Information Processing and Management 54 (2018) 1–13

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A Prospect-Guided global query expansion strategy using word embeddings

Francis C. Fernández-Reyes[a], Jorge Hermosillo-Valadez[a,*], Manuel Montes-y-Gómez[b]

[a] Centro de Investigación en Ciencias-(IICBA), Universidad Autónoma del Estado de Morelos, Av. Universidad 1001, Cuernavaca, Morelos 62209, Mexico
[b] Instituto Nacional de Astrofísica, Óptica y Electrónica, Santa María Tonantzintla, Puebla 72840, Mexico

ARTICLE INFO

Article history:
Received 10 February 2017
Revised 27 June 2017
Accepted 9 September 2017

Keywords:
Global query expansion
Word embeddings
Information retrieval
Candidate terms pooling methods

ABSTRACT

The effectiveness of query expansion methods depends essentially on identifying good candidates, or prospects, semantically related to query terms. Word embeddings have been used recently in an attempt to address this problem. Nevertheless query disambiguation is still necessary as the semantic relatedness of each word in the corpus is modeled, but choosing the right terms for expansion from the standpoint of the *un-modeled* query semantics remains an open issue. In this paper we propose a novel query expansion method using word embeddings that models the global query semantics from the standpoint of prospect vocabulary terms. The proposed method allows to explore query-vocabulary semantic closeness in such a way that new terms, semantically related to more relevant topics, are elicited and added in function of the query as a whole. The method includes candidates pooling strategies that address disambiguation issues without using exogenous resources. We tested our method with three topic sets over CLEF corpora and compared it across different Information Retrieval models and against another expansion technique using word embeddings as well. Our experiments indicate that our method achieves significant results that outperform the baselines, improving both recall and precision metrics without relevance feedback.

1. Introduction

Over the years, query expansion (QE) methods have been proposed as an effective way to address the query-document vocabulary mismatch problem in Information Retrieval (IR) tasks (Vechtomova, 2009; White & Horvitz, 2015). The aim is to enrich the query by adding semantically related words, mainly using synonyms.

Approaches to QE can be classified into global or local methods. On the one hand, global methods expand the original query independently of any retrieval result. Typically, WordNet is the standard exogenous tool of choice for selecting new terms semantically associated to the original ones (Pal, Mitra, & Datta, 2014). On the other hand, local methods use relevance feedback, whereby they perform a first retrieval whose outcome is actually used for selecting the most promising terms

[*] Corresponding author.
E-mail addresses: fcaridad@uaem.mx (F.C. Fernández-Reyes), jhermosillo@uaem.mx (J. Hermosillo-Valadez), mmontesg@inaoep.mx (M. Montes-y-Gómez).

http://dx.doi.org/10.1016/j.ipm.2017.09.001
0306-4573/© 2017 Elsevier Ltd. All rights reserved.

to be added to the initial query (Miyanishi, Seki, & Uehara, 2013; Parapar, Presedo-Quindimil, & Èülvaro Barreiro, 2014; Takeuchi, Sugiura, Akahoshi, & Zettsu, 2017). Usually, Pseudo Relevance Feedback (PRF) is preferred for cutting computational costs, as this technique automatically extracts new terms from a subset of top "k" ranked documents, obtained in a first retrieval (Colace, Santo, Greco, & Napoletano, 2015; Karisani, Rahgozar, & Oroumchian, 2016).

The above methods suffer from practical drawbacks. Usually, global methods require word sense disambiguation algorithms, as terms returned by ontological knowledge bases (such as WordNet) are polysemic. As to what local methods concerns, one typically performs a second retrieval, using terms coming from top ranked documents presumably relevant and obtained in a first retrieval. The relevance assumption about those documents could insert noise terms into the expanded query. Finally, both methods either improve mean average precision or recall metrics but usually not both.

