15.05.17

Protokoll zur Sitzung vom 15.05.2017 – Computerlinguistisches Arbeiten

2. Referat: Joseph Birkner, Ranking with neural networks derived document vectors (Institut für Informatik)

Joseph starts his presentation with a short summary of the project (IROM = 'Intelligent Recommendation of Massive Open Online Courses') and its vision: 'Ubiquitous vertical search'. The motivation for his work is that every information need should be instantly satisfied, and that this can be achieved with a vertical search engine. A more detailed goal is stated as an axiom in one of his slides: 'There is a need for efficient document representation to instantaneously rank recommended courses based on student need.'

He defines vertical search as the 'search within a specific domain', and implies that the solution for a specific domain may be applicable to other domains.

On the technical side of his work he explains that the information need is formulated as a query and a ranking algorithm then presents the top results as ranked recommendations. The algorithm can be augmented with meta data of the user, for example personal information and the search history of the user. Joseph uses a detailed graphic to visualize the relations between each component of the system.

The origin of his thesis lies within the field of neural information retrieval. He then explains that neural information retrieval is very new research field and can be categorized into two different sub-fields: Representation optimization (deep semantic structured matching) and matching optimization (deep relevance matching model). Specifically his dissertation focuses on representation optimization.

In the following slides Joseph reveals that in traditional IR documents are encoded in the tf-idf form, which intends to reflect how important a word is to a document. However, this approach has some general problems, for example the word order is ignored or a flawed word independence is present. The traditional approach also uses LSA (Latent Semantic Analysis), but Joseph decided to use a different approach with Doc2Vec, which is basically an extension to Word2Vec that learns to correlate labels and words.

In the last section Joseph gives information about his current status. Basically his work can be divided into four main parts. In the first step he needs to train the LSTM Seq-to-Seq model in tensor flow, which he has already managed to do. Currently he is about to begin with the second part of his work, which is creating an API to generate document/query vectors form trained LSTMs. The 3. Step is to evaluate the ranking performance on TREC datasets and the last part of his work is to evaluate selected features from the document vectors with heat maps.