Local Word Embeddings for Query Expansion based on Co-Authorship and Citations

André Rattinger[1,2], Jean-Marie Le Goff[1], and Christian Guetl[2]

[1] CERN, Switzerland
andre.rattinger@cern.ch, jean-marie.le.goff@cern.ch
[2] Graz University of Technology, Austria
cguetl@iicm.edu

Abstract. Word embedding techniques have gained a lot of interest from natural language processing researchers recently and they are valuable resource in identifying a list of semantically related terms for a search query. These related terms build a natural addition for query expansion, but might mismatch when the application domains use different jargon. Using the Skip-Gram algorithm of Word2Vec, terms are selected only from a specific subset of the corpus, which is extended by documents from co-authorship and citations. We demonstrate that locally-trained word embeddings with this extension provides a valuable augmentation and can improve retrieval performance. First result suggest that query expansion and word embeddings could also benefit from other related information.

Keywords: word embeddings, query expansion, co-authorship, word2vec, pseudo relevance feedback

1 Introduction

Methods to create fixed representations of words and documents have long been a staple of natural language processing (NLP) and information retrieval (IR) research. Recently neural network based method of generating those representations have gained popularity in IR. Models such as Word2Vec [10], transform the terms from a document into high-dimensional vectors. Semantically similar terms in those vector representations are close to each other and the size of the vectors is much smaller than the size of the vocabulary compared to traditional methods. This work focuses on the IR task of query expansion, and the applicability of word embeddings, even if the dataset for training is limited. Word embeddings provide a good fit for query expansion, as it can aid with the vocabulary mismatch between the query and the relevant documents. Embeddings trained on a small topically-relevant corpus promise embeddings that produce better fitting terms for a specific area. The size of the dataset can be a limitation in training and the subsequent expansion process, as the quality of word embeddings benefits from bigger datasets. We therefore propose to use documents for the expansion process that promise to be relevant by association

with the retrieved results: referenced documents and documents from co-authors. We test our approach on a small topically-relevant corpus and compare the results with a bigger more general dataset. The smaller corpus is from a specific topically-relevant subsection of research papers, computational linguistics. The bigger dataset is made up of patents from all patent classes, and is therefore very general. For an additional comparison we also perform the same test with general purpose embeddings trained on articles from the English-language edition of Wikipedia. The more detailed description of the publication and patent datasets can be found in Section 3. Section 4 describes the general approach of the local query expansion method and Section 5 describes the experimental setup for the retrieval experiments. The results of the different retrieval experiments can be found in Section 6 and Section 7 concludes the paper.

2 Related Work

A few attempts have been made at expanding queries with word embeddings. Roy et al. [13] demonstrate the effect generalization has on retrieval performance when using word embeddings. It was shown that while global methods can increase overall retrieval performance, they perform worse than co-occurrence based techniques. Diaz et al. [4] recently proposed a method for locally-trained word embeddings for query expansion, and is the closest to the work presented here. The difference between the work and our research is the application of the methods, the focus on pseudo relevance feedback and the implications of additional documents on the query expansion process. A different approach is the incorporation of word embeddings and using them to weight terms that are not part of the query [15]. This approach is similar to ours, but it uses different weighting scheme and does not operate on a local basis. Query expansion over the corpus which is indexed was previously performed and incorporated with pseudo relevance feedback [7], but with fairly big datasets which do not provide the same degree of locality. The scope of the used datasets is similar to the patent dataset presented here, however. Another approach is to use information from different local context, as was done as part of personalization of word embeddings [2]. This approach did not provide promising results as other localized methods did though.

3 Datasets

We conduct our experiments on two datasets: the ACL anthology collection and the English subset of the CLEF-IP 2011 collection. The ACL collection [12] is a small information retrieval dataset containing almost 10,000 research papers from the ACL anthology. The documents are scientific publications from the field of computational linguistics. It includes 82 research questions and their relevance assessments. As this is a small dataset for information retrieval, the information is supplemented by other documents available to us. These include other research papers from the authors as well as the work they cite in the articles contained in

the collection. With this addition, the dataset contains 33,922 research articles. The additional research papers are used in query expansion, but not for the main indexing and retrieval. They are also not part of the relevance assessments.

The English-language subset of the CLEF-IP 2011 collection [11] contains about 400,000 patents and 1,350 topics. Compared to the ACL dataset, a query is not represented by a set of terms, but by a whole patent document. The search terms have to be extracted from the document. The reason for this is that the goal in patent retrieval is to find similar documents that might invalidate a patent application. To generate the search queries, terms in the documents are weighted with tf-idf to extract the most relevant words from a document. Search queries with a length of 30 terms produced the best results and are used as a baseline for further experiments. No supplementation or addition with other documents is performed because of the size of the dataset is deemed sufficient. Citations are considered in the experiments if they are citing patents within the corpus. Patent citation promise to be valuable because they are not only added by the author, but also by the patent examiner. The CLEF-IP collection as a whole is used as a reference corpus to show the effect of query expansion on a dataset that is not as topically constrained as the ACL dataset.

