Apprentissage de représentation des documents médicaux guidé par les concepts pour la recherche d'information

Gia-Hung Nguyen, Lynda Tamine, Laure Soulier, Nathalie Bricon-Souf

Apprentissage de représentation des documents médicaux guidé par les concepts pour la recherche d'information

Gia-Hung Nguyen*, Lynda Tamine*
Laure Soulier** Nathalie Souf*

*Université de Toulouse, UPS-IRIT, 118 route de Narbonne, 31062 Toulouse, France
{gia-hung.nguyen, tamine, nathalie.souf}@irit.fr,
**Sorbonne Universités-UPMC, Univ Paris 06, LIP6 UMR 7606, 75005 Paris, France
laure.soulier@lip6.fr

Résumé. De nombreuses tâches médicales, telles que l'auto-diagnostic et le recrutement de patients pour des essais cliniques, impliquent l'utilisation d'outils d'accès à l'information. Une étape clé de ces tâches repose sur l'appariement entre documents dont l'enjeu majeur est de réduire le fossé sémantique entre les représentations de l'information brute des documents et l'interprétation issue d'experts médicaux. Guidés par les avancées récentes sur les approches neuronales, nous proposons un modèle d'apprentissage de représentation des documents permettant de capturer à la fois la sémantique des concepts médicaux validés dans une ressource externe et la sémantique issue des mots des documents. Notre modèle est évalué expérimentalement au travers de deux tâches de recherche d'information (RI) médicale différentes qui sont en l'occurrence la "RI liée à aide au diagnostic" et la "recherche clinique de cohortes".

1 Introduction

Dans le domaine de la recherche d'information (RI) médicale, le fossé sémantique désigne la différence des représentations de bas niveau des documents et l'interprétation de haut niveau de leurs contenus telle que perçue par l'expert humain. Le fossé sémantique est sous-jacent à trois enjeux fondamentaux (Edinger et al., 2012; Koopman et al., 2016) : 1) l'incompatibilité du vocabulaire lorsque deux textes expriment des sémantiques similaires sans partager les mêmes mots (e.g., *Melanome* vs. *cancer de la peau*) ; 2) l'incompatibilité de granularité liée à la généricité/spécificité des instances du texte (e.g., *anti-inflammatoire* vs *Neodex*) et 3) l'implication logique lorsque les textes comparés contiennent des indices permettant d'en déduire des implications ne pouvant pas être évaluées automatiquement (e.g., *anorexia* et *depression*). A ce jour, ce problème a été abordé en RI selon deux principales approches. La première approche consiste en l'utilisation de ressources externes, principalement pour améliorer les représentations des textes grâce à l'expansion de documents ou de requêtes (Lu et al., 2009; Dinh et Tamine, 2011). La seconde approche est basée sur des travaux récents en apprentissage profond (*deep learning*) qui exploite l'hypothèse de la sémantique distributionnelle pour l'apprentissage de représentations de documents et/ou de requêtes. Le principe de cette approche

repose sur la projection des mots dans un espace sémantique latent à partir de leur contexte (Mikolov et al., 2013).

Dans cet article, nous abordons la problématique de la représentation du document médical qui constitue une étape critique dans le processus d'appariement. Pour faire face aux problèmes de différence de vocabulaire et de granularité, nous proposons d'exploiter les approches neuronales permettant de capturer les représentations latentes de documents en intégrant les connaissances liées aux concepts médicaux spécifiés dans une ressource externe. De plus, afin de résoudre les limites liées à l'extraction du vocabulaire conceptuel (Trieschnigg, 2010) ainsi que de l'absence d'implications logiques au sein de la sémantique distributionnelle (Iacobacci et al., 2015), l'objectif de notre modèle est d'obtenir une représentation optimale des documents grâce à un raffinement utilisant à la fois des représentations brutes basées sur des concepts et celles basées sur des mots clés. Ainsi, les principales contributions de ce travail sont : 1) un modèle pour l'apprentissage de représentation des documents basés sur les réseaux de neurones et intégrant la sémantique issue d'une base de connaissances médicales ; 2) l'évaluation de l'efficacité de notre modèle sur deux tâches médicales différentes issues des campagnes d'évaluation TREC [1] : a) la recherche sur la santé en utilisant un corpus des articles scientifiques et b) la recherche clinique de cohortes utilisant un corpus de comptes rendus d'hospitalisation.

