lasigeBioTM at BioCreative VII Track 1: Text mining drug and chemical-protein interactions using biomedical ontologies*

Diana Sousa, Rodrigo Cassanheira and Francisco M. Couto
LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

***Abstract*—Identifying biomedical relations is necessary to advance our understanding of biological processes and is particularly relevant for applications in precision medicine. This work describes the participation of the lasigeBioTM team in the BioCreative VII Track 1, whose primary goal is the extraction and classification of drug and chemical-protein interactions. Our team adapted an existing neural networks system, BiOnt, that incorporates external knowledge from biomedical ontologies. To perform Track 1, we used the Gene Ontology (GO) and the Chemical Entities of Biological Interest (ChEBI) ontology. We submitted different runs taking into account the use of features such as class weights and post-processing rules. However, due to time constraints, we could not make all the improvements that we planned initially, and our results were below the mean performance of the participating teams. Still, we took the first steps towards this adaption and we are now able to continue improving this system to reach state-of-the-art performance.**

Keywords—biomedical relation extraction; text mining; deep learning; external knowledge

I. Introduction

Identifying relations between biomedical entities is a fundamental block for advancing a wide range of domains within science and medicine. However, in the continuous exploration of different hypotheses, more text is produced that any researcher or clinician can keep up with. Automatically unravelling this information from literature through text mining can guide and focus the research on biomedical relations that are genuinely relevant while minimising time and resources spent on the task.

Several systems target relations between biomedical entities in different degrees of complexity. Most systems are specific to a type of relation, such as protein-protein and chemical-protein interactions or gene-phenotype relations. However, others are suited to multiple types of relations, such as BiOnt (1), BERT (2), and BERT-derivatives (3,4). Nearly all systems also only use the information provided on the training data, overlooking openly available knowledge about the entities themselves, such as domain-specific ontologies. Ontologies such as the Gene Ontology (GO) (5), the Chemical Entities of Biological Interest (ChEBI) (6), and the Disease Ontology (DO) (7), to name a few, are important sources of biomedical information. These resources not only attribute unique identifiers to each domain entity but also define the relationships that the entities hierarchically have among themselves, among other relevant entity information.

This article describes our team's attempt at the prediction of drug and chemical-protein interactions regarding BioCreative VII Track 1 (8), through the adaptation of the BiOnt system, using the GO and ChEBI ontologies. The track is a follow up from the previous year ChemProt Biocreative VI Track (9). BiOnt is a system built using bidirectional Long Short-Term Memory (LSTM) networks, that incorporates Word2Vec word embeddings (10) and makes use of different combinations of input channels to maximize performance, including external knowledge in the form of biomedical ontologies.

We submitted five runs regarding different system parameter adjustments, including adding class weights and the use of post-processing rules. Our performance regarding the mean performance of the teams participating in Track 1 was below average, with around 0.38 difference in micro-average F1 respecting our best run. Nevertheless, we consider our approach relevant since our challenges, which we will describe in detail, were mainly regarding the preprocessing stage. BiOnt was made for a lower volume of data. Therefore, we had to optimize the preprocessing pipeline, but we did not finish all of the improvements we intended to do, mainly due to time constraints, which led to the loss of relevant information and performance.

Our main contribution is the adaption of the BiOnt system to the extraction of interactions between drugs and chemicals-proteins using the GO and ChEBI ontologies. Although our contribution to Track 1 did not yield good results, it was a starting point to adapt the state-of-the-art system BiOnt and expand it to deal with different types of entities while optimizing some of the pipeline processes. The code supporting our work and trained submitted models are available at https://github.com/lasigeBioTM/biocreativeVII.

