UIMA in the Biocuration Workflow
A coherent framework for cooperation between biologists and computational linguists

Bart Mellebeek[†]*, Carlos Rodriguez-Penagos*[†] *and Laura Furlong*[‡]
Barcelona Media Innovation Centre[†]
Research Unit on Biomedical Informatics, Universitat Pompeu Fabra[‡]
{bart.mellebeek|carlos.rodriguez}@barcelonamedia.org, lfurlong@imim.es

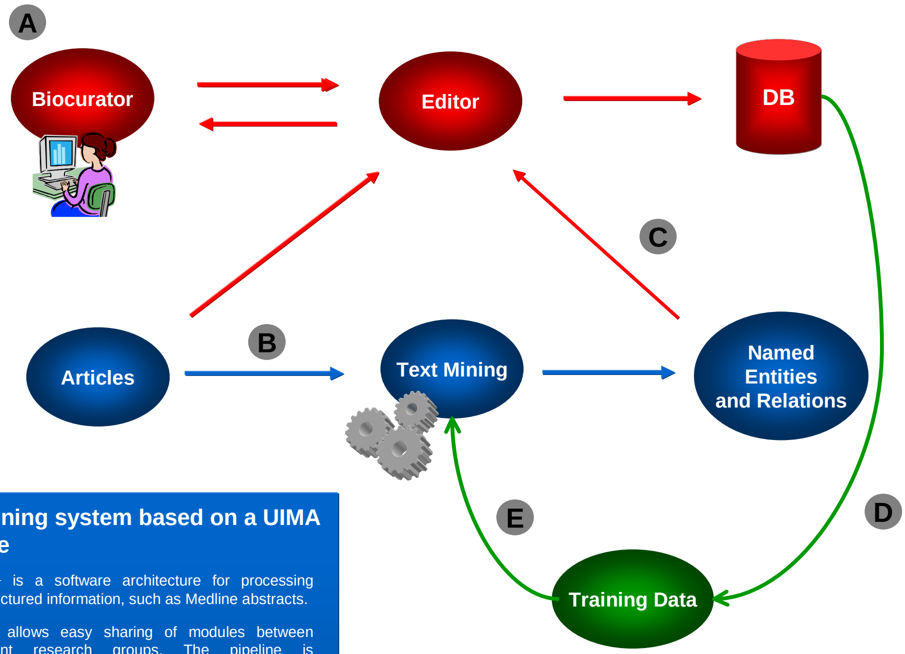
Nature Precedings : doi:10.1038/npre.2009.3171.1 : Posted 24 Apr 2009

Text Mining is useful for curation of DBs, but DBs can also be useful for Text Mining.

Goal:

To foster collaboration between Text Mining (TM) experts and biologists for the development of useful tools for database curation.

Biocuration Workflow



Text Mining system based on a UIMA Pipeline

- UIMA[1] is a software architecture for processing unstructured information, such as Medline abstracts.
- UIMA allows easy sharing of modules between different research groups. The pipeline is implemented with modules from JULIELab[2] (Jena University), from the Center for Computational Pharmacology[3] (University of Colorado), and our own modules.
- The JULIELab UIMA Type System (ontology containing concepts that TM can recognize) was extended for the identification of drugs, Single Nucleotide Polymorphisms (SNPs) and diseases.
- The pipeline consists of the following components: Medline Collection Reader, Sentence Splitter, Tokeniser, POS tagger, Linguistic Features Extractor, NER modules for the detection of diseases, genes and SNPs, and a Relation Extraction module[4].

[1] http://incubator.apache.org/UIMA
[2] http://www.julielab.de
[3] http://compbio.uchsc.edu
[4] Relation Extraction module: work in progress

Text Mining-Biocuration interaction cycle

A. Curators identify, extract and curate information from articles to populate DBs.

B. TM system identifies and extracts information from articles (see TM system based on a UIMA pipeline).

C. Information extracted by TM is provided to the curator, who decides whether it will be included in the DB.

D. Annotations from DBs can be used to create training data for automatic learning algorithms that are part of TM engines (see Boostrapping).

E. Data processing by a UIMA NLP pipeline to extract features (semantic and syntactic) required to (re)train Machine Learning based TM systems.

Processes A and B can be linked to help the biocuration process (C) and to improve quality of TM systems (D and E). Each iteration of the cycle will lead to an improvement in the quality of information provided by TM systems.

We are currently working on processes D (Bootstrapping), B (TM) and E (Data processing).

