An UIMA-based Tool Suite for Semantic Text Processing

Katrin Tomanek, Ekaterina Buyko, Udo Hahn

Jena University Language & Information Engineering Lab
StemNet - Knowledge Management for Immunology

- in life-sciences: increasing amount of knowledge stored in (unstructured) textual documents
- semantic access to this knowledge necessary
- biomedical subdomain: hematopoetic *stem cell transplantation*
  - semantic search engine for advanced document and information retrieval

- example user query:
  “get me relevant documents on human IL2Ra and CTL”
user query: “human IL2Ra” AND “CTL”

[...] on IL-2Ra-activated CD34(+) cytotoxic T-cells (CTLs). p3hr-1, the Burkit's lymphoma cell line, was [...]
user query: “human IL2Ra” AND “CTL”

[...] on **IL-2Ra**-activated CD34(+) 

**BLC-stimulated cytotoxic T-cells** showed 

[...] a more mature phenotype (low CD69, 

**CD25**, and CD62L) [...]
user query: “human IL2Ra” AND “CTL”

[...] on **IL-2Ra**-activated CD34(+) cytotoxic T-cells showed [...]. BLC-stimulated **cytotoxic T-cells** showed [...]. TNF-alpha upregulated the **interleukin 2 receptor alpha chain (Tac antigen)** on the surface of [...]. proliferation of tumor specific **CTL** [...].
UIMA in the StemNet Project

query: human IL2Ra AND CTL

NLP core system
1) comprehensive UIMA type system
   - covers the full NLP pipeline
   - five layers:
     • document meta information (bibliographic and content information)
     • document structure and style information (sentences, rhetorical zones, ...)
     • morpho-syntax (tokenisation, POS, acronyms, lemmatisation, ...)
     • syntax (shallow and full parsing information)
     • semantics (named entities, relationships, events...)
2) collection of NLP components (Analysis Engines):
   - for morpho-syntactic analysis
   - for syntactic analysis
   - for named entity recognition and normalisation/mapping

3) data import and export (Collection Reader/CAS Consumer):
   - PubMed Reader
   - Search Engine Indexer

• included tools:
   - mostly based on machine learning
   - external tools for which we have written UIMA wrappers
   - JULIE tools; have stand-alone and UIMA mode
PubMed Reader

• processes PubMed articles (XML)
• reads the following document meta-data:
  - bibliographic information: title, authors, publication date, journal name
  - content information (manually added): keywords (MeSH), list of chemicals
• writes data to CAS
  → our type system contains respective types for this kind of information
Sentence/Token Splitting, POS Tagging, Chunking

- configurable UIMA wrappers for OpenNLP tools
  - sentence splitter
  - tokeniser
  - POS tagger
  - chunker

- JULIE tools
  - sentence splitter
  - tokeniser

- available models for life-sciences:
  - trained on JULIE corpus (covers special cases and subtleties of biomedical domain)
  - trained on well-known biomedical corpora (e.g. PennBioIE)
Parsing

- UIMA wrappers for external parser implementations:
  - OpenNLP Parser (Ratnaparkhi, 1998)
    - constituency parser
  - MST Parser (McDonald, 2006)
    - dependency parser

- different linguistic paradigms supported
  - type system supports both constituency and dependency parse information
Acronym Detection

• detection and resolution of local acronyms
• implementation of M. Hearst's algorithm (Hearst 2003)
• with extension: DB lookup for unresolved acronyms
• Acronym DB generator (CAS Consumer):
  - tuples (acronym, full form), associated with spelling variants, first year of occurrence, keywords (MeSH)

[...] on IL-2Ra-activated CD34(+) cytotoxic T-cells (CTLs). p3hr-1, the Burkit's lymphoma cell line, was [...]
Named Entity Recognition

• generic named entity recognizer
• ML-based
• flexibly configurable wrt:
  - mapping: predicted labels $\rightarrow$ UIMA types
  - feature parametrization
    • user defined feature set (turn on/off, configure features)
    • CAS-specified feature information (e.g. POS tags)
• consistency preservation:
  - assures that same entity mentions within one abstract (document zone) are consistently annotated
Named Entity Mapping (1/2)

- associates identified NEs with DB entries
- in life-sciences: e.g. SwissProt

 [...] on **IL2Ra-activated CD34(+) cytotoxic T-cells (CTLs)**. p3hr-1, the Burkit's lymphoma cell line, was [...]

Named Entity Mapping (1/2)

- associates identified NEs with DB entries
- in life-sciences: e.g. SwissProt

[...] on **IL2Ra**-activated cytotoxic T-cells (CTLs) in the Burkit's lymphoma cell line, the p3hr-1, the Burkit's lymphoma cell line, was...
Named Entity Mapping (2/2)

- for gene/protein entity mentions
- principles:
  - normalization rules for bio-medical entities
    - a -> alpha
    - R -> receptor, L -> ligand
    - numbers split away
    - word order ignored
    - “IL2RA” -> “IL 2 receptor alpha”
    - “receptor of IL-4” -> “IL 4 receptor”
  - requires well-curated synonym list
JULIE Lucene Indexer

- goal: directly build search engine index from processed documents
- Lucene
  - high-performance search engine
  - fielded search and special query types (e.g. range searches)
  - open source, freely available, provides Java API
- Lucene Indexer
  - directly consumes CAS
  - tokenization as in CAS
  - currently indexed fields:
    - document meta-data (as in PubMed)
    - entity mentions + synonyms (with same offset)
- work in progress: flexible configurability
  - external mapping file (UIMA type -> Lucene field)
for further information/download of tools:

http://www.julielab.de