Kolloquium, 15. Mai 2017

Dayyan Smith

Ranking With Neural Network Derived Document Vectors

Speaker: Joseph Birkner

Advisor: Yingding Wang

Joseph Birkner is writing his bachelor thesis as part of the Intelligent Recommender of MOOCs (Irom) project at the Institute of Informatics. As the number of Massive Open Online Courses (MOOCs) grows, it becomes increasingly difficult for interested learners to find the MOOCs best equipped to address their needs. The goal of project Irom is the creation of an intelligent search engine for MOOCs. Tho achieve this the search engine will analyze each course's textual description and also take into account explicit as well as implicit feedback of users of the search engine. At its core, project Irom is trying to solve an Information Retrieval problem: There is an information need (What course is fitting for what I want to learn?) and there are documents (the courses, or to be more exact their descriptions).

Since the documents that are being searched are all of the same kind – they are all course descriptions, project Irom will offer a topical search also called vertical search. The ranking of the results in a search can be augmented by a user's metadata. Age and gender can influence the ranking algorithm, but also search history and a user's interactions with search results will be incorporated.

Neural Information Retrieval, whose beginnings lie about three years back, can be split into two subfields: Representation Optimization and Machine Optimization, that are concerned with optimizing the systems input (user query, corpus) and with making better ranking decisions respectively. Joseph's work is part of the former.

To instantaneously rank recommended courses based on student need, efficient document representations are necessary. In traditional IR a common way of encoding documents is with tf-idf. But this model is problematic: the word order is ignored and word independence is assumed.

A better representation of documents can be achieved with document vectors, which try to encode the semantic meaning of each word in a sentence as a vector. These document vectors can be created with doc2vec, which is similar to word2vec – only for document vectors.

Using Recurrent Neural Networks, a hidden representation of a document is created and this representation is then used to predict a document.

Since document vectors – like word vectors – are often very high dimensional, it is necessary to reduce their dimensionality, if you want to visualize them. For this t-SNE is helpful, which tries to keep spatial relations while reducing the dimensionality: in a 2D visualization similar documents should be close together. An easy way to visualize these vectors is Plotly, which creates an interactive diagram.

Joseph's model is currently trained and evaluated on his training data, which is provided by the Irom project, therefore one of his next steps is the evaluation of his model on TREC (Text Retrieval Conference) data. Furthermore he will create an API and evaluate select features.