Query reformulation based on word embeddings: A comparative study

Panos Panagiotou, George Kalpakis, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris

Information Technologies Institute, Centre for Research and Technology Hellas
{panagiotou,kalpakis,theodora.tsikrika,stefanos,ikom}@iti.gr

Abstract. Formulating effective queries for retrieving domain-specific content from the Web and social media is very important for practitioners in several fields, including law enforcement analysts involved in terrorism-related investigations. Query reformulation aims at transforming the original query in such a way, so as to increase the search effectiveness by addressing the vocabulary mismatch problem. This work presents a study comparing the performance of global versus local word embeddings models when applied for query expansion. Two query expansions methods are employed (i.e., CombSum and Centroid) for defining the most similar terms to each query term, based on Glove pre-trained global embeddings and local models trained on four large-scale benchmark and one terrorism-related datasets. We assessed the performance of the global and local models on the benchmark datasets based on commonly used evaluation metrics, and performed a qualitative evaluation of the respective models on the terrorism-related dataset. Our findings indicate that the local models yield promising results on all datasets.

Keywords: query expansion · word embeddings · terrorism

1 Introduction

Given the abundance of online information, the discovery of content of interest by formulating and submitting queries to search engines and social media platforms is of paramount importance for practitioners in several fields, including experts involved in crime- and terrorism-related investigations. Effective information retrieval requires though that the submitted query includes terms relevant to the vocabulary used in the sought content, so that the query and the available information are successfully matched. As this is a challenging task, automatic query reformulation, including term expansion, substitution, and reduction, can be employed so as to increase the likelihood of retrieving relevant documents higher in the rankings, even if they do not contain the terms in the original query.

The task of query reformulation usually requires the representation of the terms occurring in documents and the query in a way that effectively depicts their meaning and overall semantics; typically, vector representations are employed for this purpose. Into this direction, word embeddings have recently attracted much attention due to their effectiveness. Word embeddings are real-valued vector representations of terms that are produced by neural network-based algorithms and that rely on the co-occurrence statistics of terms in a document corpus. The word embeddings models are distinguished between global and local, based on the corpus used for their generation; the former entail the use of broad corpora covering a variety of topics, whereas the latter are based on more domain-specific corpora. The most popular word embedding algorithms are all neural network-based approaches and include Word2Vec [11], GloVe [12], and FastText [2].

In the particular case of terrorism-related material, the submission of effective queries to search engines and social media platforms is of vital significance for Law Enforcement and Intelligence Services, in terms of discovering and retrieving online content of interest for their ongoing investigations. To this end, query reformulation is an important tool that helps the investigators construct more effective queries, thus quickly reaching online content of interest that may not be discovered through the manual query formulation.

In this work, we compare the performance of global versus local embeddings models when applied for query expansion using five datasets (four benchmark and one terrorism-related dataset). In particular, we apply two query expansion methods (i.e., CombSum and Centroid) for identifying the most similar terms to each query term, using global and local word embeddings models, trained on our datasets. We focus on the Glove algorithm, where co-occurrences are calculated by moving a sliding n-words window over each sentence in the corpus. We assess the effectiveness of 100- and 300-dimensional global and local word embeddings models on the four benchmark datasets based on commonly used evaluation metrics in information retrieval, and we also perform a qualitative evaluation of the efficacy of the respective models on the terrorism-related dataset.

The remainder of this paper is organised as follows: Section 2 discusses related work and Section 3 describes word embeddings approaches and the query expansion methods. Section 4 outlines the evaluation process and Section 5 presents the experimental results. Finally, Section 6 summarises our conclusions.

2 Related Work

The effectiveness of different query expansion methods using word embeddings on the retrieval task is discussed in [10] which reports that both the CombSUM and the Centroid methods (originally proposed in [7]) for combining the word embeddings of the query terms yield similar results. In addition, recent work has shown that a retrieval process employing query expansion based on local word embeddings can outperform a solution that uses global word embeddings [5]. As far as the appropriate dimensionality of an embeddings model is concerned, it has been shown that, although there is a bias-variance trade-off in the dimensionality selection for word embeddings, the GloVe algorithm, as well as the skip-gram variation of Word2Vec (which uses a word to predict a target context), are robust to over-fitting [14]. This means that although there exists an optimal dimensionality that is dependent on the training corpus, using a greater number of dimensions is not so harmful for the performance of the aforementioned embeddings, according to experiments in Natural Language Processing tasks. In this work, we compare 100- and 300-dimensional embedding models.

