Exploration of a Threshold for Similarity based on Uncertainty in Word Embedding

Navid Rekabsaz, Mihai Lupu, Allan Hanbury

Institute of Software Technology and Interactive Systems
Vienna University of Technology
A-1040 Vienna, Austria
[last_name]@ifs.tuwien.ac.at

Abstract. Word embedding promises a quantification of the similarity between terms. However, it is not clear to what extent this similarity value can be of practical use for subsequent information access tasks. In particular, which range of similarity values is indicative of the actual term relatedness? We first observe and quantify the uncertainty of word embedding models with respect to the similarity values they generate. Based on this, we introduce a general threshold which effectively filters related terms. We explore the effect of dimensionality on this general threshold by conducting the experiments in different vector dimensions. Our evaluation on four test collections with four relevance scoring models supports the effectiveness of our approach, as the results of the proposed threshold are significantly better than the baseline while being equal to, or statistically indistinguishable from, the optimal results.

1 Introduction

Understanding the meaning of a word (semantics) and of its similarity to other words (relatedness) is the core of understanding text. An established method for quantifying this similarity is the use of *word embeddings*, where vectors are proxies of the meaning of words and distance functions are proxies of semantic and syntactic relatedness. Fundamentally, word embedding models exploit the contextual information of the target words to approximate their meaning, and hence their relations to other words.

Given the vectors representing words and a corresponding mathematical function, these models provide an approximation of the relatedness of any two terms, although this relatedness could be perceived as completely arbitrary in the language. This issue is pointed out by Karlgren et al. [9] in examples, showing that word embedding methods are too ready to provide answers to meaningless questions: *"What is more similar to a computer: a sparrow or a star?"*, or *"Is a cell more similar to a phone than a bird is to a compiler?"*. The emerging challenge here is: *how to identify whether the similarity score obtained from word embedding is really indicative of term relatedness?*

1.1 Related Work

The closest study to our work is Karlgren et al. [8], which explores the semantic topology of the vector space generated by Random Indexing. Based on their previous observations that the dimensionality of the semantic space appears different for different terms [9], Karlgren at al. now identify the different dimensionalities at different angles (i.e. distances) for a set of specific terms. It is however difficult to map these observations to specific criteria or guidelines for either future models or retrieval tasks. In fact, our observations provide a quantification on Karlgren's claim that *"*'close' *is interesting and* 'distant' *is not"* [9].

More recently, Cuba Gyllensten and Sahlgren [3] follow a data mining approach to represent the terms relatedness by a tree structure. While they suggest traversing the tree as a potential approach, they evaluate it only on the word sense induction tasks and its utility for retrieving similar words remains unanswered. They do point out however, that applying a nearest neighbour approach, where for every word we use the top k most similar words, is not theoretically justifiable. Rekabsaz et al. [17] recently showed this also experimentally in a retrieval task.

In general, different characteristics of term similarities have been explored in several studies: the concept of relatedness [10, 12], the similarity measures [11], intrinsic/extrinsic evaluation of the models [1, 4, 19, 21], or in sense induction task [3, 5]. However, there is lack of understanding on the internal structure of word embedding, specifically how its similarity distribution reflects the relatedness of terms.

1.2 Motivation

Among the recent publications using word-embeddings for information retrieval, Rekabsaz et al. [17] do a brute-force search on similarity thresholds for the typical ad-hoc search task and evaluate their results against a set of TREC test collections. The parameter scan is obviously inefficient in general and we consider their work as the main motivation for the current study of a language-specific semantic similarity threshold.

In fact, we hypothesise that the "similar" words can be identified by a threshold on similarity values which separates the semantically related words from the non-related ones. We especially want to make this threshold independent of the terms and general on word embedding model. The reason for this choice is first the computational problem of term-specific thresholds as it puts burden on practical applications. Regardless of the efficiency issues, it is still reasonable to consider a general threshold. since it considers the centrality and neighbourhood of the terms by filtering different number of similar terms for each term.

Such a threshold has the potential to improve all studies that use similar/related words in different tasks i.e. query expansion [7], query auto-completion [14], document retrieval [16], learning to rank [20], language modelling in IR [6], or Cross-Lingual IR [22]. It should be noted though, that the meaning of "similar" also depends on the similarity function. We consider here the Cosine function as it is by far the most widely used word similarity function and leave the exploration of other functions for further studies. In fact, regardless of the similarity function, a threshold that separates the semantically related terms from the rest will always be an essential element to identify.

