Enterprise Platform and Integration Concepts &
Information Systems
Master Project Winter Term 2015/2016

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Learning to Note:
Intelligent Support for Document Annotation using Semi-Supervised Learning

The goal of this master project is to develop a system to support manual annotation of documents and linking of entities to database records. Manual annotation of textual documents is often necessary for building corpora to support training and evaluation of natural language processing applications. For instance, corpora have been developed for the extraction of a variety of entities, e.g., genes/proteins, as well as relationships, e.g., protein-protein interactions. Although there are many tools for document annotation [2], they do not suggest pre-annotations based on text mining and machine learning and do not provide real-time learning.

Curation tools support extracting data from text collections for a certain topic [1]. For instance, biological databases need to extract precise information from publications, which are further stored into their databases and made available to the users via a Web interface. This is a time-consuming and complex task which requires careful reading of many publications.

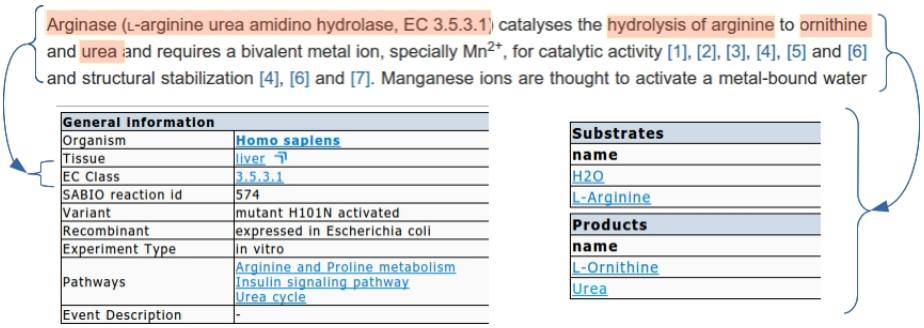

Figure 1: Annotation of chemical names in a document and the corresponding curated information in the SABIO-RK database.

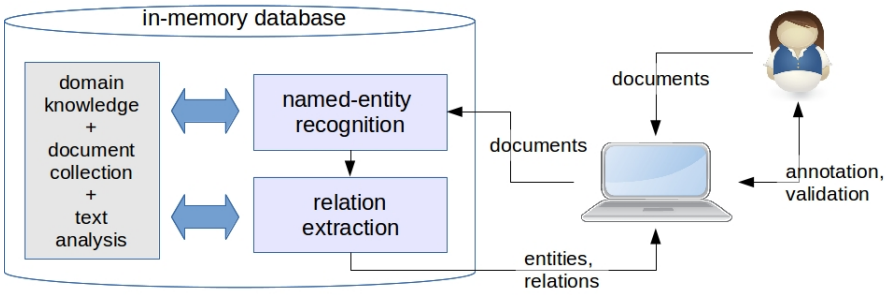

Figure 2: Architecture of an annotation and curation tool based on IMDB.

For performance purposes, the tool will be built on top of the SAP HANA in-memory database, given its potential for processing large datasets in real-time and its built-in text analysis functionalities. Interaction of the users with the system will be carried out by uploading a document or a collection of documents. The system will include a text mining pipeline for automatic processing of documents and suggestion of annotations. This pipeline will contain the following components: recognition of pre-defined entity types and extraction of pre-defined relationships between two or more entity types.

Further, ongoing annotations will be used for active learning of user preferences, for updating predictions of annotations and indicating which document to annotate next. This learning process will rely on existing machine learning algorithms implemented in the SAP HANA database, which will need to be adapted for on-line learning. Implementation of state-of-the-art on-line learning algorithms will also be considered.

Project goals

- Develop a Web application for annotation of documents and validation of data derived from text mining/machine learning
- Build a text mining pipeline for integration of named-entity recognition and relationship extraction tasks
- Evaluate the tool on benchmarks and for curation of real data
- Submit a paper describing the system and/or the methods

Technology and skills

Participants should have knowledge of SQL, of at least one programming language (preferably C++, Python or Java) and of Web development, as well as interest in database technologies, machine learning and natural language processing.

Group structure and project start

The team will consist of 3-5 students and the project will start on October 12th, 2015. There will be an initial meeting with all participants in the previous week for presentation of the details of the project and to get familiar with supervisors and colleagues.

Contact

Your are welcome to contact or visit us in room V0.01 (Villa) or room 2-02.1 (Haus E), HPI Campus II:

Dr. Mariana Neves
`mariana.neves`

Dr. Ralf Krestel
`ralf.krestel`

Prof. Felix Naumann
`felix.naumann`

Dr. Matthias Uflacker
`matthias.uflacker`

References

[1] Hirschman, L., Burns, G.A.P.C., Krallinger, M., Arighi, C., Cohen, K.B., Valencia, A., Wu, C.H., Chatr-Aryamontri, A., Dowell, K.G., Huala, E., Loureno, A., Nash, R., Veuthey, A.L., Wiegers, T., Winter, A.G.: Text mining for the biocuration workflow. Database 2012 (2012)

[2] Neves, M., Leser, U.: A survey on annotation tools for the biomedical literature. Briefings in Bioinformatics (2012)