Amélioration de la similarité sémantique vectorielle par méthodes non-supervisées

El Moatez Billah Nagoudi[1] Jérémy Ferrero[2] Didier Schwab[2]
(1) Laboratoire d'Informatique et de Mathématique LIM, Université Echahid Hamma Lakhdar d'ElOued, Algérie
(2) LIG-GETALP, Univ. Grenoble Alpes, Grenoble, France.
e.nagoudi@lagh-univ.dz, jeremy.ferrero@imag.fr, didier.schwab@imag.fr

RÉSUMÉ

Mesurer la similarité sémantique est à la base de nombreuses applications. Elle joue un rôle important dans divers domaines tels que la recherche d'information, la traduction automatique, l'extraction d'information ou la détection de plagiat. Dans cet article, nous proposons un système fondé sur le plongement de mots (*word embedding*). Ce système est destiné à mesurer la similarité sémantique entre des phrases en arabe. L'idée principale est d'exploiter la représentation des mots par des vecteurs dans un espace multidimensionnel, afin de faciliter leur analyse sémantique et syntaxique. Des pondérations dépendant de la fréquence inverse en documents et de l'étiquetage morpho-syntaxique sont appliquées sur les phrases examinées, afin d'améliorer l'identification des mots qui sont plus importants dans chaque phrase. La performance de notre système est confirmée par la corrélation de Pearson entre nos scores de similarité assignés et les jugements humains sur un corpus de référence de l'état de l'art sur des phrases en arabe.

ABSTRACT

Improved the Semantic Similarity with Weighting Vectors

Semantic textual similarity is the basis of countless applications and plays an important role in diverse areas, such as information retrieval, plagiarism detection, information extraction and machine translation. This article proposes an innovative word embedding-based system devoted to calculate the semantic similarity between sentences. The main idea is to exploit the word representations as vectors in a multidimensional space to capture the semantic and syntactic properties of words. IDF weighting and Part-of-Speech tagging are applied on the examined sentences to support the identification of words that are highly descriptive in each sentence. The performance of our proposed system is confirmed through the Pearson correlation between our assigned semantic similarity scores and human judgments on a dataset of the state of the art on arabic sentences.

MOTS-CLÉS : Similarité sémantique de phrases, Représentation vectorielle, Pondération de vecteurs, Arabe.

KEYWORDS: Semantic Sentences Similarity, Word Embedding, Vectors Weighting, Arabic.

1 Introduction

La similarité entre textes est une tâche qui trouve des applications dans divers domaines de recherche tels que la recherche d'information, la traduction automatique, la classification des textes, l'extraction

d'information ou la détection de plagiat.

Dans cet article, nous concentrons notre étude sur le problème de mesure de similarité sémantique entre les phrases en arabe. Notre étude compare plusieurs méthodes de calcul de représentations vectorielles d'une phrase à partir de vecteurs de mots, utilisant chacune différentes pondérations, dépendant par exemple de la fréquence inverse en documents (IDF) ou de la partie du discours (POS) des mots de cette phrase. L'idée sous-jacente est d'améliorer la représentation vectorielle d'une phrase en identifiant les mots qui sont les plus importants dans cette phrase.

Cet article est organisé comme suit, dans la section 2 nous présentons les modèles les plus utilisés pour représenter les mots dans un espace vectoriel. Dans la section 3, nous présentons un aperçu général sur le système proposé. Nous proposons quatre variantes de notre système de mesure de similarité dans la section 4. Enfin, la section 5 présente et discute les résultats de chacune de nos approches.

2 Modèles de représentation vectorielle continue de mots

La représentation des mots par vecteurs dans un espace multidimensionnel, en prenant en compte le contexte des mots, permet de conserver les propriétés sémantiques et syntaxiques de la langue (Mikolov *et al.*, 2013a). Cette technique se fonde sur la construction d'un modèle sémantique qui projette les termes d'une langue dans un espace dans lequel certains liens sémantiques entre ces termes peuvent être observés et mesurés. Dans la littérature, plusieurs techniques ont été proposées pour construire des modèles vectoriels (voir Schwab (2005) pour un état de l'art plus complet) mais dernièrement, ces techniques sont principalement issues des modèles de langue neuronaux (Bengio *et al.*, 2003).

