Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Open Bachelor & Master Thesis Topics

Prof. 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

08 June 2021

The DFKI SLT lab currently offers the Bachelor and Master topics shown in the table below. Most topics relate to **Computer Science** and **Computational Linguistics**, some also to **Linguistics**. Theses for other courses of study are also possible (in that case please approach us). Please note that this open call for candidates is, first and foremost, addressed to students at **TU Berlin**, **HU Berlin**, **FU Berlin** and **Universität Potsdam** (if you're interested in one of the topics but study at a different university, please contact us to see if we can work something out).

OUR EXPECTATION

- Candidates should be *self-motivated* and *ambitious*. We expect each thesis to result in a *meaningful piece of scientific work* that *we will attempt to publish at a scientific workshop or conference* after the completion of the thesis.
- Candidates should have *at least some experience with the scientific process* and how research in the area of NLP/CL/Linguistics is typically carried out.
- Practical theses in which a software, prototype or PoC is developed, will be integrated as a functional service into the European Language Grid platform.
- The thesis itself is to be written in English, ideally using LaTeX.

YOUR BENEFITS

- Especially in the initial topic definition and onboarding phase, regular meetings with one or more colleagues from the SLT team as mentors/coaches.
- Development of a thesis under the umbrella of a research project with possible integration of your work into a bigger demonstrator or prototype.
- Access to our GPU cluster for compute-heavy experiments.
- One or two presentations in joint sessions with all Bachelor/Master candidates.

WHEN CONTACTING US VIA EMAIL PLEASE …

- Provide your CV, a current Transcript and a brief motivational note.
- Indicate your experience in software development (e.g., programming languages, tools, approaches, backend vs. frontend), NLP, machine learning (neural approaches, corresponding toolkits etc.) and linguistics.
- Provide at least one piece of scientific writing (ideally several), for example, a course paper (in case of co-authorship, please indicate your contribution).
- Indicate which topic(s) you are interested in. If we invite you to a first meeting, please be prepared to give us a brief overview who you are and why you're interested in the topic you mentioned (ideally, please also read a few papers on the topic and prepare a rough idea how to tackle the corresponding topic).
- Indicate when you plan to start with the thesis and when you plan to finish.

DEADLINE

This is a rolling call. We're reviewing statements of interest on a continuous basis.

INTERESTED?

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

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	Description	Degree			Course
		Bachelor	Master	PhD	
Fine-grained analysis of geographic entities	Identification of geographical entities in a fine-grained way, not restricted to "Location" only, but also using more specific and more detailed (probably domain-specific) entities, such as "Address", "Building", "Lake", "Street", "Park", etc. This thesis topic belongs to the established area of Named Entity Recognition (NER), maybe also to Named Entity Linking (NEL) and also relates to taxonomies and knowledge graphs (including, potentially Wikidata).	X			Computer Science; Computational Linguistics
Linking and cross-referencing knowledge bases	The objective of this thesis is the development of a module capable of mapping or merging two ontologies by looking for the similarities between their classes, properties and instances. For example, if we have an ontology that refers to food and another to wine, the module has to look for the similarities between classes and instances and unite both ontologies in a single ontology. This can also be applied to the development of new ontologies, starting from several existing ontologies, where when defining some classes, properties or instances for the new ontology, complementary elements are offered to be added taken from other ontologies.	X			Computer Science; Computational Linguistics
Querying knowledge graphs using natural language	The goal of the thesis is the development of a tool that can convert natural language questions to SPARQL queries, being specially developed for concrete use cases depending on the requested knowledge base (various domains and use cases exist in our projects, for example, in the Covid-19 domain).	X			Computer Science; Computational Linguistics
Using textual entailment for semantic storytelling	Semantic storytelling aims to help content creators (e.g., journalists) to find novel stories in existing text documents, news articles etc. In this thesis, semantic storytelling is approached under the umbrella of textual entailment (TE). TE is the task of deciding, given two text fragments, whether the meaning of one text is entailed (i.e., can be inferred from) in another text. If we can observe entailment between two text fragments, these fragments can be used to put together a story.	X	X		Computer Science; Computational Linguistics
Multimedia Storytelling	News articles are typically accompanied by illustations and images, which attract the readers' attention and provide additional visual information. This thesis develops a system that automatically recommends relevant images for a given news article.	X	X		Computer Science; Computational Linguistics
Survey of NLP APIs with regard to their conceptual, technical and semantic compatibility	The goal of this thesis is to provide a comprehensive overview and comparision of the most widely used NLP APIs with regard to their conceptual, technical and semantic compatibility (see, e.g., https://rapidapi.com/collection/natural-language-processing-api). The thesis includes: identification of the most widely used, popular, advanced APIs and their providers; comparison of individual APIs concepts, capabilities (what tasks do they solve, including gap analysis); description of technical (in-)compatibility between them (protocols, formats, batch/request mode etc.); description of semantic (in-)compatibility between them (different knowledge bases, ontologies etc.). A possible extension would be to include existing platforms, tools or frameworks into the research that allow the integration of some of the above mentioned APIs (however, please see https://www.language-technology.com/textanalyticsreport).	X	X		Computer Science; Computational Linguistics
Segmentation of scientific articles from the health domain	The objective of this thesis is the segmentation of the content of scientific articles, especially from the health domain (with a focus on Covid-19), based on the textual content. This task can be conceptualised as a special case of text classification where the classes are, among others: Introduction, Background (or Related Work), Approach (or Methodology), Experiments, Results, Conclusions, etc. To carry out this thesis, we suggest to use Transformer-based models, especially SciBERT including fine-tuning the model to adapt it to the domain and type of text.	X	X		Computer Science; Computational Linguistics

