Methodology of Textual Preprocessing for Information Extraction in Scientific Papers of the Biomedical Domain

Pablo Freire Matos[1], Ricardo Rodrigues Ciferri[1], Thiago Alexandre Salgueiro Pardo[2]

[1]Department of Computer Science – Federal University of São Carlos
[2] Department of Computer Science – University of São Paulo at São Carlos
{pablo_matos, ricardo}@dc.ufscar.br, taspardo@icmc.usp.br

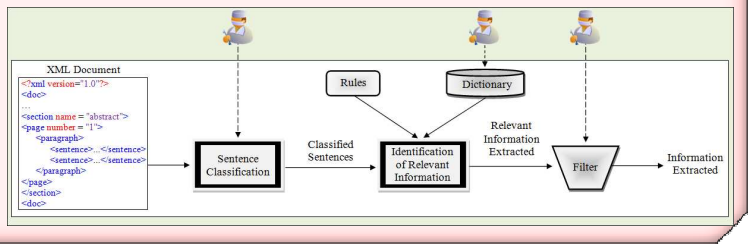
1 Context and Motivation

In the biomedical domain there are a lot of electronic documents that report experiments related to patients who were treated with medicines that can originate a positive effect or a negative effect (from the disease or from the use of a given treatment).

Nowadays, researchers and doctors are not able to process this huge number of documents to extract key information related to some issues of interest. These documents are in unstructured format, that is, in plain textual form.

Firstly, it is necessary to transform this documents from unstructured to structured format in order to submit it to an automatic knowledge discovery process. For that, a methodology of textual preprocessing is proposed which consists of two phases: **Sentence Classification** and **Identification of Relevant Information**. The most predominant approaches for knowledge extraction in the biomedical domain are used in the methodology [1]: machine learning, rule-based and dictionary-based.

Methodology of Information Extraction



2 Sentence Classification

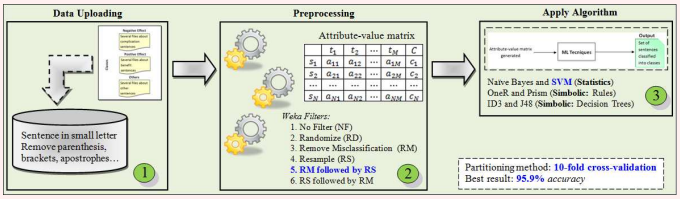
The Sentence Classification process (i.e., a supervised machine learning) is composed of three steps :

1. The "*data uploading*" which consists of manually selecting sentences from papers about negative effect, positive effect and other related to SCA [2];
2. The "*preprocessing*" that consists of constructing the attribute-value matrix (**AVM**) using the ***bag-of-words* model**. Weka filters are used to improve the entry examples;
3. The "*apply algorithm*" which uses a machine learning algorithm to classify the sentence in the respective class.

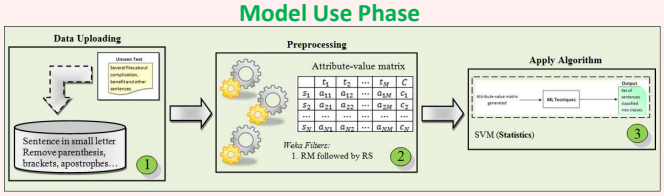
AVM configuration:

- Minimum Frequency = 2;
- Attributes: 1 to 3-grams (1 = present and 0 = absent);
- Not considered: stopwords removal and stemming.

Training and Testing Phase



Model Use Phase



In [3], sentence classification experiments is explained in details.

3 Identification of Relevant Information

After the sentence classification, it is necessary to identify the relevant information in each sentence. The sentence is tagged by a Part-Of-Speech (POS) tagger [4], e.g.: "***Six_CD patients_NNS with_IN persistently_RB abnormal_JJ TCD_NNP results_NNS developed_VBD stroke_NN ._.***"

Example of Sentences

Anoxic brain injury occurred in three patients, central nervous system hemorrhage in three, and infarction in three.

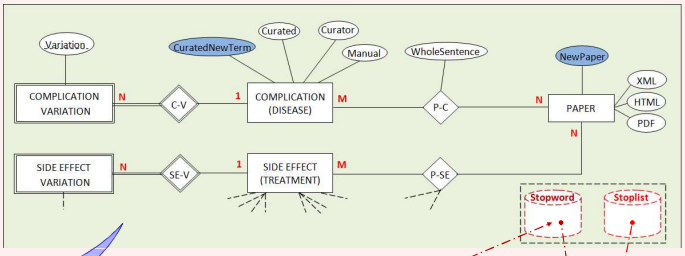
Hypersplenism, which was diagnosed in six patients, is a classical complication of SCD, but recurrent splenic sequestration episodes are unusual in children aged 6 and 7 years. Hb SC children excepted, and suggest hydroxyurea-induced hypersplenism.

During 426 patient-years of follow-up for patients with standard criteria, 3.3 acute chest syndromes, 1.3 cerebrovascular events, and 1.1 osteonecrosis per 100 patient-years were observed.

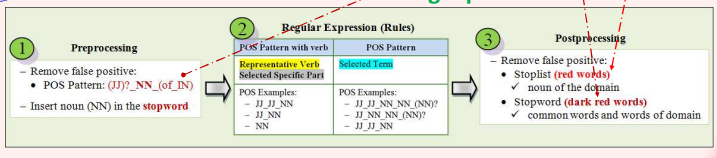
Among the specific causes were pulmonary fat embolism and 27 different infectious pathogens.

Relevant Information Representative Verb

Dictionary: Biomedical Database



Rules: Textual Processing Pipeline



4 Conclusions

We proposed a methodology of textual preprocessing for information extraction in scientific papers of the biomedical domain. Two fundamentals steps are performed to achieve the information extraction goal: sentence classification and identification of the relevant information.

Regarding the dictionary approach, there are two manual processes which remains as future work:

- Distinguish term (negative effect) from the treatment (side effect) and from the disease (complication);
- Hierarchize related terms, for instance, "*neurologic infection*" and just "*infection*", or "*acute severe anemia*" and just "*severe anemia*".

References

[1] ANANIADOU, S.; MCNAUGHT, J. (Eds.). **Text mining for biology and biomedicine.** Norwood, MA: Artech House, 2006.

[2] PINTO, A. C. S. et al. **Technical Report "Sickle Cell Anemia".** São Carlos: Department of Computer Science, Federal University of São Carlos, 2009. Available at: <http://sca.dc.ufscar.br/download/files/report.sca.pdf>.

[3] MATOS, P. F. et al. An environment for data analysis in biomedical domain: information extraction for decision support systems. In: GARCÍA-PEDRAJAS, N. (Eds.). **IEA-AIE**. 23th. Heidelberg: Springer, 2010. p. 306-316. (Lecture Notes in Computer Science; v. 6096).

[4] TOUTANOVA, K.; MANNING, C. D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: EMNLP/VLC, 38th, 2000, Hong Kong. **Proceedings…** ACL, 2000. p. 63-70.