CIS, LMU München
Fach: Kolloquium Computerlinguistisches Arbeiten SS2017
Studentin: Faridis Alberteris Azar

Protokoll zur Sitzung am 15.05.2017

Title: Ranking with Neural Netwok Derived Document Vectors.

Student: Joseph Birkner
LMU/IFI/PMS

The student Joseph Birkner incloses his bachelor‘s thesis (germ.: Bachelorarbeit) within the context of the project IROM (Intelligent Recomendation of Massive Open Online Courses). This project intends to make available Massive Open Online Courses (MOOCs) for students. Students can sign up and complete them without for free and any access requirements.

At the beginning of the presentation Birkner explains the main ideas that are taking into account for the search of this courses:

Ubiquitous Vertical Search: Every information needed can be satisfied instantly with a vertical search engine.

In order to preserve this idea, the traditional Information Retrieval (IR) models were extended to integrate word embeddings. This expansion of the traditional IR is known as Neural IR. As Joseph Birkner said, with the Neural IR it is possible to try to optimize the input of the user query to obtain better ranking decisions with vectorial representations.

After this briefly introduction in Neural IR, the student explained that the motivation for his thesis is the encoding of documents. He summarized this motivation in following axiom:

We need efficient document representations to instantaneously rank recommended courses based on student need.

The traditional representation of documents in IR is based on frequences (TF-IDF). As shown by Birkner, this way of representation has disadvantages:

- Word order is ignored
- Flawed word independence assumption (each term in the document has an independent meaning)

But if we assum that, we assum also that words do not have any semantic relation. And it is of course not true! That is why Birkner has a clearly objective in order to solve this problem: To create a semantic space for documents.

As the speaker explained, Neural IR uses the same idea of word2vec, but adapted to the documents: the system is feeded with embedding words, then the neural network is forced and the next word will be predicted (Backprop). The use of word2vec reduces the TF-IDF matrix of all documents using Gauss method. But, as the presenter mentioned, there is not an online approuch. That means that it has to be redone for all documents.

Birkner proposes a solution for this problem: Instead of feeding a fix kontext, we feed the neural network itself.

In order to achieve his aims, Birkner shows his tasks, which are listed as follows:

Task: Prototype

- Generated document vectors (embeddings) (generated 30-dimensional document vectors) (He recommends here a python tool for diagramms: plotty)

Task: Schedule:

1. Train LSTM(Long Short Term Memory) Seq2Seq Models in Tensorflows[1].
2. (As the student explaned, sequence-to-sequence model consists of two recurrent neural networks (RNNs): an *encoder* that processes the input and a *decoder* that generates the output)
3. Evaluate ranking performance on TREC (Text Retrieval Conference) datasets
4. Evaluate selected features from the document vectors with heatmaps[2]
5. Bonus: Search for constant shifts in the document space.

Birkner shows at the end of the presentation the good results achievd with processes and recommended an interesting link about the advantages of RNNs: http://karpathy.github.io/2015/05/21/rnn-effectiveness/

[1] TensorFlow is an open source software library for machine learning across a range of tasks, and developed by Google to meet their needs for systems capable of building and training neural networks to detect and decipher patterns and correlations, analogous to the learning and reasoning which humans use (Source: https://en.wikipedia.org/wiki/TensorFlow)

[2] Heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors(Source: https://en.wikipedia.org/wiki/Heat_map)