Karlsruher Institut für Technologie
Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB)
Dr.-Ing. Michael Färber
michael.faerber@kit.edu

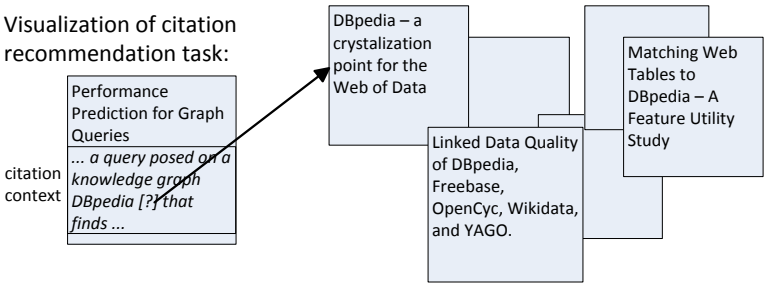
27.02.2019

Call for Master Thesis "Automatically Recommending Citations for Texts Using Neural Networks"

(in English/German)

What is the topic?

Citation recommendation refers to the task of recommending publications as citations for a given text (e.g., a sentence; see figure). The recommended publication needs to fit the citation context and substantiates the fact or concept mentioned in the citation context. Due to the vast amount of publications nowadays, citation recommendation can significantly assist researchers and students in their daily scientific work.



The goal of the proposed thesis is to develop a **citation recommendation system**. This means that a model is trained in a **supervised machine learning** setting in order to allow for selecting suitable publications for a given citation context out of the set of all available publications (typically hundreds of thousands or millions). More precisely, the work may encompass the following sub-steps:

- Learn **embeddings for all publications** represented in the given corpus of publications (and maybe for additional entities such as venues), using both *textual* information and the *links*. The learned embedding vectors for papers can then already be used for calculating similarities among papers.
- Learn **joint embeddings of both citation contexts and publications**. An example of joint learning can be found in [3]. It might be necessary to train separate models for different scientific domains, as the citation behavior might vary.
- Given the joint embeddings, a **citation recommendation system** can be built, which mainly consists of applying the learned model of joint embeddings to new incoming citation contexts.
- Evaluate the developed citation recommendation system automatically (by removing citations and re-predicting them) and manually, if necessary.

A thorough introduction to the topic and necessary literature will be provided by the supervisor.

[1] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
[2] https://www.microsoft.com/en-us/research/project/academic/articles/march-2018-graph-update/
[3] "Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation", LREC 2016. https://arxiv.org/abs/1601.01343

Which prerequisites should you have?

- Interest in *machine learning* and *data mining* (word embeddings, neural networks, etc.)
- Programming skills for machine learning (e.g., using Python).
- Willingness to work with large data sets (on servers/clusters).

Keywords: Implementation, knowledge graph, scholarly data, embeddings, neural networks, machine learning, natural language processing, big data.

Contact:
Dr. Michael Färber
michael.faerber@kit.edu