Word Embedding Causes Topic Shifting; Exploit Global Context!

Navid Rekabsaz, Mihai Lupu, Allan Hanbury*
Information & Software Engineering Group
TU WIEN
rekabsaz/lupu/hanbury@ifs.tuwien.ac.at

Hamed Zamani†
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
zamani@cs.umass.edu

ABSTRACT

Exploitation of term relatedness provided by word embedding has gained considerable attention in recent IR literature. However, an emerging question is whether this sort of relatedness fits to the needs of IR with respect to retrieval effectiveness. While we observe a high potential of word embedding as a resource for related terms, the incidence of several cases of topic shifting deteriorates the final performance of the applied retrieval models. To address this issue, we revisit the use of global context (i.e. the term co-occurrence in documents) to measure the term relatedness. We hypothesize that in order to avoid topic shifting among the terms with high word embedding similarity, they should often share similar global contexts as well. We therefore study the effectiveness of post filtering of related terms by various global context relatedness measures. Experimental results show significant improvements in two out of three test collections, and support our initial hypothesis regarding the importance of considering global context in retrieval.

KEYWORDS

Word embedding, term relatedness, global context, word2vec, LSI

ACM Reference format:
Navid Rekabsaz, Mihai Lupu, Allan Hanbury and Hamed Zamani. 2017. Word Embedding Causes Topic Shifting; Exploit Global Context!. In *Proceedings of SIGIR '17, Shinjuku, Tokyo, Japan, August 07-11, 2017,* 4 pages.
DOI: http://dx.doi.org/10.1145/3077136.3080733

1 INTRODUCTION

The effective choice of related terms to enrich queries has been explored for decades in information retrieval literature and approached using a variety of data resources. Early studies explore the use of collection statistics. They identify the global context of two terms either by directly measuring term co-occurrence in a context (i.e. document) [9] or after applying matrix factorization [3]. Later studies show the higher effectiveness of local approaches (i.e. using pseudo-relevant documents) [15]. More recently, the approaches to exploit the advancement in word embedding for IR has shown not only to be competitive to the local approaches but also that combining the approaches brings further improvements in comparison to each of them alone [12, 16, 17].

Word embedding methods provide vector representations of terms by capturing the co-occurrence relations between the terms, based on an approximation on the likelihood of their appearances in similar window-contexts. Word embedding is used in various IR tasks e.g. document retrieval [11, 12, 18], neural network-based retrieval models [4, 6, 8], and query expansion [16].

In all of these studies, the concept of "term similarity" is defined as the geometric proximity between their vector representations. However, since this closeness is still a mathematical approximation of meaning, some related terms might not fit to the retrieval needs and eventually deteriorate the results. For instance, antonyms (*cheap* and *expensive*) or co-hyponyms (*schizophrenia* and *alzheimer*, *mathematics* and *physics*, countries, months) share common window-context and are therefore considered as related in the word embedding space, but can potentially bias the query to other topics.

Some recent studies aim to better adapt word embedding methods to the needs of IR. Diaz et al. [5] suggest training separate word embedding models on the top retrieved documents per query, while Rekabsaz et al. [13] explore the similarity space and suggest a general threshold to filter the most effective related terms. While the mentioned studies rely on the context around the terms, in this work we focus on the effect of similarity achieved from global context as a complementary to the window-context based similarity.

In fact, similar to the earlier studies [9, 14], we assume each document to be a coherent information unit and consider the co-occurrence of terms in documents as a means of measuring their topical relatedness. Based on this assumption, we hypothesize that to mitigate the problem of topic shifting, the terms with high word embedding similarities also need to share similar global contexts. In other words, if two terms appear in many similar window-contexts, but they share little global contexts (documents), they probably reflect different topics and should be removed from the related terms.