In an attempt to overcome these problems, recent research in QE has been reported where *word embeddings* are used as a semantic modeling technique (ALMasri, Berrut, & Chevallet, 2016; Diaz, Mitra, & Craswell, 2016; Goodwin & Harabagiu, 2014; Roy, Paul, Mitra, & Garain, 2016). Word embeddings (WE) are distributed representations of terms, usually obtained from a neural-network that models the joint distribution of the corpus vocabulary (Bengio, Ducharme, Vincent, & Janvin, 2003). These vector representations have been used recently in different natural language processing tasks (Baroni, Dinu, & Kruszewski, 2014; Cai & de Rijke, 2016; Lebret, 2016). They consider the context of a word in the corpus, similarly to a *n-gram* model, but the context size could be higher than bi-grams or tri-grams as usually found in the literature. `Word2vec` (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) is a fairly representative method to accomplish this mapping nowadays.

These approaches essentially rely on the assumption that *each* query term can select the best candidates (or prospects) in function of their semantic closeness. This leads to the use of expansion criteria such as pooling the closest *N* terms or the search for adequate neighborhoods of query terms (e.g. K-neighbors in Roy et al. (2016)). Under this view, the query semantics is atomized and treated *locally* as prospect terms are chosen by query terms on a *one-word-at-a-time* basis. We postulate that in order to improve QE effectiveness we should model the query semantics as a whole from the perspective of prospect vocabulary terms. In this way, we expect to improve the quality of candidate terms related to the query and add new terms semantically linked to more relevant topics.

The remainder of the paper is organized as follows: Section 2 posits our research questions and contribution. Section 3 discusses related work. In Section 4, we propose the formal methods to build the new query. Section 5 describes the experimental design and setup. Our main results are presented and discussed in Section 6. We close the paper in Section 7 where we highlight the main theoretical and practical implications of our work.

2. Research questions and objective

The main question we address in this paper is how to perform QE considering the semantics of the query as a whole entity using WE.[1] Answers to this question are non-obvious since the vector representations are trained in the corpus, but query instances are nonexistent and thus factually unknown during the training. One could think about using *Paragraph Vector* (Le & Mikolov, 2014), a method conceived to model pieces of text of variable length, including sentences and documents. As it was demonstrated, this method is useful for text classification purposes. However in order to exploit it for query expansion purposes, either we would have to know the complete set of queries in advance, so as to model them together with the collection sentences or paragraphs, or we would have to re-train the model with every new incoming query. For these reasons we did not use this method to represent the query. Also, we are not seeking to estimate vector representations of the query (Sordoni, Bengio, & Nie, 2014; Zamani & Croft, 2016), as we conceive the problem in a simpler way that sheds some light on the importance of better exploiting the semantic relations between word embeddings from the corpus.

Hence, the best bet is to map each query term to its embedded representation as if the word had been seen in the corpus. Nevertheless disambiguation is still necessary because the semantic relatedness of each word in the corpus is modeled, but choosing the right terms for expansion from the perspective of the *un-modeled* query semantics remains an open issue.

This question entails other research issues of practical interest addressed by our work:

1. Which original query terms are more useful for expansion purposes; i.e. how to improve candidates semantic quality?
2. How to effectively cope with disambiguation issues without using exogenous resources?
3. How to improve both recall and precision metrics without relevance feedback?

Current QE techniques based on WE allow each query term to determine its correspondent expansion candidates. We call this a *Query-Guided* perspective. We address the problem of query expansion from the opposite perspective, by letting vocabulary terms "choose" query terms. We propose to pool candidate terms on the basis of a voting process, whereby new terms are elicited by the semantic closeness of vocabulary terms to *all* query terms. We call this a *Prospect-Guided* perspective, which aims at grasping the global sense of the query for expansion purposes.

In this paper, we propose a novel global QE method using word embeddings that considers the query as a whole, includes candidate terms pooling schemes that address disambiguation issues, and improves both recall and precision metrics in IR without relevance feedback.

[1] It is a common practice for instance to sum or average the corresponding embeddings of each term of a phrase or sentence, so that the whole phrase is mapped to a single point in the semantic space; e.g. see Vulić and Moens (2015). Arguably, the global sense of the phrase is meaningless in this space.