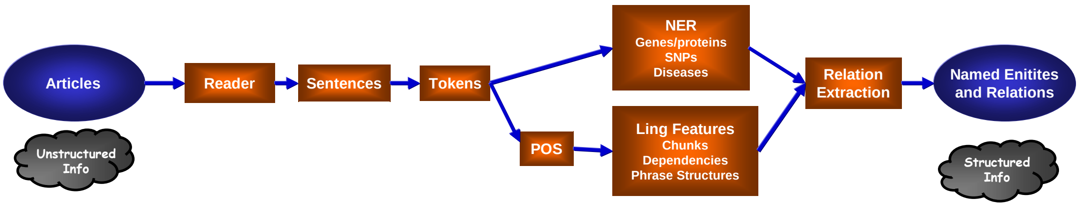
Bootstrapping

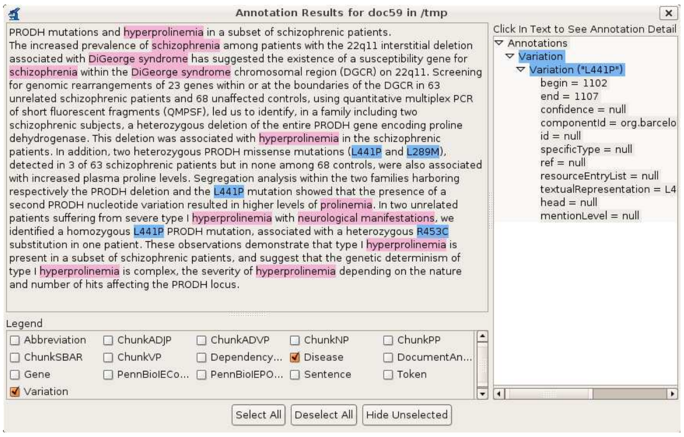
Goal: to make use of expert curated information from DBs to build a corpus of annotated documents, that in turn will be used in training/testing of TM systems.

Requirement: DB annotations have to be supported by literature (e.g. Medline articles)

1. Retrieval of annotations from DBs and supporting publication. For example, it is possible to use UniProt to extract natural variants (SNPs) associated to diseases and the supporting literature reference.
2. Automatic identification of annotated Named Entities (SNPs, diseases) in the text from the supporting publication, and extraction of sentences that express the association.
3. Further processing of the text required for finding useful features for Machine Learning algorithms.

Text Mining system based on a UIMA Pipeline




Annotation Results for doc59 in /tmp

PRODH mutations and hyperprolinemia in a subset of schizophrenic patients.
The increased prevalence of schizophrenia among patients with the 22q11 interstitial deletion associated with DiGeorge syndrome has suggested the existence of a susceptibility gene for schizophrenia within the DiGeorge syndrome chromosomal region (DGCR) on 22q11. Screening for genomic rearrangements of 23 genes within or at the boundaries of the DGCR in 63 unrelated schizophrenic patients and 68 unaffected controls, using quantitative multiplex PCR of short fluorescent fragments (QMPSF), led us to identify, in a family including two schizophrenic subjects, a heterozygous deletion of the entire PRODH gene encoding proline dehydrogenase. This deletion was associated with hyperprolinemia in the schizophrenic patients. In addition, two heterozygous PRODH missense mutations (L441P and L289M), detected in 3 of 63 schizophrenic patients but in none among 68 controls, were also associated with increased plasma proline levels. Segregation analysis within the two families harboring respectively the PRODH deletion and the L441P mutation showed that the presence of a second PRODH nucleotide variation resulted in higher levels of prolinemia. In two unrelated patients suffering from severe type I hyperprolinemia with neurological manifestations, we identified a homozygous L441P PRODH mutation, associated with a heterozygous R453C substitution in one patient. These observations demonstrate that type I hyperprolinemia is present in a subset of schizophrenic patients, and suggest that the genetic determinism of type I hyperprolinemia is complex, the severity of hyperprolinemia depending on the nature and number of hits affecting the PRODH locus.

Legend

☐ Abbreviation ☐ ChunkADJP ☐ ChunkADVP ☐ ChunkNP ☐ ChunkPP
☐ ChunkSBAR ☐ ChunkVP ☐ Dependency... ☑ Disease ☐ DocumentAn...
☐ Gene ☐ PennBioIECo... ☐ PennBioIEPO... ☐ Sentence ☐ Token
☑ Variation

Click In Text to See Annotation Detail

```
▽ Annotations
  ▽ Variation
    ▽ Variation ("L441P")
        begin = 1102
        end = 1107
        confidence = null
        componentId = org.barcelo
        id = null
        specificType = null
        ref = null
        resourceEntryList = null
        textualRepresentation = L4
        head = null
        mentionLevel = null
```

Example: SNP-disease annotations

- UIMA Annotation Viewer with occurrences of SNPs (light blue) and diseases (pink) marked up.
- Additional information (genes, linguistic features, etc) are also available in the UIMA CAS.
- All annotations have been automatically extracted using the pipeline described above.
- The annotations (linguistic, semantic) can be used as training features for the Machine Learning algorithms that are part of the Text Mining system.

Acknowledgements:This work was generated in the framework of the @neurIST and the EU-ADR projects co-financed by the European Commission through the contracts no. IST-027703 and ICT-215847, respectively. The Research Unit on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB) and a member of the COMBIOMED network.

Innovation Centre
Barcelona Media
GRIB
UNIVERSITAT POMPEU FABRA
IMIM hospitaldelmar