Reproducibility and Generalization of a Relation Extraction System for Gene-Disease Associations

Laura Menotti[1]

[1]*Department of Information Engineering, University of Padua, Padua, Italy*

Best Master Thesis Award (Extended Abstract). Understanding the interactions between genes and diseases is a great resource for improving patient care as it could provide the foundation for curative therapies, beneficial treatments, and preventative measures. This type of data is available in databases, e.g. DisGeNET [1] and BioXpress [2], in the form of Gene-Disease Associations (GDAs), that contain relationships between gene expressions and specific diseases such as cancer. Biomedical literature is a rich source of information about GDAs, that are usually extracted manually from text. Human annotations are expensive and cannot scale to the huge amount of data available in scientific literature (e.g., biomedical abstracts). Therefore, developing automated tools to identify GDAs is getting traction in the community [3]. Such systems employ Relation Extraction (RE) techniques to extract information on gene/microRNA expression in diseases from text. Once an automated text-mining tool has been developed, it can be tested on human annotated data or it can be compared to state-of-the-art systems. Indeed, it is crucial for researchers to compare newly developed systems with the state-of-the-art to assess whether they made a breakthrough. However, previous works may not be immediately reproducible for example, due to the lack of source code. At the time of writing, the state of the art is DEXTER, a rule-based system that extracts gene/microRNA expressions in diseases from biomedical abstracts [4]. DEXTER was published in *Database: The Journal of Biological Databases and Curation* in 2018 and has attracted thirteen citations so far. DEXTER takes as input biomedical abstracts from PubMed and extracts relevant information such as the correlation between genes or miRNAs and diseases. The system also classifies sentences as TypeA or TypeB based on the number of entities found; such classification is useful to integrate data in existing resources like BioXpress. In particular, TypeA sentences are comparative phrases where gene expression is contrasted between two different samples or conditions while in TypeB sentences there is no explicit comparison. Unfortunately, DEXTER's source code is not publicly available hence researchers rely only on data provided by the authors to evaluate newly developed systems.

The objective of this work is to reproduce DEXTER to provide a benchmark for RE, enabling researchers to test and compare their results to a state-of-the-art baseline. DEXTER is based on several modules, each dealing with a different part of the computation independently. While we preserved the original block structure, we decided to develop the system as an end-to-end

ITADATA2023: The 2nd Italian Conference on Big Data and Data Science, September 11–13, 2023, Naples, Italy

laura.menotti@unipd.it (L. Menotti)
https://www.dei.unipd.it/~menottilau/ (L. Menotti)
0000-0002-0676-682X (L. Menotti)

CEUR Workshop Proceedings (CEUR-WS.org)

application to foster reusability. In this way, our implementation of DEXTER can be easily run on different datasets, without extensive knowledge of the system's internal architecture. We made some changes in each component to enable a seamless integration of the different modules. While the original system is developed both in Python and Java and mainly employs the Stanford CoreNLP toolkit, our system is entirely developed in Python, exploiting the SpaCy library for linguistic annotations and dependency trees. During implementation, we faced some reproducibility issues related to the use of the SpaCy library and the PubTator annotations. SpaCy employs a different dependency parser than the Stanford CoreNLP toolkit used in the original system. This results in having different dependency trees for some sentences thus we added some patterns to match more sentences. In addition, we had to translate the original rules in a different format to comply with the new library, and this resulted in having more rules with respect to the original system. Finally, we recall that to extract disease and gene mentions we matched pre-computed annotations. With this approach, we could have problems related to text normalization, special characters, and acronyms.

To assess the accuracy of our system, we used as ground truth three datasets provided by the authors in BioXpress[1]. In particular, we evaluated our system performance based on the percentage of parsed sentences and the correctness of the results in terms of gene expression level and sentence type. Our implementation parsed 97% of the input sentences. We performed an error analysis which confirmed that discarded sentences are mostly due to missing PubTator annotations or problems related to Dependency Parsing. The system achieved an accuracy of 84% on the gene expression level. Such value raises up to 93% if input mentions are used instead of PubTator annotations. This can be related to the previously defined issues of string matching or to the different version of PubTator that we are using with respect to the original system.

In conclusion, results demonstrate that the system has been reproduced to a reasonable degree as the benefits of having an end-to-end system completely written in Python outweigh the margin of error we introduced. We released our implementation of DEXTER in a GitHub repository[2] so that anyone working on this field can use it to compare their system with the state of the art. To this end, this work has been used to validate the Collaborative Oriented Relation Extraction (CORE) system [5, 6][3], a Knowledge Base Construction (KBC) system based on the combination of distant supervision and active learning paradigms.

Acknowledgments

This work was supported by the ExaMode Project, as a part of the European Union Horizon 2020 Program under grant 825292. The author wishes to thank Prof. Gianmaria Silvello and Dr. Stefano Marchesin whose suggestions helped in improving this study.

[1] Datasets "DEXTER Glycosyltransferase Expression", "DEXTER Expression in Lung Cancer", "DEXTER miRNA Expression" from Section "BioXPress Downloads" at https://hive.biochemistry.gwu.edu/bioxpress.

[2] https://github.com/mntlra/DEXTER

[3] The knowledge base derived by CORE can be accessed via https://gda.dei.unipd.it along with a demonstration video (https://gda.dei.unipd.it/static/videos/demo.mp4)

References

[1] J. P. González, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, L. I. Furlong, The DisGeNET knowledge platform for disease genomics: 2019 update, Nucleic Acids Res. 48 (2020) D845–D855. doi:10.1093/nar/gkz1021.

[2] H. Dingerdissen, J. Torcivia-Rodriguez, Y. Hu, T. C. Chang, R. Mazumder, R. Y. Kahsay, BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery, Nucleic Acids Res. 46 (2018) D1128–D1136. doi:10.1093/nar/gkx907.

[3] S. Marchesin, G. Silvello, TBGA: a large-scale gene-disease association dataset for biomedical relation extraction, BMC Bioinform. 23 (2022) 111. doi:10.1186/s12859-022-04646-6.

[4] S. Gupta, H. Dingerdissen, K. E. Ross, Y. Hu, C. H. Wu, R. Mazumder, K. Vijay-Shanker, DEXTER: Disease-Expression Relation Extraction from Text, Database 2018 (2018). doi:10.1093/database/bay045.

[5] S. Marchesin, L. Menotti, F. Giachelle, G. Silvello, O. Alonso, Building a Large Gene Expression-Cancer Knowledge Base with Limited Human Annotations, Database (2023). doi:10.1093/database/baad061, in print.

[6] F. Giachelle, S. Marchesin, G. Silvello, O. Alonso, Searching for reliable facts over a medical knowledge base, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’23, Association for Computing Machinery, New York, NY, USA, 2023, p. 3205–3209. doi:10.1145/3539618.3591822.