To examine this hypothesis, we start by analyzing the effectiveness of each related term, when added to the query. Our approach is similar to that of Cao et al. [2] on pseudo-relevance feedback. Our analysis shows that the set of related terms from word embedding has a high potential to improve state-of-the-art retrieval models. Based on this motivating observation, we explore the effectiveness of using word embedding's similar term when filtered by global context similarity on two state-of-the-art IR models. Our evaluation on three test collections shows the importance of using global context, as combining both the similarities significantly improves the results.

*Funded by: Self-Optimizer (FFG 852624) in the EUROSTARS programme, funded by EUREKA, BMWFW and European Union, and ADMIRE (P 25905-N23) by FWF. Thanks to Joni Sayeler and Linus Wretblad for their contributions in SelfOptimizer.
†was supported in part by the Center for Intelligent Information Retrieval.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '17, Shinjuku, Tokyo, Japan
© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM.
978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080733

2 BACKGROUND

To use word embedding in document retrieval, recent studies extend the idea of translation models in IR [1] using word embedding similarities. Zuccon et al. [18] use the similarities in the language modeling framework [10] and Rekabsaz et al. [12] extend the concept of translation models to probabilistic relevance framework. In the following, we briefly explain the translation models when combined with word embedding similarity.

In principle, a translation model introduces in the estimation of the relevance of the query term t a translation probability P_T, defined on the set of (extended) terms $R(t)$, always used in its conditional form $P_T(t|t')$ and interpreted as the probability of observing term t, having observed term t'. Zuccon et al. [18] integrate word embedding with the translation language modeling by using the set of extended terms from word embedding:

$$\widehat{LM}(q,d) = P(q|M_d) = \prod_{t \in q} \left(\sum_{t' \in R(t)} P_T(t|t') P(t'|M_d) \right) \quad (1)$$

Rekabsaz et al. [12] extend the idea into four probabilistic relevance frameworks. Their approach revisits the idea of computing document relevance based on the occurrence of concepts. Traditionally, concepts are represented by the words appear in the text, quantified by term frequency (tf). Rekabsaz et al. posit that we can have a tf value lower than 1 when the term itself is not actually appear, but another, conceptually similar term occurs in the text. Based on it, they define the extended tf of a query word t in a document as follows:

$$\widehat{tf_d} = tf_d + \sum_{t' \in R(t)} P_T(t|t') tf_d(t') \quad (2)$$

However, in the probabilistic models, a series of other factors are computed based on tf (e.g. document length). They therefore propagate the above changes to all the other statistics and refer to the final scoring formulas as Extended Translation model. Among the extended models, as *BM*25 is a widely used and established model in IR, we use the extended *BM*25 translation model ($\widehat{BM25}$) in our experiments. Similar to the original papers in both models, the estimation of P_T is based on the Cosine similarity between two embedding vectors.

3 EXPERIMENT SETUP

We conduct our experiments on three test collections, shown in Table 1. For word embedding vectors, we train the word2vec skip-gram model [7] with 300 dimensions and the tool's default parameters on the Wikipedia dump file for August 2015. We use the Porter stemmer for the Wikipedia corpus as well as retrieval. As suggested by Rekabsaz et al. [13], the extended terms set $R(t)$ is selected from the terms with similarity values of greater than a specific threshold. Previous studies suggest the threshold value of around 0.7 as an optimum for retrieval [12, 13]. To explore the effectiveness of less similar terms, we try the threshold values of {0.60, 0.65..., 0.80}.

Since the parameter μ for Dirichlet prior of the translation language model and also b, k_1, and k_3 for *BM*25 are shared between the methods, the choice of these parameters is not explored as part of this study and we use the same set of values as in Rekabsaz et al. [12]. The statistical significance tests are done using the two sided paired t-test and significance is reported for $p < 0.05$. The evaluation of retrieval effectiveness is done with respect to Mean Average Precision (MAP) as a standard measure in ad-hoc IR.

Table 1: Test collections used in this paper

Name	Collection	# Queries	# Documents
TREC Adhoc 1&2&3	Disc1&2	150	740449
TREC Adhoc 6&7&8	Disc4&5	150	556028
Robust 2005	AQUAINT	50	1033461

Table 2: The percentage of the good, bad and neutral terms. #Rel averages the number of related terms per query term.

