A large-scale processing pipeline for information extraction from biomedical articles with an application to spinal cord injury treatment

In cooperation with center of neuronal regeneration (CNR) Düsseldorf

Intelligent Systems Laboratory — Winter Term 2013/2014

Raphael Dickfelder & Benjamin Paaßen & Andreas Stöckel

Supervisors: Philipp Cimiano & Matthias Hartung & Roman Klinger

Bielefeld University, Faculty of Technology

Universität Bielefeld

Abstract

Currently there are no treatments available for spinal cord injuries in humans but a large corpus of research about the effectiveness of different treatments on animals. However the sheer number of papers on the topic makes it increasingly difficult to judge which approaches might be promising to transfer to human medicine. We introduce a proof-of-concept implementation of a pipeline that extracts the relevant, semantic information from given research papers. Thereby we do first steps to a convenient access, analysis and visualization of the data available on spinal cord injury treatments in animals for medical researchers.

Information Extraction from Biomedical Literature

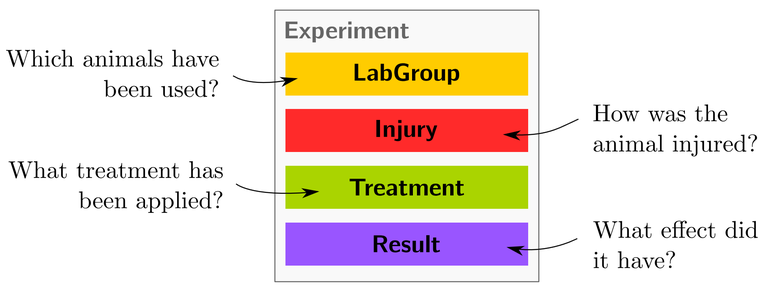

Figure 1: Relevant information we extract from the papers

Methods and Pipeline Structure

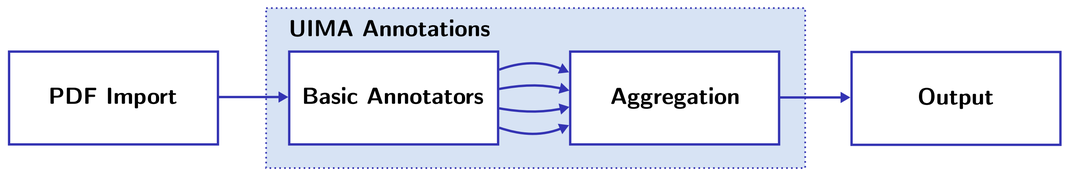

Figure 2: An illustration of the whole pipeline.

PDF-Import

- Individualized version of Apache PDFBox[5]
- Structured text as output (pages, blocks, paragraphs, strong, emphasized, etc.)

Annotations

- Based on Apache UIMA[6]
- Multiple layers (figure 3)

Basic Annotations

- Sentences and Words (JULIE Lab[7])
- Quantities (raw numbers, weights, etc.)
- Matches for node labels in our pre-defined ontologies (see below)

Aggregation

- Probability based model (see *Aggregation*)

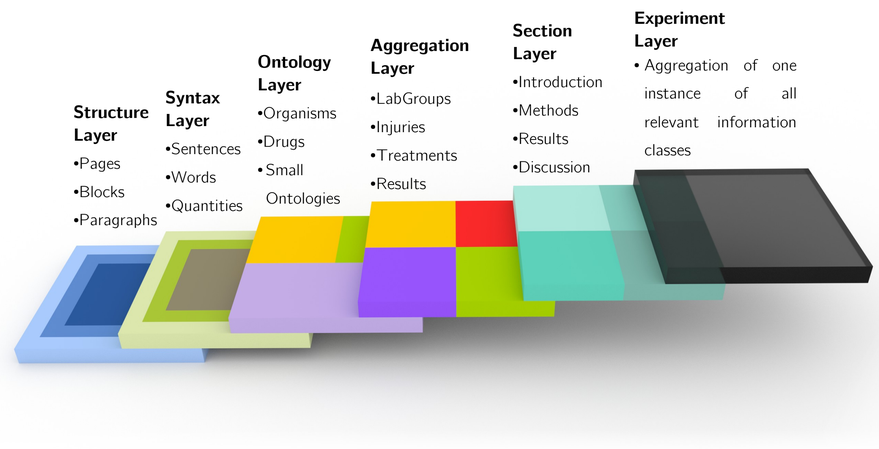

Figure 3: The different layers of the pipeline annotation process.

Ontology Annotations

Our Ontology Database supports fuzzy or strict matching for words in an ontology. Ontologies are stored as a graph structure in a relational database (with PostgreSQL via JDBC) (figure 4).

