Modèle neuronal tripartite pour la représentation de documents

Gia-Hung Nguyen[*] – Lynda Tamine[*] – Laure Soulier[**] – Nathalie Bricon-Souf[*]

** IRIT, Université de Toulouse, CNRS, INPT, UPS, UT1, UT2J, France, 118 Route Narbonne, Toulouse, France*
*** Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France*

RÉSUMÉ. *De nombreux travaux en recherche d'information (RI) ont montré que l'utilisation des sources d'évidence provenant de ressources sémantiques externes pourrait améliorer la performance de l'appariement. Par ailleurs, les approches neuronales sont devenues des modèles de référence qui permettent de capturer à partir des corpus, la sémantique latente des mots qui peut être injectée dans les modèles RI. Ce papier présente un modèle qui a pour but de réduire le fossé sémantique en RI en combinant ces deux sources d'évidence. C'est un modèle neuronal tripartite pour la représentation sémantique de documents qui exploite des connaissances explicites pour régulariser conjointement l'apprentissage de représentations des mots, des concepts et des documents. Nous montrons l'efficacité de notre modèle sur différentes tâches de RI.*

ABSTRACT. *Previous work in information retrieval (IR) have shown that using evidence, such as concepts and relations, from external knowledge resources could enhance the retrieval performance. Recently, deep neural approaches have emerged as state-of-the art models for capturing word semantics that can also be efficiently injected in IR models. This paper presents a new tri-partite neural document language framework that leverages explicit knowledge to jointly constrain word, concept, and document representation learning to tackle a number of issues including polysemy and granularity mismatch. We show the effectiveness of the framework in various IR tasks including word similarity, document similarity, and document re-ranking.*

MOTS-CLÉS : *Recherche d'information sémantique, base de connaissance, apprentissage de représentation.*

KEYWORDS: *Semantic information retrieval, knowledge resource, representation learning.*

1. Introduction

Le fossé sémantique est un problème bien connu en recherche d'information (RI) qui traduit l'écart significatif existant entre la représentation perçue par les utilisateurs des contenus textuels et les descripteurs de ces mêmes contenus, basés sur des caractéristiques de bas de niveau (ex. sac de mots) (Zhao et Grosky, 2002). Ce fossé est une des principales raisons du défaut d'appariement entre requête et document qui conduit à la dégradation des performances d'un système de RI (Crestani, 2000). Plus précisément, le défaut d'appariement peut se traduire sous trois formes génériques : 1) *non-concordance lexicale* (ou synonymie) lorsque deux mots lexicalement différents traduisent un même sens général (ex. les mot *'aperçu'* et le mot *'sommaire'* qui partagent le même sens) ; 2) *non-concordance de granularité conceptuelle* lorsque deux mots lexicalement différents représentent des granules différents reliés au même concept sans pour autant être synonymes (ex. '*chat*' et '*chien*' font référence tous deux au même concept '*animal*') ; 3) *polysémie* lorsqu'un même mot peut véhiculer des sens différents selon les contextes où il apparaît (ex. le mot *'décoller'* qui peut prendre le sens de *'détacher'* ou *'s'envoler'* selon le contexte de la phrase associée).
De très nombreux travaux en RI ont œuvré dans le sens de la réduction du fossé sémantique à l'aide d'approches qui visent l'amélioration des représentations des requêtes et/ou des documents. Une première lignée de travaux est basée sur l'exploitation des indices sémantiques explicites dérivés de ressources linguistiques (ex. WordNet, UMLS) ou graphes de connaissances (ex. DBpedia ou Freebase). L'idée sous-jacente à ces approches est d'injecter la connaissance portée par les concepts/entités et relations entre entités/concepts pour l'expansion des requêtes/documents (Xiong et Callan, 2015 ; Corcoglioniti *et al.*, 2016). Une autre lignée de travaux exploite quant à elle des indices sémantiques implicites dérivés des corpus, à savoir la sémantique distributionnelle (Harris, 1954). Cette dernière est fondée sur le calcul de la proximité sémantique entre mots sur la base des contextes partagés dans les corpus de textes. Plus spécifiquement, une approche récente qui a connu un grand succès est basée sur l'approche neuronale qui dérive le sens profond de mots (appelés *word embeddings*) par plongement lexical dans un espace latent (Mikolov *et al.*, 2013). Ces représentations vectorielles de dimension réduite sont alors utilisées dans la définition de nouveaux schémas d'appariement requête-document (Zuccon *et al.*, 2015) ou alors pour une expansion de requêtes (Zamani et Croft, 2016).
Cependant des travaux ont montré que les *word embeddings* classiques ne sont pas capables de résoudre le problème de polysémie (Iacobacci *et al.*, 2015). Des approches étendues ont alors été proposées pour pallier cela (Cheng *et al.*, 2015 ; Massimiliano *et al.*, 2017). Dans (Cheng *et al.*, 2015), les auteurs proposent d'étendre le modèle Skip-gram (Mikolov *et al.*, 2013) en identifiant les paires mot-concept (vus comme des paires de mot-sens candidat) dans un contexte donné en effectuant l'entraînement conjoint de leurs représentations latentes. Les alignements mot-concept sont établis soit avec des concepts explicites issus de ressources externes ou alors avec des concepts implicites dérivés du corpus. Dans la même perspective de résolution de la polysémie, Mancini et al. (Massimiliano *et al.*, 2017) étendent le modèle CBOW pour apprendre des représentations distinctes des différents sens d'un mot en les alignant

à des entrées d'ontologie (WordNet). Pour cela, une architecture révisée du modèle CBOW est proposée en vue d'apprendre conjointement dans le même espace à la fois le mot et les différents sens candidats associés.

Notre contribution s'inscrit dans la lignée de ces travaux dont l'objectif est d'améliorer la représentation des documents en vue de réduire le fossé sémantique, en nous intéressant spécifiquement à les intégrer dans un modèle de RI. Nous proposons un modèle neuronal qui exploite, en plus de la sémantique distributionnelle, la sémantique explicite issue d'une ressource externe en vue de remédier aux trois formes de défaut d'appariement évoquées ci-dessus. Plus précisément, le modèle permet de pallier le problème de discordance lexicale et de polysémie au travers l'apprentissage conjoint des représentations profondes des mots, concepts et documents dans le même espace latent. Cet apprentissage est déployé de façon à maximiser la qualité de la prédiction de chaque paire mot-concept dans le contexte d'un document donné. Le modèle permet également de pallier le problème de discordance de granularité conceptuelle au travers l'introduction d'une fonction de régularisation qui contraint l'apprentissage à assurer la proximité des représentations latentes des mots reliés aux mêmes concepts. Les contributions majeures de notre contribution sont les suivantes :

– Un modèle neuronal tripartite qui permet d'apprendre les représentations de documents en dérivant conjointement les représentations de mots et concepts et en considérant la contrainte de relations établies dans une ressource sémantique externe.

