Apprentissage de représentations de documents et leur exploitation en recherche d'information

Thiziri Belkacem, Taoufiq Dkaki, Jose G. Moreno, Mohand Boughanem

Apprentissage de représentations de documents et leur exploitation en recherche d'information

Thiziri BELKACEM[1] — Taofiq DKAKI[2] — Jose G. MORENO[1] — Mohand BOUGHANEM[1]

[1] *Laboratoire IRIT, Université Paul Sabatier Toulouse 3*
[2] *Laboratoir IRIT, Université Jean Jaurès Toulouse 2*
{thiziri.belkacem, mohand.boughanem, taoufiq.dkaki, jose.moreno}@irit.fr

RÉSUMÉ. *Afin de calculer la similarité document-requête, la majorité des modèles en recherche d'information (RI) représentent les documents et les requêtes sous forme de « sacs de mots » (bag of words) pondérés ou un sac de concepts, issus d'une ontologie linguistique ou construits automatiquement par des techniques de type LSI ou LDA, pour combler l'écart entre le vocabulaire utilisé par la requête et celui présenté dans les documents. D'autres approches dites word2vec proposent de modéliser les termes sous forme de vecteurs. Les approches word2vec permettent de capturer des relations au-delà de la co-occurrence, nous permettant ainsi de modéliser des relations sémantiques entre les termes. Dans cet article, nous présenterons l'état de l'art sur l'usage de ce type d'approches ainsi que notre contribution à l'exploitation de ce type d'approches dans les modèles de la RI.*

ABSTRACT. *In order to perform the document-query similarity, many information retrieval (IR) models represent documents and queries as sets of weighted key words, called « bag of words », or a bag of concepts derived from a linguistic ontology, or constructed automatically by LSI or LDA techniques, to fill the gap between the query vocabulary and the one used in the document. Recent approaches propose to model the term as an embedded vector, called word2vec approaches, allowing to capture relations beyond the co-occurrence by modelling semantic relations between the terms. In this article, we present the state of the art about this topic, as well as our contribution to integrate these approaches within IR models.*

MOTS-CLÉS : *Recherche d'information, apprentissage profond, word2vec, représentations sémantiques.*

KEYWORDS: *Information retrieval, deep learning, word2vec, semantic representations.*

1. Introduction

Pour répondre à la requête utilisateur, la majorité des modèles de RI représentent les documents et les requêtes sous forme de « sacs de mots » *(bag of words)* pondérés (Salton et Buckley, 1988) ou un sac de concepts issu d'une ontologie linguistique ou construits automatiquement par des techniques de type LSI *(Latent Semantic Indexing)* ou LDA *(Latent Dirichlet Allocation)*(Blei, 2012).

Bien que largement exploité en RI, ce type de représentation ne permet pas de capturer le sens véhiculé par les mots et les relations potentielles entre ces mots, qui sont exprimés plus précisément par la structure et l'ordre dans lequel les termes sont présentés. Afin de capturer des représentations fines, une des approches qui pourraient être explorées et qui connaît un engouement considérable, est l'exploitation d'apprentissage profond basé sur les réseaux de neurones. Ces réseaux, grâce à leur structuration multi-couches, sont capables de générer des représentations abstraites fines permettant de capturer la sémantique du contenu du document. En particulier dans le cadre du TAL (Mikolov *et al.*, 2013c), permettent d'apprendre des représentations complexes des mots et de capturer des relations plus complexes (autres que la co-occurrence) entre les termes, ce modèle dit *word2vec* permet de modéliser des relations sémantiques entre les termes via des expressions mathématiques.

