Neural Models for Improving First Phase Information Retrieval

Document Retrieval problems typically consist of two phases. The first phase tries to retrieve more number of relevant documents for a given query from a large collection of documents (typically in 10's of millions to billions of documents) whereas the second phase tries to rank the retrieved documents so that highly relevant documents appear on the top of the list. The first phase has to be simple and fast because it works on the large set of document collection. Standard models such as BM25, QL rely only on the tf-idf based term matching between query and document without considering the context of words. Recent neural ranking models have shown promising performance for the reranking task. However, there has been limited work on neural models and indexing methods to improve the first-stage retrieval. The objective of this thesis is to explore the recent contextual autoregressive models to improve the first-stage retrieval methods.

Problem Statement

Given a large collection of items — documents, passages — the aim is two fold

1) construct a neural retrieval model that exploits attention-based semantic term matching for effective recall @ k
2) explore how to make inference efficient for this family of models by using ideas from KNN-search and LSH.

Challenges

There are two key challenges what we are faced when making progress for first-stage retrieval:

1) *Measuring recall is difficult for large datasets:* The typical way in which recall is measured in existing document retrieval tasks (TREC) is by pooling multiple retrieval models followed by human judgement. But since all the pooling methods also follow the same retrieval methods, relevant documents not surfaced by the initial retrieval method are not judged. This is a hard problem because re-annotation is not possible in such a missing data setting.

2) *Efficiency constraints are strong:* Since most of the retrieval methods based on term matching and KNN search are expected to be fast, any addition of model

complexity results in unacceptable inference times. For example deeper layers are almost always expensive. Such constraints force us to consider alternate design decisions that trade-off accuracy for speed.

We will try to handle the above-mentioned challenges in the following two ways --- 1. consider a small size dataset, 2. a better retrieval model should improve the performance (map, nDCG) of the second level reranking task.

An ideal candidate should have
1. Strong background in python and deep learning
2. motivation behind learning and exploring the data
3. knowledge about basic IR concepts

Interested students are encouraged to email to Dr. Koustav Rudra at rudra(at)l3s(dot)de for scheduling a meeting.

References:

1. Context-Aware Sentence/Passage Term Importance Estimation For First Stage Retrieval [https://arxiv.org/abs/1910.10687]

2. Two tower model ICLR 2020 — https://openreview.net/forum?id=rkg-mA4FDr

3. Efficient Training on Very Large Corpora via Gramian Estimation. ICLR 2019 --- https://arxiv.org/pdf/1807.07187.pdf

4. StarSpace: Embed All The Things! AAAI 2018 --- https://arxiv.org/pdf/1709.03856.pdf