Does multilingual training improve the performance for low resource language summarization?	The WikinewsSum dataset that we created for English, German and French and that is currently being extended to additional EU languages (PT, RU, IT, PL, ES etc.) is a dataset for multilingual summarization. The summary is the Wikinews article and the document(s) to summarize are the source texts of the Wikinews article. One possible approach could be to train state of the art extractive and abstractive summarization models in different scenarios. For example, (1) the model is trained on all available entries (all languages mixed up) and a token in the input specifies on which language the summary should be; (2) the model is trained on all available entries and fine-tuned on a specific language; (3) the model is only trained on a specific language. After these experiments, we can see which scenario is better for each language, hopefully showing that multilingual training improves the performance for low resource languages.	**X**	**X**		Computer Science; Computational Linguistics
Spelling out Meaning: Automated replacement of co-referring expressions for improved text summarisation and related applications	This thesis deals with the detection of, ideally all, co-referring expressions (pronouns etc.) and flexible replacement with their corresponding referents to enable, among others, information retrieval or summarisation applications that require sentence or paragraph extraction as well as sentence or paragraph reordering and also other tasks such as anonymization of documents.	**X**	**X**		Computer Science; Computational Linguistics
Abstractive summarisation of German texts	Development of an abstractive single document or multi-document text summarisation system for German language documents. The candidate can make use of an initial code base from a previous Master project.	**X**	**X**		Computer Science; Computational Linguistics
Automated or semi-automated identification of systematic online disinformation campaigns	The goal of this thesis is the development of technical concepts for the automated identification of orchestrated online disinformation campaigns as well as the creation of a data set and development of experiments (including evaluation). This thesis should also include a thorough review and survey of the current state of the art in this field. It is possible to focus upon the Covid-19 pandemic under the umbrella of this thesis.	**X**	**X**		Computer Science; Computational Linguistics
Transformer-based Language Models for Europe's Languages	The goal of the thesis is the development of a systematic overview of Transformer-based language models for all European languages (minimally the EU 24 languages, maybe even additional ones) in terms of corpora and data sets, performance, size and coverage, availability, experiments and benchmarks. The thesis can, optionally, also tackle the topic of multilingual language models and how they relate to the individual language models. This thesis is closely aligned with our EU projects European Language Equality and European Language Grid.	**X**	**X**		Computer Science; Computational Linguistics
Transformer-based Language Models and Markup Languages	The goal of the thesis is an exploratory overview and experimental development with regard to the research question if Transformer-based language models and markup languages (especially XML-based markup languages) can be technically integrated with one another: If large amounts of XML-tagged texts are used to finetune a Transformer-based language model, can we make use of – or: integrate – the markup into the language model? Can Transformers learn and reproduce hierarchical text structures that have been marked up in the original training documents?	**X**	**X**		Computer Science; Computational Linguistics
Bias in Wikipedia or Wikidata	The goal of this thesis is the investigation of bias effects in articles of the German Wikipedia or in the knowledge base Wikidata. If the thesis concentrates on Wikipedia articles, the topic can go towards the area of Computational Linguistics and Linguistics. If the thesis concentrates on Wikidata, the topic can go towards the area of Computer Science or Information and Library Science. In both cases, there would be a close collaboration with our project partner Wikimedia Deutschland e.V.		**X**		Computer Science; Computational Linguistics; Linguistics
Text-structure-informed text summarisation	Recent abstracts of the PubMed database are inherently structured, they provide brief statements on selected aspects of a scientific publication such as, among others, Background, Methodology, Materials, Methods, Conclusions etc. The goal of this thesis is to make use of this structured information and the full text of the corresponding article to develop a novel text summarisation approach that is primarily informed by text structure.		**X**		Computer Science; Computational Linguistics

