UNIVERSITÄT BIELEFELD
Faculty of Technology

Semantic Computing Group
Moritz Blum
mblum@techfak.uni-bielefeld.de

Bachelor's or Master's Thesis

Information extraction from material science literature

We have created a small corpus of scientific literature about the material science domain. We would like to analyze and extract information from this corpus with state-of-the-art methods, and use the extracted information in further research in the field of knowledge graphs. Therefore, texts need to be aligned to one or more domain specific knowledge graphs (which ones are to be decided within the project), which means that entities (e.g., chemical elements) that occur in the KG need to be identified in the text. We call this text-data alignment, which results in an annotated corpus. Besides the identification of entities, we also need to identify values in text (e.g., the temperature of the melting point of some element). Before text-data alignment can take place, the texts need to be automatically extracted from PDF documents and cleaned.

Ideally, you should have some NLP experience, can work with big data, and basic knowledge in physics or chemistry is a plus.

Related literature

- Simone Tedeschi, Simone Conia, Francesco Cecconi, and Roberto Navigli. 2021. Named Entity Recognition for Entity Linking: What Works and What's Next. `https://aclanthology.org/2021.findings-emnlp.220.pdf`
  knowledge from materials science literature
- Mendes et al.: DBpedia Spotlight: Shedding Light on the Web of Documents (`www.dbpedia-spotlight.org`)
- `https://labs.tib.eu/falcon/falcon2/`
- Dan Jurafsky and James H. Martin: Speech and Language Processing
- `www.wikidata.org`
- `www.spacy.io`

The Semantic Computing Group researches and develops methods that enable machines to acquire, process, and understand data from natural language and knowledge graphs. Using methods from *natural language processing* and *machine learning*, we are aiming at machines that are capable of knowledge acquisition by reading various kinds of data. In particular, the group focuses on methods for information extraction, knowledge graph completion, semantic parsing, ontology learning, sentiment analysis, entity linking, as well as question answering.

More information is available at: http://sc.cit-ec.uni-bielefeld.de.

Interested? @mail to mblum@techfak.uni-bielefeld.de