthode peine à retrouver certains mots clés notamment ceux qui ne sont pas explicitement mentionnés dans les documents. Par exemple, le mot clé *dépression respiratoire*, utilisé dans l'index du couple *1202936314.txt/21688289148.txt* n'apparaît pas explicitement dans ce cas clinique ; parfois un synonyme est utilisé (par exemple, le mot clé *malformation congénitale* est utilisé pour indexer le couple *132378112.txt/2122683392.txt* tandis dans le document il est mentionné *sténose congénitale*). D'autres figurent dans le document mais sont disjoints (par exemple, *antagoniste récepteur nk1* dans le couple *12587038525.txt/22358514900.txt*). Dans certains cas, notre méthode retourne des mots clés plus généraux ; par exemple, elle prédit le mot clé *intoxication* pour le couple *11022315250.txt/21172085750.txt* alors que celui attendu est *intoxication aiguë*.

Pour l'appariement des cas cliniques et des discussions, le modèle vectoriel s'est montré performant ; plus de 88% des couples sont correctement appariés (cf. TABLE 2). En analysant les appariements incorrects, nous avons constaté que la plupart ont des scores de similarité très faibles (68% ont un score de similarité inférieur à 0,20).

6 Conclusion

Dans ce papier, nous avons présenté les méthodes que notre équipe a développées pour participer aux tâches 1 et 2 du DEFT 2019. Pour l'indexation des cas cliniques, une méthode basée sur la pondération TF.IDF a été proposée tandis que pour l'appariement des cas cliniques aux discussions, nous avons utilisé le modèle vectoriel. Comparée aux résultats globaux, notre méthode d'indexation a obtenu des performances encourageantes et nécessite d'être améliorée. Nous envisageons d'exploiter des ressources sémantiques pour surmonter des limites soulignées dans la section 5. Notre méthode d'appariement, quant à elle, a donné des résultats prometteurs que nous envisageons également d'améliorer en explorant l'utilisation des plongements lexicaux sur de gros corpus.

Remerciements

Nous remercions les organisateurs du DEFT 2019.

Références

BLEI D., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res*, p. 993–1022.

DEERWESTER S., DUMAIS S., FURNAS G., LANDAUER T. & HARSHMAN R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41 (6), p. 391–407.

DRAME K., MOUGIN F. & DIALLO G. (2016). Large Scale Biomedical Texts Classification: A KNN and an ESA-Based Approaches. *J. Biomedical Semantics* 7: 40.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS: French Corpus with Clinical Cases. In *Proc of LOUHI*, p. 122–128, Brussels, Belgium.

GRABAR N., GROUIN C., HAMON T. & CLAVEAU V. (2019). Recherche et Extraction d'information Dans Des Cas Cliniques. Présentation de la Campagne d'évaluation DEFT 2019. In *Actes de DEFT*, Toulouse, France.

HUANG M., NEVEOL A. & LU Z. (2011). Recommending MeSH Terms for Annotating Biomedical Articles. *Journal of the American Medical Informatics Association* 18 (5), p. 660–67.

LE Q. & MIKOLOV T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning* - Volume 32, p. II–1188–II–1196.

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*.

READ J., PFAHRINGER B., HOLMES G. & FRANK E. (2011). Classifier Chains for Multi-Label Classification. *Machine Learning* 85 (3): 333.

SALTON G., WONG A. & YANG C. S. (1975). A Vector Space Model for Automatic Indexing. In *ACM 18* (11), p. 613–620.

SCHUHMACHER M. & PONZETTO S. P. (2014). Knowledge-Based Graph Document Modeling. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, p. 543–552, NY, USA.