Fact checking – between fact and fiction	The goal of this thesis is a linguistic analysis of common approaches towards fact checking. First, a systematic literature overview with regard to the area of automated fact checking has to be conducted with a special focus on the used methods and data sets or knowledge bases. Second, an analysis is to be performed: of which type or nature are the "facts" that are typically "checked" by automated systems? Third, an overview is needed what type of facts or statements should be checked when it comes to coordinated misinformation or disinformation campaigns. Key question: is there a mismatch between the actual pressing need for fact checking technologies and the technical approaches currently employed? If there is a gap – how big is it?		X		Linguistics; Computational Linguistics
Ontology-driven classification of text genres	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, especially relating to the rapid production of research prototypes as well as making use of existing data sets for distant supervision scenarios. We built 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. A basic understanding of ontologies/taxonomies and their serialization formats like Turtle/RDF would be helpful.		X		Computer Science; Computational Linguistics
Automatic identification of plans and strategies	The goal of this thesis is the development of classification models that are able to detect the basic components of strategic agendas, strategic plans, roadmaps, mission descriptions etc. The underlying idea is to make use of the vast collection of documents encoded in Strategy Markup Language (StratML), https://stratml.us/drybridge/index.htm, as the main data set. In addition to the actual development, implementation and evaluation of the experiments, this thesis would also involve quite a bit of data preparation and transformation work. The candidate should have a certain level of experience of text classification approaches. An interest in text genres, text types or text classes as well as text structure patterns is necessary.		X		Computer Science; Computational Linguistics
Question answering using structured knowledge (knowledge graphs, ontologies)	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 question answering systems. It is possible to make use of the very large and commercial-grade collection of 700.000 documents provided by one of our project partners.		X		Computer Science; Computational Linguistics
Question answering using structured knowledge (query expansion, recommendations)	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. In this thesis, the candidate makes use of the very large and commercial-grade collection of 700.000 documents provided by one of our project partners (also mentioned in the topic above) in order to concentrate on aspects such as smart recommendations and query expanasion).		X		Computer Science; Computational Linguistics
Knowledge Graph Embeddings and Contextualization	The objective of this thesis is the development of techniques for the generation of embeddings for knowledge bases. In the same way that Transformer-based models such as BERT can be used to perform different tasks on natural language text, the idea is to generate a model similar to BERT, but using knowledge bases instead of natural language text. These embeddings will be used later on, either with language-specific models such as BERT or perhaps also on top of BERT, for different NLP tasks, such as Question Answering.		X		Computer Science; Computational Linguistics

