A Tri-Partite Neural Document Language Model for Semantic Information Retrieval

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

A Tri-Partite Neural Document Language Model for Semantic Information Retrieval

Gia-Hung Nguyen[1], Lynda Tamine[1], Laure Soulier[2], and Nathalie Souf[1]

[1] Université de Toulouse, UPS-IRIT, 118 route de Narbonne, 31062 Toulouse, France
[2] Sorbonne Université, CNRS - LIP6 UMR 7606, 75005 Paris, France

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

Keywords: Semantic information retrieval, knowledge source, deep learning

1 Introduction

The semantic gap is a long-standing research topic in information retrieval (IR) that refers to the difference between the low-level description of document and/or query content (in general bags of words) and the high level of their meanings [30]. The semantic gap inherently hinders the query-document matching which is the crucial step for selecting candidate relevant documents in response to a user's query. The semantic gap commonly originates from the following: 1) *Vocabulary mismatch*, also called *lexical gap*, which means that words with different shapes share the same accepted meaning (senses) (e.g., *car is a synonym of motorcar*); 2) *Granularity mismatch* which means that words with different shapes and senses belong to the same general concept (e.g., *air bag and wheel are both parts of a car*); 3) *Polysemy* which means that a word could cover different senses depending on its surrounding words in the text that represent its context (e.g., *bass* could mean a type of fish or the lowest part of harmony).

To close these gaps, the prominent approaches employed in IR focus on the improvement of query and/or document representations using explicit knowledge provided by external knowledge sources or implicit knowledge inferred from text corpora. A first line of work is based on the use of linguistic sources (e.g., WordNet) or knowledge graphs (e.g., DBpedia). The key idea of these approaches is to inject knowledge about entities/concepts and semantic relations between them (e.g., relations of synonymy or hyperonymy) into query and/or document representations [26, 5]. Another line of work particularly tackles the lexical gap in IR through distributional semantics which relies on the assumption that word senses could be inferred from their distribution in the text. Specifically, recent

approaches in this category of work aim at projecting word senses in a continuous latent space using neural language models [16] to learn distributed representations of words (also called "word embeddings") using their context. However, authors in [10] have shown that traditional word embeddings are not able to cope with the polysemy problem. Recently, some work [3, 15] have tackled this issue. For instance, Cheng et al. [3] propose to extend the skip-gram model [16] to identify the relevant word-concept pairwise given a context by jointly training the corresponding embeddings. The connection between words and concepts is set up based on either implicit senses (corpus-based) or explicit senses (inventoried in a knowledge source).

In this work, we propose a neural network-based model that can jointly cope with the three semantic gap factors mentioned above. The model is based on a semantically-oriented approach of concept/entity, word, and document embeddings which is based on the joint use of raw textual data and knowledge sources within the same embedding space. The model has a high level of generalizability in terms of use in the semantic web since 1) the learned embeddings can be integrated in different tasks such as entity linking [18], semantic annotation of unstructured or structured data [6], ontology matching by estimating levels of alignments between concepts embeddings learned using different ontologies, information extraction from texts by ranking candidate concept/entity embeddings with respect to document embedding, and, word sense disambiguation by using word embeddings as features of a supervised disambiguation method; 2) a wide range of knowledge sources (linguistic, knowledge graphs,...) can be used as evidence in the learning process of the semantic representations. The contributions of the paper are:

• We design a tri-partite neural language model that learns representations of documents, concept/entity and word representations, constrained by the pre-established relations existing in a knowledge source (Section 3).

• We experimentally show the quality and effectiveness of the learned representations for semantic IR tasks (Sections 4 and 5).

2 Related work

Traditional Neural Approaches for Learning Text Representations. Building distributed word representations (also called "word embeddings") from large corpora has received increasing attention since the introduction of the probabilistic neural network language model [2]. The *distributional hypothesis* [9] assumes that the representation of words with similar distributions should have similar meanings. For example, two efficient neural network models (i.e., word2vec) [16] use the co-occurrence of words to learn word representations. Specifically, the continuous bag-of-word (CBOW) model predicts a target word by maximizing the log-likelihood of its context words in a sliding word window while the second model (skip-gram) tries to predict the context words given the target word.These word representations have attracted lots of research from the IR community these last years with new relevance models [22, 29, 31]. Going beyond the word level, some work proposes to learn distributed

representations of text such as sentences, paragraphs, or documents [25]. A simple but efficient approach consists in inferring the document representation from embeddings of its words. A more complex approach is inspired by neural language models [11, 12]. Following the CBOW and the skip-gram frameworks [16] respectively, the Siamese CBOW model [11] and the Skip-thought [12] learn sentence representations by either predicting a sentence from its surrounding sentences or its context sentences from the encoded sentence. As an extension of word2vec, the Paragraph-Vector model [13] jointly learns paragraph (or document) and word representations within the same embedding space. This joint learning relies on the compositional assumption underlying document representation [17, 25] leading to a mutual benefit for learning the distributional semantics of both documents and words.

