An UIMA-Based Tool Suite for Semantic Text Processing

Katrin Tomanek, Ekaterina Buyko, Udo Hahn

Jena University Language and Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena

We report on recent efforts in our lab to specify a comprehensive pipeline of NLP processes within the UIMA framework. In our research, we deal with written documents, mainly taken from the biomedical domain. The goal of our analysis is to empower semantic search for information extracted from these textual resources. This is a complex task which integrates various layers of linguistic analysis on documents collected from various repositories (bibliographic databases, biomedical fact databases, and documents from the WWW). Therefore, the level of abstraction provided by UIMA specifications might be helpful in breaking the complexity barrier encountered by such semantically driven text processing systems.

The NLP tool suite we propose consists of a general-purpose TYPE SYSTEM, as well as various task- and domain-specific ANALYSIS ENGINES, COLLECTION READERS and CONSUMERS.

The UIMA TYPE SYSTEM covers several layers of annotations, such as document meta information (author, title, publication date, etc.), document structure information (different zones in a document, e.g., title, abstract, running paragraph), morpho-syntactic information (POS tags, morpho-syntactic features, chunk and syntactic parse annotation), and finally semantics (mainly entities and relations). The concrete items which appear at the semantic annotation layer, obviously, introduce some dependency related to the domain in focus of any NLP system. In our case, semantic entities and relations were drawn from two biomedical application areas, viz. gene regulation and stem cell transplantation/immunology.

The tool suite comprises ANALYSIS ENGINES for sentence splitting, tokenization, POS tagging, shallow and full parsing, acronym detection (particularly relevant in the biomedical domain), general-purpose named entity recognition, and, as a tribute to our application domain, mappings from entities-to-databases. With the exception of the acronym detector and the mapper, all components are based on general-purpose machine learning approaches so that the engines can be applied to other domains after re-training. Especially the named entity recognizer is highly configurable with respect to the labels and features to be used. We are planning to add further NLP tools in the near future, such as document classifiers, consistency checkers for semantic annotations (e.g., named entities), and relation extractors.

The input stream for our semantic text processor is handled by COLLECTION READERS. For any biomedical project, documents of major interest are abstracts from PubMed[1], which already contain some sort of meta information such as the publication date, document source, authors, title, keywords, etc. This important information is interpreted by our PUBMEDREADER, a collection reader for PubMed abstracts. Since we incorporate documents from other sources as well (e.g., conference proceedings, WWW documents), we intend to provide further specialized UIMA collection readers capable to cope with the formats they come with.

The various outcomes of text analysis processes are input to different CONSUMERS. We currently focus on semantic search as a major consumer. A crucial preprocessing step of any semantic search system is the semantic mark-up of documents that should be indexed by the search engine later on. In our research framework, the semantic mark-up comprises a large set of bio-medical (named) entities from the domain of gene regulation and stem cell transplantation (genes, proteins, organisms, variations, cytokines and their receptors, histocompatibility alleles, etc.).

Those entity mentions have to be identified in scientific documents, automatically labeled with their specific semantic type, and mapped to biological databases[2], which contain additional information. All processing which is necessary to obtain this semantic mark-up of the documents is performed by the collection of NLP tools introduced above. Particularly advantageous is the potential of the UIMA framework to alter and reconfigure the composition of different NLP pipelines according to different needs (e.g., for different task consumers or for component-wise evaluation of the effectiveness of single engines, see below).

Our semantic search engine is based on LUCENE[3]. Its index is directly built from the annotations provided by our UIMA NLP pipeline by means of the LUCENEINDEXER, a specialized UIMA consumer. Thus, the UIMA tool suite allows to read streams of relevant documents, generate mark-ups and additional meta-information, and index these documents within one single pipeline.

Part of our on-going work is to develop an evaluation suite (i.e., a set of UIMA collection readers and consumers), which will enable us to evaluate both single NLP components as well as the complete NLP pipeline within the UIMA framework. This is a particularly exciting topic as, due to the level of abstraction it allows, UIMA renders the framework to ease the investigation of the effects of alternative, say, syntactic components (taggers, chunker, parser) on the outcome of semantic processing.

Acknowledgments. This research was funded by the EC´s 6th Framework Programme (4th call) wi thin the BOOTStrep project under grant FP6-028099, and by the German Ministry of Education and Research (BMBF) via its e-Science initiative within the StemNet project (funding code: 01DS001A to 1C).

[1] `http://www.ncbi.nlm.nih.gov/entrez`

[2] e.g. UniProt `http://www.expasy.uniprot.org/`

[3] Lucene is a open source, high-performance text search engine, see also: `http://lucene.apache.org/`