L'organisation de cet article est la suivante : la Section 2 présente une synthèse des travaux de l'état de l'art en lien avec notre contribution. Le modèle d'apprentissage de représentation guidé par les concepts est détaillé dans la Section 3. La section 4 présente l'évaluation expérimentale. Enfin, nous concluons dans la Section 5 puis présentons quelques perspectives de recherche.

2 Prise en compte de la sémantique pour la RI médicale

Dans le domaine de la RI médicale, la prise en compte de la sémantique est prépondérante du fait de la grande variabilité du langage et de l'orthographe, l'utilisation fréquente d'acronymes et d'abréviations, ainsi que l'ambiguïté inhérente liées à l'interprétation des concepts selon les contextes (Trieschnigg, 2010; Koopman et al., 2016). Nous détaillons dans ce qui suit les deux lignes de travaux qui traitent du problème du fossé sémantique, l'un des facteurs critiques de l'efficacité des système de RI médicaux.

RI augmentée par les ressources externes. De nombreux travaux exploitant les ressources externes (e.g., MeSH, UMLS et SNOMED) ont été proposés pour améliorer la sémantique des textes, des documents ou des requêtes. Les principales approches s'articulent autour de l'expansion de requêtes (Lu et al., 2009; Pal et al., 2014) et/ou l'expansion de documents (Dinh et Tamine, 2011). Par exemple, Lu et al. (2009) proposent d'étendre les requêtes utilisateurs par des termes issus de la ressource MeSH en exploitant le service de *mapping* de Pubmed (ATM). Bien que l'expansion basée sur les concepts conduit à des améliorations significatives, des travaux montrent qu'elle peut cependant être améliorée par une combinaison d'approches basées sur les mots-clés (Trieschnigg, 2010) permettant de remédier à l'expressivité limitée des concepts et/ou l'imprécision de la méthode d'extraction de concepts. Une nouvelle lignée de travaux concerne l'enrichissement des modèles d'appariement entre textes à l'aide de données provenant de ressources externes (Koopman et al., 2016; Wang et Akella, 2015). Ces contributions partagent le même objectif : générer des inférences sur les associations entre les

1. Text Retrieval Conference (http ://trec.nist.gov/)

mots du document et les concepts issus de la ressource UMLS puis intégrer les degrés de ces associations dans le calcul de pertinence des documents.

Apprentissage de représentation à partir de ressources externes. Une des principales approches d'apprentissage de représentation repose sur les modèles Skip-gram et CBOW (Mikolov et al., 2013). Basés sur l'hypothèse de la sémantique distributionnelle qui guide la représentation d'un mot en fonction de son contexte, ces modèles ont été étendus pour représenter les documents (Le et Mikolov, 2014) et les concepts issus de ressources externes (Ni et al., 2016). Au-delà, plusieurs travaux se concentrent sur l'utilisation additionnelle des ressources externes afin d'intégrer les concepts et leurs relations dans la représentation latente des mots (Faruqui et al., 2015; Xu et al., 2014). Par exemple, Faruqui et al. (2015) propose une technique de lissage de représentation basée sur l'information relationnelle dérivée des ressources externes. L'intuition sous-jacente à cette approche est que des concepts adjacents dans une ressource doivent avoir des représentations distributionnelles similaires. Dans le domaine médical, un nombre croissant de travaux s'intéressent à l'apprentissage de représentation des concepts (De Vine et al., 2014; Choi et al., 2016; Liu et al., 2016), dont certains d'entre eux (De Vine et al., 2014; Liu et al., 2016) pour une tâche de RI. Par exemple, Liu et al. (2016) s'appuient sur les travaux de Yu et Dredze (2014) pour intégrer la prise en compte des relations dans la représentation du document. Le score de pertinence est ensuite calculé par une combinaison linéaire des scores de pertinence des documents obtenus sur la base des représentations uniquement basées sur les mots (*bag-of-words*) avec ceux obtenus à partir des représentations latentes intégrant la sémantique relationnelle. A un niveau de granularité plus élevé, Choi et al. (2016) proposent le modèle Med2vec qui exploite la séquence des comptes-rendus des visites des patients pour apprendre une représentation latente des concepts terminologiques et des visites.

A la différence de la plupart des travaux antérieurs (De Vine et al., 2014; Yu et Dredze, 2014; Liu et al., 2016; Choi et al., 2016) qui apprennent des représentations de concepts sans contexte, notre objectif ici est d'apprendre simultanément les représentations des documents en intégrant à la fois de la sémantique distributionnelle exprimée dans un corpus de texte et la sémantique conceptuelle exprimée dans une ressource externe. Plus spécifiquement, à la différence de la contribution décrite dans (Choi et al., 2016), nous n'inférons pas les dépendances temporelles entre les documents (à savoir la séquence des comptes-rendus de visites) et les concepts, mais nous abordons plutôt une tâche d'accès à l'information dans la mesure où notre modèle exploite la sémantique basée sur le corpus pour faire face au problème d'inférence implicite logique entre les mots qui sera exploitée dans l'appariement requête-document.

