Constraining Word Embeddings by Prior Knowledge – Application to Medical Information Retrieval

Xiaojie Liu, Jian-Yun Nie(✉), and Alessandro Sordoni

DIRO, University of Montreal, CP. 6128, succursale Centre-ville, Montreal, QC H3C3J7, Canada
{xiaojiex, nie, sordonia}@iro.umontreal.ca

Abstract. Word embedding has been used in many NLP tasks and showed some capability to capture semantic features. It has also been used in several recent studies in IR. However, word embeddings trained in unsupervised manner may fail to capture some of the semantic relations in a specific area (e.g. healthcare). In this paper, we leverage the existing knowledge (word relations) in the medical domain to constrain word embeddings using the principle that related words should have similar embeddings. The resulting constrained word embeddings are used to rerank documents, showing superior effectiveness to unsupervised word embeddings.

1 Introduction

Continuous word representations, called word embeddings, have known widespread uses in general NLP tasks [4, 6, 15, 17, 26, 27]. They offer an effective and efficient way of encoding semantic/syntactic relationships between words in semantic space, which typically relies on the distributional hypothesis that two words sharing similar contexts should be associated with similar vectors in the embedding space. Word embedding, and more generally, deep learning, has also been used in IR in recent years for different tasks: to suggest or to reformulate queries [16, 20], to extend language models [8, 24], or to determine a similarity score between queries and document titles [10, 22], questions and short answers [23] or queries and terms [29]. Although it is possible to optimize a deep network directly for the ad hoc search task as in [10, 12], this would require a large set of training data (e.g. clickthrough), which is not always available. An alternative approach is to train word embeddings on a document collection in an unsupervised manner. Word embeddings trained in this way may reflect some general syntactic or semantic relations between words in a language such as between "cat" and "kitten", but fail to capture some valid relations between words, which may have been established manually. For example, word embeddings trained on a medical collection fail to capture the strong relationship between *heart* and *cor* (a strongly related word used in prescriptions), while this relationship has been specified in the domain resource UMLS [3]. It is natural to leverage the knowledge to constrain or to adjust word embeddings so as to better fit the specific application domain.

S. Ma et al. (Eds.): AIRS 2016, LNCS 9994, pp. 155–167, 2016.
DOI: 10.1007/978-3-319-48051-0_12

The principle we use in this paper to constrain word embeddings is that related words in our prior knowledge (e.g. synonyms) should have similar embeddings.

The idea of using prior knowledge to constrain word embeddings has been used in several recent studies in NLP [4, 7, 27]. In this paper, we adapt these approaches to medical IR, and evaluate them on several test collections - OHSUMED [11] and CLEF [9, 18]. The contributions of this paper are as follows: We propose modified constrained training methods for word embeddings and show that they can bring more improvements to MIR than the original word embeddings.

The rest of the paper is organized as follows. Section 2 gives an overview of word embedding. Sections 3 and 4 present our approach to constrain word embeddings and to document reranking. Section 5 describes our experiments and analyses. Section 6 goes through the related work and Sect. 7 presents the conclusion and future work.

2 Word Embedding

In this section, we describe the standard and regularized word embeddings.

2.1 Continuous Bag-of-Words (CBOW)

Proposed by Mikolov et al. [15], the word2vec models create a vector representation for a word according to the context words frequently appearing around it. In this section, we will only describe one of the word2vec models – CBOW, which minimizes the following objective loss function:

$$L = -\sum_{t=1}^{T} \log p(w_t|w_{t\pm k}), \tag{1}$$

where T is the total number of words in the corpus and $w_{t\pm k}$ are the words in the window of size k centered at position t and excluding w_t. The probability of a word given its context is defined as:

$$p(w_t|w_{t\pm k}) = \frac{\exp(w_t^T \mathrm{c})}{\sum_{v\in V} \exp(w_v^T \mathrm{c})}, \mathrm{c} = \sum_{j=t-k, j\neq t}^{t+k} w_j, \tag{2}$$

where the context embedding c is simply the sum of the embeddings of words occurring in the text window.