Collection	Threshold 0.60				Threshold 0.80			
	#Rel	Good	Neutral	Bad	#Rel	Good	Neutral	Bad
TREC 123	8.2	7%	84%	9%	1.3	19%	68%	13%
TREC 678	8.8	9%	78%	14%	1.2	34%	48%	18%
Robust 2005	10.3	8%	77%	15%	1.1	39%	44%	17%
ALL	8.1	8%	81%	11%	1.2	27%	58%	15%

4 PRELIMINARY ANALYSIS

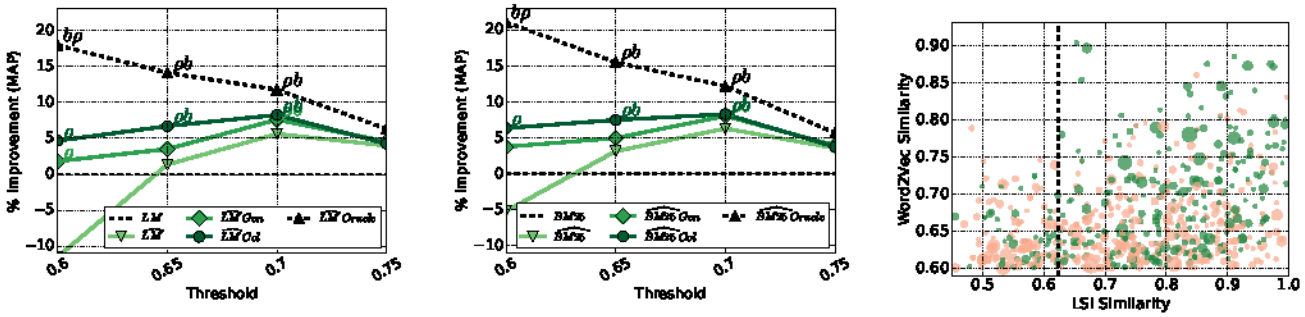
We start with an observation on the effectiveness of each individual related term. To measure it, we use the $\widehat{LM}$ model as it has shown slightly better results than the $\widehat{BM25}$ model [12]. Similar to Cao et al. [2], given each query, for all its corresponding related terms, we repeat the evaluation of the IR models where each time $R(t)$ consists of only one of the related terms. For each term, we calculate the differences between its Average Precision (AP) evaluation result and the result of the original query and refer to this value as the *retrieval gain* or *retrieval loss* of the related term.

Similar to Cao et al. [2], we define *good/bad* groups as the terms with retrieval gain/loss of more than 0.005, and assume the rest with smaller gain or loss values than 0.005 as *neutral* terms. Table 2 summarizes the percentage of each group. Due to the lack of space, we only show the statistics for the lowest (0.6) and highest (0.8) threshold. The average number of related terms per query term is shown in the `#Rel` field. As expected, the percentage of the good terms is higher for the larger threshold, however—similar to the observation on pseudo-relevance feedback [2]—most of the expanded terms (58% to 81%) have no significant effect on performance.

Let us imagine that we had a priori knowledge about the effectiveness of each related term and were able to filter terms with negative effect on retrieval. We call this approach `Oracle` post-filtering as it shows us the maximum performance of each retrieval model. Based on the achieved results, we provide an approximation of this approach by filtering the terms with retrieval loss.

Figures 1a and 1b show the percentage of relative MAP improvement of the $\widehat{LM}$ and $\widehat{BM25}$ models with and without post-filtering with respect to the original *LM* and *BM*25 models. In the plot, ignore the `Gen` and `Col` results as we return to them in Section 6. The results are aggregated over the three collections. In each threshold the statistical significance of the improvement with respect to two baselines are computed: (1) against the basic models (*BM*25 and *LM*), shown with the b sign and (2) against the translation models without post filtering, shown with the ρ sign.

