Dr. Georg Rehm
Principal Researcher and Research Fellow
DFKI GmbH
Speech & Language Technology Lab
Projektbüro Berlin
Alt-Moabit 91c
10559 Berlin

Telefon: +49 (0)30 23895-1833
Telefax: +49 (0)30 23895-1810
E-Mail: georg.rehm@dfki.de
Internet: www.dfki.de

13 September 2020

Joint Theses between TU Berlin and DFKI: Open Topics for Bachelor and Master Theses

DFKI's Speech and Language Technology lab is led by Prof. Dr. Sebastian Möller, which is why the QU lab and DFKI's SLT lab collaborate closely. One of the areas of collaboration concerns joint Bachelor and Master Theses.

The DFKI SLT lab currently offers the topics shown in the table on the following two pages, all of which are embedded in one or more of the research projects the DFKI SLT team is currently working on.

INTERESTED? If you are interested in exploring one or more of these topics further, please get in touch with Dr. Georg Rehm <georg.rehm@dfki.de> to discuss the next steps.

IMPORTANT: When contacting Dr. Georg Rehm via email (see above), please provide a brief motivational note, your CV and a current Transcript. Please also indicate your level of experience with regard to software development (e.g., languages, tools, approaches, backend vs. frontend), natural language processing and modern machine learning techniques including neural approaches (please also provide specifics).

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Firmensitz
Kaiserslautern

Weitere Standorte und Betriebsstätten:
Saarbrücken, Bremen, Osnabrück,
Oldenburg, Berlin, St. Wendel

Geschäftsführung
Prof. Dr. Antonio Krüger

Vorsitzender des Aufsichtsrats
Dr. Gabriël Clemens

Amtsgericht Kaiserslautern HRB 2313
USt-ID-Nummer DE 148 646 973
Steuernummer 19/672/50006

Stadtsparkasse Kaiserslautern
IBAN: DE60 5405 0110 0028 0004 79
BIC/SWIFT: MALADE51KLS

Topic	Project	Description	Level
OCR/OLR-based document segmentation	QURATOR	Segmentation of regular HTML documents (either single documents or multiple connected and inter-linked HTML documents) into their main building blocks such as Headline, NavigationBar, Paragraph/Textblock, Figure, Caption, Advertisement etc. using methods taken from Optical Character Recognition (OCR) and Optical Layout Recognition (OLR). Mapping of the extracted high-level document components, listed above, onto functional types such as Introduction, Discussion, Conclusions, Comparison, Elaboration, Background etc. This thesis could also include working with document grammars and document ontologies, perhaps in a Semantic Web paradigm, modelling document structures using OWL, RDF, SHACL etc.	Master, PhD
Semantic Document Segmentation	QURATOR	This thesis deals, essentially, with the same topic and research question as thesis topic 2 but it's not restricted to the application of OCR/OLR or Semantic Web methods, i.e., any type of method can be used. The goal is to perform some form of semantic segmentation as a preprocessing step for other NLP downstream tasks (classification, similarity, topic detection, relations) of long documents.	Master, PhD
Determining the credibility of online content	QURATOR	The World Wide Web Consortium (W3C) put together a list of more than 100 signals that can be used to determine the credibility of online content. The goal of this thesis is the implementation of a small number of these signals to determine the credibility of online content. As many different credibility signals in various clusters of signals exist, this topic can also be addressed in multiple theses.	Bachelor, Master
Flexible anaphor and kataphor replacement	QURATOR	Detection of anaphoric and kataphoric expressions and flexible replacement with their corresponding referents to enable, among others, information retrieval or summarisation applications that require paragraph extraction and paragraph reordering and also other tasks such as anonymization of documents.	Bachelor, Master
Identification of automatically generated or automatically paraphrased text	SPEAKER	To identify clickbait content or false content ("fake news"), one aspect can include determining if the whole text (or part of it) was automatically or manually generated (perhaps even postprocessed using an automatic paraphrasing algorithm). The goal of this thesis is to determine automatically into which category a text belongs.	Master
Fact checking – implementation of technical approaches and experiments	QURATOR	The goal of this thesis is, first, a systematic literature overview with regard to the area of automated fact checking and, second, performing experiments in the intersection between the W3C Web Credibility signals, Linked Data and Knowledge Graphs, Wikidata and Language Technology.	Master
Knowledge extraction from unstructured text	SPEAKER, QURATOR	The goal of this thesis is, first a systematic overview of recent literature (with a focus on working tools and approaches) with regard to the identification and extraction of knowledge from unstructured texts, focusing upon named entities and relations but also higher-level knowledge structures. The thesis includes the development of a working prototype for knowledge extraction from unstructured texts. This thesis is to be embedded in the wider Semantic Web and Linked Data "Knowledge Graph" paradigm, i.e., it's supposed to use representation formals such as, among others, OWL, RDF and SHACL. The thesis is supposed to establish a bridge to the Wikidata repository of knowledge items and semantic structures.	Master