Neural Approaches Empowered by Knowledge Sources for Learning Text Representations. Although distributed representations can efficiently model the semantics of words, using solely the document collection as knowledge evidence source does not allow to cope with three fundamental problems: 1) the readability of the captured word senses since the latter are not easily mappable to lexical sources leading to a limited usefulness [15]; 2) the polysemy problem since neural models fail to discriminate among different senses of a target word [10]; 3) the data sparsity problem since neural approaches based on the distributional hypothesis learn solely on corpus-based cooccurrences of words which prevents the learning of close word representations for semantically close words occurring in different word-based contexts. To tackle these problems, neural approaches investigated the joint use of both corpus-based word distributions and knowledge sources to achieve more accurate text representations [7, 14, 15, 27].

A first line of pioneer work [7, 14] have proposed to enhance the readability of the distributed representations of words learned from corpora by leveraging the *relational semantics* expressed in external knowledge sources. The intuition of those work is to bring semantically related words (via relations in a knowledge source) closer to each other in the vector space. For instance, the *retrofitting* method [7] leverages lexicon-derived relational information of words by minimizing both 1) the distance of each word with the representation of all connected words in the semantic graph and 2) its distance with the pre-trained word representation, namely its initial distributed representation.

The second and recent line of work aims at refining word embedding using relational constraints to better discriminate word senses by simultaneously learning the concept representations and inferring word senses, and accordingly tackling the polysemy issue [3, 4, 15, 21, 27]. Mancini et al. [15] simultaneously learn embeddings for both words and their senses via a semantic network based on the CBOW architecture. The originality of this work relies on the fact that words might be associated with multiple senses, allowing refining embeddings according to the polysemy issue. Unlikely, Cheng et al. [3] assume that polysemy can be captured through context words and therefore propose to compute parallel word-concept skip-grams for each context word by introducing their associated concept in the prediction. In the same mind, Yamada et al. [27] propose

a Named Entity disambiguation model that exploits word and concept embeddings learned in a two-step methodology. More particularly, word and concept latent spaces are first learned separately in skip-gram frameworks and then are aligned using word-concept anchors derived from the knowledge source.

There are two key differences between all these close previous work [3, 4, 15, 27] and ours. First, unlike these past approaches, we tackle the readability of word senses and polysemy problems by learning document representations that leverage semantics inventoried in both text corpora and knowledge sources through fine-grained elements including words and concepts in a joint learning process. Moreover, in contrast to [4] that considers the document context as a temporal feature directly injected in the objective function, we assume that there is a mutual benefit to learning simultaneously document, word, and concept embeddings to better capture the semantics at global and local levels. Second, we also tackle the data sparsity problem and show the quality of the learned representations of documents as well as related concepts and words used as auxiliary information to enhance the query-document matching while most of previous work focused on the polysemy problem within NLP tasks.

3 The Tripartite Neural Document Language Model

In this paper, we address the vocabulary mismatch, the granularity mismatch and the polysemy issues through two assumptions:

- *Multi-level context view (A1)*: we conjecture that each word conveys a unique sense within the same document with respect to a relevant concept in a knowledge source; however, a word could convey different senses and being polysemous across documents. Thus, simultaneously learning representations within a multi-level context (namely a global vs. local level for resp. document vs. word and concept contexts) allows embeddings better facing the polysemy issue.

- *Knowledge source-based context view (A2)*: constraining the learning of word-concept pairs with respect to a knowledge source structure allows obtaining close word embeddings for words sharing the same concept even if they occur in different contexts in the document. Thus, granularity mismatch is partially or completely solved based on the knowledge source context.

3.1 Neural Network Architecture

We propose a tri-partite neural document language model that jointly learns the representations of words, concepts, and documents with a prior provided by an external knowledge source. To do so, our model is an extension of the ParagraphVector model [13] which has the same generalizability property with respect to new documents. The objective function fits with: 1) assumption *A1* through component L_C which learns embeddings by making predictions from words and concepts that occur within the multi-level context; 2) assumption *A2* formalized through component L_R which regularizes the embeddings using the relational knowledge constraints. The resulting objective function is:

$$L = L_C + \beta L_R \tag{1}$$

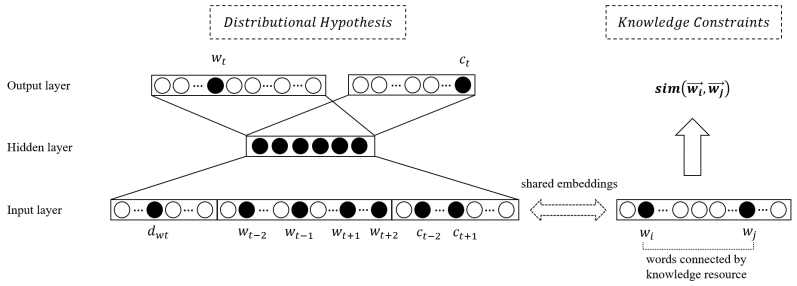

Fig. 1: Model architecture of our tri-partite neural document language model

where β is the combination coefficient which is experimentally set to the optimal value according to the validation set (Section 4.3). We detail next the training of the embeddings according to the multi-view context (L_C) and their regularization with respect to the knowledge source context (L_R). Formally, the training set consists of the set $\mathcal{S} = \{\mathcal{D}, \mathcal{W}, \mathcal{C}\}$, where $\mathcal{D}$ expresses the collection of documents d, viewed as sequences of ordered words w and ordered associated concepts c within their surrounding contexts; $\mathcal{W}$ is the word-based vocabulary of the document collection $\mathcal{D}$ and $\mathcal{C}$ is the set of concepts in the knowledge source $\mathcal{R}$ that provides knowledge about concepts and relations between concepts. Given document d, we use the automatic annotator TagMe [8] to identify the context-appropriate concept $c_i \in \mathcal{C}$, if any, associated to word $w_i \in d$ according to the mapping of its word-based surrounding context to the knowledge source R. Thus, each window considered in the model training is a sequence of words and their associated concepts (if any). We outline that in this work, we only consider single word-concept mapping within the source R and leave the mapping of multi-word concepts for future work. Figure 1 shows our learning framework on a simple training instance.

3.2 Network Training

Learning the word, concept, and document representations. In this work, we propose to extend the Distributed Memory version of the *Paragraph Vector* model [13] to learn document embeddings by jointly predicting each word-concept pairwise given their context in the document. In other words, document vectors v_d are learned so they allow predicting its belonging words and concepts, while word vectors v_w and concept vectors v_c are learned so they predict words and concepts in their surrounding context. Specifically, the objective of our joint document-word-concept training is to maximize this log-likelihood:

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

where the word sequence of document d is noted $\mathcal{W}s_d$, $w_{t \pm k}$ and $c_{t \pm k}$ refer respectively to word and concept contexts within a context window surrounding term w_t of size k, c_t is the most appropriate concept mapped to word w_t within

its context, $\frac{\gamma}{|d|}||v_d||^2$ is a $L2$ regularizer over the document vector v_d avoiding over-fitting the representation learning of long texts [1] with $|d|$ is the document length and γ is the regularization strength. The probability $p(w_t|w_{t\pm k}, c_{t\pm k}, d)$ of word w_t given its context is defined using a soft-max function:

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

where $\mathcal{W}$ is the word vocabulary of the collection, $\overline{h}_{w_t}$ is the representation of the context window taken by averaging the input vectors $v_.$ of the context words $w_{t\pm k}$ and their concepts $c_{t\pm k}$ including document d:

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

where the context window of size k includes $2k$ context words. Therefore, $4k + 1$ stands for the number of words and concepts ($+1$ for document d) in the extreme case where each word is mapped to a concept. Similarly, the probability $p(c_t|w_{t\pm k}, c_{t\pm k}, d_{w_t})$ is estimated as:

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

where $\overline{h}_{c_t}$ is the representation of the context window for concept c_t, estimated similarly to $\overline{h}_{w_t}$ (see Equation 4). With the large size of $\mathcal{W}$ and $\mathcal{C}$, Equations (3) and (5) become impractical. Following [16], we define the alternative objective functions by using the negative sampling strategy for each element $e_t \in \{w_t; c_t\}$:

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

$\sigma(x)$ is the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ and $\mathbb{E}_{e_i \sim P_n(e)}$ the expected value of $\log \sigma(-v'^\top_{e_i} \cdot \overline{h}_{e_t})$ when e_i is taken from the unigram distribution $P_n(e)$ [1].

Constraining the representation learning with a knowledge source structure. To address the granularity mismatch, we propose to capture relations between words which may not be (sufficiently) learned from the document context in the case where they do not (frequently) occur in the same contexts in documents, which is likely to be explained by data sparsity. Inspired by previous work [28], we equip the objective function with a regularization term which integrates the relational constraints from the knowledge source into word representations. The regularization will simultaneously adjust the word representations with the learning of documents in the training phase such that words that share the same concept or share related concepts have close embeddings. Formally, our objective is to maximize the similarity between any pair of words (w_i, w_j) according to the following objective function:

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

where $linkC(w_i, w_j) = 1$ if words w_i and w_j are associated to the same concept and $linkR(w_i, w_j) = 1$ if these words are associated to related concepts. $sim(w_i, w_j)$ is the cosine similarity between both word vectors v_{w_i} and v_{w_j}.