4 Local Query Expansion

Local methods for query expansion generally perform better than their global counterparts. This holds true for word embeddings as well as other techniques [4, 14]. The ACL collection represents a subset of research papers which focuses on a specific topic. This lends itself well for the training of word embeddings compared to a big general dataset. The applicability to a smaller local context can be demonstrated with a small example: Latent Semantic Indexing (LSI) [3] is a well-known NLP technique used in IR. When looking at the most similar words generated by the word embedding model for the term "latent", the global model generates terms such as "inherent", "suppresses", "innate", "inhibition" or "implicit". Some of those words provide actual synonyms in an overall context, similar to what a thesaurus would provide. The local embedding version trained on the ACL collections provides a different represenatation of the data. Similar terms to "latent" are: "plsa", "lsa", "dirichlet", "allocation", "plsi" or "probabilistic". All of these are either terms in the direct context of the LSI technique or refer to similar techniques used in NLP applications. A similar observation is made in [4], where the word "cut" is studied in a global context and compared to a local one. To train local word embeddings, a set of documents is required that reflects the local context. This is provided by the top-ranked documents in retrieval as well as the documents from references and co-authorships.

5 Experimental Setup

We use the Skip-Gram algorithm from the Word2Vec model to train the word embeddings. The embeddings are used to choose the terms that are closest to

the query, by using the cosine similarity between the projected vectors. This is done for noun phrases in the query as well as the single query terms. The most similar words to the query terms are then incorporated into the expansion model. The expansion is based on the n most relevant documents from the retrieval run, a method which is also known as pseudo relevance feedback (PRF). PRF is a proven method for expanding a query and in doing so achieving better retrieval performances. This provides a natural addition to the expansion process and helps together with word embeddings to mitigate the vocabulary mismatch problem that arises in IR when different terms are used to describe the same concepts [5]. For the evaluation both of the datasets underwent several setup steps. The setup steps are the same for both datasets with a few exceptions, notably stopword removal and tokenization.

5.1 Pre-processing

As a preparation for indexing, the corpus for both of the datasets is tokenized with a regex tokenizer and transformed to lower case. The stopwords are filtered with the SMART stopword list[3]. The stopword lists were extended with query and publication specific stopwords. Stopword removal was only performed for indexing, but not for the word embedding models, as they help by providing context for the training of the models [8]. Krovetz stemming [6] was applied to all documents to reduce the overall vocabulary size. This was beneficial in training the word embedding models, as it creates a sparser input space for the comparatively small ACL dataset.

5.2 Indexing and Initial Retrieval

We are using an extension of the Bo1 model [1], a variant of the divergence from randomness (DFR) weighting model. Bo1 was chosen because it represents a stable version in the DFR framework [1]. Weights are assigned in the following way:

$$w(t) = t_f * \log_2 \frac{1+f}{f} + \log_2 1 + f \tag{1}$$

where t_f is the frequency of the terms within the set of top ranked feedback documents, and f is the term frequency of the term in the corpus divided by the N documents indexed from the whole document collection. Bo1 is used for initial weighting and candidate term selection, which provides us with a basis for measuring the information content of the different query expansion candidates. This is an important step in ranking them. For retrieval, the reference implementation of the Inverse Document Frequency model (InL2) for weighting from the terrier retrieval platform was used [9]. The first round of retrieval provides the basis for the pseudo relevance feedback. The number of feedback documents was set to 3, which produced the best overall results.

[3] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

5.3 Word2Vec Parameters and Learning of Embeddings

The initial Word2Vec models are learned on the whole corpus for both datasets. Another model is learned from the English-language edition of Wikipedia. The initial models are learned because training a full local model is very inefficient. The model provides several parameters that can be set to improve the model results. As the ACL dataset is small, the default number of iterations is set from 5 to 20. The minimum frequency of the words appearing in the corpus was set to 8. The window size, which represents the maximum distance between the word Word2Vec looks at and the word it is trying to predict within a sentence, was set to 7. The Skip-Gram algorithm delivered superior results for the dataset compared to the continuous bag of words (CBOW) algorithm. All of those settings produced the best results combined.

5.4 Retrieval and Query Expansion

Let Q be a query issued by the user, which can also represented as a list of terms $q_1, q_2, ..., q_n$, and C be the list of candidate terms for query expansion, represented as $c_1, c_2, ...c_k$. The initial set of C is selected out of all of the terms in the first m relevant documents, which includes all terms found in the documents. The pool of candidates C is then extended by all terms that appear in the reference section of the relevant documents. In addition to this, they are extended by similar documents of their co-authors, which creates an extended list of candidate terms, and their frequencies can then be used for the weighting by Bo1. The process of adding terms from co-author documents is only executed for the ACL dataset. For the resulting set of documents, the top k terms according to their weight assigned to them from Bo1 are used for further processing. The list of terms filtered by the stopword lists all provide the basis for the lookup of similar terms $e_1^{(i)}, e_2^{(i)}, ..., e_n^{(i)}$ with the Word2Vec model. Before generating candidate terms, the Word2Vec model is retrained on the same extended dataset the candidate term lookup was performed on. Training is done with the same settings as in the initial training step described in the previous section. The lookup of terms in the model creates another list of extended candidates, which is weighted by the Bo1 model.