3 Methods

This section presents in more detail (i) word embeddings and their applicability in query expansion, and (ii) the query expansion methods employed in this paper.

3.1 Word embeddings

Word embeddings are real-valued vector representations of terms, produced by neural network-based algorithms that adopt the distributional hypothesis [8, 4], which states that words occurring in similar context tend to have similar meanings. Formally, in a word embeddings model, a term t in a vocabulary V is represented in a latent space of k dimensions by a dense vector $\boldsymbol{t} \in R^{|k|}$. In the trained word embeddings space, similar words converge to similar locations in the k-dimensional space.

The neural network-based algorithms can be applied on any available corpus of documents in order to learn the word embeddings representations of the terms that exist in the given corpus. The most typical data sources for generating new terms include: (i) large-scale external corpora that can be considered to reflect the overall term distribution in a given language, such as all Wikipedia articles in a given language [1], (ii) a document collection being searched in the current setting [13] that can be viewed as modelling term distribution in a particular domain, and (iii) documents relevant to the submitted query which are identified either interactively by the user or automatically by the system; in the former case, i.e., in the so-called relevance feedback cycle, the user pro-actively provides guidance in the form or relevant reference documents [6], while in the latter case, referred to as pseudo-relevance feedback, the top retrieved documents are assumed to be relevant [13].

If large-scale corpora, covering a sufficient number of diverse topics, are employed, the word embeddings generated on their basis are able to encode a broad context that enables their applicability in a variety of domains; we refer to such embedding models as *global* or *universal*. On the other hand, word embeddings models learned on domain-specific corpora may be more beneficial in uncovering term relationships for terms with specific interpretations in those particular domains and contexts; such embedding models are referred to as *local*.

Given a user query and a trained (global or local) word embeddings model, the goal of query expansion is to identify the top-r most relevant terms to (i) each individual query term or to (ii) the query as a whole, with r being a parameter to be defined by the user or the system. Those identified terms are then used for expanding the original query. The first option is the simplest in its implementation; the r most similar terms to each individual query term are identified using a similarity metric, such as cosine similarity, and they are added to the query. The second option requires more intricate techniques for queries that contain more than one term, but it is a more powerful solution.

3.2 Query expansion

Irrespective of the algorithm deployed to produce the word embeddings, the linguistic or semantic similarity between two terms w_i, w_j is typically measured using the cosine similarity between their corresponding embedding vectors:

$$sim(\boldsymbol{w_i}, \boldsymbol{w_j}) = cos(\boldsymbol{w_i}, \boldsymbol{w_j}) = \frac{\boldsymbol{w_i}^T \boldsymbol{w_j}}{\|\boldsymbol{w_i}\| \|\boldsymbol{w_j}\|} \tag{1}$$

Given a trained word embeddings model, the CombSUM and Centroid methods, presented in [10], are considered for the definition of the similarity of a term t (whose corresponding embedding is $\boldsymbol{t}$) to a query q consisting of M terms q_i where i=1,. . . ,M (with corresponding embeddings $\boldsymbol{q_i}$).

CombSUM method The similarity score of each of the vocabulary terms to the query is calculated separately for every query term, and then a list L_{q_i} is produced for each query term q_i, containing the top n most similar terms. Subsequently, for each of the terms t that are included in L_{q_i} the final similarity score is softmax normalised, so that it is in the form of probability $p(t|q_i) = \frac{exp(cos(\boldsymbol{q_i},\boldsymbol{t}))}{\sum_{t' \in L_{q_i}} exp(cos(\boldsymbol{q_i},\boldsymbol{t'}))}$, while $p(t|q_i) = 0$ for the terms $t \notin L_{q_i}$. Finally, the resulting term lists are fused so that the final similarity score between a query and a vocabulary term is defined as follows:

$$S_{CombSUM(t,q)} = \sum_{q_i \in q} p(t|q_i) \tag{2}$$

Centroid method The centroid method is based on the observation that the semantics of an expression can often be adequately represented by the sum of the vectors of its constituting terms. Consequently, a query q can be represented by a vector $\boldsymbol{Q_{cent}} = \sum_{q_i \in Q} \boldsymbol{q_i}$ and the similarity score between a vocabulary term and a query is defined as:

$$S_{cent(t,q)} = exp(cos(\boldsymbol{t}, \boldsymbol{Q_{cent}})) \tag{3}$$

where $\boldsymbol{t}$ denotes the L_2-normalised vector of a term t.

