CIS, LMU München

Studentin: Elena Atanasova

15.05.2017

Protokoll zur Sitzung vom 15.05.2017 – Computerlinguistisches Arbeiten

2. Vortrag: Joseph Birkner, "Ranking With Neural Network Derived Document Vectors"

BA-Betreuer: M.Sc. Wenpeng Yin LMU/IFI/PMS

The bachelor's thesis from Birkner is supervised by the teaching and research unit for programming and modeling languages, Institute of Computer Science, LMU. His supervisor is Herr Yin from the Center for Information and Language Processing (CIS).

Birkner began his presentation with an overview of the topic and presented the context of the project titled "Intelligent Recommendation of Massive Open Online Courses (IROM)". This project intends to make Massive Open Online Courses (MOOCs) available for students. Students can sign up and complete them for free and basic access requirements.

Birkner's bachelor's thesis is about developing a search engine that provides intellectual proposals for online courses. A corpus of the course descriptions, which was created at the CIS using the framework of a previous bachelor's thesis, is used for this purpose. The topic is therefore in the area of Information Retrieval (IR). Hence, one needs a search query, a database with the documents, and ranking algorithms to deliver the most appropriate result (Ubiquitous Vertical Search). Ubiquitous Vertical Search is a search within specific domain. Solutions for specific domain may be applicable to other domains.

Later, Birkner explains further about what a recommendation is. This implies Information Need and nothing but IR. Information Need is expressed through course query and user metadata, and it is satisfied by finding the relevant courses.

The motivation for Birkner's bachelor's thesis is encoding of documents. He summarized this motivation using the following axiom: "We need efficient document representations to instantaneously rank recommended courses based on student need."

First, he explains what a good document representation is:

1) Enables fast (constant-time) ranking function → Efficiency

2) Ranking seems "intelligent" to search engine user → Effectiveness

Later, Birkner showed us a document representation in traditional IR: TF-IDF and explains the disadvantages. Flawed word independence assumption and the word order is ignored. But if we consider that words do not have any semantic relation, then this is not the case! Hence, Birkner has a clearly objective in order to solve this problem: To create a semantic space for documents with help of Word2Vec and DocVec.

The prototype tasks are to generate the document vectors (embedding), 30 – dimensional document vector, train with ~1200 course descriptions and schedule tasks. His recommendations are: Train LSTM(Long Short Term Memory) Seq2Seq Models in Tensor Flow, create API, evaluate ranking performance on TREC (Text Retrieval Conference) datasets, evaluate selected features from the document vectors with heat maps and the bonus.