4 Experimental Design

The objective of our evaluation is twofold: 1) assessing the quality of document embeddings learned using our neural model and 2) measuring the impact of the learned representations on the effectiveness of IR tasks. The source code of our model and the learned embeddings will be available at `https://cloud.irit.fr/index.php/s/NQqk8fgZI7lIIGp`.

4.1 Dataset

We use the Robust04 collection[3] which is the standard news dataset used in the standard evaluation challenge TREC Robust Track 2004 including 528,155 documents and 250 topics. The title of each topic has been collected to build the set of queries. To enhance the representations with relational semantics, we exploit DBpedia as knowledge source due to its large coverage. Queries and documents are annotated by *TagMe*[4] [8], a publicly available state-of-the-art annotation tool for linking text to DBpedia entities. We use the names of DBpedia base entities to annotate the queries and documents and exploit the `gold:hypernym` relation. For the sake of simplicity with respect to the model description in Section 3, we refer to entities by concepts. The annotation of the Robust04 collection results in 1 to 3 concept-length, with 1 concept in average and documents with 31 concepts in average (over 488 words in average).

4.2 Evaluation Methodology

We evaluate our proposed tri-partite model according to three scenarios:

- **PV**: which refers to the Paragraph-Vector Model [13] from which we build our extended neural model. This scenario learns word and document representations without using any evidence from an external knowledge source.
- **S2DV**: this scenario learns document, word, and concept representations by using concepts from a knowledge source as formulated in the component L_C of the objective function L (Equation 1). But, this setting ignores the relationships established between concepts, and so, skips the regularization component L_R.
- **S2DVR**: our full proposed learning model that learns document, word, and concept representations by using both concepts and their relationships established in a knowledge source as formulated in the full objective function L.

Moreover, with respect to the experimental objectives mentioned above, we use two evaluation frameworks detailed below.

Evaluating the quality of document embeddings. Considering the primary objective of our model which consists in learning document representations,

[3] http://trec.nist.gov/data/robust/04.guidelines.html: the dataset is available for the scientific community under acceptance of a license agreement.

[4] https://tagme.d4science.org/tagme/

we first evaluate the quality of the learned document embeddings. To achieve this goal, we use the document similarity task described in the pioneer work of Mikolov et al. [13] which consists in discriminating the similarity of documents with respect to a target query. More specifically, for each query in the dataset, we create a pool of document triplet in which the two first ones are retrieved from a state-of-the IR model according to this query and the third document is randomly sampled from document rankings with respect to other queries. The underlying objective is to measure in which extent the document similarity metric (namely the standard cosine similarity) estimated using learned document representations allows to provide a more important similarity for documents issued from the same target query and a smaller similarity for documents issued from other queries. Similar to [13], we use the error rate over all the queries measuring when representations give smaller similarity for the first two documents than the third one. Obviously, the lower the error rate is, the more effective the document representation is. To evaluate the quality of our document embeddings, we compare the error rates obtained using the embeddings provided by our model to those obtained using the following document representations:

• **TF-IDF** which refers to the traditional document modeling in IR in which documents are represented through a word vector weighted using the Tf-Idf schema. This baseline aims at measuring the effect of using distributional semantics on the quality of the embeddings.

• **AWE** [25] which builds document embeddings by averaging the embedding of its words. The goal behind the comparison with this representation is to evaluate the impact of considering a multi-level context (namely concepts and documents in addition to words) on the quality of the embeddings.

Evaluating the effectiveness of embeddings within IR tasks. To evaluate the effectiveness of the obtained embeddings on IR performance, we propose two types of IR models in which those embeddings are injected. Performance effectiveness of these models is measured using standard metrics: the Mean Average Precision (MAP) and the Recall at rank 1000.

• *Document re-ranking.* This type of model consists in enhancing a basic document relevance score with an additional score based on an external evidence. To inject the learned embeddings, we combine a traditional document relevance score with a similarity score computed between the query and the document embeddings. We specifically use the model proposed in [14]:

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

where α is a combination parameter tuned using a two-fold cross-validation according to the MAP metric, $IRScore$ is the document score obtained using a traditional IR model, namely BM25, and $NeuralScore$ is the cosine similarity between the query and the document representations. While document embeddings are learned using our framework, the query embeddings are considered as "unseen documents" for which the representation is inferred from the learned model, as done in the ParagraphVector model [13].