Parmi les plus utilisés, on peut citer word2vec (Mikolov *et al.*, 2013c) ou GloVe (Pennington *et al.*, 2014). Mikolov *et al.* (2013c) utilisent un réseau de neurones récurrent (Mikolov *et al.*, 2010) pour apprendre une représentation vectorielle des mots. Les poids du réseau formé sont utilisés comme des vecteurs pour représenter les mots. Ce modèle, appelé *"Word2Vec"*, est capable de capturer des régularités sémantiques et syntaxiques. Ainsi, les angles entre les projections des mots sont corrélés aux relations qui les relient, par exemple : singulier/pluriel, féminin/masculin ou pays/capitale. Grâce à cette observation, il est possible d'exploiter ces relations avec des opérations arithmétiques simples sur ces vecteurs. Il existe deux variantes de *"Word2Vec"*, la première appelée Sac-de-Mots Continue CBOW (*Continuous Bag-Of-Words*) (Mikolov *et al.*, 2013a), et la seconde, Skip-gram, (Mikolov *et al.*, 2013b). L'objectif de l'approche CBOW est de prédire un mot à partir de son contexte d'apparition en utilisant une fenêtre de mots. Soit une suite de mots $S = w_1, w_2, \ldots, w_i$, le modèle CBOW permet de prédire les mots w_k de leurs co-textes $(w_{k-l}, , w_{k-1}, w_{k+1}, , w_{k+l})$. En revanche, la deuxième approche Skip-gram permet de prédire, pour un mot donné, le co-texte dont il est issu.

3 Description du système

Dans les travaux de Mikolov *et al.* (2013c), les approches de Collobert & Weston (2008), Turian *et al.* (2010), Mnih & Hinton (2009) et Mikolov *et al.* (2013c) sont évaluées et comparées. Ils montrent ainsi que CBOW et Skip-gram sont beaucoup plus efficaces en termes de rapidité et de précision pour

apprendre un modèle que les techniques précédentes.

C'est d'ailleurs pour cette raison que nous avons adopté dans notre travail le modèle CBOW. Par conséquent, dans le but de mesurer la similarité sémantique entre les phrases en arabe, nous avons utilisé le modèle CBOW pour représenter les mots arabes[1] (Zahran *et al.*, 2015). Ce modèle a été appris à partir de différentes sources comme le Wikipedia Arabe (WikiAr, 2006), le corpus arabe de la BBC et de CNN (Saad & Ashour, 2010) et le corpus "open parallel" (Tiedemann, 2012) représentant quelques $5,8$ milliards de mots.

Les performances du modèle CBOW dépendent essentiellement de la tâche et de la bonne définition des paramètres d'apprentissage. Après quelques tests sur des corpus échantillons, nous avons choisi de conserver les paramètres suivant : 300 pour la taille des vecteurs, 5 pour la taille de la fenêtre de contexte, 10^{-5} pour le seuil de sous-échantillonnage des mots fréquents, 10 en exemples négatifs dans l'apprentissage et 100 pour le nombre de mots les moins fréquents à supprimer.

3.1 Similarité entre mots

L'utilisation d'un modèle de représentation vectorielle continue de mots permet d'observer et de mesurer les relations sémantiques entre ces mots. Dans ce sens, nous avons utilisé le modèle CBOW (Zahran *et al.*, 2015) afin d'identifier les relations sémantiques entre deux mots w_i et w_j comme par exemple la synonymie ou le changement de nombre et de genre. La similarité entre w_i et w_j est obtenue en comparant leur représentation vectorielle v_i et v_j en utilisant la similarité cosinus. Par exemple, soit les trois mots: "الجامعة" *(l'université)*,"المساء" (*le soir*) et "الكلية" (*la fac*), la similarité entre eux sera :

$$Sim(\text{المساء}, \text{الجامعة}) = Cos(v(\text{المساء}), v(\text{الجامعة})) = 0,13$$

$$Sim(\text{الكلية}, \text{الجامعة}) = Cos(v(\text{الجامعة}), v(\text{الكلية})) = 0,72$$

L'interprétation est que les mots "الكلية" (*la fac*) et "الجامعة" (*l'université*) sont sémantiquement plus proches que les mots "المساء" (*le soir*) et "الجامعة" *(l'université)*.

3.2 Similarité entre phrases

Soit deux phrases, $S_1 = w_1, w_2, ..., w_i$ et $S_2 = w'_1, w'_2, ..., w'_j$, leurs vecteurs étant respectivement $(v_1, v_2, ..., v_i)$ et $(v'_1, v'_2, ..., v'_j)$. Pour mesurer la similarité sémantique entre S_1 et S_2, nous effectuons la somme (pondérée) des vecteurs de leurs mots. Nous avons utilisé quatre pondérations différentes durant cette somme : une pondération unitaire où tous les vecteurs ont le même poids, une pondération des mots avec leur fréquence inverse en documents, une pondération des mots dépendant de leur étiquetage morpho-syntaxique et une pondération mixte, combinaison des deux précédentes. La figure 1 illustre un aperçu global de notre procédure de calcul de similarité entre deux phrases dans notre système.