2.2 Regularized Word Embedding

Several approaches have been proposed in recent years to constrain (regularize) unsupervised word embeddings, and we describe two approaches below.

Online Training Approach. Online training approaches alter the learning objective in word embedding estimation by adding a knowledge-based regularization term [4, 26–28]. We only describe the approach by Yu and Dredze [27]. The modified loss function is as follows:

$$L = -\frac{1}{T}\sum_{t=1}^{T} \log p(w_t|w_{t\pm k}) - \frac{C}{|R|}\sum_{(w_i,w_j)\in R} \log p\left(w_i|w_j\right), \tag{3}$$

where $(w_i, w_j) \in R$ means that that two words are linked in the resource R, $|R|$ is the number of links in R, and C is a hyper-parameter controlling the strength of the regularization. Similarly to Eq. (2), the probability $p(w_i|w_j)$ is proportional to the dot product between w_i and w_j. Therefore, the regularizer sums up a similarity measure over all pairs of related words in the resource.

We observe two shortcomings of this approach. First, any pair of linked words in the resource is considered to be a constraint of equal importance (1/|R|) in the regularization. Intuitively, however, a more frequent link (or a link between two frequently used words) should play a more important role in the regularization. Second, as the two terms in the objective function sum over different elements – words in the corpus and links in the resource, Yu and Dredze have to define two sets of separate learning parameters, one for the CBOW objective and another for the regularization, which are updated separately in turn. This means that when updating the parameters of the regularization, the context of a word (considered in the first term) is no longer taken into account. The risk of this process is that the second update could undo the earlier update, making the update process quite random at the end. In this paper, we propose a solution to these problems.

Offline (Retrofitting) Approach. Offline approaches (also called retrofitting) [7] adjust word embeddings outside the original training process as follows: the new embeddings should be close to the original embeddings and respect the constraints of the external resource, i.e. minimize:

$$L = \frac{1}{V}\sum_{v=1}^{V} \left|w_v' - w_v\right|^2 + \frac{\beta}{|R|}\sum_{(w_i,w_j)\in R} \left|w_i' - w_j'\right|^2, \tag{4}$$

where w_v and w_v' are the original and the new embeddings and β a parameter.

3 Constrained Word Embedding

We propose modifications to solve the problems discussed above. A tighter regularization is used in the online method: the original CBOW cost function is combined with the requirement that if a word can be well generated from a given context, its related word should also be well generated from the same context, i.e.:

$$L = \sum_{t=1}^{T} \frac{1}{|R_t|} \sum_{w_s:(w_t,w_s)\in R} [\log p(w_t|w_{t\pm k}) - \log p(w_s|w_{t\pm k})]^2 \quad (5)$$

where $|R_t|$ is the number of words related to w_t in the resource.

A possible drawback of the above formulation is that every related word is attributed an equal weight ($1/|R_t|$). To solve this problem, we weigh each related word w_s by its relative frequency in the document collection as follows:

$$wt(w_s|w_t) = f(w_s) \Big/ \sum_{(w_t,w)\in R} f(w). \quad (6)$$

where $f(w_s)$ is the frequency of w_s in the collection. The final loss function is defined as follows:

$$L = -\sum_{t=1}^{T} [\log p(w_t|w_{t\pm k}) - \alpha \sum_{w_s:(w_t,w_s)\in R} wt(w_s|w_t)[\log p(w_t|w_{t\pm k}) - \log p(w_s|w_{t\pm k})]^2] \quad (7)$$

where α is a weighting parameter.

The above loss function solves both problems of [27]: the collection frequency of words in a relation is taken into account naturally, and the embeddings for related words are tightly related to their contexts.