– Une évaluation expérimentale qui montre la qualité des représentations latentes obtenues dans des tâches de similarité de mots et documents et leur efficacité dans une tâche de RI.

L'article est organisé comme suit : la Section 2 donne un aperçu des travaux connexes. La Section 3 détaille la description de l'architecture du réseau neuronal et les principes d'apprentissage associés. La Section 4 est dédiée à la présentation du cadre expérimental. Les résultats de l'évaluation expérimentale sont présentés et discutés en Section 5. La Section 6 conclut l'article et présente quelques perspectives.

2. Travaux connexes

2.1. *Approches neuronales traditionnelles pour l'apprentissage des représentations textuelles*

L'apprentissage de représentation de mots (*"word embeddings"*) est au cœur de nombreuses recherches ces dernières années, notamment depuis l'introduction du modèle de langage neuronal (Bengio *et al.*, 2006) qui repose sur l'*hypothèse de la sémantique distributionnelle*. Le modèle le plus connu est *word2vec* et englobe deux approches qui ont respectivement comme objectif de prédire un mot cible en fonction des mots qui co-occurrent dans une fenêtre de contexte glissante (modèle CBOW) et de prédire les mots du contexte à partir d'un mot cible (modèle Skip-gram). D'autres modèles s'intéressent à la représentation de mots, dont le modèle *Global vector* (GloVe) (Pennington *et al.*, 2014), qui exploite la co-occurrence globale des mots.
Au-delà de la granularité au niveau des mots, certains travaux proposent d'apprendre

des représentations de textes tels que des phrases, des paragraphes ou des documents (Vulić et Moens, 2015 ; Mitchell et Lapata, 2008 ; Kenter *et al.*, 2016 ; Le et Mikolov, 2014). Une approche simple consiste à inférer la représentation du document à partir des représentations des mots grâce à un opérateur d'agrégation (à savoir la moyenne) pour estimer les représentations de phrases (Mitchell et Lapata, 2008 ; Vulić et Moens, 2015). Une autre approche, plus complexe, consiste en l'extension du modèle de langage neuronal (Kenter *et al.*, 2016 ; Kiros *et al.*, 2015). Entre autres, le modèle siamois CBOW (Kenter *et al.*, 2016) et le modèle Skip-thought (Kiros *et al.*, 2015) apprennent la représentation des phrases à partir de contextes de phrases (par analogie aux contextes de mots). Orthogonalement, dans le prolongement du modèle *word2vec*, le modèle Paragraph-Vector (Le et Mikolov, 2014) apprend conjointement les représentations de paragraphes (ou de documents) et de mots dans le même espace latent. Cet apprentissage conjoint repose sur l'hypothèse de composition sous-jacente du mot à la représentation du document où il apparaît (Mitchell et Lapata, 2008 ; Vulić et Moens, 2015).

2.2. *Approches neuronales de représentation textuelle augmentées par les ressources sémantiques externes*

Bien qu'efficace pour capturer la sémantique des mots, l'apprentissage de représentation distributionnelle ne permet pas de faire face à de nombreux problèmes, dont la polysémie (Iacobacci *et al.*, 2015). Pour répondre à cet enjeu, des travaux récents ont étudié l'utilisation conjointe de sources d'évidence basées d'une part sur les corpus et d'autre part, les ressources sémantiques (Faruqui *et al.*, 2014 ; Liu *et al.*, 2016 ; Mancini *et al.*, 2016 ; Yamada *et al.*, 2016).

Une première catégorie de travaux (Faruqui *et al.*, 2014 ; Liu *et al.*, 2016 ; Mrkšic *et al.*, 2016 ; Yu et Dredze, 2014) ont proposé d'améliorer la lisibilité de la représentation distributionnelle des mots appris sur des corpus en exploitant la *sémantique relationnelle* exprimée dans les ressources sémantiques externes. Une première catégorie de travaux ont proposé d'améliorer la lisibilité de la représentation distributionnelle des mots appris sur des corpus en exploitant la *sémantique relationnelle* exprimée dans les ressources sémantiques externes. L'intuition qui guide ces travaux est que des mots liés via des relations sémantiques établies dans une ressource externe sont supposés avoir des représentations proches dans l'espace latent. Par exemple, Faruqui et al. (Faruqui *et al.*, 2014) exploitent les informations relationnelles dérivées d'une ressource sémantique afin d'obtenir des représentations qui à la fois 1) minimisent la distance entre les paires de mots connectés dans le graphe sémantique et 2) conserve la sémantique initialement apprise sur le corpus (i.e., via les représentations distributionnelles). Cette dernière contrainte est également exploitée dans les travaux de Liu et al. (Liu *et al.*, 2016) qui évaluent les représentations obtenues dans une tâche de ré-ordonnancement de document en combinant le score d'appariement obtenu à l'aide d'un modèle de RI classique avec le score de similarité neuronale. Ce dernier est calculé à l'aide d'un cosinus entre les représentations distributionnelles du document et de la requête, construits par agrégation des représentations des mots associés.

La deuxième catégorie de travaux s'intéresse à un apprentissage conjoint des éléments

du corpus (à savoir les mots) et des éléments des ressources sémantiques (à savoir les concepts). Cet apprentissage conjoint dans des espaces partagés permet de mieux discriminer le sens des mots et par conséquent, résoudre le problème de la polysémie. Pour cela, Mancini et al. (Mancini *et al.*, 2016) modélise, au travers d'une architecture du type CBOW, l'hypothèse qu'un mot peut être associé à plusieurs sens en fonction du contexte. D'une façon orthogonale, Cheng et al. (Cheng *et al.*, 2015) proposent d'estimer dans le processus d'apprentissage, la probabilité d'association d'un concept à un mot dans la fenêtre de contexte. Dans le même esprit, Yamada et al. (Yamada *et al.*, 2016) propose un modèle de désambiguïsation d'entités nommées basé sur l'alignement des espaces intrinsèques respectivement liés à la représentation de mots et à la représentation des entités. L'alignement est réalisé grâce à des ancres mot-entité identifiés dans la ressource sémantique.

