PhD Program
Knowledge Discovery in Scientific Literature

TECHNISCHE UNIVERSITÄT DARMSTADT

Knowledge Discovery in Scientific Literature

Presentation: Erik-Lân Do Dinh, Iryna Gurevych

http://www.kdsl.tu-darmstadt.de

Research Focus

Our research is about novel methods of language and knowledge processing applied to the research publications in the educational domain and to historical scientific corpora.

We research cutting-edge methods for multi-label classification, argumentation mining, focused crawling, relation extraction, temporal network analysis and figurative language analysis. KDSL applies these methods to answer research questions related to the educational research on novel as well as historic corpora. Thus, we enable new forms of intelligent information access for humanities researchers. We are part of the Research Initiative „Knowledge Discovery in the Web“ at the Technische Universität Darmstadt. Besides, we work together with the LOEWE Research Center „Digital Humanities“ (Univ. of Frankfurt, TU Darmstadt), the Institute for Computational Linguistics (Univ. of Heidelberg), and the BMBF-funded research center for Digital Humanities CEDIFOR (Univ. of Frankfurt, TU Darmstadt, German Institute for International Educational Research).

Research Program

The main topic of the PhD program is knowledge discovery in the vast amount of scientific literature ubiquitously available on the Web. This research employs methods of intelligent identification and analysis of structures in scientific texts on all scales, enabling completely new, previously unforeseen forms of access to scientific information.

The underlying data is represented by full texts of publications, including new forms of scientific publication in Web 2.0 and historical sources, together with their manifold networking via citations, authors and the linguistic and semantic knowledge extracted from the texts.

The goals of the PhD program require intensive research on data- und text-mining methods and their application to unstructured scientific information and historical corpora. The methods are for example applied to querying, indexing, discovering and preparing knowledge on the web. Implicit knowledge contained in scientific literature is thus tapped and rendered usable.

The research program “Knowledge Discovery in Scientific Literature” focuses on the educational research as the target domain. To this end, we use the data collected at the German Institute for Educational Research (DIPF) and the Technische Universität Darmstadt. We closely cooperate with the users of the developed innovative technology and humanities researchers.

Application Scenario

A researcher is interested in a research question that might be answered on basis of a data set X.

There are around 1000 publications containing work on data set X.

The system searches the Web for relevant publications to data set X, structures the results by research questions, variables, and methods. It personalizes the results and visualizes remaining research gaps.

Motivation & Goals

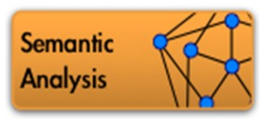
Future expert information assumes the ubiquitious availability of scientific literature on the web. It aims at intelligent added value services for scientific communities. Foundations in modern **computer science methods**, e.g. content analysis, knowledge discovery and knowledge processing need to be developed, extended and intermeshed to achieve this. Basic research on this is therefore a central goal of the program, laying foundations for digital libraries of the future.

Research Areas

- Crawling and Semantic Structuring of Scientific Publications in the Web
- Temporally Dynamic Networks of Topics and Authors in Scientific Publications
- Scalable Multilabel Classification for Educational Research
- Knowledge Extraction and Consolidation for Scientific Publications in the Educational Domain
- Domain-Adaptive Text Mining to Support Knowledge Discovery in Scientific Historical Literature
- When „Evolution“ leads the Way: Argumentation, Constructions and Metaphors in German Popular Biopolitics on the Eve of the 20th Century.
- Metaphor in Use: Analysing Figurative Speech in the Literary and Scientific Discourse of the 20th Century

Common Methods

- Corpus based work and annotation (Núñez, Gerloff)
- Multilabel classification (Nam)
- Focused crawling and relation classification (Remus)
- Co-occurrence analysis (Ma)
- Metaphor detection (Do Dinh)
- Argumentation analysis (Kirschner)