1.3 Approach

We explore the estimation of this potential threshold by first quantifying the uncertainty in the similarity values of embedding models. This uncertainty is an intrinsic characteristic of all the recent models, because they all start with some random initialization and eventually converge to a (local) solution. Therefore, even by training with the same parameters and on the same data, the created word embedding models result in slightly different word distributions and hence slightly different relatedness values. In the next step, using this observation, we provide a novel representation on the expected number of neighbours of an arbitrary term as a continuous function over similarity values, which is later used to estimate the general threshold.

In order to evaluate the effectiveness of the proposed threshold, we follow the approach previously introduced by Rekabsaz et al. [17] and test it in the context of a document retrieval task, on four different test collections, using the skip-gram with negative-sampling training word embeddings [13]. In the experiments, we apply the threshold to identify the set of terms to extend the query terms using both the Generalised Translation Model and the Extended Translation Model introduced by Rekabsaz et al. [17]. The results are compared with the optimal threshold, achieved as before by exhaustive search on the spectrum of threshold parameters. We show that in general using the proposed threshold performs either exactly the same as, or statistically indistinguishable from, the optimal threshold.

In summary, the main contributions of this paper are:

1. exploration of the uncertainty in word embedding models in different dimensions and similarity ranges.
2. introducing a general threshold for separating similar terms in different embedding dimensions.
3. extensive experiments on four test collections comparing different threshold values on different retrieval models.

The remainder of this work is structured as follows: We introduce the proposed threshold in Section 2. We present our experimental setup in Section 3, followed by discussing the results in Section 4. Section 5 summarises our observations and concludes the paper.

2 Global Term Similarity Threshold

We are looking for a threshold to separate the related terms from the rest. For this purpose, we start with an observation on the uncertainty of similarity in word embedding models, followed by defining a novel model of the expected number of neighbours for an arbitrary term, before we define our proposed threshold.

2.1 Uncertainty of Similarity

In this section we make a series of practical observations on word embeddings and the similarities computed based on them.

To observe the uncertainty, let us consider two models P and M. To create each instance, we trained the Skip-Gram with Negative-Sampling (SGNS) of the Word2Vec model with the sub-sampling parameter set to 10^{-5}, context windows of 5 words, epochs of 25, and word count threshold 20 on the Wikipedia dump file for August 2015, after applying the Porter stemmer. Each model has a vocabulary of approximately 580k terms. They are identical in all ways except their random starting point.

Figure 1a shows the distances between two terms and all other terms in the dictionary, for the two models, in this case of dimensionality 200. For each term we have approximately 580k points on the plot. As we can see, the difference between similarities calculated in the two models, appears (1) greater for low similarities, and (2) greater for a rare word (Dwarfish) than for a common word (Book). We can also observe that there are very few pairs of words with very high similarities.

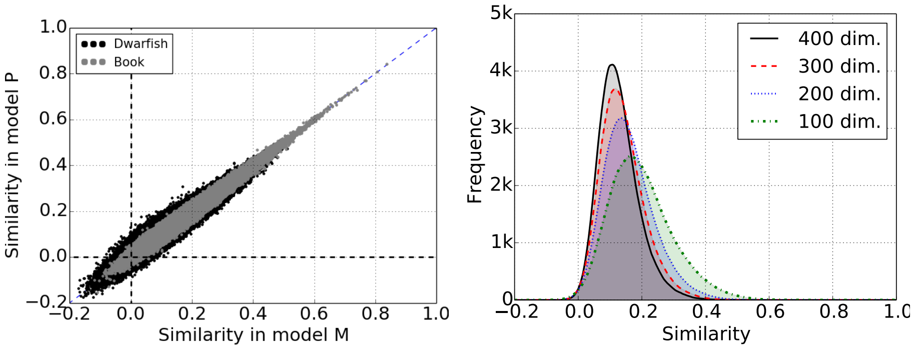


Fig. 1: (a) Comparison of similarity values of the terms *Book* and *Dwarfish* to 580K words between models M and P. (b) Histogram of similarity values of an arbitrary term to all the other words in the collection for 100, 200, 300, and 400 dimensions.

Let us now explore the effect of dimensionality on similarity values and also uncertainty. Before that, in order to generalize the observations to an arbitrary term, we had to consider a set of "representative" terms. What exactly "representative" means is of course debatable. We took 100 terms recently introduced in the query inventory method by Schnabel et al. [19]. They claim that the selected terms are diverse in frequency and part of speech over the collection terms. In the remainder of the paper, we refer to *arbitrary* term as an aggregation over the representative terms i.e. each value related to the arbitrary term is the average of the values of the representative terms.