We also propose a slightly modified version of retrofitting method by adding term weighting in it:

$$L = \sum_{v=1}^{V} [\left|w_v' - w_v\right|^2 + \beta \sum_{(w_v,w_s)\in R} wt(w_s|w_v)\left|w_v' - w_s'\right|^2] \quad (8)$$

As we will see in our experiments, our modified models can outperform the original regularized embeddings in MIR.

4 Using Constrained Embeddings for MIR

Many resources exist in the medical domain. In this paper, we use UMLS Metathesaurus [3], which is the largest resource in this area. It integrates hundreds of thesauri of different sub-domains in a uniform framework. Each concept (identified by a CUI – Concept Unique Identifier) in UMLS contains a set of expressions, which we use as synonyms. For example, the CUI C0018681 contains the expressions: {*heart, cor, hearts, cardiac, heart nos, heart structure*}. There are more types of relations defined in UMLS, but we only use synonymy relations in this paper. In addition, we only consider single-word concept expressions (i.e. *heart, cor, hearts, cardiac*), and leave

multi-word expressions to future work. This results in 302,323 synonymy relations between single words from UMLS.

Once word embeddings are trained, one faces the problem of building a representation for the whole document or query. We use a simple approach commonly used in this area, by summing up all the word embeddings in the document or the query. Cosine similarity is used to measure the similarity between the document and query embeddings. This approach is similar to that used in [15, 23, 24]. We notice, however, that a simple sum will make the global embedding of a document tuned towards frequent words which are not discriminative for IR. Therefore, we use the traditional IDF weighting to weight the embedding of a word.

Word embeddings are too noisy to be used alone to rank documents. In this paper, we use them in a re-ranking approach: we first retrieve a set of 1000 documents using a traditional baseline method (BM25 or language model); then, the results are re-ranked by the following re-ranking function:

$$s(Q,D) = \gamma BOW(Q,D) + (1-\gamma)Cosine(Q,D) \tag{9}$$

where γ is a hyper-parameter of our model, *BOW* is the score of a bag-of-word method such as BM25 or LM (language model); and *Cosine* is the cosine similarity between the query and the document embeddings. Both *BOW* and *Cosine* scores are normalized as follows:

$$NormScore = (Score - MinScore)/(MaxScore - MinScore) \tag{10}$$

where *MaxScore*, *MinScore* are the maximum and minimum scores in the list, *Score* and *NormScore* are the non-normalized and normalized scores of a document.

5 Experiments

5.1 Test Collections

The experiments are performed on the following test collections: OHSUMED [11] and CLEF-eHealth 2014 [9] and 2015 [18]. We use short queries (title field). Table 1 shows some statistics of the collections.

Table 1. Statistics of test collections

Corpus	Number of queries	Number of documents	Size
OHSUMED	106	348,566	294M
CLEF2014	50	1,095,082	6.5G
CLEF2015	66	1,095,082	6.5G

We use P@10 as the main performance indicator, and MAP and NDCG@10, which are often used on these collections, as the second indicators for OHUMED and CLEF. Two-tailed t-test ($p < 0.05$) is performed for statistical significance.

5.2 Word Embedding Training

In our experiments, we use CBOW model and negative sampling [15] to train the basic word embeddings. The CBOW program is then modified to incorporate the constraints as in Eq. (7). For all the methods tested, we set the dimension of embedding to 300, the context window size (k) to 5. This setting is common in word embedding [15] and has been shown to be reasonable in [30]. We choose 10 negative samples and we filter out words appearing less than 5 times in the collection. The collections are not preprocessed before embedding training, i.e. no stemming and stopword removal. Our intuition is that stopwords could provide useful context information for word embeddings. However, this remains to be confirmed. After training, our embedding vocabulary size is 164,434 for OHSUMED and 3,989,059 for CLEF.

