Abstract

Query suggestions are query proposals after one or more queries have been submitted. They help users refine their queries when using a search engine and they can guide the direction of the search within a session by either suggesting queries that dive deeper into the current search direction or by suggesting a change of direction into a different search space. So far, query suggestion models have focused on complex recurrent encoder-decoder architectures to solve the task. Such complex architectures require a great amount of computational power and hours to train.

This thesis proposes a novel neural model that reduces the complexity of current state-of-the-art models by using an encoder-decoder architecture that is based solely on attention. For this, it uses a Transformer model, that was first introduced for neural machine translation task, and it was shown to outperform state-of-the-art techniques (Vaswani et al., 2017).

The AOL dataset is used to compare the model proposed with current state-of-the-art query suggestion models (Sordoni et al. (2015) and Dehghani et al. (2017)). The empirical experiments show that it is possible to use a Transformer architecture for query suggestion task. Furthermore, results indicate that reducing the complexity of the architecture does not compromise the model performance. Simpler models are able to achieve good results.

This opens the door for future work to explore many different variants of Transformer models that are novel in the field of query suggestion task.

Keywords
Query Suggestion · Transformer Model · Sequence to Sequence Model