Trec-CAR 2018: A Simple Unsupervised Semantic Query Expansion Model

Robert Litschko, Federico Nanni, and Goran Glavaš

University of Mannheim
{litschko,goran,federico}@informatik.uni-mannheim.de

In this summary we present a simple and unsupervised Semantic Query Expansion model (SemQueryExp) for Complex Answer Retrieval (CAR). TREC CAR is a large-scale information retrieval shared evaluation task based on Wikipedia content. We have participated in the *Passage Ranking Task* of TREC CAR. Queries are provided as hierarchical section outlines and the goal is to retrieve the relevant paragraph, i.e., the original Wikipedia paragraphs of the respective section.

The queries consist of the page title in which the section appears, the name of the main section and one or more sub-level sections (e.g., *Thompson Capper // Early career // South African service*). A section in turn contains a number of terms that we want to use in order to expand the query with semantically related terms. Here we rely on 300-dimensional GloVe [1] word embeddings, pre-trained on Wikipedia. For each word in the query we lookup its embedding, search for the k-nearest-neighbors and add the corresponding words to the query. Distances between word embeddings are measured with the cosine similarity. The final query consists now of its original query terms as well as the words added during query expansion.

We assume that words appearing in lower sub-section levels, i.e. more specific sections, capture more relevance for the query than words appearing on higher levels such as the page title, which only describe the surrounding theme. This is encoded by assigning each query term a weight according to it's level in the hierarchical outline. If a word is in the title it receives a weight of 1, if it is in the main section it receives a weight of 2, etc. Expanded query terms are assumed to be noisy and scored with a value ranging between 0 and 1, depending on their cosine similarity. The final query consists of the union of all terms appearing in the outline and the expanded query terms, coupled with respective weights. All data is normalized by removing stop words and applying lemmatization. We execute the query on a Lucene index as a BoostedQuery using the BM25 retrieval model. We used no external data and we tuned the value of the parameter k on the benchmarkY1-train portion of the TREC CAR dataset.

References

1. 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)