Nous nous intéressons à l'exploitation de ces modèles pour apprendre les représentations des documents et des requêtes. Nous exploitons, en l'occurrence, une approche de type word2vec (Mikolov *et al.*, 2013c) pour apprendre ces représentations, dont l'objectif est de remédier au problème du sac de mots, qui ne tient pas compte de la sémantique. Nous nous intéressons particulièrement, à l'évaluation de l'impact des entrées à fournir au modèle *word2vec* (représentation du document et de la requête) et comment l'exploiter pour construire ces représentations. Nous étudions également différentes fonctions de similarité entre les représentations construites et comment adapter les modèles de RI existants tels que BM25, ML, LDA et autres, pour les représentations continues des mots, celles des documents et celles des requêtes.

2. Contexte et motivation

2.1. *Motivation*

Le plongement lexical (Vukotic *et al.*, 2015 ; Braud, 2015) des mots (word embeddings) (Mikolov *et al.*, 2013a ; Mikolov *et al.*, 2013b ; Mikolov *et al.*, 2013c) a été massivement exploités ces dernières années, notamment pour la tâche de recherche d'informations. La représentation dans un espace vectoriel des mots a permis de développer des méthodes qui prennent en compte la sémantique de ces mots. On trouvera dans (Mikolov *et al.*, 2013b) un modèle qui construit des représentations continues des mots en se basant sur leur contexte, en prenant en compte la fenêtre d'occurrence d'un mot pour construire une représentation permettant de déduire des relations sémantiques entre différents termes.

Notre travail est inspiré des approches de l'état de l'art, notamment celles qui exploitent les *words embeddings* afin d'améliorer les résultats de recherche des modèles de RI (Ai *et al.*, 2016b ; Ai *et al.*, 2016a ; Ganguly *et al.*, 2015 ; Zamani et Croft, 2016). La construction d'une représentation complexe pour un document et la prise en compte de la structure, permettent de résoudre le problème du sac de mots qui n'informe pas sur l'emplacement et la distribution des mots dans le texte. L'exploitation efficace des représentations fines des documents/requêtes qui sont construites par une projection dans un nouvel espace de représentation plus complexe, nous permettra de capturer des informations liées à la sémantique du contenu sans se référer à des ontologies de domaine. On peut trouver dans (Mikolov *et al.*, 2013b) une description des différentes informations et relations que ce type de représentation nous permet de déduire. De ce fait, l'exploitation de ces approches nous permettra de résoudre le problème d'absence de certains termes de la requête dans un document (qui est soulevé dans plusieurs des modèles classiques comme TF-IDF et LM) car elles nous permettent d'exploiter les termes sémantiquement liés à ceux de la requête, ce qui nous permettra d'améliorer les résultats de recherche.

D'après les travaux de l'état de l'art, la majorité des modèles exploitant le plongement lexical, proposent des combinaisons linéaires entre un modèle de RI classique et celui-ci basé sur l'approche word2vec(Ganguly *et al.*, 2015 ; Ai *et al.*, 2016a ; Vulić et Moens, 2015). Nous souhaitons mettre en œuvre une combinaison plus complexe des modèles de la RI et de l'approche word2vec, afin de combiner leurs performances, à cet effet, nous proposons d'exploiter les *word embeddings* afin capturer la relation qui existe entre les mots d'une requête et les termes d'un document (même si celui-ci ne contient pas tous les termes de la requête). Jusqu'à présent, cette relation reste binaire (le terme est soit présent soit absent dans le document), on trouvera dans (Ganguly *et al.*, 2015) une approche de représentation des termes du document permettant de capturer les termes similaires à ceux de la requête, mais ne prend pas compte de la relation entre les termes de la requête et ceux du document tout entier et la relation entre les termes à l'intérieure du document.