[1] https://sites.google.com/site/mohazahran/data

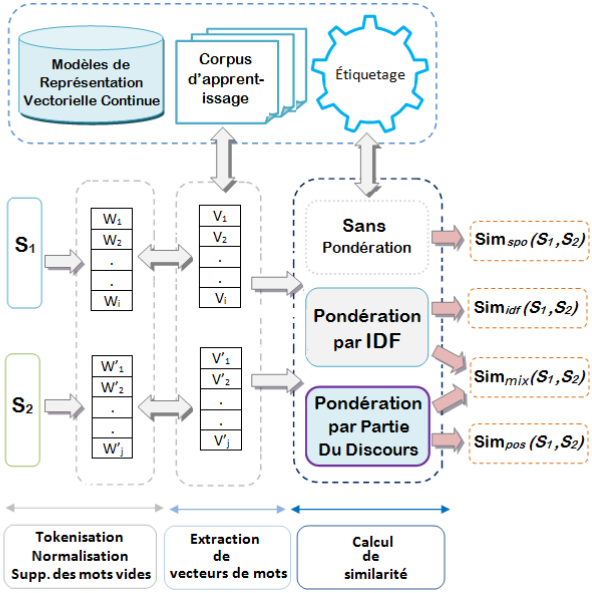

Figure 1: L'architecture du système proposé

4 Méthodes proposées

Dans ce qui suit, nous expliquons nos quatre méthodes proposées pour calculer la représentation vectorielle d'une phrase. Dans tous les cas, il s'agit d'une somme pondérée des vecteurs des mots de la phrase. Nous effectuons une somme de vecteurs de mots afin de prouver qu'une approche intuitive, optimisant une somme de vecteurs de mots en pondérant individuellement chacun des mots en fonction de leur importance dans la phrase, peut être tout aussi performante que des approches utilisant des réseaux de neurones calculant directement un plongement pour une phrase, un paragraphe ou un document, sensible elles à la tâche ou au contexte.

Nous comparons quatre pondérations, une pondération unitaire où tous les mots ont le même poids, une pondération basée sur la fréquence inverse en document des mots, une pondération basée sur la partie du discours des mots et une pondération basée sur la combinaison des deux pondérations précédentes.

Pondération unitaire Il s'agit de la méthode la plus naïve, celle généralement utilisée pour calculer la représentation vectorielle d'une phrase. Il s'agit de la somme des vecteurs des mots composants la phrase, soit $V = \sum_{k=1}^{i} v_k$ où v_k est le vecteur du kième mot de la phrase.

Cependant, nous pensons qu'il y a des mots plus importants que d'autres dans une phrase, des mots qui ont plus de chance d'être conservés même lors d'une paraphrase ou d'une reformulation, méritant donc ainsi un poids plus important dans la représentation de cette phrase. Nous faisons l'hypothèse que l'importance de ces mots peut être déterminée par leur rôle dans la phrase ou par leur rareté dans l'usage de leur langue.

Pondération avec la fréquence inverse en document La fréquence inverse en documents (IDF) est une mesure statistique qui permet d'évaluer l'importance d'un mot dans un corpus (Salton & Buckley, 1988). C'est une mesure très utilisée dans le domaine de la recherche d'information (Turney & Pantel, 2010). L'idée sous-jacente est que les mots les moins fréquents sont considérés comme plus discriminants. Notre pondération donne ainsi un poids plus important aux mots moins fréquents.

Ainsi, la représentation vectorielle d'une phrase devient alors $V = \sum_{k=1}^{i} idf(w_k) * v_k$ où v_k est le vecteur du $k^{\text{ième}}$ mot de la phrase et $idf(w)$ la fonction qui donne l'IDF du mot w.

Pour calculer les pondération IDF de chaque mot, nous avons utilisé le corpus arabe de la BBC et CNN[2] (Saad & Ashour, 2010) comme corpus d'apprentissage. La pondération IDF de chaque mot w_k est obtenue en appliquant la formule $idf(w_k) = log(\frac{S}{WS})$ où S représente le nombre total de phrases dans le corpus et WS est le nombre de phrases contenant le mot w_k.

Pondération avec l'étiquetage morpho-syntaxique Cette idée, inspirée de Schwab (2005) et plus récemment de Ferrero *et al.* (2017), consiste à donner un poids au vecteur d'un mot en fonction de son étiquette morpho-syntaxique.