Figure 1b shows frequency histograms for the occurrence of similarity values for models of different dimensionalities. As we can see, similarities are in the $[-0.2, 1.0]$ range and have positive skewness (the right tail is longer). As the dimensionality of the model increases, the kurtosis also increases (the histogram has thinner tails).

Let us first suggest a concrete definition for uncertainty: We quantify the uncertainty of the similarity between two words as the standard deviation σ of similarity values obtained from a set of identical models. We refer to identical models as the models created using the same method, parameters, and corpus. However as shown before, the similarity values of each word pair in each model are slightly different. The uncertainty of similarity between the words x and y is therefore formulated as follows:

$$\sigma_{x,y} = \sqrt{\frac{1}{|M|}\sum_{m\in M}(sim(\boldsymbol{x}_m - \boldsymbol{y}_m) - \mu)^2}, \text{ where } \mu = \frac{\sum_{m\in M} sim(\boldsymbol{x}_m - \boldsymbol{y}_m)}{|M|}.$$

where M is the set of identical models and $\boldsymbol{x}_m$ is the vector representation of term x in model m and sim is a similarity function between two vectors.

To observe the changes in standard deviation, for every dimensionality, we create five identical SGNS models ($|M| = 5$).

Figure 2a plots the standard deviation, against the similarity values, for different model dimensionalities. For the sake of clarity in visualisation, we split the similarity values into 500 equal intervals (each 2.4×10^{-4}) and average the values in each interval.

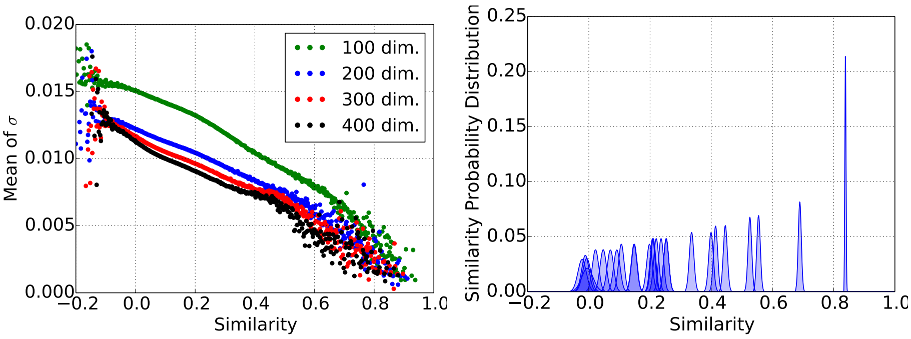


Fig. 2: (a) Standard deviation for similarity values. Points are the average over similarity intervals with equal lengths of 2.4×10^{-4} (b) Probability distribution of similarity values for the term *Book* to some other terms.

The plots are smooth in the middle and scattered on the head and tail as the majority of similarity values are in the middle area of the plots and therefore the average values are consistent. However, we can observe that overall, as the similarity increases, the standard deviation, i.e. the uncertainty, decreases.

We also observe a decrease in standard deviation as the dimensionality of the model increases. On the other hand, the differences between models decrease as the dimension increases such that the models of dimension 300 and 400 seem very similar in comparison to 100 and 200. The observation shows a probable convergence in the uncertainty at higher dimensionalities.

These observations show that the similarity between terms is not an exact value but can be considered as an approximation whose variation is dependent on the dimensionality and similarity range. We use the outcome of these observations in the following.

2.2 Continuous Distribution of Neighbours

We have demonstrated that the similarity values of a pair of terms, obtained from identical embedding models are slightly different. In the absence of additional information, we assume that these similarity values follow a normal distribution.

To estimate this probability distribution, we use the mean and standard deviation values in Section 2.1. Figure 2b shows the probability distribution of similarities for term *Book* to 25 terms in different similarity ranges[1]. As observed before, by decreasing the similarity, the standard deviation of the probability distributions increases.