2.2. *Représentation continue des mots : « Plongement lexical »*

C'est un modèle pouvant être classé dans les applications de l'apprentissage profond en traitement automatique de la langue (TAL) (Deng et Yu, 2014). Proposé par Mikolov et al, dans (Mikolov *et al.*, 2013b) et repris par Goldberg dans (Goldberg et Levy, 2014), ce modèle permet de construire des représentations continues des mots d'un texte ou d'une séquence, en se basant sur la notion de contexte auquel appartient le mot. Il permet de construire des représentations vectorielles permettant de capturer la sémantique exprimée par chacune des occurrences. Par la suite, ces représentations sont utilisées pour comparer les termes entre eux et exprimer, sous forme de distances, les relations potentielles entre les termes. Deux modèles ont été proposés, on trouvera dans la figure 1 la description de chacun des deux modèles.

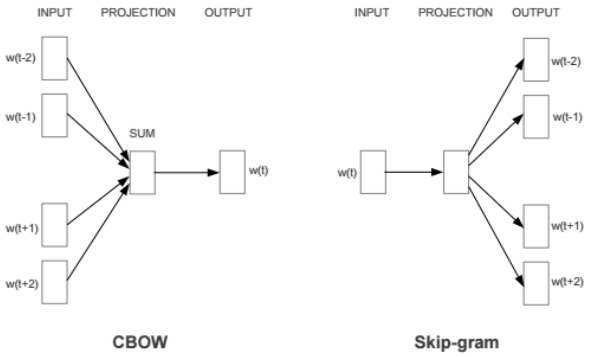

Figure 1 – Description des modèles CBOW et Skip-gram tel décrit dans (Mikolov *et al.*, 2013b)

1) *Le modèle CBOW (Continuous Bag Of Word)* : a pour objectif d'obtenir la représentation appropriée d'un mot, en se basant sur une fenêtre de mots adjacents, afin de retrouver le prochain mots du même contexte.

2) *Le modèle Skip-Gram* : l'entraînement du modèle à pour objectif d'obtenir le vecteur du mot permettant d'exprimer le même sens que le mot en cours, ie : la représentation qui permet de déduire son contexte.

Formellement, le modèle a pour objectif, pour une séquence de mots d'entraînement $w_1, w_2, ...w_T$ de maximiser la moyenne des $logs$ attribués aux fenêtres de mots k :

$$\frac{1}{T}\sum_{t=1}^{T}\left[\sum_{j=-k}^{k} logP(w_{t+j}|w_t)\right]_{(j\neq 0)} \quad [1]$$

Tel qu'à chaque mot, sont associés deux vecteurs à apprendre : U_w (vecteur d'entrée : *input*) et V_w (vecteur de sortie : *output*), puis calculer ainsi la probabilité reliant chaque mot aux mots du vocabulaire par l'équation 2.

$$P(w_i|w_j) = \frac{exp\left(U_{w_i}^T V_{w_j}\right)}{\sum_{l=1}^{V} exp\left(U_l^T V_{w_j}\right)} \quad [2]$$

Où V est la taille du vocabulaire.

3. Travaux liés

Les approches *« deep learning »* sont largement exploitées dans le traitement d'images, notamment dans la reconnaissance d'objets (Socher *et al.*, 2012b) et la classification d'images (Ciregan *et al.*, 2012). Grâce à leur efficacité, elles sont aussi bien exploitées dans la reconnaissance de la parole (Graves *et al.*, 2013) qu'en traitement

du langage naturel (TAL) (Socher *et al.*, 2012a). Dernièrement, ces techniques ont été aussi utilisées en RI, pour améliorer les résultats de recherche. Elles sont exploitées notamment dans le processus de calcul de similarité et dans la construction des représentations pour les documents et les requêtes. Ces travaux, dits *Neural Language Modeling*, partagent la caractéristique de représentation vectorielle des mots (Mikolov *et al.*, 2013b), des séquences (Severyn et Moschitti, 2015), des paragraphes et du document tout entier (Le et Mikolov, 2014).