II. Methodology

This section describes the different stages of our BiOnt adaptation to predict drug and chemical-protein interactions, including: (A) preprocessing, (B) training, and (C)

*This work was supported by FCT through funding of Deep Semantic Tagger (DeST) project (ref. PTDC/CCI-BIO/28685/2017; http://dest.rd.ciencias.ulisboa.pt/) and LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020); and FCT and FSE through funding of PhD Scholarship (ref. SFRH/BD/145221/2019).

post-processing. Fig. 1 presents the overall system architecture.

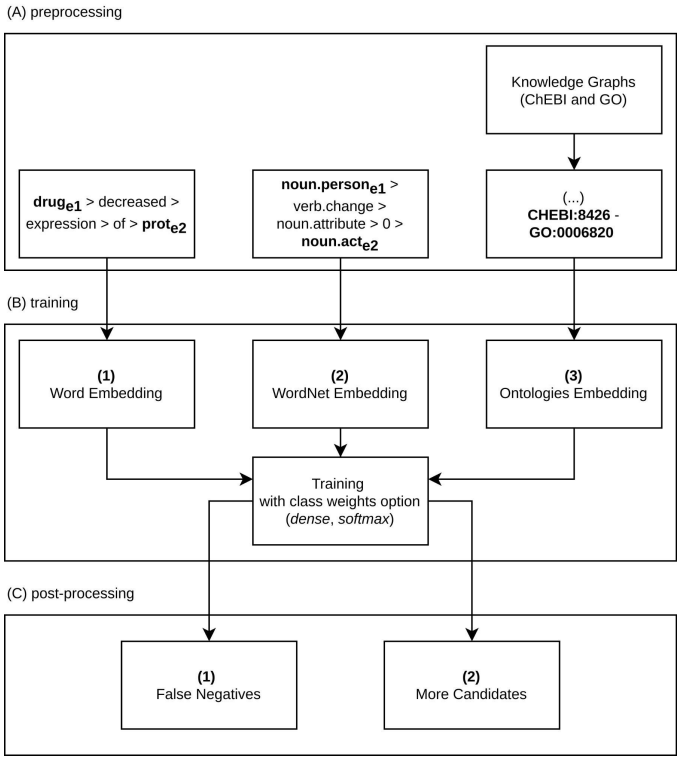

Fig. 1. System architecture diagram representing the three main stages of our BiOnt adaptation: (A) preprocessing, (B) training, and (C) post-processing. ChEBI stands for Chemical Entities of Biological Interest and GO for Gene Ontology.

A. Preprocessing

The track organizers provided both a training and a development dataset with 17288 and 3765 relations, respectively. These relations were classified as one of 14 possible positive labels, enumerated in Table II (11). For the test set, the teams were only provided with entity annotations of 10750 gold standard and background records (750 and 10000, respectively) to avoid bias towards a small test set. The goal was to predict positive labelled relations between those entities. Since we had to distinguish between positive and non-relations, we added a NO_RELATION label to all possible pairings mentioned in the same sentence but not labelled as positive in the training set. Therefore, when feeding the system with our data, the system could learn what constitutes a positive relation versus a non-relation and classify the test set accordingly. The distribution of both labels was balanced, i.e., the sum of positive labels versus the number of non-relation labels.

The first preprocessing step was tokenization followed by dependency parsing, using the SpaCy library, to obtain the Shortest Dependacy Path (SDP). Then, each element of the SDP was replaced by a generic string to reduce the effect of specific entity names on the model, and for each element, we also obtained the WordNet hypernym class (12). To finalize our baseline without the use of ontologies, we also used word2vec word embeddings (13).

Our ontological layer starts by matching each dataset entity to an ontology concept and obtaining its ancestors. While for drug/chemical entities this is straightforward using the ChEBI ontology, for gene/proteins entities is more complicated since there is not a direct match between these entities and an existing ontology. To workaround, we used GO and instead of doing a direct match between the dataset entities and the entities within the ontology we used the most representative GO term for each gene/protein entity using the protocol described in (1).