We use these probability distributions to provide a representation of the expected number of neighbours around an arbitrary term in the spectrum of similarity values: We first calculate the Cumulative Distribution Functions (CDF) of the probability distributions. We then subtract the CDF values from 1 which only reverses the direction of the distributions (from increasing left-to-right on X-axis to right-to-left). Finally, we accumulate all the cumulative distribution functions by summing all the values, shown in Figure 3a. The values on this plot indicate the number of expected neighbours that have

[1] we do not plot all the terms in the model to maintain the readability of the plot

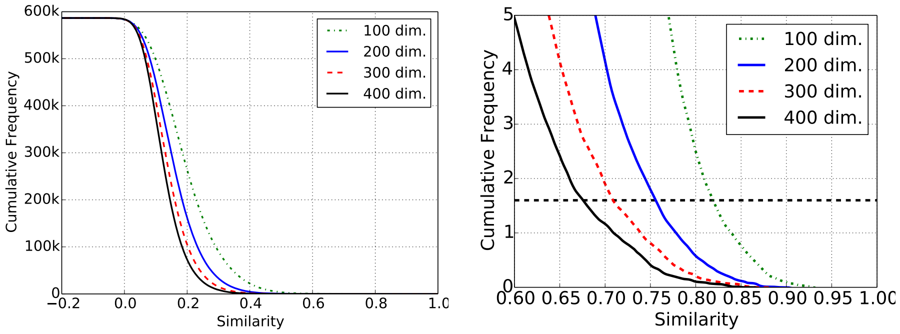


Fig. 3: (a) Mixture of cumulative probability distributions of similarities in different dimensions (b) Expected number of neighbours around an arbitrary term with confidence interval. The average number of synonyms in WordNet (1.6) is shown by the dash-line.

greater or equal similarity values to the term than the given similarity value. We can see the number of all the terms in the model (580k) in the lowest similarity value (−0.2) which then rapidly drops as the similarity increases. This representation of the expected number of neighbours in Figure 3a has two benefits: (1) the estimation is continuous and monotonic, and (2) it considers the effect of uncertainty based on five models.

As noted before, the notion of *arbitrary* term is in fact an average over the 100 representative terms. Therefore, in calculating the representation of the expected number of neighbours, we also consider the confidence interval around the mean. This interval is shown in Figure 3b. Here, the representation is zoomed on the lower right corner of Figure 3a. The area around each plot shows the confidence interval of the estimation.

This continuous representation is used in the following for defining the threshold for the semantically related terms.

2.3 Similarity threshold

Given the expected number of neighbours around the arbitrary term, represented in Figure 3a and Figure 3b, the question is "*what is the best threshold for filtering the related terms?*". In order to address the question, we hypothesise that since this general threshold tries to separate related from unrelated terms, it can be estimated from the average number of synonyms over the terms. Therefore, we transform the above question into a new question: "*What is the expected number of synonyms for a word in English?*"

To answer this, we exploit WordNet. We consider the distinct terms in the related synsets to a term as its synonyms, while filtering the terms containing multi word (e.g. Natural Language Processing, shown in WordNet in Natural_Language_Processing form) since in creating the word embedding models such terms are considered as separated terms (one word per term). The average number of synonyms over all the 147306 terms of WordNet is 1.6, while the standard deviation is 3.1.

Using the average value of the synonyms in WordNet, we define our threshold for each model dimensionality as the point where the estimated number of neighbours in Figure 3b is equal to 1.6. We also consider an upper and lower bound for this threshold based on the points on the similarity axis at which the confidence interval plots cross the horizontal line of the average value. The results are shown in Table 1.

Table 1: Proposed thresholds for various dimensionalities

Dimensionality	Threshold Boundaries		
	Lower	**Main**	Upper
100	0.802	**0.818**	0.829
200	0.737	**0.756**	0.767
300	0.692	**0.708**	0.726
400	0.655	**0.675**	0.693

In the following sections, we validate the hypothesis by evaluating the performance of the proposed thresholds with an extensive set of experiments.

3 Experimental Methodology

We test the effectiveness of our threshold in an Ad-hoc retrieval task on IR test collections by evaluating the results of applying various thresholds to retrieve the related terms.

Our relevance scoring approach is based on the query language model [15] and BM25 methods as two widely used and established methods in IR, which have shown competitive results in various domains. To use the additional information provided by word embeddings, we use the *Generalized Translation Model* and *Extended Translation Model* extensions introduced by Rekabsaz et al. [17], which build on top of the existing probabilistic models.

In the following, first we briefly explain the translation models when combined with word embedding similarity and then describe the details of our experimental setup.

3.1 Generalized and Extended Translation Model

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 (related) 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'.

Translation models in IR were first introduced by Berger and Lafferty [2] as an extension to the language model. Recently, Rekabsaz et al. [17] extend the idea of translation model into four probabilistic relevance frameworks. Their approach is based on the observation that what one wants to compute in general in IR, and in particular in a probabilistic method, is the occurrence of concepts. Traditionally, these are represented by the words present 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 present, but another, similar term occurs in the text. They call this the Generalised Translation model (GT). However, in the probabilistic models, a series of other factors are computed based on tf (e.g. document length). Propagating the above changes to all the other statistics leads to even more changes in the scoring formulas. They refer to this as the Extended Translation model (ET).