On trouvera dans (Mikolov *et al.*, 2013b) une description de l'implémentation du *word2vec*[1], exploité dernièrement par plusieurs auteurs, pour la construction des représentations continues des mots, ce modèle est décrit plus en détail dans la section (2.2). Dans (Mikolov *et al.*, 2013c), le modèle se base sur un réseau de neurones récurrent pour apprendre les relations sémantiques entre les mots en utilisant leurs représentations continues *(word embeddings)*, permettant de déduire des relations potentielles entre les mots du même contexte.

Dans (Le et Mikolov, 2014), les auteurs proposent le modèle PV permettant de construire des représentations continues des mots, dans une séquence ou d'un document tout entier. PV est basé sur l'approche de représentation des mots de (Mikolov *et al.*, 2013b) pour construire les représentations des documents complets. Dans (Ai *et al.*, 2016b), les auteurs exploitent le modèle de (Le et Mikolov, 2014) pour améliorer les résultats du modèle QL de(Ponte et Croft, 1998) en utilisant la combinaison linéaire entre les deux types de similarité (celle calculée par le modèle QL et celle du modèle PV de Mikolov), puis dans (Ai *et al.*, 2016a) ils ont analysé les limites de ce modèle dans le cadre de la tache de la RI. Ces limites se situent principalement dans la taille du document qui entraîne le sur-apprentissage des documents courts. Pour remédier à cette limite, Ai et al ont proposé une normalisation par rapport à la taille du document et une extension du modèle afin de prédire le contexte d'un terme.

On trouvera dans (Nalisnick *et al.*, 2016) une autre manière d'exploiter le modèle *word2vec* de Mikolov dans la tache de RI, dans ce papier les auteurs utilisent les deux espaces de projection *input* et *output* produits par le modèle *word2vec* et montrent que la projection des requêtes et des documents sur les espaces input et ouput, respectivement, donne de meilleurs résultats que l'exploitation d'un espace de projection unique. Dans (Ganguly *et al.*, 2015), les auteurs ont analysé l'absence de certain termes de la requête dans le document et l'expression de leur topique autrement ; ce qu'ils ont appelé « mutation des termes ». Ils proposent une extension du modèle de langue généralisé (Hiemstra, 2001 ; Ponte et Croft, 1998 ; Zhai et Lafferty, 2004) avec le modèle word2vec de Mikolov et proposent un modèle générique permettant d'exploiter les termes similaires à la requête pour augmenter la probabilité qu'un document soit pertinent, lorsqu'il traite du sujet de la requête, sans se restreindre aux documents qui contiennent que les termes de la requête.

L'exploitation des modèles continues ou plongement lexical, afin de représenter le contenu d'un document ou d'une requête, implique l'exploitation d'une fonction

1. https://code.google.com/archive/p/word2vec/

de combinaison des vecteurs des différents termes pour construire le vecteur global du document ou de la requête, la majorité des travaux cités ci-dessus exploitent l'approche de sommation dite *AWE* (Average Word Embedding) qui consiste en la sommation ou la moyenne des vecteurs des termes du document ou de la requête. Dans (Zamani et Croft, 2016), les auteurs proposent une nouvelle méthode de combinaison des vecteurs des termes en se basant sur le modèle de la requête. Les auteurs ont proposé une construction du vecteur du document et celui de la requête en exploitant des transformations softmax ou sigmoïde de la mesure de cosinus qui est utilisée généralement pour le calcul de similarité vectorielle, ils ont montré que l'approche *AWE* n'est qu'un cas particulier du modèle proposé.

4. Modèle continue pour la RI

Le modèle BM25 (Robertson et Walker, 1994 ; Robertson *et al.*, 1995) est l'un des modèles de RI qui ont donné de très bons résultats de recherche, en se basant sur la représentation du contenu de document sous forme de sac de mots. Étant donné que ce type de représentation suppose que les termes sont indépendants les uns des autres, des relations potentielles pouvant exister entre les termes nous permettra de déduire ou d'identifier des concepts communs et des relations sémantiques entre la requête et le document.