B. Training

The BiOnt system produces a neural network model with integrated ontological information. Thus, our preprocessing pipeline for the DrugProt corpus culminated in three different information channels fed to the training system: (1) Word Embeddings, (2) WordNet Classes, and (3) Concatenation of Ontology Ancestors.

For word embeddings, we used word2vec pre-trained on the English Wikipedia, which represents each specific word as a vector that expresses the semantic similarity between different words. The data is inputted as in the following example, taking into account the SDP:

$$\textbf{drug}_{e1} > \text{decreased} > \text{expression} > \text{of} > \textbf{prot}_{e2}$$

The WordNet channel consisted of the hypernyms of each word and the data was inputted taking into account the part-of-speech tags of each word and the grammatical relations between the words of the SDP:

$$\textbf{noun.person}_{e1} > \text{verb.change} > \text{noun.attribute} >$$

$$0 > \textbf{noun.act}_{e2}$$

Finally, in the third and last channel, concatenation of ontology ancestors, we define the chain of ancestors within each respective ontology to each entity as represented in Fig. 2. Each sequence of ancestors is represented as a one-hot vector and then transformed into a dense vector in the embedding layer.

All models were trained using Adam as the mini-batch gradient descent optimization algorithm, a learning rate of 1e-4, categorical cross-entropy as loss function, and a dropout rate of 0.5 for every layer except the penultimate and output layers.

On some models, we applied class weights due to the significant imbalance in the number of instances for each label. To keep the dataset representative of itself and for application on real-world data, instead of normalizing the number of classes for each label, we measure the class weight logarithmic by order of magnitude to avoid damaging bigger classes.

C. Post-processing

In post-processing, we focused our efforts on defining rules which could catch more positive pairs. Since, from early

experiences using the development dataset as a test set, we realized we had missed multiple positive pairs, primarily due to the preprocessing step.

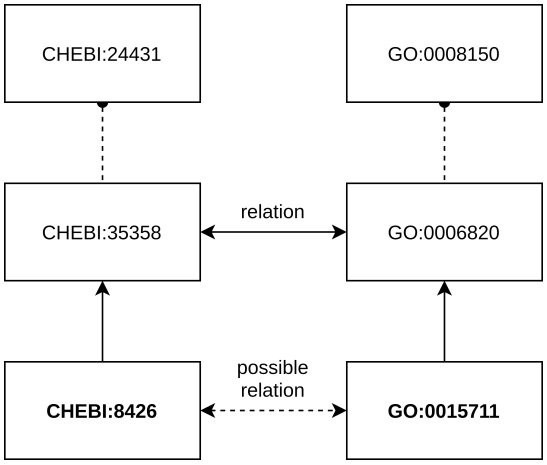

Fig. 2. Example of a possible relation between two entities reinforced by a relation between their ontology ancestors.

The nature of the BiOnt system would not allow for vectorization of a sentence where the entities considered overlapped. Thus, when entities, such as "Glutathione peroxidase$_{e1}$" (GENE-N) and "Glutathione$_{e2}$" (CHEMICAL), were considered for a relation, the system decided to discard the sentence due to overlapping entities. One possibility that we thought was adding a repeating sentence with different entities tagged in the training set. For instance, one sentence would have e_1 and the other e_2, which we did not test due to time constraints. This issue brought us to another fault of the BiOnt system: the lack of optimization of the preprocessing step, which became widely apparent when using larger datasets than the ones previously explored, as the one provided for the BioCreative task at hand.

Coming back to our post-processing stage, having the last faults in mind, we decided to implement rules that captured more positive relations. These relations could be of two types: (1) positive relations tagged with NO_RELATION and (2) candidate positive relations, which were discarded by BiOnt's vectorization issue in the preprocessing phase. For (1), our system assigned a positive label if the pair considered was already tagged with a positive label within the same article. While for (2), we first recovered annotations discarded in the preprocessing stage and then performed (1). The last step was discarding all NO_RELATION pairs, keeping only the positive labels for evaluation.