• *Query expansion.* This type of model consists in rewriting the initial query by exploiting an external evidence. In our setting, we use evidence issued from

relevant words and/or concepts based on the assumption that relevance could be captured by computing similarities between query embeddings in one side and word/concept embeddings in the other side. To do so, we rely on the state-of-the art model proposed in [29] in which queries are expanded using each element e (namely words and/or concepts) with the highest neural similarity score $p(e|\hat{q})$:

$$p(e|\hat{q}) = \frac{\sigma(\hat{e}, \hat{q})}{Z} \tag{9}$$

where $\hat{q}$ and $\hat{e}$ are respectively the embeddings of query q (learned as explained above) and word/concept element e, $\sigma(\cdot, \cdot)$ denotes the exponential cosine similarity of two vectors and Z is the normalization factor calculated by summing $\sigma(\hat{e'}, q)$ over all terms e' in the vocabulary (namely all words over all documents or all concepts extracted from all words). Then, this neural probability is linearly interpolated with the maximum likelihood estimation $p_{mle}(e|q)$ of the original query (namely, term-based count probability) as follows:

$$p(e|q*) = \alpha p_{mle}(e|q) + (1 - \alpha)p(e|\hat{q}) \tag{10}$$

The top m elements (words and/or concepts) with the highest probabilities $p(e|q*)$ are used to expand the initial query.

For comparative effectiveness purpose, we inject the learned representation obtained using the PV, S2DV, and S2DVR scenarios within each of the models described above. In addition, we compare the effectiveness of those models to a traditional baseline IR model that does not rely on a neural approach. To keep fair comparison, we choose a semantic baseline IR model, noted **LM-QE** [20]. The latter performs a language-based query expansion with semantically related concepts. Using such baseline additionally ensures the comparability of results with scenarios S2DV and S2DVR since all these models are likely faced to the problem of word sense disambiguation that could degrade retrieval performance as already shown in past work [19].

4.3 Experimental Setting

For the distributional-based model configurations (PV, SD2V, SD2VR), we set the dimension of embeddings to 300 and empirically select the window size $k = 8$. After removing non-alphanumeric words, we only keep words with frequency in the corpus higher than 5. The initial learning rate is set to 0.02 and decreased linearly during the SGD training process. We use the negative sampling technique where the negative sample is set to 5. The β parameter in Equation 1 is set up to 10^{-5}. We test the regularization strength γ in Equation 2 from 0.1, 1, and 10 as suggested in [1], the best performance is obtained with $\gamma = 0.1$. In practice, it is worth mentioning that since our model is based on the ParagraphVector learning mechanism, the integration of concepts in the input vector simply increases the training time linearly function of the vocabulary size. The complexity of the model is likely impacted by the regularization term. The inference to new documents or queries is not time-consuming. All the retrieval models are performed using the Indri[5] search engine.

[5] https://www.lemurproject.org/indri.php

Table 1: Comparative results for the document similarity task measuring the quality of our document embeddings. %Chg: error rate reduction w.r.t. SD2VR.

Model	Error rate	%Chg
TF-IDF	7.2%	-12.5%
Avg-WE	9.6%	-34.37 %
PV	7.9%	-20.25 %
SD2V	8.3%	-24.09 %
SD2VR	6.3%	

5 Results

5.1 Analyzing the Quality of Document Embeddings

We analyze here the quality of document embeddings using the document similarity task described in Section 4.2. Table 1 illustrates the obtained results in terms of error rate for each scenario (PV, SD2V, SD2VR) in comparison to each baseline (TF-IDF and AWE). From a general point of view, we could say that our full model *S2DVR* obtains better results than all the scenarios and baselines. We can observe that using relational information between concepts for learning document embeddings lowers the error rate of -24.09% (from 8.3% for the *SD2V* scenario to 6.3% for the *S2DVR* one). One could infer from this observation that there is a synergic effect for representing documents while learning jointly the implicit relations between words in the text and the explicit relations between its associated concepts as inventoried in a knowledge source. This statement clearly argues toward the effectiveness of our model to cope with both the polysemy and granularity mismatch problems. Second, we can see that learning document embedding by leveraging both concepts and relations allows building document representations better suited for capturing the document semantics. Specifically, by comparing our best scenario (*SD2VR*) with respect to the different baselines, we could suggest the following statements:

• Our full model *SD2VR* decreases the error rate of -12.5% with respect to the *TF-IDF* baseline. This result is consistent with previous work [13] that argues toward the benefit behind learning document representations by leveraging the distributional semantics. We can also see that the error rate for the scenario *SD2VR* based on a multi-level of distributional semantics (6.3%) is lower than the one obtained by the *AWE* baseline (9.6%) which estimates document representations at the word level. This result is also consistent with prior work [13]. In this spirit, we show that the error rate obtained by the PV baseline (7.9%) is lower than the one obtained by the *AWE* baseline (9.6%).