6 Results

In this section, we present the results of both datasets and different configurations for query expansion. The results of both datasets have low retrieval performance in terms of the main metrics used, which can also be found in reference works considered to provide baselines [11, 12]. The datasets are challenging because of the low number of relevant documents for each query, as can be seen in Table 1. The following notation is used for the result tables:

– **Baseline** represents the retrieval without any query expansion method applied.

- **QE global** denotes the global query expansion approach with a general purpose query expansion model trained on a dataset from the English-language edition of Wikipedia.
- **QE local** is the locally-trained model.
- **QE local ext.** is the locally-trained model with the extension of reference documents and documents from co-authors.

Table 1. Comparison of the datasets in terms of vocabulary size, documents and relevance judgments.

Name	Topics	Vocab Size	Indexed Docs	Avg. Relevant Docs
ACL	82	329,490	9,793	23.67
CLEF-IP	1,350	2,648,818	420,193	7.2

Table 2 shows the performance in terms of Mean Average Precision (MAP), Precision at 5 and Precision at 10 for the retrieved documents through the different query expansion methods. All of the query expansion methods outperform the baseline result, but global query expansion does this by a very slight margin. Local embeddings perform generally better, with the local version reaching the highest retrieval results.

Table 2. Results comparing the local query expansion method against the baseline for the ACL research paper collection. The best results in a column are bold. Significance testing has been performed with the paired t-test and is denoted by *.

Method	MAP	P@5	P@10
Baseline	0.1497	0.2268	0.1683
QE global	0.1502	0.2268	0.1732
QE local	0.1623*	**0.2347**	0.1805
QE local ext.	**0.1713***	0.2314	**0.1822**

Table 3 shows the results for the CLEF-IP dataset. The local method does not score higher by the same margin as it does in the ACL dataset.

7 Conclusion and Discussion

In this paper we showed the implication of local query expansion by using documents from references and co-authors. The inclusion of those documents provides further information for term selection for the models on two datasets with low baseline performances. Extending the approach to generate a bigger list of candidates that are potentially relevant improved retrieval performance for the ACL

Table 3. Results comparing the local query expansion method against the baseline for the CLEF-IP patent collection.

Method	MAP	P@5	P@10
Baseline	0.0914	0.0630	0.0446
QE local	0.0916	0.0631	0.0448
QE local ext.	**0.0923**	**0.0636**	**0.0455**

dataset. For the CLEF-IP patent dataset only slight improvements can be observed and no statistic significance could be shown. One potential issue might be caused by the pre-training of the word embeddings. As training local embeddings is very costly and inefficient, retraining on a previously created dataset can speed up the training. The results might indicate that a certain level of topical relevance needs to be achieved for this approach to be effective, even if the system was trained on a relevant corpus. The addition of supplementary information from references and co-authors might not be as beneficial for datasets with better overall performance, as the number of retrieved documents that can be used reliably as a source for pseudo relevance feedback is greater. The retrieval results might be improved by switching the weighting of candidate terms from a distribution based method (Bo1) to association based term selection, which is used as a basis for other work in word embedding query expansion [15, 4]. Future work may help to shed more light on the implication of different weighting models as well as how topically restrained embeddings have to be in order to achieve the best results.

Bibliography

[1] Giambattista Amati. *Probability models for information retrieval based on divergence from randomness*. PhD thesis, University of Glasgow, 2003.

[2] Nawal Ould Amer, Philippe Mulhem, and Mathias Géry. Toward word embedding for personalized information retrieval. In *Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval*, 2016.

[3] Scott Deerwester. Improving information retrieval with latent semantic indexing. 1988.

[4] Fernando Diaz, Bhaskar Mitra, and Nick Craswell. Query expansion with locally-trained word embeddings. *arXiv preprint arXiv:1605.07891*, 2016.

[5] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[6] Robert Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 191–202. ACM, 1993.

[7] Saar Kuzi, Anna Shtok, and Oren Kurland. Query expansion using word embeddings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1929–1932. ACM, 2016.

[8] Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*, 2016.

[9] Craig Macdonald, Richard McCreadie, Rodrygo LT Santos, and Iadh Ounis. From puppy to maturity: Experiences in developing terrier. *Proc. of OSIR at SIGIR*, pages 60–63, 2012.

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[11] Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. Clef-ip 2011: Retrieval in the intellectual property domain. In *CLEF (notebook papers/labs/workshop)*, 2011.

[12] Anna Ritchie. *Citation Context Analysis for Information Retrieval*. PhD thesis, University of Cambridge, UK, 2008.

[13] Dwaipayan Roy, Debjyoti Paul, Mandar Mitra, and Utpal Garain. Using word embeddings for automatic query expansion. *arXiv preprint arXiv:1606.07608*, 2016.

[14] Xing Wei and W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2006.

[15] Hamed Zamani and W Bruce Croft. Embedding-based query language models. In *Proceedings of the 2016 ACM international conference on the theory of information retrieval*, pages 147–156. ACM, 2016.