Ainsi, la représentation vectorielle d'une phrase devient alors $V = \sum_{k=1}^{i} Pos\_weight(Pos_{w_k}) * v_k$ où v_k est le vecteur du $k^{\text{ième}}$ mot de la phrase et $Pos\_weight(Pos_w)$ est la fonction qui retourne le poids de la partie du discours du mot w.

Alors que Schwab (2005) et Ferrero *et al.* (2017) utilisent des approches supervisées, nous reprenons ici l'aproche de Lioma & Blanco (2009) qui est entièrement non-supervisée. Le principe est d'étiqueter morpho-syntaxiquement le corpus de test en utilisant l'analyseur morpho-syntaxique de Gahbiche-Braham *et al.* (2012), puis de calculer pour chacune des parties du discours sa fréquence inverse en documents et d'en déduire le poids à attribuer à chaque partie du discours. La pondération pos de chaque partie du discours pos_i est obtenue en appliquant la formule $idf(pos_i) = log(\frac{S}{PS})$ où S représente le nombre total de phrases dans le corpus et PS est le nombre de phrases contenant la partie du discours pos_i.

Pondération mixte Cette dernière pondération utilise à la fois la pondération IDF et la pondération en parties du discours. Ainsi, avec cette pondération, la représentation vectorielle d'une phrase devient alors $V = \sum_{k=1}^{i} idf(w_k) * Pos\_weight(Pos_{w_k}) * v_k$ où $Pos\_weight(Pos_w)$ est la fonction qui retourne le poids de la partie du discours correspondant au mot w et $idf(w)$ la fonction qui donne l'IDF du mot w.

5 Expériences et résultats

Principe Afin d'évaluer les performances de nos méthodes, nous avons utilisé quatre corpus de tests tirés de la sous-tâche 1 de la tâche 1 (détection de similarité textuelle sémantique sur des paires en langue arabe) de la campagne d'évaluation SemEval 2017 (Agirre *et al.*, 2017)[3], comprenant les trois corpus d'essai suivant :

[2] https://sourceforge.net/projects/ar-text-mining/files/Arabic-Corpora/

[3] http://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools

- *Microsoft Research Paraphrase Corpus* (MSR-Paraphrase) composé de 510 paires ;

- *Microsoft Research Video Description Corpus* (MSR-Video) composé de 368 paires ;

- *WMT2008 Development Dataset* (SMT-europarl) composé de 203 paires ;

ainsi que le corpus d'évaluation composé de 250 paires.

Dans chacun de ces corpus, cinq annotateurs ont donné un score entre 0 et 5 aux paires de phrases, un score de 0 indiquant que le sens des deux phrases est totalement différent et un score de 5 indiquant que les deux phrases ont exactement le même sens. Le score final de chaque paire est la moyenne des scores des cinq annotateurs.

Évaluation Les performances de nos quatre approches ont donc été évaluées sur plus de 1338 paires de phrases. Nous avons calculé la corrélation de Pearson entre les scores de similarité sémantique résultants de nos méthodes et les jugements humains. Les résultats sont présentés dans le Tableau 1.

Corpus	MSRpar	MSRvid	SMTeuroparl	STS 2017	Global
Pondération unitaire	0,6745	0,7233	0,6233	0,5957	0,6653
Pondération avec IDF	0,7432	0,7820	0,7110	0,7309	0,7467
Pondération avec POS	0,7446	0,7951	0,7317	0,7425	0,7562
Pondération mixte	**0,7523**	**0,8276**	**0,7460**	**0,7646**	**0,7745**
Baseline de SemEval 2017	/	/	/	0,6045	/
BIT (gagnant de SemEval 2017)	/	/	/	0,7543	/

Table 1: Performances de nos quatre méthodes ainsi que ceux des organisateurs et des vainqueurs de SemEval 2017.

Ces résultats montrent que sur chacun des corpus considéré, la pondération unitaire est nettement moins performante que les autres. La pondération par fréquence inverse en documents et celle basée sur les parties du discours obtiennent des résultats comparables (respectivement $+8,14\%$ et $+9,09\%$ par rapport à la pondération unitaire), même si la seconde est légèrement meilleure. Enfin, la combinaison de ces deux pondérations donne les meilleurs résultats ($+10,92\%$ vis-à-vis de la pondération unitaire).