3 Modèle de représentation guidé par les concepts

3.1 Formalisation du problème

L'objectif de notre approche est d'améliorer les représentations de documents médicaux pour les tâches de RI médicale (e.g., l'expansion de requête) en combinant les connaissances sémantiques médicales et les informations textuelles brutes. Notre modèle est guidé par les intuitions suivantes : 1) la prise en compte des concepts dans le processus d'apprentissage de représentation des documents, en plus des mots, devrait permettre de construire des représentations de documents sémantiques qui remédient aux limites des processus d'extraction de concepts ; 2) la représentation optimale d'un document dans un espace latent de faible di-

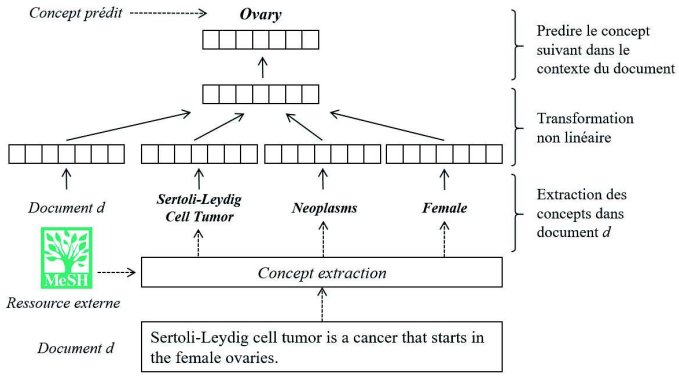

FIG. 1: Architecture du modèle *conceptualDoc2vec*

mension nécessite la proximité des représentations issues de façon indépendante des sources d'évidence basées sur la ressource et celles basées sur les corpus de documents.

Formellement, chaque document d est modélisé sous la forme de deux éléments : $d = \{\mathcal{W}_d, \mathcal{C}_d\}$, où $\mathcal{W}_d$ et $\mathcal{C}_d$ représentent respectivement les ensembles de mots ordonnés w_i et de concepts ordonnés c_j dans le document d, à savoir $\{w_1, ..., w_i, ..., w_n\}$ et $\{c_1, ..., c_j, ..., c_m\}$. En exploitant le modèle *Paragraph Vector* (PV-DM) (Le et Mikolov, 2014), le document d est représenté par un vecteur latent $\hat{d}^{(PV-DM)}$. Notre premier objectif consiste à construire la représentation du document d dans l'espace des concepts $\hat{d}_i^{(cd2vec)}$ qui capture la sémantique explicite exprimée dans une ressource externe. Pour cela, nous proposons le modèle *cd2vec (conceptualDoc2vec)* (Section 3.2). Notre second objectif est d'optimiser la représentation du document d afin d'obtenir un vecteur latent $\hat{d}$ permettant de rapprocher les représentations basées sur les concepts $\hat{d}_i^{(cd2vec)}$ et celles sur le texte brut $\hat{d}^{(PV-DM)}$. Ce problème peut être formulé par la fonction objectif suivante :

$$\Psi(D) = \sum_{d \in D} \psi(d) = \sum_{d \in D} \left[(1-\beta) \times \|d - \hat{d}^{(cd2vec)}\|^2 + \beta \times \|d - \hat{d}^{(PV-DM)}\|^2 \right] \quad (1)$$

où D est la collection de documents, $||x - y||$ la distance euclidienne entre les vecteurs de représentation x et y, et β correspond au coefficient de pondération, défini expérimentalement.