Notre contribution se distingue des travaux précédents selon deux principaux axes. En premier, c'est à notre connaissance le premier modèle d'apprentissage de représentation des documents qui exploite à la fois la sémantique latente disponible dans les corpus de documents (Le et Mikolov, 2014) et les connaissances de granularité plus fine en intégrant un apprentissage conjoint des mots et des concepts (Cheng *et al.*, 2015 ; Choi *et al.*, 2016 ; Mancini *et al.*, 2016 ; Yamada *et al.*, 2016). De plus, à la différence de (Choi *et al.*, 2016) qui considère les documents comme contexte temporel directement injecté dans le processus d'apprentissage des représentations d'éléments de granularité fine, nous supposons qu'un apprentissage conjoint des documents, des mots et des concepts permettrait de mieux capturer la sémantique aux niveaux global et local. En second, nous montrons la qualité des représentations apprises des documents ainsi que des concepts et des mots, utilisés comme informations auxiliaires pour améliorer l'appariement entre les documents et les requêtes (Zamani et Croft, 2016), à la différence des travaux précédents qui les utilisent principalement dans des tâches de traitement automatique de la langue.

3. Modèle neuronal tripartite

3.1. *Architecture du réseau de neurones*

Les modèles de langue neuronaux classiques permettent de résoudre la problématique du fossé sémantique liée principalement à la discordance lexicale. Cependant, ils sont incapables de faire face aux problèmes de discordance de granularité conceptuelle et de polysémie. Dans cet article, nous abordons ces trois problèmes en formalisant les hypothèses suivantes :

- *La prise en compte d'un contexte multi-niveaux (H1)* : chaque mot peut être assigné à un sens unique, c'est-à-dire à un concept pertinent identifié dans une ressource sémantique, au sein d'un même document alors qu'il pourrait relever de sens différents et être polysémique s'il est analysé sur un ensemble de documents. Ainsi, nous supposons que l'apprentissage simultané de représentations dans un contexte à plusieurs niveaux (à savoir, un niveau global pour des contextes de documents et un niveau local pour des contextes de mots et de concepts) permet d'affiner les représentations pour mieux résoudre le problème de polysémie.

- *La prise en compte d'un contexte basé sur les ressources sémantiques (H2)* : nous supposons que le problème lié à la discordance de granularité conceptuelle peut être

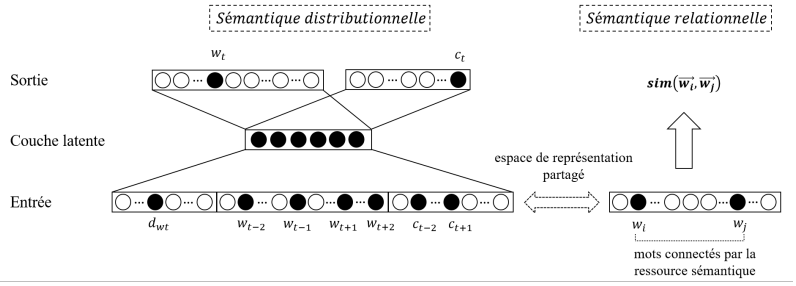

Figure 1. *Architecture du modèle neuronal tripartite*

partiellement ou complètement résolu en considérant la connaissance établie dans une ressource sémantique externe portant sur les relations entre mots. Selon le même principe que l'hypothèse distributionnelle, notre hypothèse repose sur le fait que des mots reliés au même concept devraient avoir des représentations latentes proches. Ainsi, notre modèle propose une approche de représentation des documents augmentée par une ressource sémantique, permettant conjointement de dériver la représentation des mots et des concepts sous-jacents. Formellement, l'apprentissage repose sur un ensemble $\mathcal{S} = \{\mathcal{D}, \mathcal{W}, \mathcal{C}\}$, où $\mathcal{D}$ fait référence à la collection de documents d, modélisés individuellement comme une séquence de mots ordonnés w_i et de concepts ordonnés c_i ; ces derniers étant associés aux mots w_i en fonction de leurs contextes dans une fenêtre textuelle ; $\mathcal{W}$ représente le vocabulaire (c'est-à-dire les mots) des documents de la collection $\mathcal{D}$ et $\mathcal{C}$ correspond à l'ensemble des concepts associés au vocabulaire et identifiés dans la ressource sémantique $\mathcal{R}$. Cette dernière fournit des connaissances au travers de concepts et de relations entre concepts. Nous soulignons que dans ce travail, nous considérons seulement les associations mot-concept pour des mots simples (uni-grammes) et laissons les associations entre concepts et mots composés pour de futurs travaux. La Figure 1 illustre l'architecture du modèle neuronal sur une instance d'apprentissage. La fonction objectif répond aux deux hypothèses énoncées précédemment : 1) la composante L_C (Hypothèse *H1*) qui apprend la représentation de plusieurs niveaux de granularité (documents, mots et concepts) en fonction de la prédiction des mots et concepts qui occurrent dans une fenêtre de contexte multi-niveaux ; 2) la composante L_R (Hypothèse *H2*) qui régularise l'apprentissage de relations en prenant en compte les contraintes de sémantique relationnelle. Ainsi, la fonction objectif est formalisée comme suit :

$$L = L_C + \beta L_R \quad [1]$$

où β est un coefficient de combinaison. Nous détaillons ci-dessous l'apprentissage de représentation en fonction du contexte multi-niveau (L_C) et de la régularisation par rapport au contexte basé sur une ressource sémantique (L_R).