4 Evaluation

This section describes the experiments performed in order to assess the performance of the global and local word embeddings models in query expansion.

4.1 Experiments on Benchmark Datasets

The first set of our experiments was performed on benchmark datasets. In particular, we used the ClueWeb2009 Category B corpus[1] which has been extensively used by the TREC conference[2]. The corpus consists of 50,220,423 English-language Web pages which cover a wide range of subjects. These benchmark datasets are widely employed in order to assess the effectiveness of information retrieval and acquisition methods and thus allow us to determine the methods that are likely to provide the best results in an operational setting.

We used the topics of TREC 2009, 2010, 2011, and 2012 Web Tracks as queries in our evaluation experiments; each of these TREC tracks consists of a set of 50 topics (queries). Initially, we retrieved the top-1000 documents for each of the queries of a query set and then used the superset of those documents to train a local embeddings model that corresponds to this query set. In fact, two GloVe models were produced by each query set, that differ in the dimensions of the embeddings. Specifically, we trained 100-dimensional and 300-dimensional embeddings. When applying the local models for the query expansion process in the retrieval experiments, we use those models that correspond to the query set where the specific query belongs.

The process of building each of the local GloVe models involved an initial step of extracting the main content of each retrieved Web page and removing its boilerplate using the python implementation of boilerpipe (Kohlschütter et al., 2010). We have also experimented with the exact vocabulary for which the embeddings were built. More specifically, in an attempt to deal with the problem of mis-spelled terms, we considered embedding models where terms existing in only one Web document were not taken into account by the learning process. Indeed, those models completely outperformed models that included the complete set of words in the collection. In addition, the exclusion of terms with a frequency of less than a threshold of five has led to improved retrieval performance in most of the cases. Therefore, we consider those local models built with this specific process for our further analysis.

As for the global embeddings, we used two of the GloVe embeddings trained on the union of Wikipedia 2014 and Gigaword 5 datasets, specifically, the 100-dimensional and the 300-dimensional models[3].

For each combination of query and embedding model, we performed retrieval using both the expansion methods presented above. In addition, we variated the number k of the expansion terms; $k = 5, 10, 25, 50$. For the retrieval process, we used the Indri search engine[4]. The initial query was combined with the expansion terms. Moreover, it was associated with a weight of 0.8, while the set of expansion terms was given a weight of 0.2. In total, we have conducted thirty-two experiments for each query, i.e., all the combinations of four embedding models (i.e., local and global GloVe-based models of 100 and 300 dimensions), two

[1] https://lemurproject.org/clueWeb09.php/

[2] https://trec.nist.gov/

[3] https://nlp.stanford.edu/projects/glove/

[4] http://boston.lti.cs.cmu.edu/Services/

expansion methods (i.e., CombSUM and Centroid methods), and four different values for the parameter k.

We tested the performance of the retrieval processes using four commonly used evaluation metrics, namely MAP (Mean Average Precision), $P@k$ (Precision at k corresponds to the number of relevant results among the top k retrieved documents), $nDCG@k$ (Discounted Cumulative Gain), and $ERR@k$ (Expected Reciprocal Rank). For all the models with parameter k, we use $k = 20$. Both $nDCG$ [9] and ERR [3] are designed for situations of non-binary notions of relevance and ERR is an extension of the classical reciprocal rank.

4.2 Experiments on a Terrorism-Related Dataset

The second set of our experiments was performed on a terrorism-related dataset consisting of 329 Web pages containing text in English. This set was collected by domain experts and consists of Web pages referring to the religion of Islam and islamism, to the Islamic State (ISIS), as well as pages containing news and references related to the region of Middle East (i.e. Israel, Palestine, Saudi Arabia etc.). The content of the Web pages was downloaded and scraped. Similarly with the experiments on the benchmark datasets, we have employed the boilerpipe algorithm in order to remove content such as navigational elements, templates, advertisements, etc. The local embedding models were produced using the derived dataset based on the Glove algorithm.