Semantic storytelling: filling gaps	Semantic storytelling aims to help content creators (e.g., journalists) to find novel stories in existing texts. This topic focuses on the generation of natural language text that connects previously fragmented texts into one coherent story. Given two text fragments A and C, a new fragment B should be generated based on the semantic information provided by A and C or the context of A and C. This work will utilize the latest Transformer models and information from discourse markers.		**X**		Computer Science; Computational Linguistics
Layman Language Translation: Paraphrasing German text	The objective of this thesis is the development of a paraphrasing module that works in German and perhaps also in English and that allows text simplification. One possible application domain could be the simplification of legal documents so that they are more comprehensible for the general public or the simplification of medical documents (for example, papers on Covid-19) into layperson language that can be understood by the general public. One of the techniques commonly used to implement paraphrasing is machine translation, but language is not translated between two different natural languages, but between two different styles or registers of the same language. This approach can be explored to check its feasibility for the text simplification task (from, say, "regular language" to "simplified language").		**X**		Computer Science; Computational Linguistics
Linked Data infrastructure including knowledge graphs and annotations	The objective of this thesis is the development of a Linked Data platform, combined with document management (supporting both text and multimedia files). Quite a few linked data platforms exist, but not that many allow the synchronized handling of documents. Some like Trellis or Apache Marmota are rather advanced, but still have some gaps that we want to address with this thesis.		**X**		Computer Science; Computational Linguistics
Generation and population of knowledge graphs	The goal of this thesis is, first a systematic overview of recent literature (with a focus on working tools and best practice 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 (various domains and use cases exist in our projects). 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.		**X**		Computer Science; Computational Linguistics
Development of an ontology for event detection based on a multilingual verb synonym lexicon	The goal of this thesis is to conceptualize and create an ontology of event types that can be used as a target knowledge base in named entity recognition and relation extraction-like processing scenarios and also used as a human-readable and human-understandable database for all types of events, processes and states. It should be based on the Czech-English synonym verb lexicon SynSemClass that we are currently extending with German entries. For more Information on the SynSemClass Lexicon, please refer to the paper: https://www.aclweb.org/anthology/2020.globalex-1.2.pdf		**X**		Computer Science; Computational Linguistics
Do intertextual argumentative relations or coherence relations exist?	Texts on the same topic and, roughly, of the same genre or class of text types, but written by *different* authors exhibit certain commonalities. Can we say that intertextual argumentative relations or coherence relations exist between different segments taken from such texts written by *different* authors? How can such relations be detected in an automated way? How can we extract the relevant segments? How can we rearrange them, for example, under the umbrella of an approach for multi-document summarisation or semantic storytelling?		**X**	**X**	Computational Linguistics; Linguistics

Semantic segmentation of HTML documents with optical methods	The goal of this thesis is to develop a system that is able to perform, in a maximally robust way, semantic segmentation of regular HTML documents (either single documents or multiple connected and inter-linked HTML documents) into their main document building blocks such as Headline, NavigationBar, Paragraph/Textblock, Figure, Caption, Advertisement etc. using methods taken from Optical Character Recognition (OCR) and/or Optical Layout Recognition (OLR). After the identification of the document building blocks (probably to be performed by processing a screenshot of the page), the identified information is to be mapped back onto the DOM tree representation of the document so that the information about the different building blocks is readily available in the DOM tree as additional elements. (The processing of building blocks can also be performed together with information taken or extracted from the DOM tree and from the document itself, i.e., making use of hybrid methods.) One optional step could include the mapping of the extracted high-level building blocks (or document components) onto functional and more semantic or rhetorical types such as, among others, Introduction, Discussion, Conclusions, Comparison, Elaboration, Background etc. This thesis could include working with document grammars and document ontologies, perhaps in a Semantic Web and Linked Data paradigm, modelling document structures using OWL, RDF, SHACL etc. The overall idea is to combine the extracted visual/optical, structural and content information to add an additional semantic or text-structural layer to an HTML document that can be made use of for the automated processing of the document.		**X**	**X**	Computer Science; Computational Linguistics
Semantic Segmentation of HTML documents	This thesis deals, essentially, with the same research question as the previous topic 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.		**X**	**X**	Computer Science; Computational Linguistics
From named event detection to un-named event detection	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. This topic focuses on the detection of named events, i.e., events that have had such a large impact that they are commonly referred to by using their name. One key question is if it's possible to recognise named events in a precise way and to use these results to learn corresponding models to detect un-named events.		**X**	**X**	Computer Science; Computational Linguistics
Neural RST parsing	The goal of this thesis is the development of an RST (Rhetorical Structure Theory) parser for English language documents using neural technologies, based on available RST corpora (GUM corpus RST Discourse Treebank).		**X**	**X**	Computer Science; Computational Linguistics
Can typical text structure patterns be embedded, too?	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 is possible to automatically learn embeddings that relate to typical text or text structure patterns, i.e., 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 and scientific curiosity. An interest in text genres, text types or text classes as well as text structure patterns is necessary.		**X**	**X**	Computer Science; Computational Linguistics