III. Experiments

We performed several experiences using different features, including the addition of ontolgy ancestors, class weights, and post-processing rules. Table I presents the performance metrics for the five best runs considering the usage of the different features: ancestors (*A*), class weights (*CW*), and post-processing rules (*PR*). The overall teams' micro-averaged metrics obtained for the task were 0.6430 for precision, 0.6291 for recall, and 0.6196 for F1. Thus, our results are quite low in comparison, which we considered being due to the issues raised in the previous section. Nevertheless, our best run (run 1) shows that using the three features simultaneously yields the best results, with each individual feature having a positive effect on recall.

TABLE I. System Best Runs (*A* stands for Ancestors, *CW* for Class Weights, and *PR* for Post-processing Rules)

Run	Features			Metrics		
	A	*CW*	*PR*	*Precision*	*Recall*	*F1*
1	x	x	x	0.3690	**0.1865**	**0.2478**
2	-	-	-	**0.4818**	0.1255	0.1991
3	-	x	-	0.3266	0.1630	0.2175
4	-	-	x	0.3427	0.1849	0.2394
5	-	x	x	0.4025	0.1650	0.2340

Even though adding post-processing rules improved the results, particularly on recall, it was not the difference we expected. The rule where we assigned positive labels to entities that we were already labelled as positive within the same document added some false positives to the results. We were expecting this because not all entities mentioned in the same sentence are implicitly related, which is predominant in article titles, for instance. One example where this happens is the article 23123662. This article states in several sentences that chemical entities *copper*, *platinum*, and *silver* have a positive relation of type SUBSTRACT with the gene *rCtr1*. However, these relations hold for some sentences of the article but not for others or even within the same sentence, leading our generalization to fail:

The uptake of ***copper***$_{T26}$ *by both cultured rat DRG neurons and HEK/****rCtr1***$_{T42}$ *cells was saturable and inhibited by cold temperature, silver and zinc, consistent with it being mediated by* ***rCtr1***$_{T30}$.

<T26, T42, SUBSTRACT>
<T26, T30, NO_RELATION>

This discrepancy points to our naive approach to the dataset, which had complex annotation rules and guidelines that we should have considered in the system planning.

Table II presents the metrics for the different positive relation types considered for our best run (run 1). Our system did not label any candidate relation with the labels (1) AGONIST-INHIBITOR, (2) PRODUCT-OF, (3) SUBSTRACT_PRODUCT-OF, and (4) AGONIST-ACTIVATOR. These labels correspond to three of the types of relation with fewer training examples: 13 (1), 25 (3), and 29 (4). However, label (2) had 921 training examples, and therefore, it was surprising that our system could not label this type of relation. Upon inspecting the results of our experiments, we realized that this was probably due to the complexity and variety of the entities. Most of those entities

were composed of multiple words (three or more) divided by apostrophes with mid entity capitalization, which our system was not sensitive to maintain in the preprocessing stage. The system performed best on the labels (1) ACTIVATOR, (2) ANTAGONIST, (3) INHIBITOR, and (4) PART-OF. These labels correspond to some of the ones who had more training examples: 1429 (1), 972 (2), 5392 (3), and 886 (4). However, the best-performing labels also correspond to relations that are between repeating and non complex entities. The differences between best and worst-performing labels reinforce the need to optimize the preprocessing step to handle a more diverse group of entities (e.g., with different string and length variations).

