BioRED track lasigeBioTM submission: Relation extraction using Domain Ontologies with BioRED

Sofia I. R. Conceição [1*], Diana F. Sousa [1] , Pedro M. Silvestre [1] , Francisco M. Couto[1]

[1]LASIGE, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal

*Corresponding author: E-mail: sconceicao@lasige.di.fc.ul.pt

Abstract

Biomedical relation extraction is a crucial task for extracting valuable knowledge from unstructured scientific literature. This paper discusses our team's involvement in the BioCreative VIII Track 1: BioRED in both sub-tasks. Our primary focus was on the relation extraction (RE) task, utilizing the K-RET system in combination with Gene Ontology, Chemical Entities of Biological Interest, Human Phenotype Ontology, Human Disease Ontology and NCBITaxon Ontology. The objective was to evaluate whether the use of external knowledge could enhance the performance of the relation extraction task, both for entity relationships and for detecting novel information. Our results in both tasks were below the average and we were not able to discern the impact of the introduced external knowledge. However, it was observed that for our model, a cleaner dataset is needed for improved performance and the necessity for a larger number of example instances, as our model struggled to identify low-represented labels.

Introduction

The core of the biomedical Relation Extraction (RE) task is the characterization and identification of relations between biomedical concepts in literature. RE is essential for facilitating advancement in a number of biomedical disciplines (1). The BioCreative VIII Track 1 BioRED (Biomedical Relation Extraction Dataset) leverages already-existing relation extraction tasks through the use of a multiple entity and multiple relation pairs dataset, BioRED (2). The task requires that the systems recognize the asserted relationships as well as determine whether or not they are novel findings that are not available elsewhere. The BioRED corpus was developed to fill the gap in the biomedical corpora by providing numerous entity types and their relationships taking into account document-level relations (2). This Track is divided into two sub-tasks: *Sub-task 1* involved locating every relationship involving the annotated entities in the abstracts, and *Sub-task 2* involved developing an end-to-end system based on paper abstracts, recognizing the pertinent entities, normalizing them to a database, and finally asserting and classifying the relations.

This work describes the participation of our team lasigeBioTM at the BioCreative VIII Track 1: BioRED (Biomedical Relation Extraction Dataset) Track for both sub-tasks.

Our approach mainly focuses on the RE task by using K-RET (3), a system that employs knowledge from external sources, in this case, ontologies. K-RET can make use of any BERT-based pre-trained model and uses knowledge from the knowledge base in triples, expanding the original sentence into a knowledgeable sentence tree (3). Journal publications in the biomedical field are abundant in domain-specific terms that are difficult to fully comprehend without prior knowledge of the topic (4). Despite the fact that there are several RE models, not all of them make use of the relevant domain information accessible in knowledge bases. To

obtain more accurate predictions, this external knowledge may be necessary for comprehending the complexities and richness of biomedical literature. Some models have already demonstrated that including domain knowledge contributes to a better performance of the models in RE task (3, 5-7).
This paper describes our participation in BioCreative that aimed at assessing whether using a system that makes use of external knowledge would benefit the relation extraction task either for the relationships between the entities or for novelty. Our results and other details are available at https://github.com/lasigeBioTM/biocreativeVIII_Track1.

Material and Methods

The BioRED corpus was divided into 500 PubMed articles for the training set and 100 PubMed articles for the validation set. Regarding its characteristics, this corpus is annotated with six distinct biomedical concepts and nine possible relations between them and also information about relation novelty.
Furthermore, 400 Pubmed articles that are not part of the original BioRED corpus were annotated for the test set. The test set was hidden between approximately 10,000 non-relevant documents including 60 lacking titles and abstracts, 76 with titles but no abstracts, 34 with errors, and other abstracts outside the biomedical scope.

Train Set Bias

In the context of the training dataset, seven abstracts exhibited no discernible relations at all. Among the remaining 493 abstracts, we explored the distribution of various relation types and their respective contribution to the overall relation count. The most representative label in relation type was "Association" with 2752 instances (51.54%), followed by 'Positive Correlation' with 1441 (26.99%) and next by 'Negative Correlation' with 979 (18.33%). The remaining labels had very few instances, namely 'Bind' with 80 instances (1.50%), 'Cotreatment' with 41 (0.77%), 'Comparison' with 33 (0.62%), 'Drug Interaction' with 11 (0.21%) and lastly 'Conversion' with 3 (0.06%).

Sub-task 1 - Relation Extraction

This sub-task consisted of identifying all the relationships involving the annotated entities in the abstracts.
For the relation extraction task, we used K-RET: knowledgeable biomedical relation extraction system (3). K-RET makes use of external domain knowledge in the form of ontologies to enhance BERT-based systems. This system allows a flexible integration of the knowledge allowing the use of diverse sources, where to apply it and the handling of multi-token entities.
We used the pre-trained *allenai/scibert_scivocab_uncased* (8) and independently fine-tuned it for the association labels and for the novelty labels.
As external knowledge sources for K-RET, we used Gene Ontology (GO) (9-10) for 'GeneOrGeneProduct', Chemical Entities of Biological Interest (ChEBI) (11) for 'ChemicalEntity', Human Phenotype Ontology (HPO) (12) and Human Disease Ontology (DO) (13) for 'DiseaseOrPhenotypicFeature' and NCBITaxon Ontology (14) for 'OrganismTaxon'.
For fine-tuning, since the development set did not have any relationships (Sub-task 1) and annotations (Sub-task 2) it was used as test set and the original train set was randomly split into 90% for training and 10% for development.

The training was performed during 20 epochs with a batch size of 8 maintaining the parameters settings from the SciBERT model on a Tesla T4 GPU. This process was equal to the model for relationship labels (K-RET-E20) and the model to predict novelty labels (K-RET-E20-Novelty).
The resulting models were employed in both Sub-tasks for the RE task. For prediction, due to resource and time limitations, the final test set was only evaluated for 3 epochs.

