UTEMPL - The UIMA based Text Mining Pipeline

Jannik Strötgen, Verena Meyer, Roman Klinger and Juliane Fluck

Fraunhofer Institute for Algorithms and Scientific Computing
Schloss Birlinghoven, 53754, Sankt Augustin, Germany

Fraunhofer Institute Algorithms and Scientific Computing

Overview

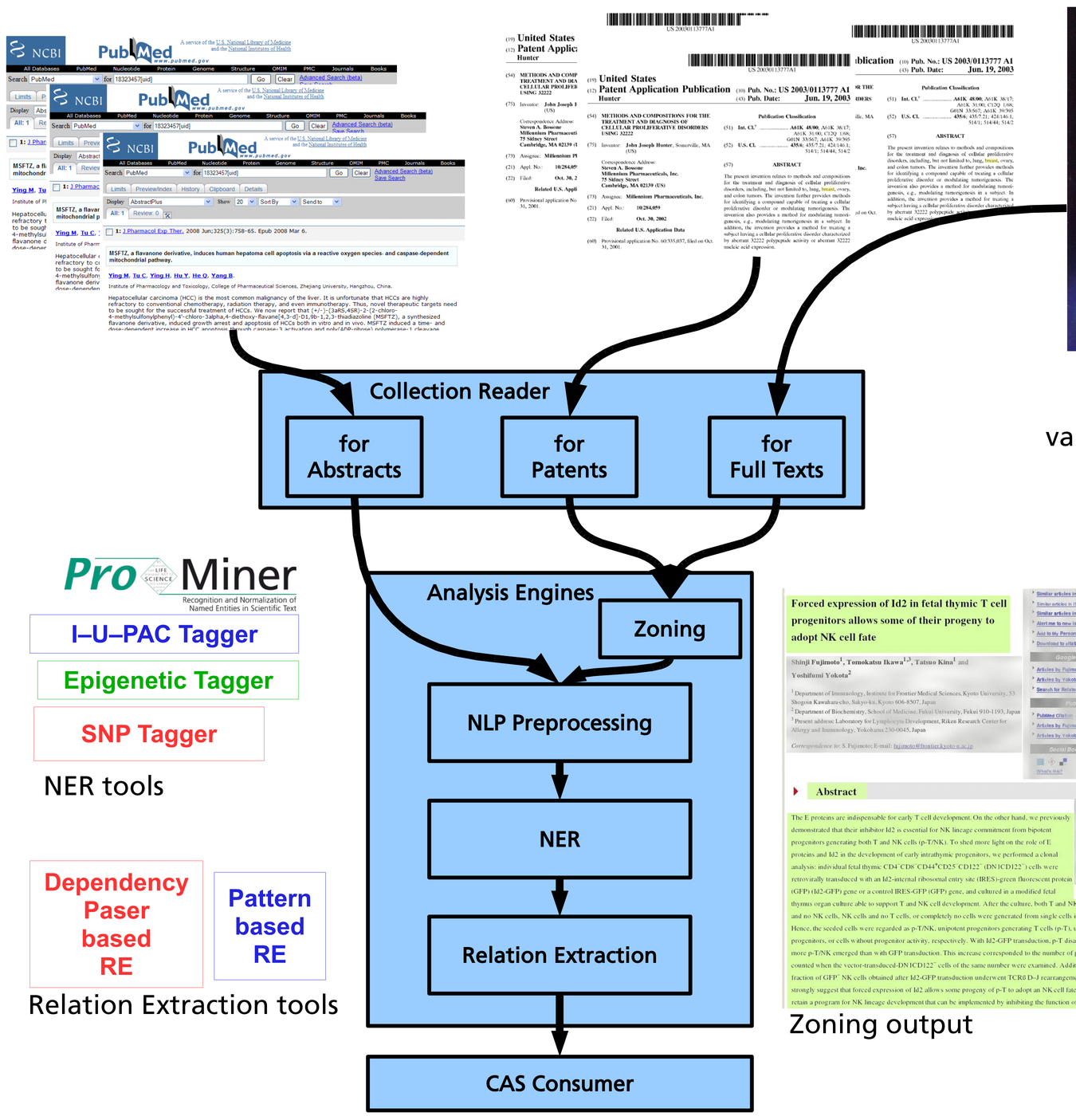
UTEMPL is a flexible tool for information extraction and text mining:

- based on UIMA
- easy use of several input formats:
  - Medline abstracts
  - Journal articles (HTML or XML)
  - Patents
  - ...
- standardized interfaces: integration of
  - NLP tools from different developers
  - new analysis engines for new tasks
- high exchangeablity
- combination and selection of various internal and external applications