As reported by Rekabsaz et al. [13], for the thresholds less than 0.7 the retrieval performance of the translation models (without post filtering) decreases as the added terms introduce more noise. However, the models with the `Oracle` post filtering continue to improve the baselines further for the lower thresholds with high margin. These demonstrate the high potential of using related terms from word embedding but also show the need to customize the set of terms for IR. We propose an approach to this customization using the global-context of the terms in the following.



(a) LM (b) BM25 (c) Retrieval Gain/Loss

Figure 1: (a,b) The percentage of relative MAP improvement to the basic models, aggregated on all the collections. The b and ρ signs show the significance of the improvement to the basic models and the extended models without post filtering respectively (c) Retrieval gain or loss of the related terms for all the collection. The red (light) color indicate retrieval loss and the green (dark) retrieval gain.

5 GLOBAL-CONTEXT POST FILTERING

Looking at some samples of retrieval loss, we can observe many cases of topic shifting: e.g. Latvia as query term is expanded with Estonia, Ammoniac with Hydrogen, Boeing with Airbus, and Alzheimer with Parkinson. As mentioned before, our hypothesis is that for the terms with high window-context similarity (i.e. word2vec similarity) when they have high global context similarity (i.e. co-occurrence in common documents), they more probably refer to a similar topic (e.g. USSR and Soviet) and with low global context similarity to different topics (e.g. Argentina and Nicaragua).

To capture the global context similarities, some older studies use measures like Dice, Tanimoto, and PMI [9]. Cosine similarity has been used as well, considering each term a vector with dimensionality of the number of documents in the collection, with weights given either as simple incidence (i.e. 0/1), or by some variant of TFIDF. Cosine can also be used after first applying Singular Value Decomposition on the TFIDF weighted term-document matrix, resulting in the well known Latent Semantic Indexing (LSI) method [3] (300 dimensions in our experiments). To compute these measures, we consider both the collection statistics and Wikipedia statistics, resulting in 12 sets of similarities (Dice, Tanimoto, PMI, Incidence Vectors, TFIDF Vectors, LSI Vectors)×(collection, Wikipedia). We refer to these similarity value lists as global context features.

Let first observe the relationship between LSI and word2vec similarities of the terms. Figure 1c plots the retrieval gain/loss of the terms of all the collections based on their word2vec similarities as well as LSI (when using test collection statistics). The size of the circles shows their gain/loss values as the red color (the lighter one) show retrieval loss and green (the darker one) retrieval gain. For clarity, we only show the terms with the retrieval gain/loss of more than 0.01. The area with high word2vec and LSI similarity (top-right) contains most of the terms with retrieval gain. On the other hand, regardless of the word2vec similarity, the area with lower LSI similarity tend to contain relatively more cases of retrieval loss. This observation encourages the exploration of a set of thresholds for global context features to post filter the terms retrieved by word embedding.

To find the thresholds for global context features, we explore the highest amount of total retrieval gain after filtering the related terms with similarities higher than the thresholds. We formulate it by the following optimization problem:

$$\underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{N} \mathbb{1}\Big[\bigcap_{j=1}^{F} x_j > \theta_j\Big] g_i \tag{3}$$

where $\mathbb{1}$ is the indicator function, N and F are the number of terms and features respectively, Θ indicates the set of thresholds θ_j, x_j the value of the features, and finally g refers to the retrieval gain/loss.

We consider two approaches to selecting the datasets used to find the optimum thresholds: *per collection*, and *general*. In the per collection scenario (`Col`), for each collection we find different thresholds for the features. We apply 5-fold cross validation by first using the terms of the training topics to find the thresholds (solving Eq. 3) and then applying the thresholds to post filter the terms of the test topics. To avoid overfitting, we use the bagging method by 40 times bootstrap sampling (random sampling with replacement) and aggregate the achieved thresholds.

In the general approach (`Gen`), we are interested in finding a 'global' threshold for each feature, which is *fairly* independent of the collections. As in this approach the thresholds are not specific for each individual collection, we use all the topics of all the test collections to solve the optimization problem.