3.2. *Mécanismes d'apprentissage du réseau*

3.2.1. *Apprentissage de représentation des documents, des mots et des concepts.*

Afin d'apprendre la représentation des documents de façon conjointe à l'apprentissage de représentations des mots et concepts identifiés dans le contexte du document,

nous proposons d'étendre le modèle de représentation des documents *Paragraph Vector* (Le et Mikolov, 2014). Plus particulièrement, les représentations des documents (également appelés vecteurs de documents) v_d sont apprises en fonction de leurs mots et concepts, et ce, en maximisant la prédiction des vecteurs de mots v_w et de concepts v_c en fonction de leur contexte. Ainsi, la fonction objectif de l'apprentissage conjoint document-mot-concept maximise la log-vraisemblance suivante :

$$L_C = \sum_{d \in \mathcal{D}} \sum_{w_t \in \mathcal{W}_d} [\log p(w_t | w_{t\pm k}, c_{t\pm k}, d) + \log p(c_t | w_{t\pm k}, c_{t\pm k}, d) - \frac{\gamma}{|d|} ||v_d||^2] \quad [2]$$

où l'ensemble des mots du document d est noté $\mathcal{W}_d$. k correspond à la taille de la fenêtre de contexte liée à un mot cible w_t, c_t est le concept associé au mot w_t en fonction de son contexte, $\frac{\gamma}{|d|}||v_d||^2$ est une régularisation $L2$ du vecteur document v_d qui permet de limiter le sur-apprentissage lié à l'apprentissage des textes longs (Ai *et al.*, 2016) avec $|d|$ correspondant à la longueur du document et γ est le coefficient de régularisation. La probabilité $p(w_t | w_{t\pm k}, c_{t\pm k}, d)$ du mot w_t étant donné son contexte est définie par une fonction soft-max :

$$p(w_t | w_{t\pm k}, c_{t\pm k}, d) = \frac{\exp(v_{w_t}^\top \cdot \overline{h}_{w_t})}{\sum_{w' \in \mathcal{W}} \exp(v_{w'}^\top \cdot \overline{h}_{w_t})} \quad [3]$$

où $\mathcal{W}$ correspond au vocabulaire de la collection. $\overline{h}_{w_t}$ représente la représentation du contexte moyennant les vecteurs v des mots dans le contexte $w_{t\pm k}$ et des concepts dans le contexte $c_{t\pm k}$ et incluant le vecteur document v_d. Cette représentation $\overline{h}_{w_t}$ est estimée ainsi :

$$\overline{h}_{w_t} = \frac{1}{4k+1} \left(v_d + \sum_{-k \leq j \leq k, j \neq 0} \left(v_{w_{t+j}} + v_{c_{t+j}} \right) \right) \quad [4]$$

De façon similaire, la probabilité $p(c_t | w_{t\pm k}, c_{t\pm k}, d_{w_t})$ du concept c_t en fonction de son contexte est estimée comme suit :

$$p(c_t | w_{t\pm k}, c_{t\pm k}, d) = \frac{\exp(v_{c_t}^\top \cdot \overline{h}_{c_t})}{\sum_{c' \in V_c} \exp(v_{c'}^\top \cdot \overline{h}_{c_t})} \quad [5]$$

où $\overline{h}_{c_t}$ est la représentation du vecteur de contexte lié au concept c_t, estimé avec la même méthode que $\overline{h}_{w_t}$ (voir Équation 4).

Étant donné la taille importante des ensembles $\mathcal{W}$ et $\mathcal{C}$, les probabilités décrites dans les formules (3) et (5) sont difficiles à estimer. Guidé par des précédents travaux (Mikolov *et al.*, 2013), nous exploitons les stratégies d'échantillonnage négatif ("*negative sampling*") pour définir des fonctions objectif alternatives pour chaque élément $e_t \in \{w_t; c_t\}$:

$$p(e_t | w_{t\pm k}, c_{t\pm k}, d) = \log \sigma(v_{e_t}'^\top \cdot \overline{h}_{e_t}) + \sum_{i=1}^{n} \mathbb{E}_{e_i \sim P_n(e)} \left[\log \sigma(-v_{e_i}'^\top \cdot \overline{h}_{e_t}) \right] \quad [6]$$

où $\sigma(x)$ correspond à la fonction sigmoid $\sigma(x) = \frac{1}{1+e^{-x}}$ et $\mathbb{E}_{e_i \sim P_n(e)}$ est la valeur attendue de $\log \sigma(-v'^{\top}_{e_i} \cdot \overline{h}_{e_t})$ quand e_i est tiré de la distribution uniforme pondérée $P_n(e)$, comme réalisé par Ai et al. (Ai *et al.*, 2016).

3.2.2. *Régularisation par rapport à des connaissances a priori issues d'une ressource sémantique externe.*

Pour remédier à la problématique de la discordance de granularité conceptuelle, nous proposons de raffiner l'apprentissage de représentations, développé précédemment, en incorporant la sémantique relationnelle établie dans une ressource sémantique externe. L'intuition sous-jacente est d'intégrer dans le processus d'apprentissage, des contraintes de relations entre des mots qui peuvent ne pas être (suffisamment) mises en évidence dans les contextes des documents utilisés pour l'apprentissage basé sur l'analyse distributionnelle ; c'est particulièrement le cas lorsque des mots, pourtant sémantiquement reliés, occurrent peu fréquemment dans les mêmes contextes en partie en raison de la diversité du vocabulaire. Inspirés par les travaux précédents (Yu et Dredze, 2014), nous proposons de régulariser la fonction objectif afin d'intégrer les contraintes relationnelles dans les représentations de mots. La régularisation ajustera les représentations des mots, et simultanément l'apprentissage de représentation des documents, de sorte que les mots qui partagent le même concept aient des représentations proches. De façon formelle, notre objectif est de maximiser la similarité entre les mots (w_i, w_j) reliés dans la ressource sémantique au travers le terme de régularisation suivant :

$$L_R = \sum_{(w_i,w_j) \in \mathcal{W} \times \mathcal{W} \setminus linkC(w_i,w_j)=1 \ or \ linkR(w_i,w_j)=1} sim(w_i, w_j) \qquad [7]$$

où $linkC(w_i, w_j) = 1$ si les mots w_i et w_j partagent le même concept et $linkR(w_i, w_j) = 1$ si les mots sont associés à des concepts reliés. $sim(w_i, w_j)$ correspond à la similarité cosinus entre les vecteurs de mots v_{w_i} et v_{w_j}.

4. Cadre d'évaluation expérimentale

L'objectif de notre évaluation expérimentale est double : 1) évaluer la qualité des représentations de documents apprises par notre modèle neuronal et 2) mesurer l'impact de ces représentations sur l'efficacité de différentes tâches de RI.

4.1. *Jeux de données*

Deux jeux de données sont utilisés pour mener l'évaluation expérimentale.

– Robust04[1] qui est un jeu de données de nouvelles (*news*) fourni dans le cadre de la campagne d'évaluation TREC Robust Track 2004 comprenant 528,155 documents et 250 requêtes sous forme de titres.