Event detection	QURATOR	The goal of this thesis is, first a systematic overview of recent literature (with a focus on working tools and approaches) with regard to the identification and extraction of events from unstructured texts. The thesis includes the development of a working prototype for event extraction from unstructured texts based on the application of new approaches, derived from the literature review, probably deep learning-based techniques together with other NLP methods (named entity recognition, entity linking, relation extraction, etc.) or knowledge graphs like Wikidata and EventKG.	Bachelor, Master
Interoperability and workflow management in Natural Language Processing	QURATOR, ELG	The goal of this thesis is the further development of a prototypical workflow management engine developed by DFKI, for example, by including a graphical user interface that would allow the configuration of the workflow manager as well as the generation, execution, modification and deletion of workflows in the manager. The thesis is embedded in the wider topic of interoperability, i.e., the interoperability of multiple NLP tools and services that do not necessarily use the same annotation formats internally. In a more ambitious and advanced form of the thesis, the research can also include the implementation of an abstract and shared semantic space that is used to bridge between different annotation formats.	Bachelor, Master
Ontologies and Text Classification	QURATOR	The goal of this thesis is to explore technical approaches how to combine an ontology that contains genre or text type related knowledge with an actual classification system that is able to recognise or identify the genre of documents to be processed. There are various ways how to approach this topic, which requires a certain amount of interest in the topic of text genres (Textsorten, Texttypen, Textklassen) and also creativity on the side of the candidate. We build a basic ontology that provides a structured vocabulary to describe different kinds of document elements (mainly based on the Document Components Ontology, see http://www.semantic-web-journal.net/system/files/swj1016_0.pdf). The candidate could make use of this ontology for further research. This is why a basic understanding of ontologies/taxonomies and their serialization formats like Turtle/RDF would be helpfull, but it is not a must.	Master
Text Structure Embeddings	QURATOR	There are various types of embeddings, i.e., word embeddings, sentence embeddings, paragraph embeddings etc. The goal of this thesis is to examine if it possible to automatically learn embeddings that relate to typical text patterns, i.e., typical patterns that are used in typical ways of putting together an argumentation or the highly conventionalised structures of genres such as weather reports, football match reports or stock exchange reports. This topic requires in-depth knowledge of deep learning approaches and a high degree of creativity.	Master, PhD
Long Document Classification	QURATOR	The new Transformer architecture (BERT etc.) has led to significant improvements for various NLP tasks. However, Transformers are usually limited to short texts. This thesis explores how Transformer models can be set up and modified in such a way so that they work with longer documents for classification tasks.	Master
Question answering using structured knowledge (knowledge graphs, ontologies)	SPEAKER	Typical Question Answering systems are based on the usage of vast amounts of unstructured texts as training data or knowledge bases, in which the answers to given questions are then searched using deep learning architectures and learned models. The goal of this thesis is to explore the integration of structured knowledge graphs or ontologies and if they are able to improve the performance of current state of the art systems.	Bachelor, Master
Linked Data, NER, RE	SoNAR	From coarse-grained social relations explicitly addressed through detailed bibliographic metadata to fine-grained and detailed semantic and discourse relations based on full-text analysis.	Master