5.3 Retrieval Results

BM25 (with the default setting) and LM (language model with Dirichlet smoothing with $\mu = 2000$) are used as the basic retrieval methods to retrieve 1000 candidates for reranking. In order to test the effectiveness of CBOW, we also use the standard CBOW model alone (i.e. γ in Eq (9) is set to 0). The original and modified online and offline constrained word embeddings are used to rerank the documents as in Eq. (9). We use 2-fold cross-validation to set hyper-parameters (α, β, Υ) of the models for each collection. We report the performance of different methods in Table 2.

We observe that the traditional CBOW alone (line **c**) leads to poor retrieval effectiveness. This could be explained by the noisy nature of word embedding for a whole document. However, when it is combined with a traditional IR method (**d** and **e**), we observe significant improvements. Similar observations have been made in [30].

Next, we observe that our online method (lines **g** and **i**) outperforms significantly CBOW and Yu's method when combined with BM25 or LM. This confirms that the constraints imposed by UMLS relations are helpful in training better word embeddings for MIR. We also see that the method of Yu does not always produce better results than CBOW, and the differences between Yu and CBOW are not statistically significant. This result could be explained by our earlier observation that the loosely tied constraint used by Yu does not necessarily lead to better word embeddings.

Retrofitting has shown better performance in several NLP tasks [7] than the method of Yu and Dredze. This is also confirmed in our results (lines **j** and **l** vs. lines **f** and **h**). However, the differences are not statistically significant. Our modified offline method (**k** and **m**) makes larger improvements. The differences with the original CBOW are statistically significant on CLEF collections. The only change between the original retrofitting (**Faruqui**) and our modified version (**Offline**) is the weighting of embeddings we added. This suggests the usefulness of embedding weighting in IR.

Table 2. Retrieval results of different methods (Significant difference with a method is marked by a letter corresponding to that method)

	OHSUMED		CLEF2014		CLEF2015	
	P@10	MAP	P@10	DCG@10	P@10	NDCG@10
(a) BM25	0.4390	0.2922	0.6720	0.6876	0.3561	0.3217
(b) LM	0.3752	0.2325	0.7280	0.7200	0.3712	0.3276
(c) CBOW ($\gamma = 0$)	0.1631	0.0401	0.0490	0.0596	0.0530	0.0616
(d) CBOW + BM25	0.4610[a]	0.2986	0.7056[a]	0.7085[a]	0.3727[a]	0.3461[a]
(e) CBOW + LM	0.4438[b]	0.2745[b]	0.7470[b]	0.7327[b]	0.3909[b]	0.3560[b]
(f) Yu + BM25	0.4600	0.2990	0.7120	0.7060	0.3682	0.3460
(g) Online + BM25	**0.4771[df]**	**0.3005**	0.7315[df]	0.7320[df]	0.3864[df]	0.3647[df]
(h) Yu + LM	0.4467	0.2778	0.7490	0.7340	0.3909	0.3557
(i) Online + LM	0.4581[eh]	0.2793	0.7580[eh]	0.7460[eh]	0.4086[eh]	0.3682[eh]
(j) Faruqui + BM25	0.4695	0.3001	0.7200	0.7250	0.3818	0.3593
(k) Offline + BM25	0.4715[d]	0.3001	0.7296[dj]	0.7300[d]	0.3848[d]	0.3596[d]
(l) Faruqui + LM	0.4470	0.2778	0.7520	0.7420	0.3955	0.3665
(m) Offline + LM	0.4486	0.2781	0.7530[e]	0.7440[e]	0.3970[e]	0.3666[e]

The online and offline constraint methods lead to similar results, with a slight advantage (not statistically significant) to the online method. This suggests that both constrained methods could be reasonably used to incorporate prior knowledge.

The above comparison shows the benefit of constrained word embeddings. To better understand the effect of constraining embeddings, we analyze a specific example of word "heart", a common medical term. The most similar words, based on word embeddings trained on OHSUMED with different methods, are shown in Table 3.

