Enhancing Information Retrieval with Adapted Word Embedding

Navid Rekabsaz
Vienna University of Technology
Favoritenstrasse 9/11
1040, Vienna, Austria
rekabsaz@ifs.tuwien.ac.at

Keywords

IR Models; word embeddings; similarity; related terms

ABSTRACT

Recent developments on word embedding provide a novel source of information for term-to-term similarity. A recurring question now is whether the provided term associations can be properly integrated in the traditional information retrieval models while preserving their robustness and effectiveness. In this paper, we propose addressing the question of combining the term-to-term similarity of word embedding with IR models. The retrieval models in the approach are enhanced by altering the basic components of document retrieval, i.e. term frequency (tf) and document frequency (df). In addition, we target the study of the meaning of the term relatedness of word embedding models and its applicability in IR. This research topic consists of first explore of reliable similarity thresholds of word embedding vectors to indicate "related terms" and second, identification of the linguistic types of the terms relatedness. In the following, we explain the mentioned research topics in more detail.

In Information Retrieval, terms are still the fundamental building blocks for establishing topical relevance relationships between documents and queries. The "basic" models to approximate this relationship, namely Vector Space Model (VSM), Probabilistic Retrieval Framework [4], and Language Model (LM) [3], have maintained a respectable command of the research and practice of IR. They are all fundamentally based on term frequency (tf) as a representation of importance of a term within a document and also a representation of the specificity of a term, usually realized by document frequency (df). These IR models are basically based on the underlying assumption of term independence.

As the first research topic, we challenge this assumption by conducting the idea of integrating word embedding into various IR models. As the basis of our approach, we consider the terms as concepts and define the relation between them based on the term-term similarity of word embedding models. In fact, the occurrence of a concept in a document is not only the frequency of the term (representing the concept) but also the partial frequency of the related terms. Considering this approach, we investigate the enhancement of various IR models, namely VSM, probabilistic, and language models with word embedding while preserving their robustness, and reliability.

Extending IR models with the embedded knowledge of word embedding particularly requires a deep understanding of its basic building block: term-term relatedness.

An issue in the word embedding methods is that they and their corresponding mathematical functions can provide approximation on the relatedness of any two terms, although this relatedness could be perceived as completely-meaningless in the language. Karlgren et al. [1] point it out by 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?"*.

In addition, as the relatedness of word embedding methods fundamentally consider the "close" terms as the terms with similar contexts, they generalize all different term relations e.g. antonyms, hypernymy, hyponyms, co-hyponyms, synonyms, antonyms, etc. into one notion of relatedness. However, using all these types bias the search to unrelated topics (e.g. searching for dog instead of cat!). Kruzewski and Baroni [2] notice this issue by an example: *animal*, *puppy*, and *cat* are all closely related to *dog*, but if you tell me that Fido is a dog, I will conclude that he is an animal, that he is not a cat, and that he might or might not be a puppy.

Therefore, as the second research direction, we address the mentioned issues by (1) exploring a threshold under which we would no longer consider two terms to be sufficiently related to the same concept (2) understanding the underlying notions of the terms' relations and adopting them for IR models.

Considering the mentioned research questions, the main goal of this Ph.D. is providing stable, reliable, and reusable information retrieval models, enhanced with adapted word embedding methods. The performance of the scoring models should be proofed by testing it on various IR domains i.e. ad hoc, news, health, patent, and social image sharing.

Acknowledge: This PhD is partly funded by two FWF projects MUCKE (I 1094-N23) and ADMIRE (P 25905-N23).

1. REFERENCES

[1] J. Karlgren, A. Holst, and M. Sahlgren. Filaments of meaning in word space. In *Proc. of ECIR*. 2008.

[2] G. Kruszewski and M. Baroni. So similar and yet incompatible: Toward automated identification of semantically compatible words. In *Proc. of NAACL*, 2015.

[3] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. of SIGIR*, 1998.

[4] S. Robertson and H. Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

SIGIR '16 July 17-21, 2016, Pisa, Italy

ACM ISBN 978-1-4503-4069-4/16/07.

DOI: http://dx.doi.org/10.1145/2911451.2911475