6 RESULTS AND DISCUSSION

To find the most effective set of features, we test all combinations of features using the per collection (`Col`) post-filtering approach. Given the post-filtered terms with each feature set, we evaluate the $\widehat{LM}$ and $\widehat{BM25}$ models. Our results show the superior effectiveness of the LSI feature when using the test collections as resource in comparison with the other features as well as the ones based on Wikipedia. The results with the LSI feature can be further improved by combining it with the TFIDF feature. However, adding any of the other features does not bring any improvement and therefore, in the following, we only use the combination of LSI and TFIDF features with both using the test collections statistics.

The evaluation results of the original $\widehat{LM}$ and $\widehat{LM}$ with post filtering with the general (`Gen`) and per collection (`Col`) approaches are shown in Figure 2. The general behavior of $\widehat{BM25}$ is very similar and therefore no longer shown here. As before, statistical significance against the basic models is indicated by b and against the translation models without post filtering, by ρ.

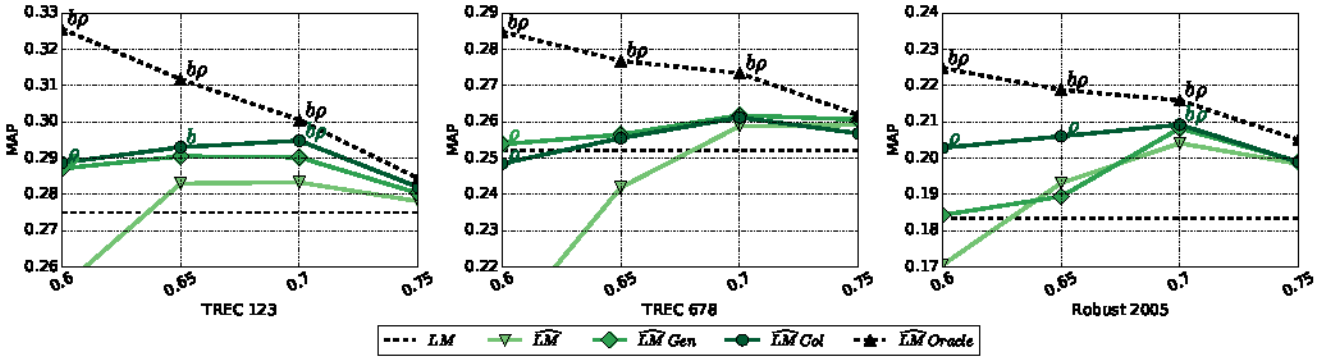


Figure 2: Evaluation results of $\widehat{LM}$ with/without post filtering. The b and ρ signs show the significance of the improvement to LM and $\widehat{LM}$ without post filtering respectively.

The results show the improvement of the $\widehat{LM}$ models with post-filtering in comparison with the original $\widehat{LM}$. The models with post-filtering approaches specifically improve in lower word embedding thresholds, however similar to the original translation models, the best performance is achieved on word embedding threshold of 0.7. The results for both $\widehat{LM}$ and $\widehat{BM25}$ models with word embedding threshold of 0.7 are summarized in Table 3. Comparing the post-filtering approaches, `Col` shows better performance than `Gen` as with the optimum word embedding threshold, it achieves significant improvements over both the baselines in two of the collections.

Let us look back to the percentage of relative improvements, aggregated over the collections in Figures 1a and 1b. In both IR models, while the `Col` approach has better results than `Gen`, their results are very similar to the optimum word embedding threshold (0.7). This result suggests to use the `Gen` approach as a more straightforward and general approach for post filtering. In our experiments, the optimum threshold value for the LSI similarities (as the main feature) is around 0.62 (shown in Figure 1c by vertical line).

