Exploring Deep Learning for Query Expansion

Mohannad ALMasri, Catherine Berrut, and Jean-Pierre Chevallet

Université Grenoble Alpes
{mohannad.almasri,catherine.berrut,jean-pierre.chevallet}@imag.fr
LIG laboratory, MRIM group, Grenoble, France

User queries are usually too short to describe the information need accurately. Important terms can be missing from the query, leading to a poor coverage of the relevant documents. To solve this problem, automatic query expansion techniques leveraging on several data sources and employ different methods for finding expansion terms [2]. Selecting expansion terms is challenging and requires a framework capable of adding interesting terms to the query.

Different approaches have been proposed for selecting expansion terms. Pseudo-relevance feedback (PRF) assumes that the top-ranked documents returned for the initial query are relevant, and uses a sub set of the terms extracted from those documents for expansion. PRF has been proven to be effective in improving retrieval performance [4].

Corpus-specific approaches analyze the content of the whole document collection, and then generate correlation between each pair of terms by co-occurrence [6], mutual information [3], etc. Mutual information (MI) is a good measure to assess how much two terms are related, by analyzing the entire collection in order to extract the association between terms. For each query term, every term that has a high mutual information score with it is used to expand the user query.

Other approaches like semantic vectors and neural probabilistic language models, propose a rich term representation in order to capture the similarity between terms. In these approaches, a term is represented by a mathematical object in a high dimensional semantic space which is equipped with a metric. The metric can naturally encode similarities between the corresponding terms. A typical instantiation of these approaches is to represent each term by a vector and use a cosine or distance between term vectors in order to measure term similarity [1][7][8].

Recently, several efficient Natural Language Processing methods, based on Deep Learning, are proposed to learn high quality vector representations of terms from a large amount of unstructured text data with billions of words [5]. This high quality vector representation captures a large number of term relationships. We propose to investigate these term vector representations in query expansion. We then experimentally compare this approach with two other expansion approaches: pseudo-relevance feedback and mutual information.

First step, term vector representations are learned from a large amount of unstructured text data using Deep Learning. Several syntactic and semantic relationships are captured within these term vectors [5]. Each term is represented by a vector of a predefined dimension, a real-valued vector of a predefined dimension, 600 dimensions for example. Cosine or distance similarity could be used to measure the strength of semantic similarity between two term vectors, and to list the top similar terms for a given query term.

Second step, we collect top similar terms for each query term using their vector representations. Then, top similar terms for each query term are used to expand the user query.

Experiments are conducted on four CLEF medical collections:

- Image2010, Image2011, Image2012: contain short documents and queries.
- Case2011: contains long documents and queries.

Table 1, shows some statistics about them, *avdl* and *avql* are average length of documents and queries, respectively. These medical collections provide a huge amount of medical text that we need in the training phase.

Table 1. Test collections statistics.

Corpus	**#d**	**#q**	**avdl**	**avql**
Image2010	77,495	16	62.12	3.81
Image2011	230,088	30	44.83	4.0
Image2012	306,530	22	47.16	3.55
Case2011	55,634	10	2594.5	19.7

We use language modeling framework to evaluate expanded queries. Two smoothing methods of language models are tested: Jelinek-Mercer and Dirichlet. We use word2vec to generate deep learning vectors [5]. The word2vec tool takes a text corpus as input and produces the term vectors as output.

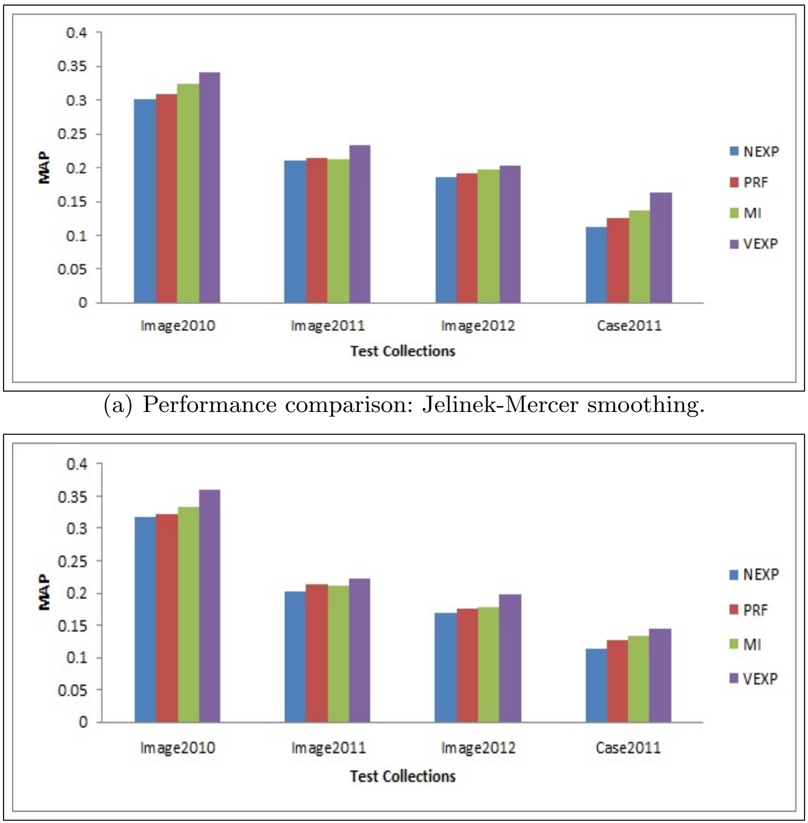
We compare three expansion methods: expansion using term vector representations (VEXP), Pseudo-relevance feedback (PRF), and Expansion using mutual information (MI). We use language models with no expansion as a baseline (NEXP).

Results are summarized in Figure 1. Experimental results show that the retrieval effectiveness can be significantly improved over the ordinary language models and pseudo-relevance feedback. We find that term vector representations, extracted using deep learning, are promising source for query expansion.

Deep learning vectors are learned from hundreds of millions of words, in contrast to PRF which is obtained from top retrieved document and MI which is calculated on the collection itself. Deep learning vectors are not only useful for collections that were used in the training phase, but also for other collections which contain similar documents.

References

1. Yoshua Bengio, Holger Schwenk, Jean-Sebastien Sencal, Frederic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. volume 194 of *Studies in Fuzziness and Soft Computing*, pages 137–186. Springer Berlin Heidelberg, 2006.



(a) Performance comparison: Jelinek-Mercer smoothing.

(b) Performance comparison: Dirichlet smoothing.

Fig. 1. Performance comparison using MAP on test collections: Image2010, Image2011, Image2012, Case2011.

2. Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50, January 2012.
3. Jiani Hu, Weihong Deng, and Jun Guo. Improving retrieval performance by global analysis. In *ICPR 2006.*, pages 703–706, 2006.
4. Victor Lavrenko and W. Bruce Croft. Relevance based language models. SIGIR '01, pages 120–127, New York, NY, USA, 2001. ACM.
5. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, 2013.
6. Helen J. Peat and Peter Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *J. Am. Soc. Inf. Sci*, 1991.
7. Midori Serizawa and Ichiro Kobayashi. A study on query expansion based on topic distributions of retrieved documents. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, pages 369–379. Springer Berlin Heidelberg, 2013.
8. D. Widdows and T. Cohen. The semantic vectors package: New algorithms and public tools for distributional semantics. In *ICSC*, pages 9–15, 2010.