En ce qui concerne la campagne d'évaluation SemEval 2017 (Agirre *et al.*, 2017), la dernière pondération serait arrivée en tête avec une corrélation de $0,7646$ contre 0.7543 pour l'équipe gagnante, BIT (Wu *et al.*, 2017). À l'heure où nous écrivons ces lignes, nous n'avons pas pu lire en détail ce travail mais nous savons qu'il s'agit d'une approche utilisant intensivement WordNet alors que notre méthode n'exploite pas de telles données et pourrait être utilisée avec des langues qui n'en disposent donc pas.

6 Conclusion

Dans cette article, nous avons présenté un système d'estimation de la similarité sémantique entre phrases en arabe basée sur l'utilisation de représentations continues de mots. Nous avons montré qu'identifier les mots les plus importants d'une phrase permettait d'améliorer fortement les perfor-

mances d'un tel système. En effet, une pondération des vecteurs des mots qu'elle se fasse en se basant sur la fréquence inverse en documents ou en fonction de la partie du discours est bien plus efficace que si la pondération est la même pour chaque mot. Les performances des systèmes non-supervisés proposés sont confirmées par la corrélation de Pearson avec des jugements humains sur quatre corpus différents. Une pondération combinant à la fois la fréquence inverse en documents et la parties du discours des mots aurait remporté la récente tâche 1 de la campagne d'évaluation SemEval 2017. Les pistes de travail que nous poursuivons actuellement cherchent à combiner ces résultats à ceux d'autres techniques classiques dans le domaine (fingerprint, n-gram, etc.) ou à ceux d'autres ressources (bases lexicales pour enrichir des phrases de synonymes, génériques, voire traductions dans d'autres langues).

Références

AGIRRE E., CER D., DIAB M., INIGO LOPEZ-GAZPI & SPECIA L. (2017). Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of Semeval 2017*.

BENGIO Y., DUCHARME R., VINCENT P. & JAUVIN C. (2003). A neural probabilistic language model. *Journal of machine learning research*, **3**(Feb), 1137–1155.

COLLOBERT R. & WESTON J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, p. 160–167: ACM.

FERRERO J., BESACIER L., SCHWAB D. & AGNÈS F. (2017). Using word embedding for cross-language plagiarism detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, p. 415–421, Valencia, Spain: Association for Computational Linguistics.

GAHBICHE-BRAHAM S., BONNEAU-MAYNARD H., LAVERGNE T. & YVON F. (2012). Joint segmentation and pos tagging for arabic using a crf-based classifier. In *LREC*, p. 2107–2113.

LIOMA C. & BLANCO R. (2009). Part of speech based term weighting for information retrieval. In *European Conference on Information Retrieval*, p. 412–423: Springer.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *In: ICLR: Proceeding of the International Conference on Learning Representations Workshop Track*, p. 1301–3781.

MIKOLOV T., KARAFIÁT M., BURGET L., CERNOCKỲ J. & KHUDANPUR S. (2010). Recurrent neural network based language model. In *Interspeech*, volume 2, p.3̃.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, p. 3111–3119.

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013c). Linguistic regularities in continuous space word representations. In *Hlt-naacl*, volume 13, p. 746–751.

MNIH A. & HINTON G. E. (2009). A scalable hierarchical distributed language model. In D. KOLLER, D. SCHUURMANS, Y. BENGIO & L. BOTTOU, Eds., *Advances in Neural Information Processing Systems 21*, p. 1081–1088. Curran Associates, Inc.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, p. 1532–1543.

SAAD M. K. & ASHOUR W. (2010). Osac: Open source arabic corpora. In *6th ArchEng Int. Symposiums, EEECS*, volume 10.

SALTON G. & BUCKLEY C. (1988). Term-weighting Approaches in Automatic Text Retrieval. In *Information Processing and Management*, volume 24, p. 513–523, Tarrytown, NY, USA: Pergamon Press, Inc.

SCHWAB D. (2005). *Approche hybride-lexicale et thématique-pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte*. PhD thesis, Université Montpellier II.

TIEDEMANN J. (2012). Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, p. 2214–2218.

TURIAN J., RATINOV L. & BENGIO Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, p. 384–394: Association for Computational Linguistics.

TURNEY P. D. & PANTEL P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, **37**, 141–188.

WIKIAR (2006). Arabic wikipedia corpus, http://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/, (accessed january 21,2017).

WU H., HUANG H., JIAN P., GUO Y. & SU C. (2017). Bit at semeval-2017 task 1: Using semantic information space to evaluate semantic tex- tual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*.

ZAHRAN M. A., MAGOODA A., MAHGOUB A. Y., RAAFAT H., RASHWAN M. & ATYIA A. (2015). Word representations in vector space and their applications for arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics*, p. 430–443: Springer.