• In addition to these findings, we can see that our full model *SD2VR* scenario allows to drastically decline the error rate of -20.25% with respect to the PV baseline. This confirms our intuition about the benefit of integrating the relational semantics inventoried in knowledge sources while learning distributed representations of documents. We further this quantitative analysis with a comparative qualitative analysis between PV and *SD2VR*. The results highlighted from a 2D-visualization of queries and document embeddings obtained using both models corroborate the previous statements. Figure 2 shows an example

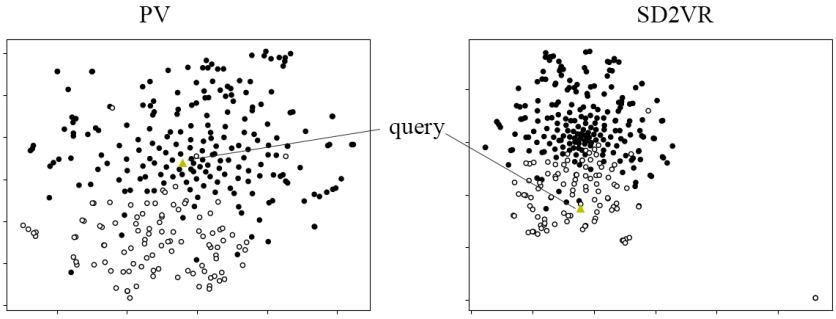

Fig. 2: A t-SNE projection of document embeddings issued from the *PV* (left side) and the *SD2V* models (right side) for the query 443 (yellow triangle). Relevant documents are white, irrelevant documents are black.

of query output, namely query 443, and its document set including the ground truth and other irrelevant documents. We can see that for both models, there is a distinction between two clusters of documents (the cluster of relevant documents in white, the cluster of irrelevant documents in black). However, when looking at the *SD2VR* model, we can see that the query is better located within the relevant cluster, in comparison to the *PV* model, where the query is centered between both clusters with a trend toward the cluster of irrelevant documents.

5.2 Evaluating the effectiveness of learned embeddings in IR tasks

Table 2 presents the results obtained for document re-ranking and query expansion tasks according to the different scenarios (PV, SD2V, SD2VR) and the retrieval baseline (LM-QE). Below, we discuss the results obtained within each of both tasks and then we conduct a cross-analysis of the main emerging results.

Document re-ranking. From a general point of view, we can see from Table 2 two main statements:

• Our full model scenario (SD2VR) significantly overpasses the semantic IR model *LM-QE* with an improvement rate reaching +17.73%. This suggests, according to formula 8, that the injection of the neural similarity scores in the relevance document score computation enhances the ranking performance. By comparing to the LM-QE baseline, we can conjecture that this is probably due the use of the deep semantic representation of documents. This is confirmed by the results obtained using the *PV* scenario compared to the LM-QE baseline. However, as we can see, considering additional evidence issued from the knowledge source seems to hinder the document re-ranking performance since the performance results achieved with both *SD2V* and *SD2VR* are lower than the *PV* scenario. A deep analysis of this observation is reported in the cross-analysis.

• The comparison of our both scenarios *SD2V* and *SD2VR* shows that the full version of our model *SD2VR* slightly increases the search effectiveness. Similarly to findings risen from the analysis of the document similarity task, this result suggests that relations in knowledge sources provide useful relational knowledge for enhancing the quality of learned word representations.

Table 2: Effect of the embeddings on retrieval effectiveness for both IR tasks in terms of MAP and Recall. Bold values express results higher than baselines.

	IR Models		MAP	Recall
Semantic IR baseline	LM-QE		0.2110	0.6593
Document re-ranking	PV		0.2507	0.6895
	SD2V		0.2379	**0.6834**
	SD2VR		0.2384	**0.6841**
Query expansion	PV	word	0.2460	0.6804
	SD2V	word	0.2443	**0.6891**
		concept	**0.2497**	**0.6897**
		both	**0.2461**	**0.6894**
	SD2VR	word	0.2451	**0.6886**
		concept	**0.2516**	**0.6892**
		both	0.2489	**0.6890**

Query expansion task. Table 2 presents the retrieval performance results obtained within the query expansion task by considering three configurations (See formula 9-10): expanding with words only, expanding with concepts only, expanding with both words and concepts. As can be seen from Table 2, the comparison of the performance results achieved using our proposed model (either in terms of MAP and Recall) outlines: 1) the superiority of the neural model scenarios (*PV* (0.2460), *SD2V* (0.2497), *SD2VR* (0.2516)) over the semantic baseline IR model *LM-QE* (0.2110) with respect to all query expansion configurations. This result highlights the ability of the word/concept embedding-based similarity to select good expansion elements, likely due to the quality of the embeddings themselves; 2) expanding queries with 'concepts only' seems to be the most successful retrieval scenario. More particularly, we observe that expanding queries with concepts identified on the basis of concept embeddings used in our both scenarios (SD2V, SD2VR) overpasses the *PV* scenario which is the strongest baseline in both evaluation tasks. For instance, the query 683 "Czechoslovakia breakup" is expanded with terms "use_chapel" and "targy" for the *PV* setting while the SD2VR setting allows to extend the query with concepts $\$\#!Czechoslovak\_Socialist\_Republic$ and $\$\#!Dissolution\_of\_Czechoslovakia$ which are more related to the query topic. This result could be considered as consistent with previous work [24] that argue toward the joint use of words and concepts to perform effective query expansion since, in our proposed model, concept embeddings are learned in a joint learning process of words and concepts.

