Reaching a modular, domain-agnostic and containerized development in Biomedical Natural Language Processing systems.

Javier Corvi*, Josep Lluis Gelpí*†, Salvador Capella-Gutierrez*
*Barcelona Supercomputing Center, Barcelona, Spain
†University of Barcelona, Barcelona, Spain
E-mail: {javier.corvi, josep.gelpi, salvador.capella} @bsc.es

***Keywords*—Natural Language Processing (NLP), Text mining, Toxicology, Preclinical, Biomaterials.**

I. Extended Abstract

The last century saw an exponential increase in scientific publications in the biomedical domain. Despite the potential value of this knowledge; most of this data is only available as unstructured textual literature, which have limited their systematic access, use and exploitation. This limitation can be avoided, or at least mitigated, by relying on text mining techniques to automatically extract relevant data and structure it from textual documents. A significant challenge for scientific software applications, including Natural Language Processing (NLP) systems, consists in providing facilities to share, distribute and run such systems in a simple and convenient way. Software containers can host their own dependencies and auxiliary programs, isolating them from the execution environment. In addition, a workflow manager can be used for the automated orchestration and execution of the text mining pipelines.

Our work is focused in the study and design of new techniques and approaches to construct, develop, validate and deploy NLP components and workflows with sufficient genericity, scalability and interoperability allowing their use and instantiation across different domains. The results and techniques acquired will be applied in two main uses cases: the detection of relevant information from preclinical toxicological reports, under the eTRANSAFE project [1]; and the indexation of biomaterials publications with relevant concepts as part as the DEBBIE project.

A. PretoxTM

The treatment-related findings detected on the test subjects after the administration of the compound are the kind of relevant information included in the toxicology reports that is valuable for the drug development process. Examples of sentences with treatment-related findings are: *“The decrease in food consumption and body weight of the animals from the mid dose onwards is regarded as evidence of general toxicity.”* and *”At dose level 3, absolute and relative liver weights were increased in male rats.”*.

PretoxTM is a preclinical text mining system that extracts treatment-related findings from toxicological studies and present this information for toxicology experts validation. PretoxTM is being developed as part of eTRANSAFE, a research project funded within the Innovative Medicines Initiative (IMI), which aims at developing integrated databases and computational tools that support the translational safety assessment of new drugs.

In order to develop, train and validate text mining models for the detection of treatment-related findings, a gold standard corpus of preclinical studies annotated by toxicology experts was developed. The formal annotation guideline and the training material are available at the preclinical corpus development materials.

Figure 1.A shows the components of the PretoxTM pipeline. After the extraction of relevant paragraphs from the toxicology reports and the pre-processing steps; the PretoxTM sentence classifier is executed. The classifier was trained with the preclinical toxicology corpus and predicts if a sentence is relevant (contains information on toxicology findings) or not. The model is a maximum entropy classifier, was developed in JAVA and it uses the Stanford CoreNLP framework. The classifier obtains a performance of 0.91 F1-score. Following, the Named Entity Recognition (NER) and Relation Extraction (RE) components are executed to detect concepts and related them to form a toxicological finding. The NER is based on different controlled terminologies such as CDISC SEND (Standard for Exchange of Nonclinical Data). The last step of the pipeline is the mapping of the detected findings to the Study Report Domain (SR-Domain) format, an specific format that is being developed by another partner of the eTRANSAFE project.

The PretoxTM Web App (Figure 1.B) was developed to present the extracted information to the experts in order to provide a user-friendly interface to visualise and validate the findings.

B. DEBBIE

Biomaterials are natural or synthetic materials used for constructing artificial organs, fabricating prostheses, or replacing tissues. The DEBBIE (Database of Biomaterials and their Biological Effect) project aims at integrating biomaterials extracted metadata by storing in its database indexed articles with relevant concepts.

Figure 2 shows the architectural overview of the DEBBIE text mining pipeline. The first step in the pipeline entails the retrieval and standardization of published abstracts from PubMed. Following, is the recognition of relevant abstracts using the DEBBIE SVM model; which performs multiclass classification of abstracts to determine if they are relevant (either clinical or non-clinical studies) or not relevant to the field of biomaterials. The SVM DEBBIE model obtains

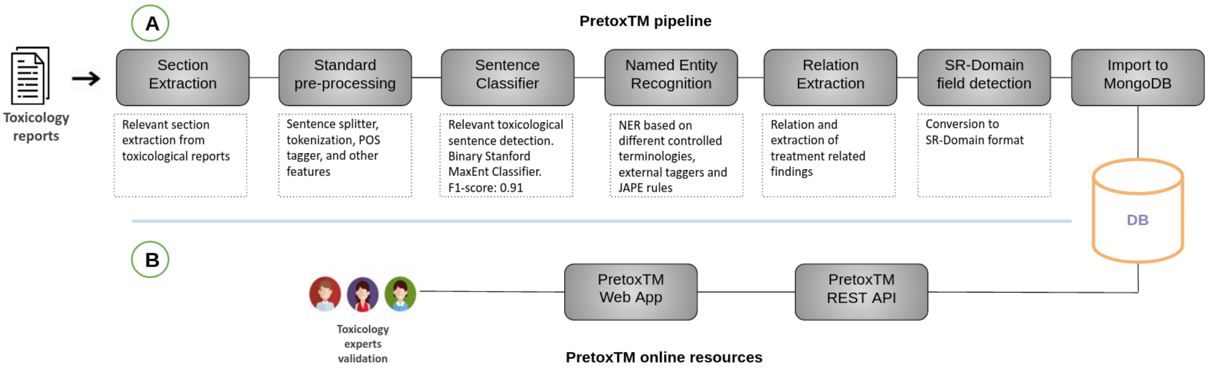
A
PretoxTM pipeline
Toxicology reports
Section Extraction
Standard pre-processing
Sentence Classifier
Named Entity Recognition
Relation Extraction
SR-Domain field detection
Import to MongoDB
Relevant section extraction from toxicological reports
Sentence splitter, tokenization, POS tagger, and other features
Relevant toxicological sentence detection. Binary Stanford MaxEnt Classifier. F1-score: 0.91
NER based on different controlled terminologies, external taggers and JAPE rules
Relation and extraction of treatment related findings
Conversion to SR-Domain format
DB
B
Toxicology experts validation
PretoxTM Web App
PretoxTM REST API
PretoxTM online resources

