Darmstadt Knowledge Processing Repository Based on UIMA

Iryna Gurevych, Max Mühlhäuser, Christof Müller,
Jürgen Steimle, Markus Weimer, Torsten Zesch
Ubiquitous Knowledge Processing Group, Telecooperation Division
Darmstadt University of Technology
http://www.ukp.tu-darmstadt.de

February 9, 2007

Introduction

The Ubiquitous Knowledge Processing (UKP) Group at Darmstadt University of Technology pursues the vision of using information management, information retrieval, and text mining technologies to create innovative applications, such as intuitive information access in Web 2.0 (O'Reilly, 2005) and eLearning. Thereby, semantic information processing technologies are utilized to transform unstructured information into structured knowledge for different media types, including text and handwriting.

In order to support the interoperability of components created in various research projects of UKP, we decided to build upon Unstructured Information Management Architecture (UIMA) (Ferrucci and Lally, 2004), released as an open-source project by IBM in 2006. The mid-term goal is to provide a collection of software components for semantic information processing based on UIMA, called Darmstadt Knowledge Processing Software Repository (DKPro). DKPro should support semantic information processing along several dimensions, such as:

- Media types (text, speech, handwriting, multimedia, etc.)
- Domains (eLearning, semantic web services, etc.)
- Natural languages (English, German, etc.)

Projects building upon UIMA

Currently, UIMA is deployed in two projects: "Semantic Information Retrieval" (SIR),[1] and "Automatic Quality Assessment and Feedback in eLearning 2.0" (AQUA).[2]

The SIR project aims at improving information retrieval by incorporating lexical semantic relationships between words or concepts. The lexical semantic relationships are determined using knowledge sources, such as WordNet (Fellbaum, 1998), GermaNet (Kunze, 2004), or Wikipedia[3]. The knowledge is used to augment the search space to retrieve documents that do not literally contain query terms, but strongly related terms. Another goal of the SIR project is to enable user input in the form of natural language texts.

Within the AQUA project, we investigate two types of discourse in eLearning resulting from user generated content: (1) online discussions as found on Yahoo or Google Groups; and (2) electronic notes, either typed or handwritten, taken on scientific presentations. We develop methods to automatically assess the quality and the communicative function of this eLearning discourse. The results of automatic quality assessment will be used to provide useful feedback to authors and to improve automatic content summarization.

[1] http://www.ukp.tu-darmstadt.de/projects/sir
[2] http://www.ukp.tu-darmstadt.de/projects/aqua
[3] http://www.wikipedia.org

Darmstadt Knowledge Processing Repository (DKPro)

So far, we created a set of general purpose and project specific knowledge processing components. Table 1 gives an overview of the available components as well as components, which will be implemented in the near future (those are written in *italics*).

The SIR project uses UIMA based components for extracting important query terms from natural language queries used in information retrieval as well as creating index files from text corpora. Our preprocessing pipeline contains tokenizer, sentence splitter, stemmer or lemmatizer (depending on the system configuration), stopword tagger, PoS-tagger and indexer.

Our current work in the AQUA project focuses on using machine learning to predict the quality of forum postings. We use the components integrated on basis of UIMA to annotate these posts and compile feature vectors from these annotations. We then export these feature vectors to an ARFF-file. This facilitates experiments using state of the art machine learning toolkits like WEKA (Witten and Frank, 2005) and YALE[4]. In the future, we plan to extend this to deal with hand written notes.

Both projects may seem very different at the first glance, but many UIMA based components can be shared between projects, such as tokenizer, lemmatizer, or PoS-taggger. Thus, UIMA proved to support the collaborative creation and use of natural language processing software components. We are looking forward to see how it will facilitate exchange and re-use of components on a broader scale.

Acknowledgments

This work was carried out in the project "Semantic Information Retrieval from Texts in the Example Domain Electronic Career Guidance" (SIR) and as part of the Graduate School "Feedback Based Quality Management in eLearning" funded by the German Research Foundation under the grants GU 798/1-2 and GK 1223, respectively.

[4] `http://yale.sourceforge.net`

References

Fellbaum, C. (1998). *WordNet An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ferrucci, D. and Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Gospodnetic, O. and Hatcher, E. (2005). *Lucene in Action*. Manning Publications Co.

Kunze, C. (2004). *Lexikalisch-semantische Wortnetze*, chapter Computerlinguistik und Sprachtechnologie, pages 423–431. Spektrum Akademischer Verlag.

O'Reilly, T. (2005). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. URL `http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web%-20.html`.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., and Lioma, C. (2006). Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*.

Patwardhan, S. and Pedersen, T. (2006). Using WordNet Based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, Trento, Italy.

Schmid, H. (1995). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

Schmid, H. (2004). Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2 edition.

Zesch, T., Gurevych, I., and Mühlhäuser, M. (2007). Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In *Biannual Conference of the Society for Computational Linguistics and Language Technology*, pages 213–221, Tuebingen, Germany.

Type	Component	Functionality
Linguistic Preprocessing	Tokenizer	Tags tokens.
	Sentence splitter	Tags sentence boundaries.
	Stopword tagger	Tags tokens that are found in a stoplist.
	List tagger	Tags lists and enumerations.
	Paragraph tagger	Tags paragraphs based the document structure.
	Separator tagger	Tags content separators, e.g., ‘‘------’’.
	Language detector	Based on heuristics using language specific dictionaries.
Morphological analysis	Stemmer	Wrapper for the Snowball stemmer (`http://snowball.tartarus.org`).
	Lemmatizer	Uses the lemmatizing capabilities of TreeTagger (Schmid, 1995).
	Compound splitter	Splits German compounds based on a linguistically motivated rule set (credits to Nils Ott).
Syntactic analysis	Part-of-speech tagger	Wrapper for TreeTagger (Schmid, 1995).
	Parser	Wrapper for BitPar (Schmid, 2004).
Lexical analysis	Swear word tagger	Tags swear words based on a dictionary.
	Spelling error tagger	Tags spelling errors based on `aspell` dictionaries.
String analysis	URL tagger	Tags occurrences of URLs in a text, e.g., `http://www.ukp.tu-darmstadt.de`.
	Path tagger	Tags `UNIX` paths.
	Code tagger	Tags text parts that are programming code.
Semantic analysis	*Named Entity Recognizer*	Tags named entities using a hybrid system (rules & gazetteers).
	Sentiment Detector	Detects sentiment expressions in English and links them to the evaluated entity.
	Word Sense Disambiguator	Tags word senses using the algorithm by Patwardhan and Pedersen (2006).
Web forum analysis	Topic similarity	Computes topic similarity between a forum and a post based on the vector space model.
	Quote annotator	Tags explicit quotes, e.g. lines starting with “`>`” in emails.
Data Import	Wikipedia reader	Imports Wikipedia articles by means of the Wikipedia API’s query interface (Zesch et al., 2007). Wikipedia API is available at `http://www.ukp.tu-darmstadt.de/software/WikipediaAPI`
	Forum reader	Imports forum discussions into the UIMA pipeline.
Data Export	Indexer	Creates indexes for Lucene (Gospodnetic and Hatcher, 2005) and Terrier (Ounis et al., 2006) from a corpus.
	ARFF export	Exports feature vectors for machine learning tools.

Table 1: List of components in Darmstadt Knowledge Processing Repository. Components in *italics* are work in progress.