Knowledge Based Text Representations for Information Retrieval

Chenyan Xiong

Abstract

The successes of information retrieval in recent decades were built upon bag-of-words representations. Effective as it is, bag-of-words is only a shallow text understanding; there is a limited amount of information for document ranking in the word space. This dissertation goes beyond words and builds knowledge based text representations, which embed the carefully curated information from knowledge bases, and provide richer and structured evidence for more advanced information retrieval systems.

This thesis research first builds query representations with entities associated with the query. Entities’ descriptions are used by query expansion techniques that enrich the query with explanation terms. Then we present a framework that represents a query with entities that appear in the query, are retrieved by the query, or frequently show up in the top retrieved documents. A latent space model is developed to jointly learn the connections from query to entities and the ranking of documents, modeling the external evidence from knowledge bases and internal ranking features cooperatively. To further improve the quality of relevant entities, a defining factor of our methods, we introduce learning to rank to entity search and retrieve better entities from knowledge bases. In the document representation part, this thesis research also moves one step forward with a bag-of-entities model, in which documents are represented by their automatic entity annotations, and the ranking is performed in the entity space.

This proposal includes plans to improve the quality of relevant entities with a co-learning framework that learns from both entity labels and document labels. We also plan to develop a hybrid ranking system that combines words and entities together with their uncertainties considered. At last, we plan to enrich the text representations with connections between entities. We propose several ways to infer entity graph representations for texts, and to rank documents using their structure representations.