Information Extraction with the Darmstadt Knowledge Processing Software Repository

Iryna Gurevych and Mark-Christoph Müller
Ubiquitous Knowledge Processing (UKP) Lab
Computer Science Department, Technische Universität Darmstadt

Current Natural Language Processing (NLP) systems feature high-complexity processing pipelines that require the use of components at different levels of linguistic and application specific processing. These components often have to interface with external e.g. machine learning and information retrieval libraries as well as tools for human annotation and visualization. At the UKP Lab, we are working on the Darmstadt Knowledge Processing Software Repository (DKPro) (Gurevych et al., 2007a; Müller et al., 2008) to create a highly flexible, scalable and easy-to-use toolkit that allows rapid creation of complex NLP pipelines for semantic information processing on demand. The DKPro repository consists of several main parts created to serve the purposes of different NLP application areas.

- DKPro **core** components are general purpose analysis components. Core components are readers for generic text and XML files, and annotators for standard preprocessing tasks like tokenization, sentence splitting, POS-tagging and lemmatization,[1] stop word removal, parsing[2], and others. The core also includes annotation consumers, e.g. one that can produce output in the format used by the general-purpose annotation tool MMAX2 (Müller & Strube, 2006).

- DKPro **information retrieval** components supply functionality for all phases of information retrieval, including indexing, retrieval, and (qualitative and quantitative) evaluation. The components use existing information retrieval frameworks, viz. Lucene and Terrier[3]. Evaluation components are based on the standard TREC evaluation tools[4].

- DKPro components for **text mining** include readers for importing text from specialized sites like FAQs, forums like e.g. Nabble, social Q/A sites like YahooAnswers, and Technorati. Since these texts are often highly subjective, annotators for detecting opinion- or sentiment-related properties are also included.

- DKPro components for **processing user generated discourse** are tailored towards tackling the problems that come with noisy, error-ridden, and ill-formed input found in forums, blogs, and other community web sites. They include components for spell checking and -correcting and for annotating e.g. swear words and artifacts like smileys.

So far, the DKPro repository has been successfully employed as the technical framework in several research projects at the UKP Lab:

- *Semantic Information Retrieval* (funded by the DFG) for information retrieval in the domain of electronic career guidance (Gurevych et al., 2007b; Müller et al., 2008), computing semantic relatedness of words (Zesch et al., 2008), and constructing lexical semantic graphs (Schwager, 2008);

[1] Based on e.g. TreeTagger (`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`).

[2] Based on e.g. the Stanford parser (`http://nlp.stanford.edu/software/lex-parser.shtml`) or the BITPAR parser (`http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/BitPar.html`).

[3] `http://lucene.apache.org/`, `http://ir.dcs.gla.ac.uk/terrier/`

[4] `http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz`

- *Question Answering for eLearning* (funded by the DFG) for question answering by mining FAQs, question paraphrase recognition (Bernhard & Gurevych, 2008), automatic quality assessment (Weimer et al., 2007), and comparative analysis of user generated discourse (Shen, 2008);

- *Theseus-TEXO* (funded by the BMWI) and *Sentiment Analysis for eLearning* (funded by the DFG) as a common architecture for community mining, e.g. opinion and trend mining in customer reviews or blogs (Toprak, 2007; Qu, 2007; Ferreira et al., 2008);

Parts of the DKPro repository will be released to the general public in the near future. In particular, general-purpose and selected information retrieval components will be made available at the UKP Lab website. Further topic-specific toolkits, such as the community mining toolkit and the toolkit for processing user generated content, will be made available as the projects evolve in the future.

Acknowledgements: The Darmstadt Knowledge Processing Software Repository is a joint project of the UKP Lab. Crucial parts of this work were carried out by Christof Müller, Lizhen Qu and Torsten Zesch. Substantial parts of the DKPro repository were also contributed by Niklas Jakob, Cigdem Toprak, Kateryna Ignatova, Delphine Bernhard and several undergrad students.

References

Bernhard, Delphine & Iryna Gurevych (2008). Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications, ACL 2008*. Columbus, Ohio, USA. To appear.

Ferreira, Liliana, Niklas Jakob & Iryna Gurevych (2008). A comparative study of feature extraction algorithms in customer reviews. In *Proceedings of the Second IEEE International Conference on Semantic Computing*. Santa Clara, CA, USA. To appear.

Gurevych, Iryna, Christof Müller & Torsten Zesch (2007a). Teaching "Unstructured Information Management: Theory and Applications" to Computational Linguistics Students. In *Proceedings of the First Workshop on Unstructured Information Management Architecture at Biannual Conference of the Society for Computational Linguistics and Language Technology*. Tübingen, Germany.

Gurevych, Iryna, Christof Müller & Torsten Zesch (2007b). What to be? - electronic career guidance based on semantic relatedness. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 1032–1039. Prague, Czech Republic: Association for Computational Linguistics.

Müller, Christof, Torsten Zesch, Mark-Christoph Müller, Delphine Bernhard, Kateryna Ignatova, Iryna Gurevych & Max Mühlhäuser (2008). Flexible UIMA Components for Information Retrieval Research. In *Proceedings of the LREC 2008 Workshop 'Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP'*, Marrakech, Morocco, May 31, 2008.

Müller, Mark-Christoph & Michael Strube (2006). Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn & Joybrato Mukherjee (Eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt a.M., Germany: Peter Lang.

Qu, Lizhen (2007). *Using Semantic Knowledge to improve Information Search in Web 2.0*. Diploma Thesis, Universität Karlsruhe.

Schwager, Florian (2008). *Automated Analysis of Lexical Cohesion*. Diploma Thesis, Technische Universität Darmstadt.

Shen, Zhi (2008). *A toolkit for the automatic pre-processing and analysis of user generated discourse*. Diploma Thesis, Technische Universität Darmstadt. To appear.

Toprak, Cigdem (2007). *Sentiment Detection in Natural Language Texts based on a case study: Analysis of professional profiles*. Diploma Thesis, Technische Universität Darmstadt.

Weimer, Markus, Iryna Gurevych & Max Mühlhäuser (2007). Automatically Assessing the Post Quality in Online Discussions on Software. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume, Proceedings of the Demo and Poster Sessions*, pp. 125–128. Prague, Czech Republic: Association for Computational Linguistics.

Zesch, Torsten, Christof Müller & Iryna Gurevych (2008). Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of AAAI*. To appear.