– SQuAD (Rajpurkar *et al.*, 2016) qui est jeu de données adapté à une tâche question-réponse formé d'articles Wikipédia ainsi qu'un ensemble de questions formulées par des *crowdworkers*. La réponse à chaque question est un passage (segment

1. http ://trec.nist.gov/data/robust/04.guidelines.html

de texte) d'un article wikipédia. Nous avons adapté ce corpus à une tâche de RI *ad-hoc* en retenant 490 requêtes associées aux titres d'articles wikipédia soumises à un corpus formé de 20963 passages. On retient la vérité terain donnée dans le jeu de données SQUAD.

Pour enrichir les représentations par la sémantique relationnelle, nous exploitons DBpedia en tant que ressource sémantique externe en raison de sa large couverture. Les requêtes et les documents de Robust sont annotés par *TagMe* (Ferragina et Scaiella, 2010), un outil d'annotation pour lier du texte aux entités DBpedia. Les requêtes et les documents de SQUAD sont annotés en utilisant directement les liens wikipédia vers les entités DBpedia. Pour les deux jeux de données, nous utilisons précisément les noms des entités de DBpedia pour annoter les requêtes et les documents et exploiter la relation `gold: hypernym`. Par rapport à la description du modèle de la section 3, nous référons simplement aux entités par concepts.

4.2. *Méthodologie d'évaluation*

Nous évaluons notre modèle tripartite proposé selon trois scénarios :

– **PV** : le modèle Paragraph-Vector (Le et Mikolov, 2014) à partir duquel nous construisons notre modèle neuronal étendu. Ce scénario permet d'évaluer l'impact de la prise en compte des concepts et des relations dans l'apprentissage de la représentation puisque ce scénario est basé que des contextes de mots et de documents.

– **S2DV** : notre modèle d'apprentissage proposé sans l'étape de régularisation L_R. L'objectif de ce scénario est donc d'évaluer l'impact de l'intégration des seules relations verbales implicites dans le processus d'apprentissage basé sur l'analyse distributionnelle.

– **S2DVR** : notre modèle d'apprentissage complet.

En outre, pour répondre aux objectifs de l'évaluation expérimentale mentionnés ci-dessus, nous utilisons deux cadres d'évaluation détaillés ci-dessous.

4.2.1. *Évaluer la qualité des représentations distributionnelles.*

Compte tenu du premier objectif de notre modèle qui consiste à apprendre les représentations de documents en dérivant conjointement les représentations de mots et concepts, nous évaluons d'abord la qualité de ces représentations entraînées. De ce fait, nous utilisons deux tâches :

– *Similarité de mots* : Nous évaluons les représentations de mots sur trois jeux de données différents et standards qui ont été largement utilisés pour mesurer la similarité des mots. Le premier est le jeu de données WS-353 (noté WS) (Finkelstein *et al.*, 2001) qui contient 353 paires de mots anglais. Le deuxième est le jeu de données RG-65 (Rubenstein et Goodenough, 1965) qui contient 65 paires de noms. Nous utilisons également le jeu de données MEN (Bruni *et al.*, 2012) formé de 3 000 paires de mots qui apparaissent au moins 700 fois dans un grand corpus web. Dans chaque jeu de données, tous les paires de mots ont une note de similarité attribuée par des assesseurs humains. Nous calculons la similarité cosinus entre les vecteurs des deux représentations associées aux paires de mots puis reportons le coefficient de corréla-

tion de Spearman entre : 1) le classement obtenu des paires de représentations de mots issues notre modèle et 2) le classement de ces même paires de mots obtenu grâce aux scores attribués par les experts humains.

– *Similarité de documents* : qui consiste à discriminer la similarité des documents par rapport à une requête donnée comme décrite dans (Le et Mikolov, 2014). Plus précisément, pour chaque requête du jeu de données, nous créons un *pool* de triplets de documents dont les deux premiers sont renvoyés par un modèle RI à partir de cette requête et le troisième est échantillonné au hasard à partir des documents renvoyés pour d'autres requêtes. L'objectif sous-jacent est de mesurer dans quelle mesure la similarité des documents (c'est-à-dire la mesure de cosinus) estimée à l'aide des représentations de documents apprises permet de fournir une mesure de similarité plus importante pour les documents issus de la même requête et une similarité plus faible pour les documents issus d'autres requêtes. De façon analogue à (Le et Mikolov, 2014), nous utilisons le taux d'erreur sur toutes les requêtes mesurées lorsque la similarité entre les deux premiers documents est plus petite que celles du premier et troisième documents. Nous comparons les taux d'erreur obtenus en utilisant les représentations de documents fournies par notre modèle à celles obtenues par la représentation de documents **AWE** (Vulić et Moens, 2015). Cette représentation de documents est obtenue en moyennant ses vecteurs de mots (*word embeddings*). Le but cette comparaison est d'évaluer l'impact de la prise en compte d'un contexte multi-niveaux (concepts et documents en plus des mots) sur la qualité des représentations

4.2.2. *Évaluer l'efficacité des représentations dans des tâches de RI.*

En accord avec le second objectif d'évaluation, nous mesurons l'efficacité des représentations apprises sur les performances d'une tâche de RI en les injectant dans un modèle de réordonnancement, comme proposé dans (Liu *et al.*, 2016) :

$$RSV(Q, D) = \alpha \cdot IRScore(Q, D) + (1 - \alpha) \cdot NeuralScore(Q, D) \qquad [8]$$

où α est le coefficient de combinaison déterminé par une double validation croisée selon la métrique MAP, $IRScore$ est le score de document obtenu à l'aide d'un modèle de RI classique (à savoir BM25) et $NeuralScore$ (la similarité cosinus entre les représentations de la requête et du document apprises à l'aide de notre modèle). De plus pour des raisons de comparabilité de l'efficacité, nous injectons la représentation obtenue par des scénarios PV, S2DV et S2DVR. Nous comparons l'efficacité de notre modèle au modèle de référence BM25 en mesurant l'efficacité à l'aide des mesures standards : la précision moyenne (MAP) et le rappel au rang 1000,