4.1. *Extension de la BM25 avec les représentations continues*

La classe des modèles probabilistes (Probabilistic Relevance Framework PRF) est un cadre formel pour le tri des documents basé sur un travail effectué dans les années 1970-1980 (Robertson et Jones, 1976). Ce cadre a conduit au développement de l'un des algorithmes de RI les plus réussis : BM25. On trouvera dans (Robertson et Walker, 1994 ; Robertson *et al.*, 1995) la description, le développement et les résultats obtenus par ce modèle. On trouvera aussi dans (Singhal *et al.*, 1996) le paramétrage de la BM25 et sa normalisation afin de prendre en compte la taille d'un document.

Soit la forme suivante de la BM25 telle décrite dans (Lv et Zhai, 2011) : Le document D et la requête Q étant vus comme sacs de mots pondérés, le score de pertinence du document vis à vis de la requête est calculé par :

$$score(D,Q) = \sum_{q_i \in Q} unit_{score}(D, q_i) \quad [3]$$

Avec :

$$unit_{score}(D, q_i) = IDF(q_i).\frac{f(q_i, D).(k_1+1)}{f(q_i, D) + k_1.\left(1 - b + b.\frac{|D|}{avgdl}\right)} \quad [4]$$

et

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \quad [5]$$

Dans cette approche, le document et la requête sont représentés dans l'espace des termes V. Dans cet espace, les termes sont considérés comme des entités indépendantes. Les différentes relations sémantiques qui peuvent exister entre les termes ne sont donc pas prises en compte. En se plaçant dans un autre espace que celui-ci, l'espace $V \times V$ où des relations entre les différents termes du vocabulaire peuvent être modélisées par une fonction de similarité prédéfinie, nous proposons une adaptation de ce modèle à la nouvelle représentation des termes et des documents. Nous pouvons calculer le score de similarité entre un document D et une requête Q par l'équation 6 :

$$score^*(D,Q) = \sum_{d_j \in D} \sum_{q_i \in Q} unit_{score}(D,d_j) \times [term_{sim}(d_j,q_i)]^{\alpha} \quad [6]$$

Avec $unit_{score}$ peut être calculé par la formule 4, le paramètre α est utilisé pour diminuer l'influence de la similarité entre les termes de la requête et ceux du document sur le calcul de score et $term_{sim}(.,.)$ est une fonction qui calcule la similarité entre les termes et pouvant être définis de la manière suivante :

$$term_{sim}(d_j,q_i) = cos(\overrightarrow{d_j}, \overrightarrow{q_i}) \quad [7]$$

Où : $\overrightarrow{d_j}$ et $\overrightarrow{q_i}$ représentent le vecteur du terme d_j du document et du terme q_i de la requête respectivement.

5. Expérimentation et résultats

5.1. *Protocole d'expérimentation*

Afin d'évaluer l'approche proposée, nous avons opté pour des expérimentations en utilisant une collection de documents TREC (AP88 disk-2) constituée de 79919 documents et les requêtes de tests correspondantes (251-300). Afin de construire le nouvel espace de représentation des documents et des requêtes, nous avons entraîné le modèle word2vec de Mikolov (Mikolov *et al.*, 2013a) utilisant la librairie *gensim*(Rehurek et Sojka, 2010). Nous avons entraîné le word2vec en utilisant le modèle *CBOW* et une fenêtre de contexte de 5 mots pour produire l'espace de projection constitué des vecteurs de dimension 300 après 15 itérations. Dans un premier temps nous avons transformé la collection en un corpus de phrases (*corpus sentences*) qui sont tokenizées, en utilisant la librairie *NLKT*[2] de gensim, par la suite nous avons éliminé les mots vides en utilisant les *stopwords* de cette librairie.