3.2 Apprentissage de représentation conceptuelle des documents

Guidé par le modèle *Paragraph Vector* (PV-DM) (Le et Mikolov, 2014) qui apprend la représentation de textes courts à partir du texte brut, nous proposons le modèle *conceptualDoc2vec* qui produit la représentation sémantique distributionnelle des concepts sous-jacents au texte. De façon similaire au modèle PV-DM, notre modèle d'apprentissage de représentation conceptuelle des documents *conceptualDoc2vec* repose sur un objectif de prédiction de concept à partir d'un contexte, permettant ainsi d'apprendre la représentation des concepts et du document $\hat{d}^{(cd2vec)}$. L'architecture de notre modèle incluant une illustration *conceptualDoc2vec* est illustrée dans la Figure 1. Plus formellement, le modèle *conceptualDoc2vec* a pour objectif de maximiser la log-vraisemblance suivante :

$$\varphi = \sum_{c_j \in \mathcal{C}_d} logP(c_j \mid c_{j-W} : c_{j+W}, d) \quad (2)$$

où c_j est j^e concept de l'ensemble de concepts ordonnés $\mathcal{C}_d$. W exprime la taille de la fenêtre de contexte, $c_{(j-W)} : c_{(j+W)}$ représente l'ensemble de concepts de positions comprises entre $j - W$ et $j + W$ dans le document d, sans inclure le concept c_j.

La probabilité $P(c_j \mid c_{j-W} : c_{j+W}, d)$ est définie par une fonction soft-max :

$$P(c_j \mid c_{j-W} : c_{j+W}, d) = \frac{exp((\overline{v}_j^W)^\intercal . v_{c_j})}{\sum_{c_k \in \mathcal{C}_d} exp((\overline{v}_k^W)^\intercal . v_{c_k})} \quad (3)$$

où v_{c_j} est la représentation du concept c_j, et $\overline{v}_j^W$ correspond à la moyenne des représentations des concepts dans la fenêtre $[j - W; j + W]$, incluant le document d.

Nous résolvons ensuite, à l'aide de la méthode de descente de gradient stochastique, le problème d'optimisation (Équation 1) qui infère la représentation optimale des documents d afin de rapprocher la représentation apprise sur du texte brut et celle apprise sur la base des concepts.

4 Evaluation expérimentale

4.1 Protocole d'évaluation

Tâches. Nous évaluons les représentations des documents sur deux tâches de RI médicale : 1) **Tâche 1 : RI liée à l'aide au diagnostic :** Cette tâche fait référence à la situation où le médecin souhaite accéder à des articles scientifiques comme aide à l'établissement d'un diagnostic/pronostic ou la suggestion d'un traitement en réponse à un cas médical. Nous utilisons la collection standard OHSUMED constituée de 348566 documents et de 63 requêtes MEDLINE. Cet ensemble de données constitue une collection standard utilisée à grande échelle pour la RI médicale (Stokes et al., 2009). Un exemple de requête est "*adult respiratory distress syndrome*" ; 2) **Tâche 2 : recherche clinique de cohortes :** Cette tâche consiste à identifier les patients ou les groupes de patients qui correspondent aux besoins d'une étude comparative. Nous utilisons la collection standard issue de la campagne d'évaluation majeure TREC [2], en l'occurrence *TREC Med* dans laquelle les requêtes spécifient des ensembles de maladies/conditions et des traitements/interventions, exprimés par les médecins en langage naturel. Cette collection comprend plus de 17000 rapports de visites médicales et 35 requêtes. Un exemple de requête est "*find patients with gastroesophageal reflux disease who had an upper endoscopy*".

Appariement de documents basé sur l'expansion de requêtes. Cette technique (notée $\mathbf{Exp_d}$) consiste à enrichir la formulation de chaque requête initiale en ajoutant les éléments (termes ou concepts) les plus représentatifs issus des premiers documents ordonnés en réponse à la requête. Plus précisément, un score de pertinence est estimé pour chaque paire élément (mot et/ou concept)-document sur la base de leur représentation dans l'espace latent produite par le processus d'apprentissage décrit par l'équation (1) et ce, en utilisant l'algorithme $CombSum$. Les concept et/ou mots ayant les meilleurs scores sont ensuite rajoutés à la requête.

Modèles de référence. Nous utilisons deux modèles de l'état de l'art d'expansion de requêtes : 1) **Rocchio**, un modèle d'expansion de requête basé sur les termes (Rocchio, 1971) ;

2. http ://trec.nist.gov/

TAB. 1: Comparaison de l'efficacité de notre modèle Exp_d sur deux tâches de RI médicale.