We can first observe that CBOW is able to find some strongly related words without using UMLS: *hearts*, *cardiovascular*, *cardiorespiratory*. The words *synergist*, *acyanotic* and *ventricular* are also concepts often used in association with *heart*. However, *ouvrier* (name of an author) and *thrive* are not strongly related to *heart*.

Table 3. The most similar words to "heart".

CBOW		Online		Offline	
Cardiac	0.4891	Cardiac	0.5205	Cardiac	0.7960
Synergist	0.4494	Hearts	0.5030	Cor	0.6957
Hearts	0.4276	Cor	0.4939	Synergist	0.5030
Cardiovascular	0.4096	Synergist	0.4690	Hearts	0.4738
Acyanotic	0.3987	Cardiovascular	0.4156	Biventricular	0.4721
Ouvrier	0.3934	Cerebrovascular	0.4149	Cyanotic	0.4720
Multiorgan	0.3931	Acyanotic	0.3985	Cardiorespiratory	0.4714
Ventricular	0.3837	Ventricular	0.3979	Ventricular	0.4651
Cardiorespiratory	0.3829	Cardiorespiratory	0.3969	Acyanotic	0.4585
Thrive	0.3766	Biventricular	0.3831	Circulatory	0.4552

UMLS contains three synonym words to *heart*: *hearts*, *cor* and *cardiac*, which are incorporated in the constrained embeddings. As we can see, these words have been added or promoted (with higher similarities) in the list using constrained methods. First, we observe that CBOW is unable to discover alone the similar word *cor*, which is often used in prescriptions for heart diseases. The prior domain knowledge provides complementary means to link this word. This is part of the benefit we expected from using prior knowledge for embedding training.

Second, we can also observe that in addition to the synonyms, other strongly related words such as *biventricular* and *cyanotic* have also been promoted in the constrained embeddings. In fact, requiring synonym embeddings to be closer also makes the embeddings of their related words closer. In this specific example, even if we do not expect to find the word *cor* in the relevant documents to *heart* in OHSUMED, the words related to *cor* such as *cyanotic* could be found in them. This indirect constraint effect can affect many more words than just synonyms.

We do not see clear differences between the lists of the Online and Offline methods. Both method are capable of finding some strongly related words.

5.4 Parameter Sensitivity

The methods we propose contain some hyper-parameters (α, β, Υ), which we set by cross validation in the previous results. In this section, we examine the sensitivity of retrieval effectiveness to these parameters. We will show the variation of P@10 on OHUSMED and CLEF2015 (CLEF2014 is very similar to CLEF2015).

Figures 1, 2, 3, and 4 show that the retrieval effectiveness (P@10) varies depending on the setting of α and β. The impact of parameters depends on the test collection (OHSUMED and CLEF), and on the basic retrieval model used (BM25 and LM). Globally, the setting of parameters α and β tends to have a larger impact on CLEF than on OHSUMED. This can be explained by the nature of documents in the collections: OHSUMED contains documents written by professionals while CLEF contains web pages crawled from the Web. The domain knowledge is naturally better encoded in OHSUMED than in CLEF. So, using domain knowledge as constraint will make smaller impact on word embeddings in OHSUMED than in CLEF.

We can also see that it is preferable to set these parameter to smaller values when combined with LM than with BM25. This could indicate that less regularization is preferred with LM. Further analyses are needed to understand the reason.

It is difficult to compare directly the parameters α and β because they are used in different constraint processes. We can still observe the general trend that β is preferably set to a large value than α. This may mean that the offline method may need a larger regularization than the online method to adjust word embeddings.

On the parameter Υ (Fig. 5), we observe more consistent behavior on different collections (the variations on other collections and retrieval models are very similar). The best setting is always around 0.5–0.6.