Le nouvel espace de représentation est constitué alors d'un vocabulaire de 67469 termes uniques. Chaque requête est projetée dans le nouvel espace de représentation afin de construire sa nouvelle représentation, qui sera appariée avec celles des documents, selon le modèle exploité. Les premiers résultats obtenus (section 5.2) par le modèle proposé sont encourageant et montrent que l'approche pourra aboutir à une amélioration de l'état de l'art.

2. http ://www.nltk.org/

Tableau 1 – Résultats préliminaires obtenus par le modèle d'extension de la BM25 avec les représentations continues

	MAP	P@10
TF-IDF	0.112	0.158
TF-IDF_vec $\alpha = 7$	**0.136**	**0.176**
BM25	0.167	0.222
BM25_vec $\alpha = 7$	**0.179**	**0.230**
BM25*	0.195	**0.266**
BM25_vec* $\alpha = 6$	**0.197**	0.262

5.2. *Résultats et discussion*

Les premiers résultats obtenus par l'approche proposée sont représentés dans le tableau 1. Afin d'évaluer la performance de notre approche, nous comparons les résultats obtenus par notre modèle à ceux des modèles classiques de la recherche d'information, notamment la $BM25$ et l'approche $tf - idf$ en utilisant le même corpus de test. Les résultats mis en caractère gras sont les meilleurs, les lignes marquées par l'astérisque (*) montrent les résultats obtenus en utilisant les requêtes complètes. Le paramètre α représente l'ordre d'atténuation du paramètre de similarité $term_{sim}$ dans l'équation 6, les valeurs de ce paramètre ont été fixées après un ensemble de tests. Celles rapportées dans ce tableau sont les valeurs ayant donné les meilleurs résultats.

D'après les résultats du tableau 1, nous pouvons remarquer que l'approche proposée donne de bons résultats sur la collection de test utilisée ; ce qui montre l'apport de l'exploitation des représentations continues, car elles permettent de capturer des caractéristiques sémantiques, nous permettant de prendre en considération les termes sémantiquement similaires ou du même contexte que ceux de la requête. Cette approche nous a permis d'obtenir une amélioration de 7.1%, ce qui est encourageant pour étudier l'approche et effectuer plus de tests.

6. Conclusion

Notre contribution est une extension de la BM25 pour la recherche d'information en exploitant le plongement lexical des terme. L'approche que nous proposons permet d'exploiter les relations potentielles entre les termes en se basant sur la notion de contexte et nous a permis d'aboutir à une amélioration par rapport aux modèles de l'état de l'art. Le travail à venir va se focaliser sur la validation des résultats du modèle proposé. Pour ce faire, nous allons effectuer d'autres tests sur des collections plus larges, analyser et évaluer l'impacte de la combinaisons des vecteurs des termes afin

de construire la représentation du document tout entier ainsi que l'exploitation de ce type de représentations avec les modèles de l'état de l'art en RI.

7. Bibliographie

Ai Q., Yang L., Guo J., Croft W. B., « Analysis of the Paragraph Vector Model for Information Retrieval », *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, ACM, New York, NY, USA, p. 133-142, 2016a.

Ai Q., Yang L., Guo J., Croft W. B., « Improving Language Estimation with the Paragraph Vector Model for Ad-hoc Retrieval », *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, ACM, New York, NY, USA, p. 869-872, 2016b.

Blei D. M., « Probabilistic Topic Models », *Commun. ACM*, vol. 55, n° 4, p. 77-84, April, 2012.

Braud C., Automatically Identifying Implicit Discourse Relations using Annotated Corpora and Raw Data, Theses, Universite Paris Diderot-Paris VII, December, 2015.

Ciregan D., Meier U., Schmidhuber J., « Multi-column deep neural networks for image classification », *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, p. 3642-3649, 2012.

Deng L., Yu D., « Deep Learning », *Signal Processing*, vol. 7, p. 3-4, 2014.

Ganguly D., Roy D., Mitra M., Jones G. J., « Word embedding based generalized language model for information retrieval », *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, p. 795-798, 2015.