	RI liée à l'aide au diagnostic			recherche clinique de cohortes		
Modèle	MAP	P@20	R@20	MAP	P@20	R@20
LM-QE	0.0265	0.0686	0.0288	0.0793	0.1091	0.0519
Rocchio	0.0925	0.2262	0.0917	0.2096	0.2603	0.1701
$Exp_{\hat{d}^{PV-DM}}$	0.1017	0.2556	0.1086	0.3254	0.3971	0.2278
$Exp_{\hat{d}^{cd2vec}}$	0.0956	0.2365	0.0980	0.2255	0.2676	0.1319
Exp_d	0.1020	0.2556	0.1086	0.2996	0.3426	0.1989

2) **LM-QE**, un modèle de langue appliquant une expansion de requêtes par les concepts (Pal et al., 2014). Nous utilisons les paramètres par défaut.
Afin de tester la qualité de nos représentations, nous avons effectué une expansion de requête sur la base des représentations suivantes : 1) $\mathbf{Exp}_{\hat{\mathbf{d}}^{\mathbf{cd2vec}}}$, la représentation conceptuelle des documents (Section 3.2) ; 2) $\mathbf{Exp}_{\hat{\mathbf{d}}^{\mathbf{PV-DM}}}$, la représentation des documents basée sur le texte brut du modèle PV-DM.

Métriques. Nous avons utilisé des mesures d'évaluation standard, à savoir : la *MAP (Mean Average Precision)* ainsi que *P@20* et *R@20* correspondant respectivement à la précision et au rappel sur la liste des 20 premiers documents retournés par la requête.

4.2 Résultats

Nous présentons dans ce qui suit les résultats de l'évaluation comparative de notre modèle sur deux tâches de RI médicale : l'aide au diagnostic et la recherche clinique de cohortes. Le Tableau 1 présente l'impact de nos représentations Exp_d combinant le texte brut et les concepts issus des ressources sémantiques pour l'expansion de requêtes en termes de *MAP, P@20* et *R@20*. De façon générale, nous pouvons observer que les modèles d'expansion basés sur les représentations latentes ($Exp_{\hat{d}^{PV-DM}}$, $Exp_{\hat{d}^{cd2vec}}$ et Exp_d) permettent d'obtenir de meilleures performances dans le cas des deux tâches comparativement aux modèles de l'état de l'art, à savoir $LM-QE$ et *Rocchio* utilisant respectivement des représentations brutes des concepts et des mots. Par exemple, l'expansion basée sur les représentations latentes du texte $\hat{d}^{PV-DM}$ atteint de meilleurs résultats (MAP=$0,2996$) que l'expansion basée sur le texte brut *Rocchio* (MAP=$0,2096$). En particulier, notre modèle d'expansion Exp_d obtient des résultats significativement supérieurs à ceux du modèle $LM-QE$ sur l'ensemble des deux tâches et selon les trois métriques. Ces observations mettent en évidence le fait que les modèles basés sur les représentations latentes peuvent améliorer l'expansion de la requête avec l'aide de la sémantique latente des mots et/ou des concepts. Il est intéressant de noter également que la comparaison des espaces latents utilisés pour notre modèle d'expansion de requête renforce notre intuition quant à la complémentarité des termes et des concepts. En effet, notre représentation optimisée d permet d'obtenir des résultats légèrement supérieurs que ceux obtenus à partir des représentations latentes basées sur le texte $\hat{d}^{PV-DM}$ ou basées sur les concepts $\hat{d}^{cd2v}$. Ce résultat montre que notre représentation des documents permet de dépasser, d'une part, les enjeux liés à l'ambiguïté au niveau brut dans les textes et, d'autre part, les limites liées à l'extraction de concepts à partir de textes. Le Tableau 2 illustre un exemple d'expansion de requête (requête 131 de la collection TREC Med) où notre modèle Exp_d bénéficie à la fois

TAB. 2: Exemple de termes/concepts ajoutés à la requête 131 de Trec Med

Requête	*patients underwent minimally invasive abdominal surgery*
Concepts extraits	*Patients ; General Surgery ;*
Ajoutés par $Exp_{\hat{d}^{PV-DM}}$	*myofascia ; ultrasonix ; overtube*
Ajoutés par $Exp_{\hat{d}^{cd2vec}}$	*Mesna ; Esophageal Sphincter, Upper ;* Ganglioglioma
Ajoutés par Exp_d	*umbilical ; ventral ; biliary-dilatation*

des sources d'évidence termes et concepts pour la représentation afin d'identifier des éléments d'expansion plus pertinents que les autres scénarios $Exp_{\hat{d}^{PV-DM}}$ et $Exp_{\hat{d}^{cd2vec}}$. Plus particulièrement, même si la signification de haut niveau du terme "*abdominal surgery*" peut être capturée par le concept MeSH ("*General Surgery*"), notre modèle Exp_d est capable d'identifier des mots candidats pertinents de plus bas niveau pour l'expansion des requêtes ("*ventral*", "*biliary dilatation*"). De même, le modèle $Exp_{\hat{d}^{PV-DM}}$ identifie des mots candidats moins significatifs, tels que "*myofascia*". Cette observation renforce notre intuition quant à l'utilité de combiner des représentations latentes de textes et de concepts pour améliorer les tâches de RI, ce résultat étant conforme aux travaux antérieurs (Trieschnigg, 2010).