Cross-analysis and discussion. One general result that we can infer from the above experiments is that retrieval performance depends heavily on the nature of the embeddings exploited and/or the nature of the retrieval task. This finding is consistent with the general feeling in the IR community that points on the variability of performance levels of semantic IR models [23]. More specifically, the analysis of the performance results of document re-ranking and query expansion tasks reveals that the embeddings learned using our tri-partite neural

Table 3: Qualitative analysis of search effectiveness for the re-ranking task with respect to the document length criteria in terms of concept number.

Query set	*#queries*	*Avg_concept_qrels*	*Avg_concept_docs*
Q^-	117 (46.99%)	78.56	66.80
$Q^=$	7 (2.81%)	69.92	50.24
Q^+	125 (50.20%)	64.77	62.25

Table 4: Examples of (top ranked document-query pairs belonging to $\mathcal{Q}^+$ and $\mathcal{Q}^-$. Terms in bold are query terms

Q+ Q412 airport security	
Document FT941-4175 34 concepts relevant	...In spain \$#!*Spain*, another european union country \$#!*Nation_state* facing terrorist campaign, only armed police \$#!*Police* have responsibility \$#!*Moral_responsibility* for **security** \$#!*Security* in **airports**. In Heathrow \$#!*Heathrow_Airport*, since BAA \$#!*Heathrow_Airport_Holdings* was privatised, [...]
Q- Q314: Marine Vegetation	
Document LA091189-0098 60 concepts irrelevant	[...] the **marine** \$#!*Ocean* craft, the same color \$#!*Color* as surrounding **vegetation**, was not easy to spot, singley said. [...]Because the plane route was not known and radar \$#!*Radar* was unable to track it, **Marine** officials relied upon civilian reports in the search \$#!*The_Search_*(2014_*film*) [...]

model are likely to be more effective to capture auxiliary knowledge to enhance the query representation than to improve a query-document relevance score. At a first glance, this observation could be explained by the fact that query expansion leverages the complete multi-level context including words, concepts, and documents, while the document re-ranking only leverages document representations. In addition, although document representations are used for both tasks, the document latent space serves to learn the representation of short queries in the query expansion task while this space serves to learn long documents in the document re-ranking task. Our intuition is that the accuracy of the alignment performed by our model between document vectors to both word and concept vectors might be sensitive to the disambiguation error introduced in the concept-based document annotation stage. To get a better insight on this intuition, we studied the relationship between the length of both documents/queries, in terms of concept numbers, and performance. This study revealed that only the document level is significant. More particularly, we performed a qualitative analysis aiming at measuring to what extent documents with higher number of concepts are less likely to be selected by our knowledge-based retrieval model, and most specifically by the document re-ranking model. We first identify query sets for which our best model *SD2VR* performs worse (Q^-), identically ($Q^=$), or better (Q^+) in terms of MAP than the PV model (with a margin ranging between $+/-5\%$). This baseline is particularly interesting since it does not involve concepts and so, abstracts the problem of word sense disambiguation. Second, we analyze the average number of identified concepts in relevant documents, namely the ground truth, (noted *Avg_concept_qrels*) and in top selected documents (noted *Avg_concept_docs*). Table 3 presents the obtained results. We can

see that for worse queries (Q^-) the number of concepts in documents belonging to the ground truth is higher than the one for other query sets (namely 78.26 vs. 69.92 for $Q^=$ and 64.77 for Q^+). This suggests that our model is less able to catch the semantics of documents including a high number of concepts. To get a better insight on this phenomena, we depict in Table 4 an example of query extracted from both query sets Q^+ and Q^- and one associated top retrieved document obtained using our SD2VR scenario. We can see that the document retrieved by the query extracted from Q^+ is annotated with a few concepts that are semantically close to the query topic than those identified in the document retrieved for the query belonging to Q^- that are less-topically focused ($\$\#!Color$) or erroneously annotated ($\$\#!The\_Search\_(2014\_film)$). This observation corroborates our possible explanation related to the relationship that might exist between 1) the improper alignment of the document representations with word/concept vectors during the representation learning process, and 2) disambiguation error rate; particularly for documents that entail a high number of concepts. However, further investigation is needed.

6 Conclusion

In this paper, we introduce a new neural tri-partite document model powered with evidence issued from external knowledge sources to overcome the crucial semantic gap issue in IR. The key idea is to leverage explicit relational semantics, namely concepts and relations, provided in knowledge sources to enhance distributional-based document representations that could be injected in a retrieval model. The framework extends the ParagraphVector model by jointly learning document, word, and concept representations in a same distributional semantic space. The experimental evaluation shows the effectiveness of our framework for different IR settings. An interesting future work would be the generalization of the representation learning to multi-word concepts through compositional neural representations. The analysis of the sensitivity of these representations to disambiguation errors introduced by word sense disambiguation algorithms would also be worth of interest.