As a final point, comparing the two IR models shows that despite the generally better performance of the $\widehat{LM}$ models, the $\widehat{BM25}$ models gain more. We speculate that it is due to the additional modification of other statistics (i.e. document length and IDF) in the $\widehat{BM25}$ model and therefore it is more sensitive to the quality of the related terms. However an in-depth comparison between the models is left for future work.

7 CONCLUSION

Word embedding methods use (small) window-context of the terms to provide dense vector representations, used to approximate term relatedness. In this paper, we study the effectiveness of related terms, identified by both window-based and global contexts, in document retrieval. We use two state-of-the-art translation models to integrate word embedding information for retrieval. Our analysis shows a great potential to improve retrieval performance, damaged however by topic shifting. To address it, we propose the use of global context similarity, i.e. the co-occurrence of terms in larger contexts such as entire documents. Among various methods to measure global context, we identify LSI and TFIDF as the most effective in eliminating related terms that lead to topic shifting. Evaluating the IR models using the post-filtered set shows a significant improvement in comparison with the basic models as well as the translation models with no post-filtering. The results demonstrate the importance of global context as a complementary to the window-context similarities.

Table 3: MAP of the translation models when terms filtered with word embedding threshold of 0.7 and post filtered with the `Gen` and `Col` approach.

Collection	Model	Basic	Tran.	Tran.+Gen	Tran.+Col
TREC 123	LM	0.275	0.283	0.290	**0.295** $b\rho$
	BM25	0.273	0.285	0.288	**0.290** b
TREC 678	LM	0.252	0.259	**0.262**	0.261
	BM25	0.243	0.255	**0.257**	0.256 b
Robust 2005	LM	0.183	0.204	0.208	**0.209** $b\rho$
	BM25	0.181	0.203	0.207	**0.209** $b\rho$

REFERENCES

[1] Adam Berger and John Lafferty. 1999. Information Retrieval As Statistical Translation. In *Proc. of SIGIR*.

[2] Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. 2008. Selecting good expansion terms for pseudo-relevance feedback. In *Proc. of SIGIR*.

[3] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* (1990).

[4] Mostafa Dehghani, Hamed Zamani, A. Severyn, J. Kamps, and W Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *Proc. of SIGIR*.

[5] Fernando Diaz, Bhaskar Mitra, and Nick Craswell. 2016. Query expansion with locally-trained word embeddings. *Proc. of ACL* (2016).

[6] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proc. of CIKM*.

[7] Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[8] Bhaskar Mitra, F. Diaz, and N. Craswell. 2017. Learning to Match Using Local and Distributed Representations of Text for Web Search. In *Proc. of WWW*.

[9] Helen J Peat and Peter Willett. 1991. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American society for information science* (1991).

[10] Jay M. Ponte and W. Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. In *Proc. of SIGIR*.

[11] Navid Rekabsaz, Ralf Bierig, Bogdan Ionescu, Allan Hanbury, and Mihai Lupu. 2015. On the use of statistical semantics for metadata-based social image retrieval. In *Proc. of CBMI Conference*.

[12] Navid Rekabsaz, Mihai Lupu, and Allan Hanbury. 2016. Generalizing Translation Models in the Probabilistic Relevance Framework. In *Proc. of CIKM*.

[13] Navid Rekabsaz, Mihai Lupu, and Allan Hanbury. 2017. Exploration of a threshold for similarity based on uncertainty in word embedding. In *Proc. of ECIR*.

[14] G Salton and MJ MacGill. 1983. Introduction to modern information retrieval. *McGraw-Hill* (1983).

[15] Jinxi Xu and W Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proc. of SIGIR*.

[16] Hamed Zamani and W Bruce Croft. 2016. Embedding-based query language models. In *Proc. of ICTIR*.

[17] Hamed Zamani and W Bruce Croft. 2017. Relevance-based Word Embedding. In *Proc. of SIGIR*.

[18] Guido Zuccon, Bevan Koopman, Peter Bruza, and Leif Azzopardi. 2015. Integrating and evaluating neural word embeddings in information retrieval. In *Proc. of Australasian Document Computing Symposium*.