Text Representation, Retrieval, and Understanding with Knowledge Graphs

Chenyan Xiong
Carnegie Mellon University
cx@cs.cmu.edu

September 25, 2018

Abstract

This dissertation aims to improve text representation, retrieval, and understanding with knowledge graphs. Previous information retrieval systems were mostly built upon bag-of-words representations and frequency-based retrieval models. Effective as they are, word-based representations and frequency signals only provide shallow text understanding and have various intrinsic challenges. Utilizing entities and their structured semantics from knowledge graphs, this dissertation goes beyond bag-of-words and improves search with richer text representations, customized semantic structures, sophisticated ranking models and neural networks.

This thesis research starts by *enriching query representations* with entities and their textual attributes. It first presents query expansion methods that better represent the query with words from entity descriptions. Then it develops a supervised latent-space ranking model that connects query and documents through related entities from the knowledge graph. It also provides a novel supervised related entity finding technique in the entity search setup.

Then this dissertation presents our *entity-oriented search* framework that represents query and documents with *entity-based text representations* and matches them in the entity space. We construct a *bag-of-entities* model that represents texts using automatically linked entities with a customized linking strategy. Ranking with bag-of-entities can be done either solely with discrete match—as in classic retrieval models—or by our Explicit Semantic Ranking approach that soft matches the query and documents with continuous knowledge graph embeddings. The entity-based text representations are then combined with word-based representations in a *word-entity duet* representation method. In the duet, query and documents are represented by both bag-of-words and bag-of-entities; the ranking of them goes through both in-space matches and cross-space matches which together incorporates various types of semantics from knowledge graphs. The duet framework also introduces a hierarchical ranking model that learns the linking of entities and the ranking of documents jointly from relevance labels.

This thesis research concludes with a neural entity salience estimation model that provides a deeper text understanding capability. We developed a Kernel Entity Salience Model that better estimates the importance of entities in text with distributed representations and kernel-based interactions. Not only does it improve the salience estimation accuracy, it can also

be used to estimate the importance of query entities in documents, which provides effective ranking features that transfer the model's deeper text understanding capability to improve retrieval.

With the effective usage of entities, their structured semantics, customized semantic grounding techniques and novel machine learning models, this dissertation formulates a new entity-oriented search paradigm that overcomes the limitation of bag-of-words and frequency based retrieval. The better text representation, retrieval, and understanding ability provided by this dissertation is a solid step towards the next generation of intelligent information systems.