Advantages

- easily adaptable to new tasks
- allows to compare and evaluate different tools or methods
- efficient adaptation of existing workflows

UTEMPL workflow



various textual resources

ProMiner
Recognition and Normalization of Named Entities in Scientific Text

I–U–PAC Tagger

Epigenetic Tagger

SNP Tagger

NER tools

Dependency Paser based RE

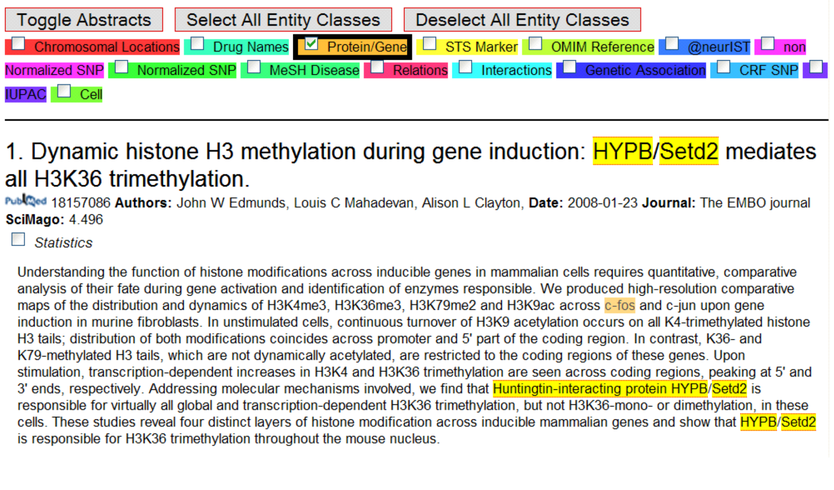
Pattern based RE

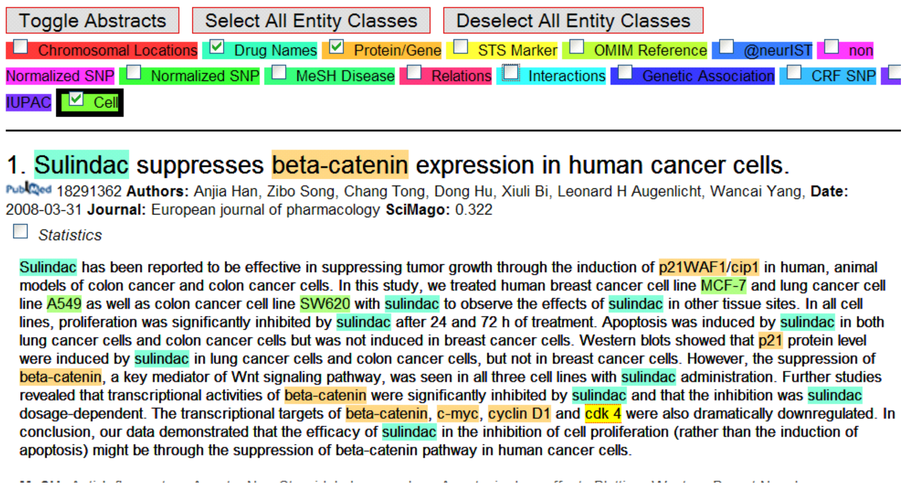
Relation Extraction tools

Zoning output

Processing tools

- Text Zoning
- Sentence Splitter
- several Tokenization Tools
- PoS Tagger
- various CRF based Named Entity Recognizer
- ProMiner containing various dictionaries and regular expression
- Pattern based relation extraction
- Dependency Parser based relation extraction





References

[1] D. Hanisch, K. Fundel, H. T. Mevissen, and R. Zimmer J. Fluck. ProMiner: Organism-specific protein name detection using approximate string matching. *Proceedings of the BioCreative Challenge Evaluation Workshop 2004*, 2004.

[2] R. Klinger, C. Kolárik, J. Fluck, M. Hofmann-Apitius, and C. M. Friedrich. Detection of IUPAC and IUPAC-like names. *BMC Bioinformatics*, 24(13):i268–i276, 2008.

[3] UIMA Homepage. http://incubator.apache.org/uima/. 2008.

Contact

Juliane Fluck
Fraunhofer SCAI
Schloss Birlinghoven
53754, Sankt Augustin,
juliane.fluck@scai.fraunhofer.de
http://www.scai.fraunhofer.de