Fig. 1. A) PretoxTM pipeline components for the extraction for treatment-related findings. B) PretoxTM Web App components for expert validation.

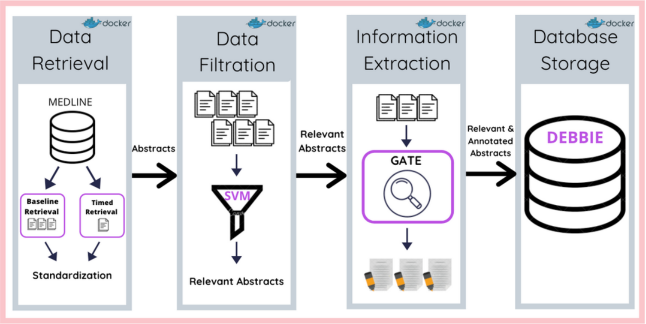

Fig. 2. DEBBIE automated text mining pipeline.

a performance of 0.93 F1-score. For the recognition of relevant biomaterials concepts, DEBBIE relies on the Biomaterials Annotator [2], a Named Entity Recognition (NER) system based on lexical resources specially tailored for the biomaterials literature [3]. The Biomaterials Annotator performs semantic mapping of the annotations by, not only recognizing the category of an entity, but also linking it to the appropriate entry in a well established resource. The Biomaterials Annotator shows a performance of 0.75-F1 score using a strict approach -exact matching of the term- and 0.79 F1-score in the partial-relaxed approach.

In order to access, visualize and retrieve the extracted information, a Web App (https://debbie.bsc.es/search/) was developed to allow the scientific community and general public to search the DEBBIE database using simple keyword queries, and receive results in the form of an intuitive dashboard.

C. Conclusion and Future Steps

Generic and/or reusable NLP components were developed, encapsulated into software containers and integrated in the text mining workflows of PretoxTM and DEBBIE systems. A first version of the PretoxTM system to extract and visualise treatment-related findings from toxicology reports is already functional. Different state-of-the-art NLP techniques will be explored to enhance the NER and RE components of the PretoxTM pipeline using the preclinical annotated corpus to train the models. As part of the DEBBIE project, the Biomaterials Annotator was published and different components were developed to build the DEBBIE text mining pipeline and the Web App.

II. Acknowledgment

The author would like to thank Osnat Hakimi, who was the leader of DEBBIE project, a Bosch-Aymerich tenure tracked fellow at the Universitat Internacional de Catalunya and a visitor to the BSC. J.C. is partly supported by eTRANSAFE (IMI2 Joint Undertaking, grant agreement No 777365 and support from the European Union's Horizon 2020 and EFPIA).

References

[1] F. Pognan, T. Steger-hartmann, C. Díaz, N. Blomberg, F. Bringezu, K. Briggs, G. Callegaro, S. Capella-gutierrez, E. Centeno, J. Corvi, P. Drew, W. C. Drewe, J. M. Fernández, L. I. Furlong, E. Guney, J. A. Kors, M. A. Mayer, M. Pastor, J. Piñero, J. M. Ramírez-anguita, F. Ronzano, P. Rowell, J. Saüch-pitarch, A. Valencia, B. van de Water, J. van der Lei, E. van Mulligen, and F. Sanz, "The etransafe project on translational safety assessment through integrative knowledge management: Achievements and perspectives," *Pharmaceuticals*, 2021.

[2] J. Corvi, C. Fuenteslópez, J. Fernández, J. Gelpi, M.-P. Ginebra, S. Capella-Guitierrez, and O. Hakimi, "The biomaterials annotator: a system for ontology-based concept annotation of biomaterials text," in *Proceedings of the Second Workshop on Scholarly Document Processing*. Online: Association for Computational Linguistics, Jun. 2021, pp. 36–48. [Online]. Available: https://aclanthology.org/2021.sdp-1.5

[3] O. Hakimi, J. L. Gelpi, M. Krallinger, F. Curi, D. Repchevsky, and M. Ginebra, "The Devices, Experimental Scaffolds, and Biomaterials Ontology (DEB): A Tool for Mapping, Annotation, and Analysis of Biomaterials Data," *Advanced Functional Materials*, vol. 30, no. 16, p. 1909910, apr 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.201909910

Javier Corvi received the BCS degree from Univertity of La Plata in 2011 and the MSc degree in Bioinformatics from University of Quilmes in 2020. Since 2018 he is a Bioinformatics researcher at the Spanish National Bioinformatics Institute (INB) Coordination Node, Life Sciences Department, Barcelona Supercomputing Center. He is currently doing a PhD in Biomedicine at the University of Barcelona focused on the development of modular, domain-agnostic and containerized Biomedical NLP systems.