Klyueva Olga
CIS, LMU München
15.05.1917

Joseph Birkner: Ranking with neural network derived document vector.

The motivation for the work is to build a search engine for online Courses, based on IRON project. IRON project contains of online courses description.
First, it was described, how the traditional search engine works. The theme is within the frame of Information Retrieval, that's why the relevant terminology for the work were mentioned: search engine with ranking algorithm; metadata, which contains general information about user and his search history. Metadata helps the ranking algorithm to make better decisions.

The work is based on the new domain of Information Retrieval: Neural Information Retrieval. Neural Information Retrieval divided into 2 subfields. The first subfield is Representation Optimization. It is the optimization of the input, of the representation of user's query and of the documents in the corpus that are fed into ranking algorithm. This helps the algorithm to make better ranking decisions. The other subfield is called deep relevant matching Models, which tries to make better decisions with representations that are given. In the work, the attention is concentrated on the representation optimization. There are two main conditions for a good document representation:
1. The document must be small enough for the ranking algorithm to operate on it.
2. The algorithm must gather enough information from the document to rank it.

Then it was said about traditionally way of document encoding in Information Retrieval. It was mentioned the method which is called TR IDF. TR IDF method tries to represent the document as a frequency distribution over all terms of the document.
There is a general problem with TR IDF method. These are two assumptions about terms as features of a document.
1. Word order doesn't matter so much about semantic characteristics of the document
2. Each term in the document carries an independent dimension of meaning, it is taken as separate entities.
It was given the example that shows the problem with terms as features of a document. There were mentioned two name of courses and descriptions to them. After normalizing both of the descriptions, there were two identical frequency distributions of the terms in the TR IDF representation.
The system is trained with neural network and generates the embedded representation for words. Neural Network comes up with features that optimally represent a word. It's a unsupervised learning technique. It can be trained on unlabeled data. In general, the neural network is forced to predict the document from the hidden representation that it generated. It's called Doc2Vec or Seq2seq approach.

At the end of the presentation, tasks to achieve were mentioned:
1. Train LSTM (long shot term memory) Seq2Seq Models.
2. Create API to generate document representation from trained LSTM
3. Evaluation ranking performance on TREC datasets, data sets with documents as well as users' queries. There is a label that this document is relevant for this query. These labels can be used to evaluate the retrieval system.