An UIMA Annotation Type System for a Generic Text Mining Architecture

Udo Hahn, Ekaterina Buyko, Katrin Tomanek

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

Scott Piao, Yoshimasa Tsuruoka, John McNaught, Sophia Ananiadou

National Centre for Text Mining & School of Computer Science
University of Manchester

Meta data annotation languages for language resources have already a long-standing tradition (e.g., TEI[1], XCES[2], Penn TreeBank[3], Dublin Core Metadata Initiative[4]). Their goal, however, has never been to comprehensively cover the complete NLP pipeline as needed for complex tasks such as information extraction or text mining. Still, this work contains valuable information which should be incorporated into larger specification efforts for NLP systems.

This is where UIMA comes in, which serves as a platform for the integration of NLP components and the deployment of complex NLP pipelines. Crucial for the integration of single components are their input/output specifications provided with the component-specific wrappers. These specifications should previously be defined in the so called UIMA annotation type system.

We present here the type system for a generic text mining architecture as developed within the context of two projects, *viz.* StemNet (BMBF)[5] and BOOTStrep (EU)[6]. The general goal of both projects is to provide human language technologies for automatically identifying relevant knowledge for researchers in the life sciences in bio-medical texts, especially knowledge in the domain of stem cell transplantation (StemNet) and gene regulation (BOOTStrep). The type system should be apt to cover the full circle of NLP analysis, i.e., from the analysis of the formal document structure and its meta information to the core NLP analysis steps such as tokenisation, part-of-speech tagging and parsing and, finally, the results of semantic analysis (named entity recognition, relation extraction).

[1] http://www.tei-c.org
[2] http://www.cs.vassar.edu/XCES/
[3] http://www.cis.upenn.edu/ treebank
[4] http://dublincore.org
[5] http://www.stemnet.de
[6] http://www.bootstrep.eu

The type system we designed consists of six layers: *Document Meta, Document Structure, Style, Morpho-Syntax, Syntax and Semantics.* The *Document Meta* layer describes the bibliographical and content information about a complete document. The bibliographical information can often be retrieved from the header of the analysed document and contains, e.g., publication date, authors, title and copyright information. The description of its content often comes with a list of keywords and will be represented at this layer, as well. We clearly distinguish here between domain-independent information such as language, title, document type and domain-dependent information as relevant for text mining in the bio-medical domain.

The *Document Structure* and *Style* layers contain information about the organisation and layout of the analysed documents. Sentences, paragraphs and rhetorical zones (such as title, abstract, etc.) are represented at this layer. The *Morpho-Syntax* layer represents the results of the morpho-syntactic analysis such as tokenisation, part-of-speech tagging. Furthermore, the representations for abbreviations, acronyms and their expanded forms are designed in appropriate types. The results of lemmatisation, stemming and decomposition of words can be represented at this layer, as well. The annotations from shallow and full parsing can be represented at the *Syntax* layer. The appropriate types permit the representation of dependency- and constituency-based parsing results.

The *Semantics* layer comprises currently the representation of named entities, particularly for the bio-medical domain, and will soon be extended with the representation of relationships between entities and events. The entity types are hierarchically organised, with the super type *Entity*, which links annotated (named) entities to the ontologies and databases through appropriate features. The subtypes are currently being developed in the bio-medical domain and cover, e.g., genes, proteins, organisms, diseases, etc. This hierarchy can easily be adapted to entities from other domains.

The design of our type system allows the annotation of the entire cycle of NLP analysis, parallelism in the annotation (e.g., tagging may proceed with different part-of-speech tagsets), semantic annotation control through the restriction of type values (e.g., using the particular Penn Tagset), and the connection to the external resources such as ontologies, lexica and databases. The type system is designed as domain-independent as possible and is easily extensible. We provide domain- and application-independent core specifications which are then complemented by task- and application-specific extensions of the type system — for the bio-medical domain given the context of our projects.

We strive for the elaboration of a common standard UIMA type system. The advantages of such a standard include an easy exchange and integration of different NLP analysis engines, the facilitation of sophisticated evaluation studies (where, e.g., alternative components for NLP tasks can be plugged in and out at the spec level), and the reusability of single NLP tools developed in various academic research and industry labs.

Acknowledgments. This research was funded by the EC´s 6th Framework Programme (4th call) within 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).