In both translation models, they use word embedding to generate the $R(t)$ set by selecting the terms with the similarity value of greater than a given threshold to the query term t. In the following experiments we will show that the analytically obtained threshold described in the previous section is optimal for the ad-hoc retrieval task.

3.2 Experiment Setup

We evaluate our approach on four test collections: TREC-6, TREC-7, and TREC-8 of the AdHoc track, and TREC-2005 HARD track. Table 2 summarises the statistics of the test collections. For pre-processing, we apply the Porter stemmer and remove stop words using a small list of 127 common English terms.

Table 2: Test collections used in this paper

Name	Collection	# Documents
TREC 6	Disc4&5	551873
TREC 7 and 8	Disc4&5 without CR	523951
HARD 2005	AQUAINT	1033461

In order to compare the performance of the thresholds, we test a variety of threshold values for each model. The thresholds cover a set of values on both sides of our introduced thresholds: for 100 dimension $\{0.67, 0.70, 0.74, 0.79, 0.81, 0.86, 0.91, 0.94, 0.96\}$, 200 dimension $\{0.63, 0.68, 0.71, 0.73, 0.74, 0.76, 0.78, 0.82\}$, 300 dimension $\{0.55, 0.60, 0.65, 0.68, 0.70, 0.71, 0.73, 0.75\}$, and 400 dimension $\{0.41, 0.54, 0.61, 0.64, 0.66, 0.68, 0.70, 0.71, 0.75\}$.

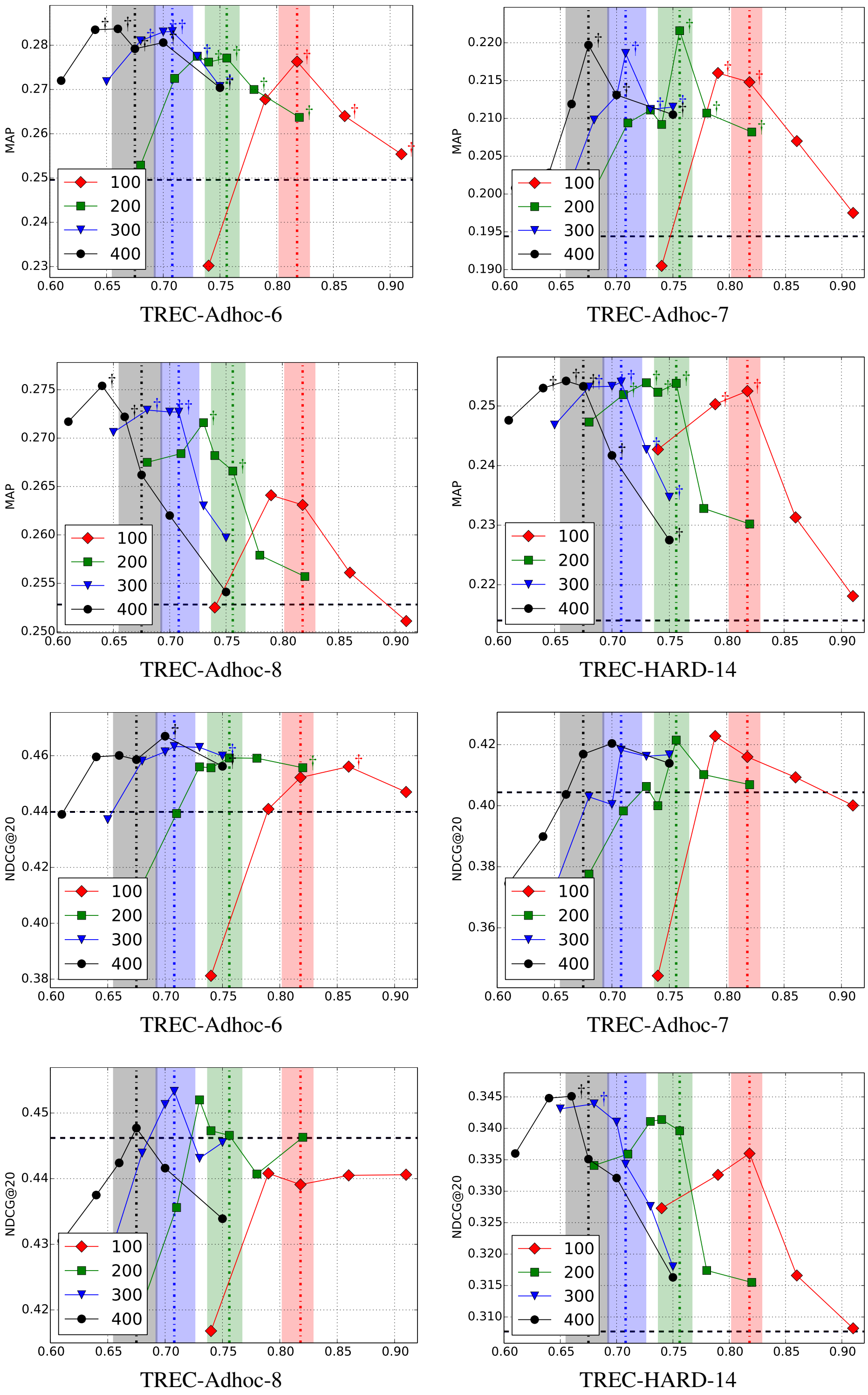
We set the basic models (language model or BM25) as baseline and test the statistical significance of the improvement of the translation models with respect to their basic models (indicated by the symbol $\dagger$). Since the parameter μ for Dirichlet smoothing of the translation language model and also b, k_1, and k_3 for BM25 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. [17]. The statistical significance test are done using the two sided paired t-test and statistical significance is reported for $p < 0.05$.