Supervisors and Associated Researchers

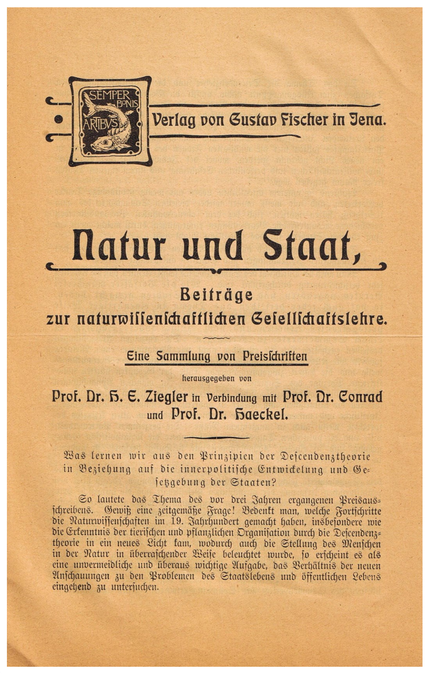
Dr. Sabine Bartsch (Literary Studies, English Linguistics)
Prof. Dr. Chris Biemann (Computer Science, Language Technology)
Dr. Judith Eckle-Kohler (Computer Science, Ubiquitous Knowledge Processing)
Prof. Dr. Johannes Fürnkranz (Computer Science, Knowledge Engineering)
Prof. Dr. Petra Gehring (Philosophy, Language and Technical Philosophy)
Prof. Dr. Iryna Gurevych (Computer Science, Ubiquitous Knowledge Processing)
Prof. Dr. Nina Janich (Literary Studies, German Linguistics)
Prof. Dr. Andrea Rapp (Literary Studies, Digital Philology)
Prof. Dr. Karsten Weihe (Computer Science, Algorithmics)

PhD Program

The supervision of young researchers in the PhD program strongly relies on close contacts between members of the program, regular joint meetings, co-supervision by professors and senior researchers from multiple disciplines and a lively exchange in the research and qualification program “Language and Knowledge Engineering”.
Furthermore, the program strives to publish research findings at leading scientific conferences and provides its software freely accessible as an open source product on the basis of the DKPro framework.

Data & Corpora

Our data collections include modern corpora in the field of educational research, such as **FIS Bildung Literaturdatenbank** and **peDOCS**, containing references and full text articles. We also employ historical educational corpora from scientific literature, like **Natur und Staat** (1903 - 1911).



Verlag von Gustav Fischer in Jena.

Natur und Staat,

Beiträge
zur naturwissenschaftlichen Gesellschaftslehre.

Eine Sammlung von Preisschriften
herausgegeben von
Prof. Dr. H. E. Ziegler in Verbindung mit Prof. Dr. Conrad und Prof. Dr. Haeckel.

Was lernen wir aus den Prinzipien der Descendenztheorie in Beziehung auf die innerpolitische Entwickelung und Gesetzgebung der Staaten?

So lautete das Thema des vor drei Jahren ergangenen Preisausschreibens. Gewiß eine zeitgemäße Frage! Bedenkt man, welche Fortschritte die Naturwissenschaften im 19. Jahrhundert gemacht haben, insbesondere wie die Erkenntnis der tierischen und pflanzlichen Organisation durch die Descendenztheorie in ein neues Licht kam, wodurch auch die Stellung des Menschen in der Natur in überraschender Weise beleuchtet wurde, so erscheint es als eine unvermeidliche und überaus wichtige Aufgabe, das Verhältnis der neuen Anschauungen zu den Problemen des Staatslebens und öffentlichen Lebens eingehend zu untersuchen.

Participants

UBIQUITOUS KNOWLEDGE PROCESSING

Language Technology

Knowledge Engineering

Algorithmics

Language and Technical Philosophy

Digital Philology

This work has been supported by the
Deutsches Institut für Internationale Pädagogische Forschung
Technische Universität Darmstadt, Fachbereich Informatik.

Fachbereich Informatik

DIPF Bildungsforschung und Bildungsinformation