Given the small vocabulary size of this dataset (i.e., consisting of 7,651 distinct terms), in order to produce the local Glove models, we experimented with the *window size* and the number of *epochs*. After experimental tuning, we produced models with window sizes of 5 and 10, as well as $epochs = 50$ trained on 100 dimensions; we refer to the derived models as *local-wind5* and *local-wind10*, respectively. For our experiments we used the 100-dimensional and the 300-dimensional global word embeddings employed at the experiments on the benchmark datasets; we refer to these models as *global100d* and *global300d*, respectively. In order to compare the efficacy of the retrieval process of the global versus the local word embeddings models on the terrorism-related dataset, we extracted the top three terms generated by the two global and the two local word embeddings variations for a number of terrorism-related search terms.

5 Results

This section presents the evaluation results of the experiments on both the benchmark datasets and the terrorism-related dataset.

5.1 Benchmark Datasets

Following the experimental setting on the benchmark datasets, we computed the mean performance for each combination of an embeddings model with a query expansion process when applied to a query set, using the four evaluation metrics. We took this approach of analysis to better present and interpret the results.

Figures 1, 2, 3, and 4 present those mean performances, comparing the efficacy of the local models versus the global ones, for each query set and evaluation metric. Each plot depicts the results obtained with both the 100- and 300-dimensional models, in order to analyse the effect of the dimensionality of the models and the interdependence of the model's origin and dimensionality. In each plot, the eight different points corresponding to each embeddings model (i.e., local or global) represent different combinations of the expansion method and the number of expansion terms used. Specifically, each point in the plots represents the average performance of experiments that use the same model, expansion method, and number of expansion terms. Blue points represent the 100-dimensional models and orange the 300-dimensional ones.

At a first level of analysis, the local models outperform the global ones when measured by the $ERR@20$ metric for the query sets of TREC 9 and 10, by MAP and $P@20$ for TREC 11, and by $ERR@20$ for TREC 12. On the other hand, the global models perform better than the local ones when measured by MAP for the queries of TREC 10. As far as the dimensionality is concerned, the

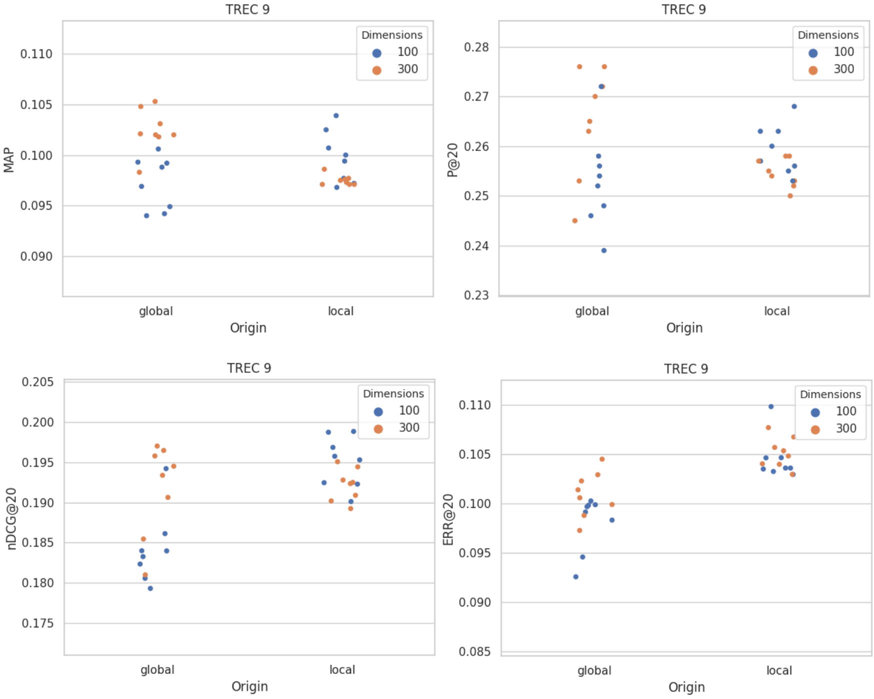

Fig. 1. The average performances of the local and global models for the query set of TREC 9, for the metrics MAP, P@20, nDCG@20 and ERR@20 on the retrieval task.