TABLE II. METRICS BY RELATION TYPE FOR RUN 1

Relation-Type	Precision	Recall	F1
ACTIVATOR	0.5033	0.2305	0.3162
AGONIST	0.2895	0.1089	0.1583
AGONIST-INHIBITOR	0.0	0.0	0.0
ANTAGONIST	0.5610	0.3007	0.3915
DIRECT-REGULATOR	0.2734	0.2657	0.2695
INDIRECT-DOWNREGULATOR	0.2561	0.1382	0.1795
INDIRECT-UPREGULATOR	0.3411	0.1588	0.2167
INHIBITOR	0.5108	0.2474	0.3333
PART-OF	0.5556	0.1096	0.1832
PRODUCT-OF	0.0	0.0	0.0
SUBSTRACT	0.1410	0.076	0.099
SUBSTRACT_PRODUCT-OF	0.0	0.0	0.0
AGONIST-ACTIVATOR	0.0	0.0	0.0

IV. CONCLUSIONS AND FUTURE WORK

This work presented the lasigeBioTM team's approach to BioCreative VII Task 1 to extract and classify interactions between drug and chemical-protein entities. Our team took the preceding steps of adaptation of the deep learning system BiOnt to deal with larger data volumes and overlapping entities participating in different relations. Our results were below the mean performance of the participating teams due to issues relating to the preprocessing stage. However, we demonstrated the positive impact of using external information in the form of the biomedical ontologies GO and ChEBI, class weights, and post-processing rules.

There is work to be done for our results to be up to par with the top-performing systems. However, we plan to substantially improve our approach by resolving the issues stated and applying rules to capture difficult-to-label relations.

REFERENCES

1. Sousa,D. and Couto,F.M. (2020) BiOnt: Deep Learning Using Multiple Biomedical Ontologies for Relation Extraction. Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II, Vol. 12036, pp. 367-374.

2. Devlin,J., Chang,M.W., Lee,K. and Toutanova,K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT, Vol. 2019, pp. 4171-4186.

3. Beltagy,I., Lo,K. and Cohan,A. (2019) SciBERT: A Pretrained Language Model for Scientific Text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing EMNLP-IJCNLP, pp. 3615-3620.

4. Lee,J., Yoon,W., Kim,S., Kim,D., Kim,S., So,C.H. and Kang,J. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, Vol. 36, pp. 1234-1240.

5. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J. M., ... and Sherlock,G. (2000) Gene ontology: tool for the unification of biology. Nature genetics, Vol. 25, pp. 25-29.

6. Degtyarenko,K., De Matos,P., Ennis,M., Hastings,J., Zbinden,M., McNaught,A., ... and Ashburner,M. (2007) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic acids research, Vol. 36, pp. D344-D350.

7. Schriml,L.M., Arze,C., Nadendla,S., Chang,Y.W.W., Mazaitis,M., Felix,V., ... and Kibbe,W.A. (2012) Disease Ontology: a backbone for disease semantic integration. Nucleic acids research, Vol. 40, pp. D940-D946.

8. Krallinger,M., Rabal,O., Miranda-Escalada,A. and Valencia,A. (2021) DrugProt corpus: Biocreative VII Track 1 - Text mining drug and chemical-protein interactions (1.2) [Data set]. Zenodo.

9. Krallinger,M., Rabal,O., Akhondi,S.A., Pérez,M.P., Santamaría,J., Rodríguez,G.P., Tsatsaronis,G., Intxaurrondo,A., López,J.A., Nandal,U.K., Buel,E.M., Chandrasekhar,A., Rodenburg,M., Lægreid,A., Doornenbal,M.A., Oyarzábal,J., Lourenço,A. and Valencia,A. (2017) Overview of the BioCreative VI chemical-protein interaction Track.

10. Church,K.W. (2017) Word2Vec. Natural Language Engineering, Vol. 23(1), pp. 155-162.

11. Miranda,A., Mehryary,F., Luoma,J., Pyysalo,S., Valencia,A. and Krallinger,M. (2021) Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. Proceedings of the seventh BioCreative challenge evaluation workshop.

12. Ciaramita,M. and Altun,Y. (2006) Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 594–602.

13. Mikolov,T., Sutskever,I., Chen,K., Corrado,G. and Dean,J. (2013) Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. NIPS'13. USA: Curran Associates Inc, pp. 3111–3119.