4.3. *Détails d'implémentation*

Pour obtenir les représentations de documents issus des jeux de données SQuAD et Robust, nous avons entraîné les configurations (PV, SD2V, SD2VR) sur les documents de chaque corpus, à savoir Wikipédia et Robust. Nous ajoutons aussi un scénario basé sur Robust pré-entraîné (*Robust-Pre*) où les vecteurs de mots sont initialisés par ceux appris sur Wikipédia. Pour les configurations (PV, SD2V, SD2VR), nous avons défini la dimension des vecteurs à 300 et nous avons choisi empiriquement la taille de fenêtre $k = 8$. Après avoir supprimé les mots non alphanumériques, nous n'avons retenu que

Tableau 1. *Résultats comparatifs pour la tâche de similarité des mots en termes de mesures de corrélation*

	Robust			Wikipédia			Robust-Pre		
	MEN	RG65	WS	MEN	RG65	WS	MEN	RG65	WS
PV	**0,48**	**0,40**	**0,35**	0,63	0,58	0,52	**0,53**	0,49	**0,33**
SD2V	0,44	0,30	0,24	**0,71**	**0,70**	0,58	0,51	0,49	0,29
SD2VR	0,43	0,37	0,27	**0,71**	**0,70**	**0,59**	**0,53**	**0,51**	0,30

les mots dont la fréquence dans le corpus est supérieure à 4. Le taux d'apprentissage initial est configuré à 0,02 puis diminué linéairement pendant le processus d'entraînement SGD. La valeur du paramètre β de l'équation 1 est choisie empiriquement selon chaque tâche comme suggéré dans (Massimiliano *et al.*, 2017). Nous varions la force de régularisation γ dans l'équation 2 avec les valeurs 0.1, 1 et 10 comme suggéré dans (Ai *et al.*, 2016) ; la meilleure performance est obtenue avec $\gamma = 1$. Enfin, tous les modèles de RI sont déployés à l'aide du moteur de recherche Indri[2].

5. Résultats

5.1. *Analyse de la qualité des représentations distributionnelles*

Afin de mesurer la qualité des représentations, notre évaluation repose sur les tâches de similarité des mots et des document décrites dans la section 4.2.

5.1.1. *Etude quantitative*

Nous nous intéressons en premier lieu au niveau mot. Les résultats (estimés au travers d'une mesure de corrélation de Spearman) sont présentés dans le tableau 1. Les résultats obtenus à partir des représentations de mots apprises sur Robust montrent que le modèle PV obtient de meilleurs résultats que les scénarios de notre modèle. Par exemple, pour les jeux de données MEN, PV obtient un coefficient de corrélation de $0,48$ alors que nos scénarios atteignent au maximum $0,44$. Une explication pourrait être liée au bruit de l'annotation des concepts sur la collection Robust effectuée automatiquement par l'outil *TagMe*. Afin de valider cette hypothèse, nous avons entraîné nos scénarios sur le corpus Wikipédia où l'annotation mot-concept est modérée par l'humain, et donc supposée moins (voir très peu) bruitée. Nous remarquons dans un premier temps que les mesures de corrélation sont supérieures à celles obtenues sur Robust pour tous les scénarios (PV, SD2V et SD2VR). Ceci peut être expliqué par la taille plus importante de la collection Wikipedia ainsi que les caractéristiques de la collection Wikipédia incluant des connaissances plus génériques, et donc plus adaptées pour des tâches de similarité de mots. En effet, certains documents de la collection Robust, notamment les documents relatifs aux "dossiers fédéraux" (`Federal Register 94`), sont caractérisés par des contenus plus structurés avec des phrases nominales très courtes et de nombreux acronymes ; gênant possiblement la capture de sémantique distribuée. De plus, nous observons que les scénarios de notre modèle (SD2V et SD2VR) obtiennent de meilleurs résultats que le modèle de référence (PV).

2. https ://www.lemurproject.org/indri.php

Tableau 2. *Résultats comparatifs pour la tâche de similarité des documents en termes de taux d'erreurs*

	Robust	SQuAD	Robust-Pre
AWE	93,1%	95,4%	91,3%
PV	**6,9%**	30,5%	9,9%
SD2V	11,5%	33,3%	13,8%
SD2VR	14,7%	**6,2%**	**9,8%**

Ce résultat renforce notre hypothèse sur les faibles performances de notre modèle lorsque la qualité de l'annotation ne peut être assurée. Dans cette optique, nous avons évalué la qualité des représentation sur une configuration Robust-Pre, où les représentations sont initialisées sur la collection Wikipédia pour chacun des modèles et ensuite raffinées sur la collection Robust. Nous pouvons constater des améliorations pour tous les modèles PV, SD2V, SD2VR sur l'ensemble des jeux de données MEN, RG65 et WS. De plus, nous pouvons constater qu'un de nos scénarios (SD2VR) obtient les plus grands coefficients de corrélation sur MEN et RG65 ($0,53$ et $0,51$), démontrant ainsi l'apport de la sémantique dans la représentations des termes. La comparaison entre SD2V et SD2VR suggère que ce sont les relations qui permettent de compenser, voire de dépasser, les problèmes d'annotation des concepts (avec SD2V qui semble sous-optimal) et de capturer plus efficacement la sémantique latente des termes.

L'analyse de la qualité des représentations documents est présentée dans le tableau 2. Elle oppose les scénarios (PV, SD2V, SD2VR) par rapport au modèle de référence AWE qui estime la représentation des documents en moyennant les représentations de termes inclus dans les documents. A noter que nous avons appris nos représentations sur les configurations Robust, Wikipedia et Robust-Pre et que la tâche de similarité des documents est construite sur les cadres d'évaluation standard issus de Robust (pour les configurations Robust et Robust-Pre) et SQuAD (pour la configuration Wikipedia dans la mesure où ce dernier repose sur la collection Wikipédia). En premier lieu, nous pouvons constater que le modèle AWE ne permet pas de discriminer les documents car il obtient des taux d'erreurs dépassant les 90%. Ce résultat suggère que le vecteur moyen des mots d'un document peut être bruité par certains mots génériques ou polysémiques, ne permettant pas de capturer la sémantique globale des documents, plus particulièrement pour les documents longs. Ensuite, en se focalisant sur les modèles qui apprennent directement la représentation de documents (à savoir PV, SD2V et SD2VR), nous observons des résultats qui corroborent les résultats obtenus pour la tâche de similarité de mots : 1) l'apprentissage de représentation des documents est sensible à la qualité d'annotation et donc plus performant sur les jeux de données SQuAD et Robust-Pre et, 2) lorsque l'annotation est fiable ou compensée par un pré-apprentissage, notre modèle obtient de meilleurs résultat que le scénario PV (par exemple, 6.2% vs. 30.5% pour resp. SD2VR et PV sur le jeux de données SQuAD reposant sur la collection Wikipedia). Egalement, nous observons que le scénario SD2V dépasse le scénario PV pour la tâche de similarité des documents, ce qui n'était pas le cas pour la tâche de similarité des mots. Ce résultat suggère que notre modèle est plus