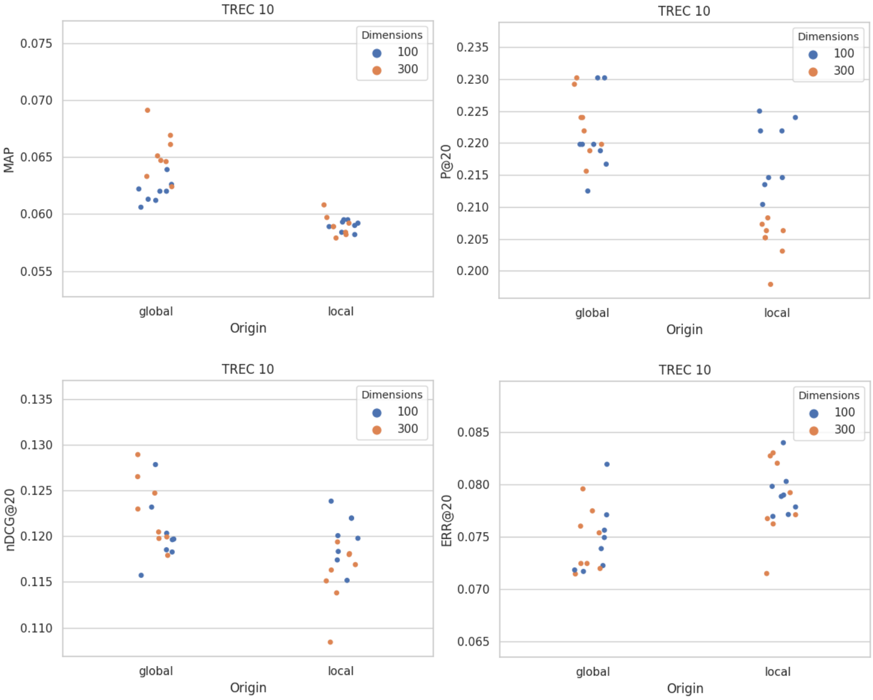

Fig. 2. The average performances of the local and global models for the query set of TREC 10, for the metrics MAP, P@20, nDCG@20 and ERR@20 on the retrieval task.

300-dimensional models outperform the 100-dimensional ones in MAP, P@20, and $nDCG$@20 for the queries of TREC 12. Overall, the results of Wilcoxon signed-rank test (non-parametric statistical hypothesis test) imply that both the origin of the model and its dimensions are important parameters in the retrieval process, since a modification in our choice for any of those parameters yields statistically significant change in the performance.

At a second level, we observe that the optimal decision regarding the origin of an embeddings model and its dimensionality are interdependent in many cases. On the one hand, there are cases where the comparison of the local models versus the global ones gives a specific outcome when considering only the 100-dimensional models, but a different one when considering only the 300-dimensional models. As an example, consider the MAP for the TREC 9 queries; the 100-dimensional local models are better than the 100-dimensional global models, but the opposite is observed for the 300-dimensions. Similarly, when observing from the dimensionality point of view, in many cases it is clear that the origin of the model also affects the outcome. For example, in TREC 9 and according to all metrics, the 300-dimensional global models outperform the 100-