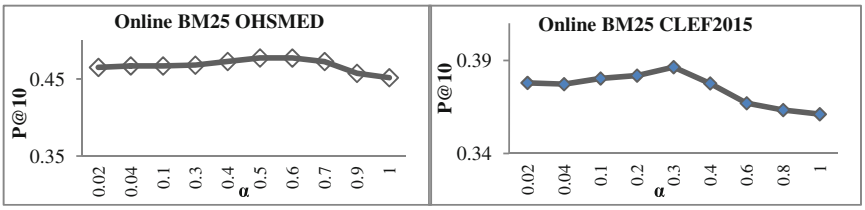

Fig. 1. Sensitivity of α in Online method (combined with BM25)

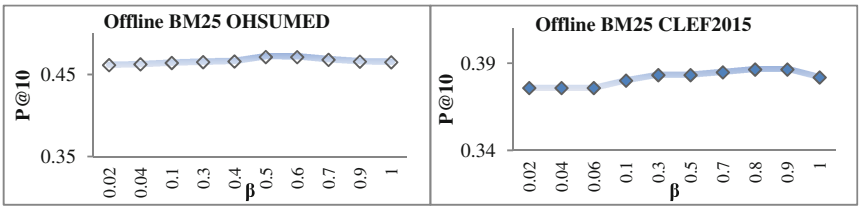

Fig. 2. Sensitivity of β in Offline method (combined with BM25)

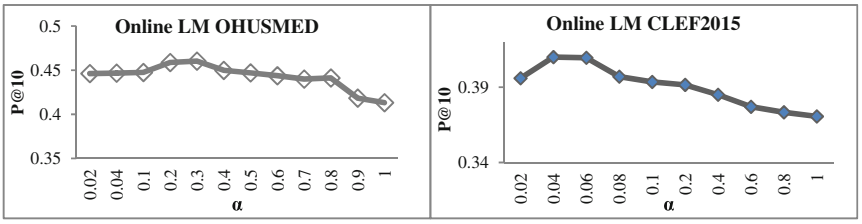

Fig. 3. Sensitivity of α in Online (combined with LM)

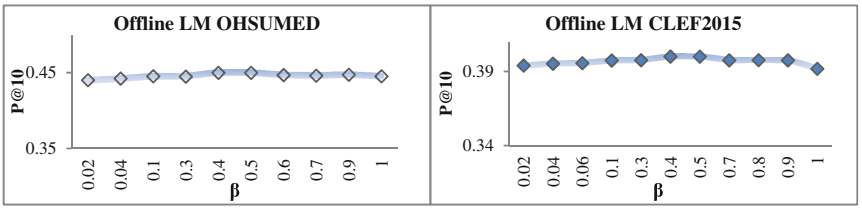

Fig. 4. Sensitivity of β in Offline method (combined with LM)

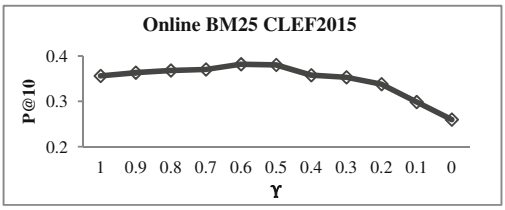

Fig. 5. Sensitivity to the re-rank parameter ϒ on CLEF2015

6 Related Work

6.1 Medical Information Retrieval

A number of studies have attempted to exploit the existing resources in medical area such as UMLS. Two categories of approaches have been proposed in the literature.

The first approach is based on concepts: One first identifies concepts from documents and queries using a concept identification tool such as MetaMap [1]; then documents and queries are matched through their concepts and related concepts. Although improvements using concepts have been observed on some test collections [12, 13, 25, 31], namely in TREC Medical Record Track, which is a different IR task than the one considered in this paper, the improvements on the test collections considered in this paper have been limited and unstable [21]. An important reason lies in the relatively low accuracy of concept identification: about 70–80 % concepts identified are correct, and a number of concepts are unidentified [21].

A second method performs query expansion using the relations stored in a thesaurus [2, 14]. Typically, an additional ranking score is generated from synonyms and related terms of the query terms, and this score is combined with that of the original score. Concept phrases can also be used in this method.