Goldberg Y., Levy O., « word2vec Explained : deriving Mikolov et al.'s negative-sampling word-embedding method », *arXiv preprint arXiv :1402.3722*, 2014.

Graves A., Mohamed A.-r., Hinton G., « Speech recognition with deep recurrent neural networks », *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, p. 6645-6649, 2013.

Hiemstra D., *Using language models for information retrieval*, Taaluitgeverij Neslia Paniculata, 2001.

Le Q. V., Mikolov T., « Distributed Representations of Sentences and Documents. », *ICML*, vol. 14, p. 1188-1196, 2014.

Lv Y., Zhai C., « Lower-bounding term frequency normalization », *Proceedings of the 20th ACM international conference on Information and knowledge management*, ACM, p. 7-16, 2011.

Mikolov T., Chen K., Corrado G., Dean J., « Efficient estimation of word representations in vector space », *arXiv preprint arXiv :1301.3781*, 2013a.

Mikolov T., Le Q. V., Sutskever I., « Exploiting similarities among languages for machine translation », *arXiv preprint arXiv :1309.4168*, 2013b.

Mikolov T., Yih W.-t., Zweig G., « Linguistic Regularities in Continuous Space Word Representations. », *HLT-NAACL*, vol. 13, p. 746-751, 2013c.

Nalisnick E., Mitra B., Craswell N., Caruana R., « Improving Document Ranking with Dual Word Embeddings », *Proceedings of the 25th International Conference Companion on*

World Wide Web, WWW '16 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, p. 83-84, 2016.

Ponte J. M., Croft W. B., « A language modeling approach to information retrieval », *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, p. 275-281, 1998.

Rehurek R., Sojka P., « Software framework for topic modelling with large corpora », *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Citeseer, 2010.

Robertson S. E., Jones K. S., « Relevance weighting of search terms », *Journal of the American Society for Information science*, vol. 27, n° 3, p. 129-146, 1976.

Robertson S. E., Walker S., « Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval », *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Springer-Verlag New York, Inc., p. 232-241, 1994.

Robertson S. E., Walker S., Jones S., Hancock-Beaulieu M. M., Gatford M. *et al.*, « Okapi at TREC-3 », *NIST SPECIAL PUBLICATION SP*, vol. 109, p. 109, 1995.

Salton G., Buckley C., « Term-weighting approaches in automatic text retrieval », *Information processing & management*, vol. 24, n° 5, p. 513-523, 1988.

Severyn A., Moschitti A., « Learning to rank short text pairs with convolutional deep neural networks », *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, p. 373-382, 2015.

Singhal A., Buckley C., Mitra M., « Pivoted document length normalization », *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, p. 21-29, 1996.

Socher R., Bengio Y., Manning C. D., « Deep learning for NLP (without magic) », *Tutorial Abstracts of ACL 2012*, Association for Computational Linguistics, p. 5-5, 2012a.

Socher R., Huval B., Bath B. P., Manning C. D., Ng A. Y., « Convolutional-Recursive Deep Learning for 3D Object Classification. », *NIPS*, vol. 3, p. 8, 2012b.

Vukotic V., Claveau V., Raymond C., « IRISA at DeFT 2015 : Supervised and Unsupervised Methods in Sentiment Analysis », *DeFT, Défi Fouille de Texte, joint à la conférence TALN 2015*, Actes de l'atelier DeFT, Défi Fouille de Texte, joint à la conférence TALN 2015, Caen, France, June, 2015.

Vulić I., Moens M.-F., « Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings », *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, p. 363-372, 2015.

Zamani H., Croft W. B., « Estimating Embedding Vectors for Queries », *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, ACM, New York, NY, USA, p. 123-132, 2016.

Zhai C., Lafferty J., « A study of smoothing methods for language models applied to information retrieval », *ACM Transactions on Information Systems (TOIS)*, vol. 22, n° 2, p. 179-214, 2004.