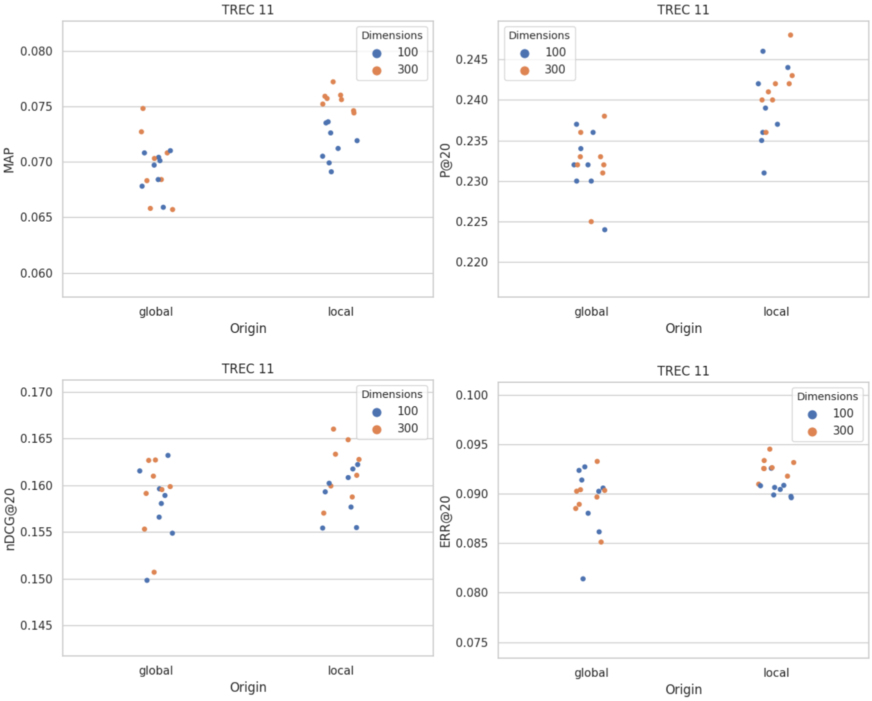

Fig. 3. The average performances of the local and global models for the query set of TREC 11, for the metrics MAP, P@20, nDCG@20 and ERR@20 on the retrieval task

dimensional global ones, but among the local models the 100-dimensional are better, according to MAP, $P@20$, and $nDCG@20$.

As far as the expansion method is concerned, we paired experiments that share the same query, type of embeddings model, and number of expansion terms, but differ on the expansion method. The Wilcoxon signed-rank test between those pairs has shown that choosing among the investigated expansion methods does not elicit a statistically significant change in the performance of the retrieval, for any of the evaluation metrics considered. This outcome is on par with the findings of (Kuzi, 2016) and it is important, especially when we consider the efficiency of the two expansion methods, since the centroid method is much more preferable than the CombSUM method in terms of execution time.

5.2 Terrorism-Related Dataset

Table 5.2 presents the most relevant words to a number of search terms provided by domain experts based on their relevance to the terrorism domain, after employing the four embedding models used in this experimental set up.

The results illustrate the differences and complementarity between the local and global word embeddings on the presented search terms. While global

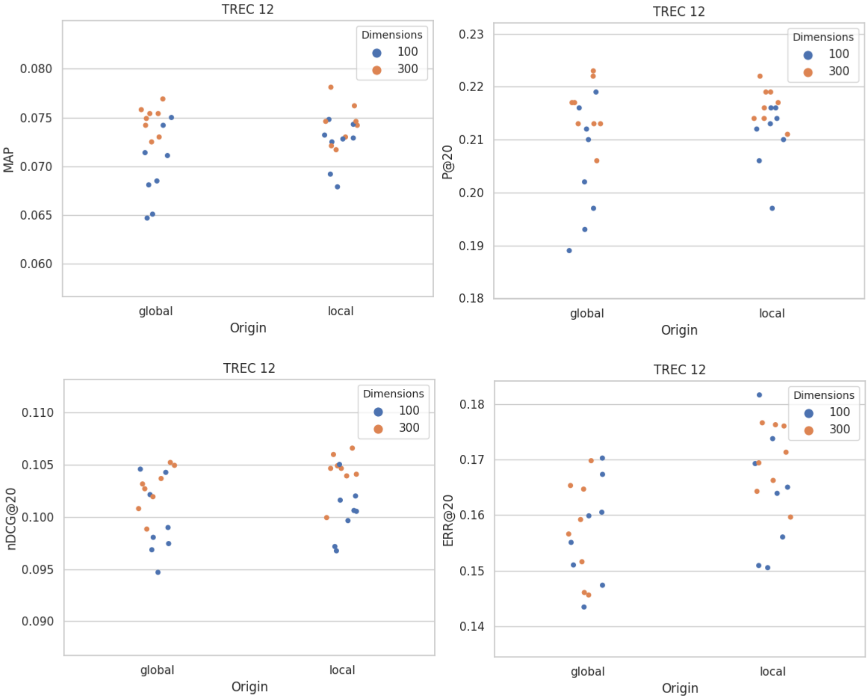

Fig. 4. The average performances of the local and global models for the query set of TREC 12, for the metrics MAP, P@20, nDCG@20 and ERR@20 on the retrieval task.

word embeddings capture the overall context, local word embeddings provide interpretations relevant to the particular domain.