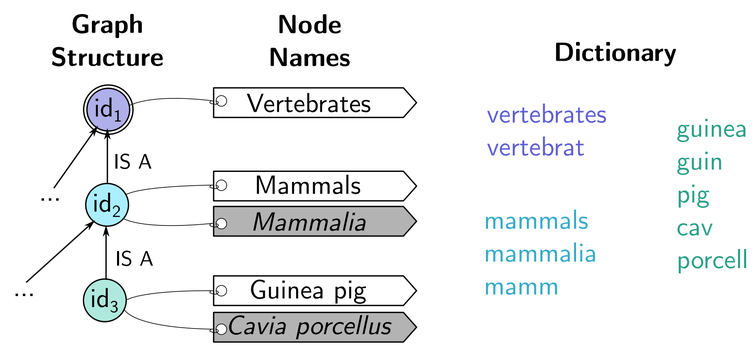

Figure 4: Schematic overview of the ontology database

Aggregation

Forementioned annotations are aggregated to instances of the four relevant information classes (see *Main Objectives*). The probability of an aggregation $P_{Aggr}(a, b)$ of two annotations a and b is calculated using a custom **semantic-syntactic probabilistic aggregation model**. We define:

$$P_{Aggr}(a, b) := P_{Syn}(a, b) \cdot P_{Sem}(a, b)$$

$$P_{Syn}(a, b) := e^{-d(a,b)/2} \quad \text{where } d \text{ is the syntactic distance in the text}$$

$$P_{Sem}(a, b) \quad \text{semantic domain knowledge}$$

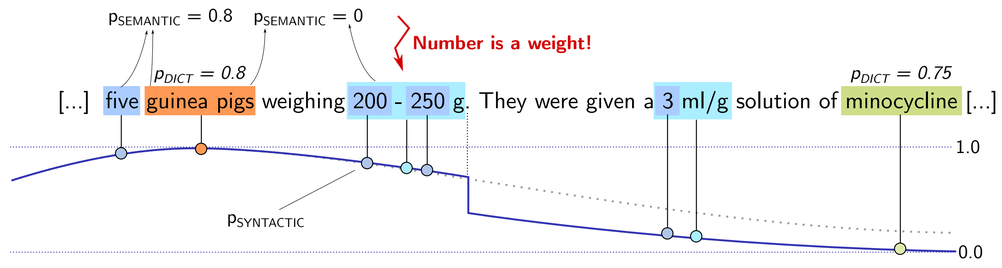

Figure 5: Aggregation example for a laboratory animal group: The semantic probability model prevents a number already in the *weight* slot from being used to specify the number of animals in the laboratory animal group.

Results

Materials and methods

Adult, Long–Evans female rats (Simonsen Laboratories, Gilroy, CA, USA, n = 56) were used for this study. All procedures were conducted in compliance with [...]

[...]

Surgical procedures

Spinal cord contusion injuries. After anesthesia induction with sodium pentobarbital (Nembutal, intraperitoneally 0.1 ml/ 100 g body weight), a laminectomy was made at the T9–T10 vertebral level, exposing dura, followed by a 25-mm contusion injury using the NYU IMPACTOR device and MASCIS protocols.
Rats were given Cefazolin (0.02 cc subcutaneously) twice daily for the first [...] Conta et al., 2008

Figure 6: Excerpt from a test run on an actual paper

Discussion and Outlook

In a rough, qualitative examination injury- and laboratory animal group-annotations were found to be acceptable, while treatment- and result-annotations remain problematic.

We will assess the abilities of our system in more detail in a quantitative study and improve weak points by making use of machine learning techniques in the second semester. We hope to improve the treatment annotation quality by using MeSH [3] instead of Drugbank, which also provides information about the application field of each drug. This additional domain knowledge allows for improvements in assessing the semantic aggregation probability. Further improvements are planned to the import module and the Ontology Database.

References

[1] Fred J. Damerau. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176, March 1964.

[2] Drugbank. `http://www.drugbank.ca/`. [Online; accessed 2013 to 2014].

[3] MeSH. `http://www.nlm.nih.gov/mesh/`. [Online; accessed 2014].

[4] National Center for Biotechnology Information. `http://www.ncbi.nlm.nih.gov/`. [Online; accessed 2013 to 2014].

[5] Apache PDFBox - A Java PDF Library. `http://pdfbox.apache.org/`. [Online; accessed 2013 to 2014].

[6] Apache UIMA. `http://uima.apache.org/`. [Online; accessed 2013 to 2014].

[7] Julie Lab. `http://www.julielab.de/Resources/NLP+Tools/Download/UIMA+Collection+Reader-p-96.html`. [Online; accessed 2013 to 2014].

Acknowledgements

Special thanks to our supervisors for their support during the project and to the CNR team for their valuable insight into the domain as well as providing test data.