Protokoll zur Sitzung vom 15.05.2017

Kurzvorstellung von der Bachelorarbeit von Joseph Birkner

BA-Betreuer: François Bry, Yingding Wang

Thema: Ranking with Neural Network Derived Document Vectors

In this meeting Joseph gave us a brief description of his thesis, which is devoted to ranking with neural network derived document vectors. Joseph is working on his thesis within the framework of the research project of the Institute of Informatics at LMU, which is called Irom, or Intelligent Recommender of MOOCs (Massive Open Online Courses). The main aim of project is to create an intelligent MOOCs search engine, helping students to find the best fulfilling course. As you know, recommendations imply the information need, which Information Retrieval occupies with. Stating briefly, the process is following: the user has an information need and words it as a query; than this query goes through domain-specific ranking algorithm and after some computing, the system returns ranked recommendations to the user like a result. Neural Information Retrieval, which Joseph uses for his work, is quite a new field and has some subfields. One of them is Representation Organization, where you try to optimize the query. Another one is Matching Optimization, which tries to make better ranking decisions using Deep Relevance Matching Model.

The main purpose of Joseph's work is encoding of documents, so they need efficient document representations to instantaneously rank recommendations based on students' need. There is a traditional model IF-IDF, which represents document as a frequency distribution. But this model has some weaknesses: firstly, the word order is ignored and secondly, every word is independent although there are some words which could be close to each other. To achieve his objective, semantic space for documents, Joseph uses two implementations: Word2Vec (One-Hot Term Vector) and Doc2Vec, which reduces TF-IDF matrix of all documents and uses Gauss-Method to find principal components for low-dimensional document vectors.

Joseph actually creates documents with RNN, using a large document corpus to train Seq2Seq and then using trained encoder to extract feature vectors from documents. So he has two tasks: Prototype and Schedule. He has already generated document vectors and created 30-dimensional document vectors. He presented his prototype-diagram in the meeting, which looked like confetti and was created with Plotty. Next steps, Joseph should do in the nearest future, are:

- train LSTM Seq2Seq Models in Tensor Flow (using different combinations of training data);
- generate documents and query vectors from trained LSTMs;
- retrieve ranked sets with query among document sets;
- evaluate ranking performance on TREC datasets;
- evaluate selected features from the document vectors with heat maps.