NLP04 Information Extraction from Research Papers for DIGIS

Background
The objective is to devise approaches for the automated extraction of geochemical data and metadata from research papers and implement them prototypically pipeline for the geochemical data infrastructure DIGIS.

We extract specific mentions of methods from papers. This information can be part of the paper or included in tables or figures. The structure depends on the journal.

Goal
• A prototype for extraction specific information from research papers

Tasks
• Compare existing approaches for information extraction for a given set of papers
• Implement a prototype
• Draft a data pipeline

Digital Geochemistry Infrastructure for GEOROC 2.0

Daniel Kurzawe
kurzawe@sub.uni-goettingen.de

Mathias Göbel
goebel@sub.uni-goettingen.de

We are hiring! Student assistants and PhD candidates!
Background

The non-statement view (or structuralistic theory concept) uses set theory to describe a scientific theory through its internal structure and in conjunction with larger theory networks. This philosophical framework allows a generic theory description.

There are several publications about structural reconstructions of scientific theories, e.g. Newton particle mechanics. Due to its set-theoretical nature, a (semi-)automatic approach for such a reconstruction might be possible. This project explores this approach.

Goal

• Extraction theory components theory and transformation into a structural theory description

Tasks

• Explore concept for a semi-automatic reconstruction process
• Mapping semantic and concepts
• Build a theory network
• Implement a parser and transformer for a specific domain

Daniel Kurzawe
kurzawe@sub.uni-goettingen.de
NLP23 Textual Criticism and Plagiarism

Background
“The identification of textual variants, or different versions, of either manuscripts or of printed books” (Wikipedia) is a major task in philology entitled “textual criticism”. The analysis of a single text in different variants starting from the very first sketch up to the latest authorized version is provided with a historical-critical edition. Before the digital age, these editions used an obnoxious amount of signs marking and categorizing these differences, like the Leiden convention has standardized. However, newer visualization technologies provide more and more interactive views to these editions.

What are the shared approaches of plagiarism detection and textual criticism? Can they benefit from each other?

Goal
• Investigate if/how a software for plagiarism detection can be utilized to deal with a set of documents that represent the same text. Comparison of both methods.

Tasks
• Input material selection (assisted)
• Data conversion
• Usage of PD software/visualization, publish result using web technology
• Workflow to automatize main steps and to scale up (deal with as many editions as possible)