5 Conclusion et Perspectives

Dans cet article, nous traitons le problème du fossé sémantique sous-jacent à l'accès à l'information médicale en exploitant à la fois la sémantique des textes bruts et des ressources externes et ce, dans le but de produire des représentations de documents d'un haut niveau d'expressivité. Pour cela, nous proposons une fonction d'optimisation qui permet d'obtenir une représentation optimale des documents basée sur l'apprentissage de représentation des textes bruts et une extension du modèle PV-DM intégrant les concepts. Nous montrons expérimentalement l'efficacité des représentations de documents apprises grâce à l'expansion de requêtes sur deux tâches de RI médicale. Ce travail présente cependant certaines limites. Par exemple, pour identifier les mots activés et les concepts réinjectés dans une méthode d'expansion de requête, nous supposons que l'espace latent construit à l'aide de notre modèle est proche des espaces de représentation initiaux (espace latent des termes et espace latent des concepts). Cela pourrait être amélioré à l'avenir en construisant un modèle unifié pour l'apprentissage joint des représentations latentes à la fois sur les termes et les concepts avec une fonction de coût qui les relie.

Références

Choi, E., M. T. Bahadori, E. Searles, C. Coffey, et J. Sun (2016). Multi-layer representation learning for medical concepts. *KDD*, 1495–1504.

De Vine, L., G. Zuccon, B. Koopman, L. Sitbon, et P. Bruza (2014). Medical semantic similarity with a neural language model. In *CIKM*, pp. 1819–1822.

Dinh, D. et L. Tamine (2011). Combining Global and Local Semantic Contexts for Improving Biomedical Information Retrieval. In *ECIR*, pp. 375–386.

Edinger, T., A. Cohen, S.Bedrick, K. A. K, et W. H. W (2012). Barriers to retrieving patient information from electronic health record data : Failure analysis from the trec medical records track. In *AMIA Annual Symposium*, pp. 180–188.

Faruqui, M., J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, et N. A. Smith (2015). Retrofitting word vectors to semantic lexicons. In *NAACL*.

Iacobacci, I., M. T. Pilehvar, et R. Navigli (2015). Sensembed : Learning sense embeddings for word and relational similarity. In *ACL*, pp. 95–105.

Koopman, B., G. Zuccon, P. Bruza, L. Sitbon, et M. Lawley (2016). Information retrieval as semantic inference : A graph inference model applied to medical search. *Information Retrieval 19*(1-2), 6–37.

Le, Q. V. et T. Mikolov (2014). Distributed representations of sentences and documents. In *ICML*, pp. 1188—-1196.

Liu, X., J.-Y. Nie, et A. Sordoni (2016). Constraining word embeddings by prior knowledge – application to medical information retrieval. In *AIRS*.

Lu, Z., W. Kim, et W. J. Wilbur (2009). Evaluation of query expansion using MeSH in PubMed. *Information Retrieval 12*(1), 69–80.

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

Ni, Y., Q. K. Xu, F. Cao, Y. Mass, D. Sheinwald, H. J. Zhu, et S. S. Cao (2016). Semantic documents relatedness using concept graph representation. In *WSDM*.

Pal, D., M. Mitra, et K. Datta (2014). Improving query expansion using wordnet. *JASIST 65*(12), 2469–2478.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In *The SMART retrieval system*, pp. 313–323.

Stokes, N., Y. Cavedon, et J. Zobel (2009). Exploring criteria for succesful query expansion in the genomic domain. *Information retrieval 12*, 17–50.

Trieschnigg, D. (2010). *Proof of Concept : Concept-based Biomedical Information Retrieval*. Ph. D. thesis, University of Twente.

Wang, C. et R. Akella (2015). Concept-based relevance models for medical and semantic information retrieval. In *CIKM*, pp. 173–182.

Xu, C., Y. Bai, J. Bian, B. Gao, G. Wang, X. Liu, et T.-Y. Liu (2014). Rc-net : A general framework for incorporating knowledge into word representations. In *CIKM*.

Yu, M. et M. Dredze (2014). Improving lexical embeddings with semantic knowledge. In *ACL*, pp. 545–550.