In the previous experiments on the test collections we consider, query expansion approaches have been found more effective than concept-based matching [21]. All the top performing systems at CLEF 2014 and 2015 have used query expansion approaches [9].

To position our methods with respect to the existing approaches, we show the top three results in CLEF 2014 and CLEF 2015 in Table 4. For CLEF 2014, our results are comparable to those of the best team [21], which used MetaMap and all concept expressions in UMLS to perform phrase-based retrieval and query expansion. On CLEF 2015 [18], our results are clearly below the best participating system. However, this best system leveraged Google search results, and this gave a considerable advantage to the system. It is unfair to compare our results with that system. Our methods compare favorably to the other participating systems that do not use Google results. Overall, our methods compare favorably to the state of the art in MIR.

Table 4. Comparison with the best CLEF results

System	CLEF2014		CLEF2015	
	P@10	NDCG@10	P@10	NDCG@10
Best Team 1	0.7560	0.7445	*0.5394*	*0.5086*
Best Team 2	0.7540	0.7406	0.3864	0.3464
Best Team 3	0.7400	0.7301	0.3803	0.3465
Online	**0.7580**	**0.7460**	**0.4086**	**0.3682**

6.2 Word Embeddings for IR

Several studies in IR used word embeddings. [24] used word embedding in cross-language IR task. The goal was to train word embeddings in the same representation space for words in both languages. In [23], word embeddings (CBOW) are used to generate an additional feature to be embedded in a learning to rank framework to rank short answers to a question. Zuccon et al. [30] tested the effectiveness of word embeddings in IR as well as the impact of different parameters. They made similar observation that word embeddings can significantly improve IR effectiveness. De Vine et al. [5] compared several similarity measures for medical IR, and found the one based on word embeddings outperforms the others.

All the above studies showed that the semantic features captures in word embeddings are useful for IR. However, none of the above studies used constrained word embedding. In this paper, we showed that constrained word embeddings can further improve IR effectiveness.

7 Conclusion

In this paper, we explored the utilization of constrained word embedding for IR in a specialized domain. Our assumption is that constrained word embeddings can better fit the application domain and lead to better retrieval results. This is confirmed by our experiments.

Our methods to constrain word embeddings are adapted from the existing studies. In our experiments, we showed that the modifications we made lead to better retrieval results than their original versions. In particular, our modifications corrected two important problems in the original online training method and we added embedding weighting. The modifications resulted in significant changes in IR effectiveness.

We did not observe a large difference between the online and offline methods to incorporate prior knowledge. More investigations are needed to determine the best method to incorporate domain knowledge in word embeddings.

Our investigation has been limited to synonym word, while there are many other types of relation in domain resources (e.g. hierarchical relations). Such relations have been used in MIR [31] and in applications of word embedding in NLP [26]. It would be interesting to extend our study to cover more types of relation.

We only focused on single-word concepts in this study and used a very simple method to build a representation for the entire document and query. It will be

interesting to investigate how an appropriate phrase embedding [6, 19], as well as a representation for the entire document and query, could be built for IR. These are some interesting topics for our future work.

Acknowledgement. This work is partly supported by an NSERC Discovery research grant.