Tableau 3. *Exemples illustratifs - Mots les plus similaires à un document dans l'espace de représentation*

ID et thèmes du document	PV	SD2V	SD2VR
FT924-286	galaxies	galaxies	**subatomic**
(physics, astronomy,	quasars	quasars	planetary
cosmology, nuclear)	planets	**pulsars**	pulsars
	universe	subatomic	celestial
	cosmic	cosmic	**relativistic**
FR940926-2-00016	bingdu	**nonpelagic**	**tursiops**
(marine mammals,	drachmei	yaobin	**eumetopias**
endangered species, ocean	ismls	guanghong	**zalophus**
legislation)	rhoger	docha	outerlimits
	seneker	hxcdd	**vitulina**

efficace pour capturer la sémantique globale effectuée au niveau des documents que la sémantique locale relevant des mots. En effet, la représentation des documents prend en compte l'ensemble des concepts, et éventuellement des relations pour SD2VR, permettant de réaliser des inférences de plus haut niveau lors de l'apprentissage et donc de raffiner la qualité de représentation des documents. En résumé, ces constats démontrent la qualité des représentations obtenues à partir de notre modèle tripartite, sous contrainte de prendre en compte le niveau de la qualité d'annotation du corpus d'apprentissage.

5.1.2. *Etude qualitative*

Afin d'approfondir ces résultats, nous avons effectué une analyse qualitative croisée basée sur les représentations des mots et des documents. Pour cela, nous identifions pour chaque document les mots les plus proches dans l'espace latent de représentation. Après observation manuelle, deux tendances principales se dégagent et sont illustrées dans le tableau 3. La première met en évidence la capacité du modèle PV à identifier les mots les plus similaires reliés au sujet du document (document FT924-286). L'apprentissage de représentation guidé par les ressources sémantiques externes (SD2V et SD2VR) permet alors de raffiner la représentation en identifiant l'ensemble des thématiques du document (par exemple "subatomic" reflétant le thème "nuclear" du document) ou des thématiques orthogonales (e.g., "relativistic"). Dans le deuxième cas, le contenu des documents n'étant pas assez verbeux et/ou avec de nombreux termes techniques (comme le cas des dossiers fédéraux), le modèle de l'état de l'art PV ne permet pas d'identifier les termes les plus similaires d'un document. Par contre, notre modèle permet d'identifier des termes reliés au sujet. Dans notre exemple (document FR940926-2-00016), les termes "nonpelagie", "tursiops", "eumetopias", "zalophus" et "vitulina" correspondent à des espèces d'animaux marins qui ont pu être identifiés grâce à l'apport de la ressource sémantique externe, et particulièrement les relations entre les concepts recensés dans la ressource.

Tableau 4. *Effet de différents scénarios sur l'efficacité de la recherche en termes de MAP et de Rappel. Les valeurs en gras expriment des résultats supérieurs aux valeurs de référence (BM25).*

	Robust		SQUAD		Robust-Pre	
	MAP	Rappel	MAP	Rappel	MAP	Rappel
BM25	0,2510	0,6898	0,5872	0,8190	0,2510	0,6898
PV	0,2505	0,6894	0,5871	0,8189	0,2504	**0,6902**
SD2V	0,2507	0,6896	**0,5874**	**0,8191**	0,2508	0,6897
SD2VR	**0,2517**	**0,6905**	**0,5892**	**0,8208**	**0,2539**	**0,6928**

5.2. *Évaluer l'efficacité des représentations dans une tâche de RI*

Le tableau 4 présente les résultats obtenus pour la tâche de ré-ordonnancement de documents pour les différents scénarios (PV, SD2V, SD2VR) et le modèle de référence classique (BM25). D'un point de vue général, nous pouvons observer que notre modèle tripartite (particulièrement le SD2VR) permet d'améliorer les résultats d'ordonnancement. Spécifiquement, pour le Robust, SD2VR atteint une valeur de MAP à $0,2517$ par rapport à celle de PV $0,2510$. Cette amélioration est plus importante lorsque nous pré-entraînons les vecteurs de mots dans le Robust-Pre, $0,2539$ du SD2VR à $0,2510$ du PV. Nous notons cependant que les améliorations sont très faibles (de l'ordre du millième). Ce constat a déjà été observé dans les précédents travaux s'intéressant à l'intégration des représentations neuronales dans les tâches de RI (Arora *et al.*, 2017 ; Diaz *et al.*, 2016) et est constaté également lorsque nous analysons le scénario PV (modèle de représentation des documents de l'état de l'art (Le et Mikolov, 2014)). En effet, le modèle PV est moins efficace que le modèle de référence BM25 sur les trois jeux de données (par exemple avec des valeurs de MAP égales à $0,2505$ vs. $0,2510$ pour resp. BM25 et PV sur le jeu de données Robust). Plus particulièrement, l'analyse des résultats du tableau 4, permet de déduire les principales conclusions suivantes :

– En comparant les résultats sur différents jeux de données, nous pouvons observer les mêmes conclusions que précédemment sur la qualité du processus d'annotation : les configurations neuronales (surtout SD2V et SD2VR) obtiennent une meilleure performance sur Robust quand elles sont initialisées avec des représentations pré-entraînées. Nos modèles exploitant la sémantique des concepts associés aux mots ainsi que leur relations dans la ressource de connaissance.

– Nos scénarios SD2V et SD2VR dépassent le PV en termes de MAP et Rappel dans tous les trois jeux de données. Cette observation souligne l'effet positif de la combinaison de mots et concepts dans l'apprentissage de représentations distributionnelles des documents et la capacité de notre modèle à dépasser les limites d'intégration de la sémantique distributionnelle dans des tâches de RI. De plus, le modèle SD2VR qui a la meilleure performance, dépasse le modèle de référence classique BM25, permettant de souligner le rôle important de la régularisation par les relations extraites de la ressource. Ce résultat est cohérent avec les résultats obtenus pour les tâches de similarité qui montrent que les relations permettent de capter la sémantique globale

(voir Section 5.1) et par conséquent, d'augmenter l'efficacité de l'appariement des documents avec la requête.