Consider for instance the term “karbala” that is relevant to “martyrdom” according to the *local-wind10* model. This term most likely refers to the Battle of Karbala that was fought on October 680 between the army of the second Umayyad caliph Yazid I and a small army led by Husayn ibn Ali, the grandson of the Islamic prophet Muhammad; Husayn and his companions are widely regarded as martyrs by both Sunni and Shi’a Muslims[5]. It is thus evident in this case that the local models provide related terms within the particular context of interest, while the global models provide more universally related terms, and in particular terms with the same root as the term “martyrdom”.

Furthermore, the local models output “syria” as a term relevant to “war”, while the global models have a preference over more general terms. The same also applies to the outputs for the search term “believers”, such as “thabit” vs. “adherents”; the former is indeed related to the particular context of interest, while the latter is virtually a synonym to the search term “believers” and therefore could be considered in any context, and not only in this specific one.

[5] https://en.wikipedia.org/wiki/Battle_of_Karbala

Table 1. Top-3 most similar terms to the search terms based on global and local word embeddings.

Search term	local-wind5	local-wind10	global100d	global300d
Allah	messenger blessings merciful	exalted messenger blessings	god almighty unto	god almighty bless
almighty	god blessings allah	exalted attributes accept	allah god bless	allah god merciful
apostates	victory software fight	al-raqqah arabulus alab	infidels unbelievers traitors	infidels unbelievers heretics
believers	thabit rah rabbi	camp thabit learned	christians adherents catholics	christians adherents nonbelievers
jihad	terrorism tomorrow intro	terrorism compared converting	militant hamas islamic	militant hamas islamic
martyrdom	hell paradise interview	karbala thinking al-saleheen	martyr resurrection martyrs	martyr martyred martyrs
soldiers	perish fiqhnamaz libya	tools rabab peaceful	troops army policemen	troops army policemen
war	crime syria crimes	crime syria crimes	conflict invasion military	conflict battle civil

Finally, there are cases, where the local models yield possibly unrelated terms; however, this may be attributed to the very small size of the domain-specific dataset on which those models were built.

6 Conclusions

In this work, we compared the performance of global versus local word embeddings models for the task of query expansion based on four large-scale benchmark datasets and one domain-specific dataset related to terrorism. With regards to the benchmark datasets, our findings indicate that local models outperform global ones for the majority of the experiments run and the metrics employed. At the same time, it is evident that there is an interdependency among the origin of a model and its dimensionality. Regarding the terrorism-related dataset, we found that the local models delivered relevant words to a number of terrorism-related search terms, despite the small size of the corpus. The domain could benefit from larger domain-specific corpora for building embeddings models that can better capture the semantic relationships in a relevant vocabulary.

Acknowledgements

This work was supported by the TENSOR (H2020-700024) and the CONNEXIONs (H2020-786731) projects, both funded by the European Commission.

References

1. Balog, K., Weerkamp, W., De Rijke, M.: A few examples go a long way: constructing query models from elaborate query formulations. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 371–378. ACM (2008)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
3. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 621–630. ACM (2009)
4. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)
5. Diaz, F., Mitra, B., Craswell, N.: Query expansion with locally-trained word embeddings. arXiv preprint arXiv:1605.07891 (2016)
6. Efthimiadis, E.N.: Interactive query expansion: A user-based evaluation in a relevance feedback environment. Journal of the American Society for Information Science **51**(11), 989–1003 (2000)
7. Fox, E.A., Shaw, J.A.: Combination of multiple searches. NIST special publication SP **243** (1994)
8. Harris, Z.S.: Distributional structure. Word **10**(2-3), 146–162 (1954)
9. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS) **20**(4), 422–446 (2002)
10. Kuzi, S., Shtok, A., Kurland, O.: Query expansion using word embeddings. In: Proceedings of the 25th ACM international on conference on information and knowledge management. pp. 1929–1932. ACM (2016)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
12. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
13. Xu, J., Croft, W.B.: Quary expansion using local and global document analysis. In: Acm sigir forum. vol. 51, pp. 168–175. ACM (2017)
14. Yin, Z., Shen, Y.: On the dimensionality of word embedding. In: Advances in Neural Information Processing Systems. pp. 887–898 (2018)