References

1. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of AMIA Symposium, pp. 17–21 (2001)
2. Babashzadeh, A., Huang, J., Daoud, M.: Exploiting semantics for improving clinical information retrieval. In: SIGIR (2013)
3. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. **32**, D267–D270 (2004)
4. Bian, J., Gao, B., Liu, T-Y.: Knowledge-powered deep learning for word embedding. ECML-PKDD, pp. 132–148 (2014)
5. De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., Bruza, P.: Medical semantic similarity with a neural language model. In: CIKM (2014)
6. Dinu, G., Baroni, M.: How to make words with vectors: phrase generation in distributional semantics. In: Proceedings of ACL, pp. 624–633
7. Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A.: Retrofitting word vectors to semantic lexicons. In: NAACL (2015)
8. Ganguly, D., Roy, D., Mitra, M., Jones, J.F.: A word embedding based generalized language model for information retrieval. In: SIGIR, pp. 795–798 (2015)
9. Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G.J.F.: ShARe/CLEF eHealth evaluation lab 2014, task 3: user-centred health information retrieval. In: CLEF 2014 Online Working Note, pp. 43–61 (2014)
10. Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: CIKM, pp. 2333–2338 (2013)
11. Hersh, W., Buckley, C., Leone, T.J., Hickam, D.: OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: SIGIR, pp. 192–201 (1994)
12. Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., Lawley, M.: Information retrieval as semantic inference: a graph inference model applied to medical search. Inf. Ret. **19**(1), 6–37 (2016)
13. Limsopatham, N., Macdonald, G., Ounis, I.: Inferring conceptual relationships to improve medical records search. In: Proceedings of Conference on Open Research Areas in IR, pp. 1–8 (2015)
14. Martinez, D., Otegi, A., Soroa, A., Agirre, E.: Improving search over electronic health records using UMLS-based query expansion through random walks. J. Biomed. Inf. **51**, 100–106 (2014)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
16. Mitra, B.: Exploring session context using distributed representations of queries and reformulations. In: SIGIR, pp. 3–12 (2015)
17. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: EMNLP, pp. 1532–1543 (2014)

18. Palotti, J., Zuccon, G., Goeuriot, L., Kelly, L., Hanbury, A., Jones, G.J.F., Lupu, M., Pecina, P.: CLEF eHealth evaluation lab 2015, task 2: retrieving information about medical symptoms. In: CLEF 2015 Online Working Notes, pp. 32–55 (2015)
19. Socher, R., Manning, C.D., Ng, A.Y.: Learning continuous phrase representations and syntactic parsing with recursive neural networks. In: Deep Learning and Unsupervised Feature Learning Workshop – NIPS (2010)
20. Sordoni, A., Bengio, Y., Vahabi, H., Lioma, C., Simonsen, J.G., Nie, J.-Y.: A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In: CIKM (2015)
21. Shen, W., Nie, J.-Y., Liu, X.-J.: An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM@CLEF2014eHealthTask3. User-centred health information retrieval. In: Proceedings of CLEF 2014 (2014)
22. Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: A latent semantic model with convolutional-pooling structure for information retrieval. In: CIKM, pp. 101–110 (2014)
23. Severyn, A., Moschitti, A.: Learning to rank short text pairs with convolutional deep neural networks. In: SIGIR, pp. 373–382 (2015)
24. Vulic, I., Moens, M.-F.: Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: SIGIR, pp. 363–372 (2015)
25. Wang, Y., Liu, X., Fang, H.: A study of concept-based weighting regularization for medical records search. In: ACL (2014)
26. Xu, C., Bai, Y., Bian, J., Gao, B., Wang, G., Liu, X., Liu, T.-Y.: RC-NET: a general framework for incorporating knowledge into word representations. In: CIKM (2014)
27. Yu, M., Dredze, M.: Improving lexical embeddings with semantic knowledge. In: ACL, pp. 545–555 (2014)
28. Zeiler, M.D., Fergus, R.: Stochastic pooling for regularization of deep convolutional neural networks. arXiv preprint arXiv:1301.3557 (2013)
29. Zheng, G., Callan, J.: Learning to reweight terms with distributed representations. In: SIGIR (2015)
30. Zuccon, G., Koopman, B., Bruza, P., Azzopardi, L.: Integrating and evaluating neural word embeddings in information retrieval. In: Proceedings of Australasian Document Computing Symposium (2015)
31. Zuccon, G., Koopman, B., Nguyen, A., Vickers, D., Butt, L.: Exploiting medical hierarchies for concept-based information retrieval. In: Proceedings of Australasian Document Computing Symposium (2012)