The evaluation of retrieval effectiveness is done with respect to Mean Average Precesion (MAP) and Normalized Discounted Cumulative Gain at cut-off 20 (NDCG@20), as standard measures in Adhoc information retrieval. Similar to Rekabsaz et al. [17] and in order to make the results comparable with this study, we consider MAP and NDCG over the condensed lists [18].

4 Results and Discussion

The evaluation results of the MAP and NDCG@20 measures of the BM25 Extended Translation (BM-ET) model on the four test collections, with vectors in 100, 200, 300, and 400 dimensions are shown in Figure 4. Due to lack of space, we only show the detailed results of the BM-ET model as it has shown the best overall performance among the other translation models in Rekabsaz et al. [17]. For each dimension, our threshold and its boundaries (the interval between the lower and upper bound in Table 1) are shown with vertical lines. The baseline (basic BM25) is shown in the horizontal line. Significant differences of the results to the baseline are marked by the † symbol.

The plots show that the performance of the translation models are highly dependent on the choice of the threshold value. In general, we can see a trend in all the models: the results tend to improve until reaching a peak (optimal threshold) and then converges to the baseline. Based on this general behaviour, we can assume that including the terms whose similarity values are less than the optimal threshold introduces noise and deteriorates the results while using the cutting point greater than the optimal threshold filters the related terms too strictly. We test the statistical significance between the results of



TREC-Adhoc-6

TREC-Adhoc-7

TREC-Adhoc-8

TREC-HARD-14

TREC-Adhoc-6

TREC-Adhoc-7

TREC-Adhoc-8

TREC-HARD-14

Fig. 4: MAP (above) and NDCG@20 (below) evaluation of the BM25 Extended Translation model on TREC-6, TREC-7, TREC-8 Adhoc, and TREC-2005 HARD for different thresholds (X-axes) and word embedding dimensions. Significance is shown by †. Vertical lines indicate our thresholds and their boundaries in different dimensions. The baseline is shown by the horizontal line. To maintain visibility, points with very low performance are not plotted.