References

1. Ai, Q., Yang, L., Guo, J., Croft, W.B.: Analysis of the paragraph vector model for information retrieval. In: ICTIR. pp. 133–142. ACM (2016)
2. Bengio, Y., Schwenk, H., Senécal, J.S., Morin, F., Gauvain, J.L.: Neural probabilistic language models. In: Innovations in Machine Learning (2006)
3. Cheng, J., Wang, Z., Wen, J.R., Yan, J., Chen, Z.: Contextual text understanding in distributional semantic space. In: CIKM. pp. 133–142 (2015)
4. Choi, E., Bahadori, M.T., Searles, E., Coffey, C., Sun, J.: Multi-layer representation learning for medical concepts. KDD pp. 1495–1504 (2016)
5. Corcoglioniti, F., Dragoni, M., Rospocher, M., Aprosio, A.P.: Knowledge extraction for information retrieval. In: ESWC. vol. 9678, pp. 317–333 (2016)
6. Efthymiou, V., Hassanzadeh, O., Rodriguez-Muro, M., Christophides, V.: Matching web tables with knowledge base entities: From entity lookups to entity embeddings. In: d'Amato, C., Fernandez, M., Tamma, V., Lecue, F., Cudré-Mauroux, P., Sequeda, J., Lange, C., Heflin, J. (eds.) ISWC 2017. pp. 260–277 (2017)

7. Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A.: Retrofitting word vectors to semantic lexicons. In: NAACL
8. Ferragina, P., Scaiella, U.: Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: CIKM. pp. 1625–1628. ACM (2010)
9. Harris, Z.S.: Distributional structure. Word **10**(2-3), 146–162 (1954)
10. Iacobacci, I., Pilehvar, M.T., Navigli, R.: Sensembed: Learning sense embeddings for word and relational similarity. In: ACL. pp. 95–105 (2015)
11. Kenter, T., Borisov, A., de Rijke, M.: Siamese cbow: Optimizing word embeddings for sentence representations. In: ACL 2016. pp. 941–951 (2016)
12. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: NIPS. pp. 3294–3302 (2015)
13. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: ICML. pp. 1188–1196 (2014)
14. Liu, X., Nie, J.Y., Sordoni, A.: Constraining word embeddings by prior knowledge–application to medical information retrieval. In: Information Retrieval Technology. pp. 155–167. Springer (2016)
15. Mancini, M., Camacho-Collados, J., Iacobacci, I., Navigli, R.: Embedding words and senses together via joint knowledge-enhanced training. In: CoNLL (2017)
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
17. Mitchell, J., Lapata, M.: Vector-based models of semantic composition. In: ACL. pp. 236–244 (2008)
18. Moreno, J.G., Besançon, R., Beaumont, R., D'hondt, E., Ligozat, A., Rosset, S., Tannier, X., Grau, B.: Combining word and entity embeddings for entity linking. In: ESWC. pp. 337–352 (2017)
19. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. **41**(2), 10:1–10:69 (Feb 2009)
20. Pal, D., Mitra, M., Datta, K.: Improving query expansion using wordnet. Journal of the Association for Information Science and Technology **65**(12), 2469–2478 (2014)
21. Rastogi, P., Poliak, A., Durme, B.V.: Training relation embeddings under logical constraints. In: KG4IR@SIGIR (2017)
22. Rekabsaz, N., Mitra, B., Lupu, M., Hanbury, A.: Toward incorporation of relevant documents in word2vec. In: Neu-IR@SIGIR (2017)
23. Richardson, R., Smeaton, A.F.: Using wordnet in a knowledge-based approach to information retrieval (1995)
24. Trieschnigg, D.: Proof of Concept: Concept-based Biomedical Information Retrieval. Ph.D. thesis, University of Twente (2010)
25. Vulić, I., Moens, M.F.: Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: SIGIR. pp. 363–372. ACM (2015)
26. Xiong, C., Callan, J.: Query expansion with freebase. In: ICTIR. ACM (2015)
27. Yamada, I., Shindo, H., Takeda, H., Takefuji, Y.: Joint learning of the embedding of words and entities for named entity disambiguation. pp. 250–259 (2016)
28. Yu, M., Dredze, M.: Improving lexical embeddings with semantic knowledge. In: ACL. pp. 545–550 (2014)
29. Zamani, H., Croft, W.B.: Estimating embedding vectors for queries. In: ICTIR. pp. 123–132. ACM (2016)
30. Zhao, R., Grosky, W.I.: Narrowing the semantic gap-improved text-based web document retrieval using visual features. IEEE transactions on multimedia **4**(2) (2002)
31. Zuccon, G., Koopman, B., Bruza, P., Azzopardi, L.: Integrating and evaluating neural word embeddings in information retrieval. In: ADCS. p. 12. ACM (2015)