6. Conclusion

Nous avons présenté dans cet article un modèle neuronal tripartite permettant d'apprendre les représentations des documents, conjointement avec les mots et concepts issus d'une ressource sémantique externe. De plus, l'apprentissage est contraint par les relations établies dans cette ressource afin d'assurer la lisibilité des représentations obtenues et au delà, réduire le fossé sémantique en RI. Les expérimentations ont montré l'impact positif de l'usage des concepts et des relations à la fois pour des tâches de similarité de mots, de documents et de RI. Cependant, nous avons également observé que le modèle proposé est sensible à la qualité de l'annotation conceptuelle. Une perspective intéressante serait d'intégrer dans le processus d'apprentissage un alignement approximatif mot-concepts candidats qui permettrait de réduire l'effet négatif d'une annotation bruitée sur la qualité des représentations apprises.

7. Bibliographie

Ai Q., Yang L., Guo J., Croft W. B., « Analysis of the paragraph vector model for information retrieval », *ICTIR*, ACM, p. 133-142, 2016.

Arora P., Foster J., Jones G. J. F., « Query Expansion for Sentence Retrieval Using Pseudo Relevance Feedback and Word Embedding », *in* G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, N. Ferro (eds), *CLEF*, p. 97-103, 2017.

Bengio Y., Schwenk H., Senécal J.-S., Morin F., Gauvain J.-L., « Neural probabilistic language models », *Innovations in Machine Learning*, 2006.

Bruni E., Boleda G., Baroni M., Tran N.-K., « Distributional semantics in technicolor », *ACL*, p. 136-145, 2012.

Cheng J., Wang Z., Wen J.-R., Yan J., Chen Z., « Contextual Text Understanding in Distributional Semantic Space », *CIKM*, p. 133-142, 2015.

Choi E., Bahadori M. T., Searles E., Coffey C., Sun J., « Multi-layer Representation Learning for Medical Concepts », *KDD*p. 1495-1504, 2016.

Corcoglioniti F., Dragoni M., Rospocher M., Aprosio A. P., « Knowledge Extraction for Information Retrieval », *ESWC*, vol. 9678, p. 317-333, 2016.

Crestani F., « Exploiting the Similarity of Non-Matching Terms at Retrieval Time », *Information Retrieval*, vol. 2, n° 1, p. 27-47, Feb, 2000.

Diaz F., Mitra B., Craswell N., « Query Expansion with Locally-Trained Word Embeddings », *ACL*, 2016.

Faruqui M., Dodge J., Jauhar S. K., Dyer C., Hovy E., Smith N. A., « Retrofitting Word Vectors to Semantic Lexicons », *NAACL*, 2014.

Ferragina P., Scaiella U., « Tagme : on-the-fly annotation of short text fragments (by wikipedia entities) », *CIKM*, ACM, p. 1625-1628, 2010.

Finkelstein L., Gabrilovich E., Matias Y., Rivlin E., Solan Z., Wolfman G., Ruppin E., « Placing search in context : The concept revisited », *WWW*, ACM, p. 406-414, 2001.

Harris Z. S., « Distributional structure », *Word*, vol. 10, n° 2-3, p. 146-162, 1954.

Iacobacci I., Pilehvar M. T., Navigli R., « SensEmbed : Learning Sense Embeddings for Word and Relational Similarity », *ACL*, p. 95-105, 2015.

Kenter T., Borisov A., de Rijke M., « Siamese cbow : Optimizing word embeddings for sentence representations », *arXiv preprint arXiv :1606.04640*, 2016.

Kiros R., Zhu Y., Salakhutdinov R. R., Zemel R., Urtasun R., Torralba A., Fidler S., « Skip-thought vectors », *NIPS*, p. 3294-3302, 2015.

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

Liu X., Nie J.-Y., Sordoni A., « Constraining word embeddings by prior knowledge–application to medical information retrieval », *Information Retrieval Technology*, p. 155-167, 2016.

Mancini M., Camacho-Collados J., Iacobacci I., Navigli R., « Embedding Words and Senses Together via Joint Knowledge-Enhanced Training », *arXiv preprint arXiv :1612.02703*, 2016.

Massimiliano M., Jose C.-C., Ignacio I., Roberto N., « Embedding words and senses together via joint knowledge-enhanced training », *CoNLL*, p. 100-111, 2017.

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

Mitchell J., Lapata M., « Vector-based Models of Semantic Composition. », *ACL*, p. 236-244, 2008.

Mrkšic N., OSéaghdha D., Thomson B., Gašic M., Rojas-Barahona L., Su P.-H., Vandyke D., Wen T.-H., Young S., « Counter-fitting Word Vectors to Linguistic Constraints », *NAACL-HLT*, p. 142-148, 2016.

Pennington J., Socher R., Manning C., « Glove : Global Vectors for Word Representation », *EMNLP*, p. 1532-1543, 2014.

Rajpurkar P., Zhang J., Lopyrev K., Liang P., « Squad : 100,000+ questions for machine comprehension of text », *arXiv preprint arXiv :1606.05250*, 2016.

Rubenstein H., Goodenough J. B., « Contextual correlates of synonymy », *Communications of the ACM*, vol. 8, n° 10, p. 627-633, 1965.

Vulić I., Moens M.-F., « Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings », *SIGIR*, ACM, p. 363-372, 2015.

Xiong C., Callan J., « Query expansion with Freebase », *ICTIR*, ACM, p. 111-120, 2015.

Yamada I., Shindo H., Takeda H., Takefuji Y., « Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation », p. 250-259, 2016.

Yu M., Dredze M., « Improving Lexical Embeddings with Semantic Knowledge », *ACL*, p. 545-550, 2014.

Zamani H., Croft W. B., « Estimating embedding vectors for queries », *ICTIR*, ACM, p. 123-132, 2016.

Zhao R., Grosky W. I., « Narrowing the semantic gap-improved text-based web document retrieval using visual features », *IEEE transactions on multimedia*, vol. 4, n° 2, p. 189-200, 2002.

Zuccon G., Koopman B., Bruza P., Azzopardi L., « Integrating and Evaluating Neural Word Embeddings in Information Retrieval », *ADCS*, ACM, p. 12, 2015.