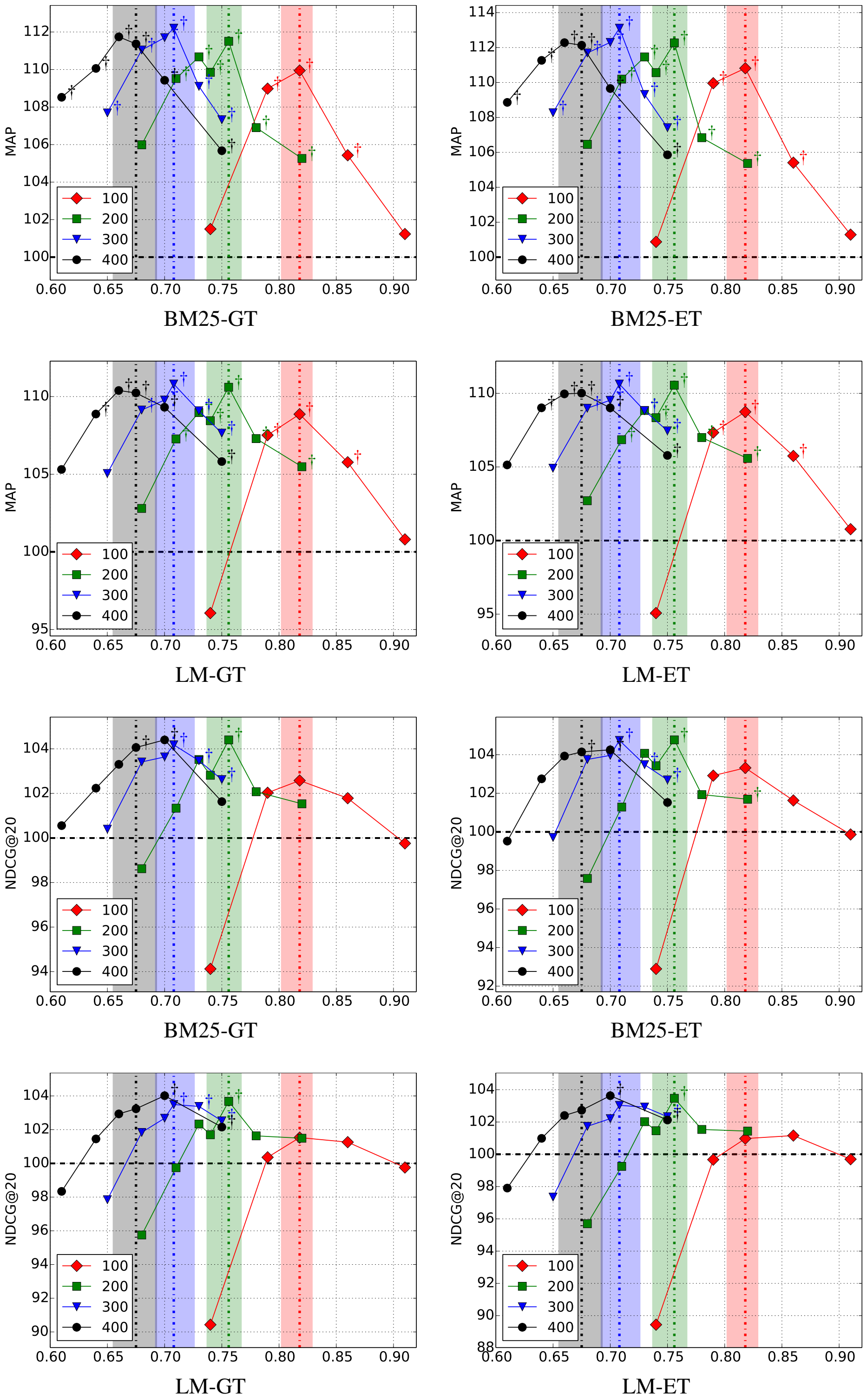


Fig. 5: Percentage of improvement of the relevance scoring models BM25 and Language Model (LM), combined with the Generalized Translation (GT) and Extended Translation (ET) models with respect to the baselines (standard LM and BM25) with the MAP (above) and NDCG@20 (below) evaluation measures for different thresholds, and word embedding dimensions, aggregated over all the collections.

the optimal and proposed threshold in all the experiments (both evaluation measures, all relevance scoring models, collections, and dimensions), observing no significant difference in any of the cases.

In order to have an overview of all the models, we calculate the gain of each relevance scoring model for different thresholds and dimensionalities over its corresponding baseline and average the gains on the four collections. The scoring models are BM25 and Language Model (LM), combined with the Generalized Translation (GT) and Extended Translation (ET) models. The results for MAP and NDCG are depicted in Figure 5. In all the translation models, our threshold is optimal for dimensions 100, 200, and 300. In dimension 400, the significance test between their results does not show any significant difference. These results justify the choice of the proposed threshold as a generally stable and effective cutting-point for identifying related terms.

To observe the effect of the proposed threshold, let us take a closer look at the terms, filtered as related terms. Table 3 shows some examples of the retrieved terms when using the word embedding model with 300 dimensions with our threshold (same as optimal in this dimension for all the translation models). As expected, the examples show the strong differences in the number of similar words for various terms. The mean and standard deviation of the number of similar terms for all the query terms of the tasks is 1.5 and 3.0 respectively. Almost half of the terms are not expanded at all. We can observe the similarity between this calculated mean and standard deviation and the aggregated number of synonyms we observed in WordNet in Section 2.3—mean of 1.6 and standard deviation of 3.1. It appears that although the two semantic resources (WordNet and Word2Vec) cast the notion of similarity in different ways and their provided sets of similar terms are different, they correspond to a similar distribution of the number of related terms.

Table 3: Examples of similar terms, selected with our threshold

book: publish, republish, foreword, reprint, essay
eagerness: hoping, anxious, eagerness, willing,wanting
novel: fiction, novelist, novellas, trilogy
microbiologist: biochemist, bacteriologist, virologist
shame: ashamed
guilt: remorse
Einstein: relativity
estimate, dwarfish, antagonize: no neighbours

5 Conclusion

We have analytically explored the thresholds on similarity values of word embedding to select related terms. Based on empirical observations on various models trained on the same data, we have introduced a method to identify the minimal cosine similarity value between two term vectors, allowing practical use of similarity values. The proposed threshold is estimated based on a novel representation of the neighbours around an arbitrary term, taking into account the variance of similarity values, captured from the values generated by different instances of identical models.

We extensively evaluate the application of the introduced threshold on four information retrieval collections using four state-of-the-art relevance scoring models. The results show that the proposed threshold is identical to the optimal threshold (obtained by parameter scan) in the sense that its results on ad-hoc retrieval tasks are either equal to or statistically indistinguishable from the optimal results.

Acknowledgement

This work is funded by: Self-Optimizer (FFG 852624) in the EUROSTARS programme, funded by EUREKA, the BMWFW and the European Union, and ADMIRE (P 25905-N23) by FWF. Thanks to Joni Sayeler and Linus Wretblad for their contributions in the SelfOptimizer project.

References

1. M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proc. of ACL Conference*, 2014.
2. A. Berger and J. Lafferty. Information Retrieval As Statistical Translation. In *Proc. of SIGIR*, 1999.
3. A. Cuba Gyllensten and M. Sahlgren. Navigating the semantic horizon using relative neighborhood graphs. In *Proc. of EMNLP*, Lisbon, Portugal, 2015.
4. L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, and P. Bruza. Medical semantic similarity with a neural language model. In *Proc. of CIKM*.
5. K. Erk and S. Padó. Exemplar-based models for word meaning in context. In *Proc. of ACL*, 2010.
6. D. Ganguly, D. Roy, M. Mitra, and G. J. Jones. Word Embedding based Generalized Language Model for Information Retrieval. In *Proc. of SIGIR*, 2015.
7. M. Grbovic, N. Djuric, V. Radosavljevic, F. Silvestri, and N. Bhamidipati. Context-and content-aware embeddings for query rewriting in sponsored search. In *Proc. of SIGIR*, 2015.
8. J. Karlgren, M. Bohman, A. Ekgren, G. Isheden, E. Kullmann, and D. Nilsson. Semantic topology. In *Proc. of CIKM Conference*, 2014.
9. J. Karlgren, A. Holst, and M. Sahlgren. Filaments of meaning in word space. In *Proc. of ECIR Conference*, 2008.
10. D. Kiela, F. Hill, and S. Clark. Specializing word embeddings for similarity or relatedness. In *Proc. of EMNLP*, 2015.
11. B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In *Proc. of CIKM*, 2012.
12. G. Kruszewski and M. Baroni. So similar and yet incompatible: Toward automated identification of semantically compatible words. In *Proc. of NAACL*, 2015.
13. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
14. B. Mitra. Exploring session context using distributed representations of queries and reformulations. In *Proc. of SIGIR*, 2015.
15. J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. of SIGIR*, 1998.
16. N. Rekabsaz, R. Bierig, B. Ionescu, A. Hanbury, and M. Lupu. On the use of statistical semantics for metadata-based social image retrieval. In *Proc. of CBMI Conference*, 2015.
17. N. Rekabsaz, M. Lupu, and A. Hanbury. Generalizing translation models in the probabilistic relevance framework. In *Proc. of CIKM*, 2016.
18. T. Sakai. Alternatives to bpref. In *Proc. of SIGIR*, 2007.
19. T. Schnabel, I. Labutov, D. Mimno, and T. Joachims. Evaluation methods for unsupervised word embeddings. In *Proc. of EMNLP*, 2015.
20. A. Severyn and A. Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *Proc. of SIGIR*, 2015.
21. Y. Tsvetkov, M. Faruqui, W. Ling, G. Lample, and C. Dyer. Evaluation of word vector representations by subspace alignment. In *Proc. of EMNLP*, 2015.
22. I. Vulić and M.-F. Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proc. of SIGIR*, 2015.