Sub-task 2 - End-to-end system

The purpose of this Sub-task was to establish an end-to-end system given the abstract, able to identify the relevant entities, normalise them to a database and lastly assert and classify the relationships.
Our system consisted of using HunFlair for Named-Entity Recognition, dictionaries for Named-Entity Linking and K-RET for relation extraction.

Named-Entity Recognition and Named-Entity Linking

For Named-Entity Recognition (NER), we used the HunFlair tool (15). HunFlair is a NER tagger that covers several entity types. The following HunFlair NER models with BioRED Correspondence were applied: Chemical: 'ChemicalEntity', Gene: 'GeneOrGeneProduct', Species: 'OrganismTaxon', Disease: 'DiseaseOrPhenotypicFeature' and CellLine: 'CellLine'.
Regarding "SequenceVariant" we used the following REGEX pattern = r'rs\d+(\s|,|.|))'.
In the Named-Entity Linking (NEL) task, we created dictionaries for each entity resource file with name + identifier directly processing .tsv files or using obonet to process .obo files.
The entities in dictionaries were matched with entities found by the HunFlair NER tool, allowing for an edit Levenshtein distance of 2. We used '-' for the identifier if there was no match, following the same representation as in the original dataset.
The following databases were used for the creation of the dictionaries Comparative Toxicogenomics Database (16) for ChemicalEntity and DiseaseOrPhenotypicFeature, NCBI Taxon (14) for OrganismTaxon, Cellosaurus (17) for CellLine and NCBI Gene (18) for GeneOrGeneProduct. For the RE task, we used the resulting models from Sub-task 1, K-RET-E20 and K-RET-E20-Novelty to perform the predictions on the test set also using the same parameters.

Unofficial Runs

The organization gracefully gave us the possibility to submit five additional runs for each sub-task. We submitted three additional runs for Sub-task 1 and two additional runs for Sub-task 2.
For both sub-tasks, we maintained the novelty model used in the official submission, K-RET-E20-Novelty, and only used new models regarding the relationship labels.
For Sub-task 1 we fine-tuned the SciBERT model in a BioRED version that did not contain "NAN" examples for 20 epochs with proportional label weight (K-RET-E20-Clean). This model was used in the predictions of unofficial run 1. Additionally, using the same fine-tunning dataset we trained for 15 epochs with the following label weights Association: 0.485, Positive Correlation: 2.0, Negative Correlation: 2.5, Bind, Comparison, Conversion, Cotreatment and Drug Interaction: 3.0 for run 2 (K-RET-E15-Weights-Clean).
Lastly, our unofficial run 3 used the official submission fine-tuned model K-RET-E20 with a weight of 3.0 in all labels except for 'Negative Correlation' and 'No' in novelty during prediction.

For Sub-task 2 in run 1, we finalized the official run with the full dataset and corrected offset values from the NER stage. As for run 2, it was evaluated with the K-RET-E20-Clean.

Results and Discussion

Concerning Sub-task 1, our results were below the average and median performance reported for the task.

The task average F-score was 0.6703 regarding Entity Pair, 0.4774 in Entity Pair + Relation Type, 0.4923 in Entity Pair + Novelty and 0.3522 at Entity Pair + Relation Type + Novelty. The median F-score was 0.735 for Entity Pair, 0.5317 in Entity Pair + Relation Type, 0.564 in Entity Pair + Novelty and 0.4073 for Entity Pair + Relation Type + Novelty.

In the official submission, only one run per sub-task was submitted. For Sub-task 1 our model was only able to predict two labels, 'Association' and 'Negative Correlation'. The scores were low, achieving a low 0.324 F-score regarding Entity Pair, 0.0727 in Entity Pair + Relation Type, 0.1289 in Entity Pair + Novelty and 0.0296 at Entity Pair + Relation Type + Novelty.

As for Sub-task 2 only a partial test set with annotations and asserted relationships was submitted due to restricted computational resources. Predictions for this task had to be delivered before the release of the test set for sub-task 1 since it had entity annotations. So, we only had one week to annotate and perform RE. Since we had limited computational resources, it was not feasible to fully complete the assignment.

The task average F-score was 0.7687 for NER, 0.633 for Normalization, 0.2862 for Entity Pair, 0.2139 in Entity Pair + Relation Type, 0.2182 in Entity Pair + Novelty and 0.1625 at Entity Pair + Relation Type + Novelty. The median F-score was 0.7858 for NER, 0.6681 for Normalization, 0.3447 for Entity Pair, 0.2540 in Entity Pair + Relation Type, 0.2678 in Entity Pair + Novelty and 0.1979 for Entity Pair + Relation Type + Novelty. Our run results were 0.0001 for NER, 0.1226 for Normalization, 0.0315 for Entity Pair, 0.0072 in Entity Pair + Relation Type, 0.0105 in Entity Pair + Novelty and 0.0026 at Entity Pair + Relation Type + Novelty.

Regarding not having submitted the full test set, errors regarding the offsets in the NER task were detected which explains the poor results of our run as well as the problems detected for RE in task 1, since we used the same model for prediction. The NER errors where a result of the entities offsets being expected taking into account both title and abstract with no space in between them.

Unofficial Runs

For sub-task 1 we used a less noisy version of the BioRED train set. Our system was only capable of identifying the three most representative labels, 'Association', 'Positive Correlation' and 'Negative Correlation' in Run 1 and Run 2. In Run 3 similar to official Run 1 it was only capable of identifying 'Association' and 'Negative Correlation', indicating that the use of a cleaner dataset had a positive impact. In run 1 we obtained an F-score of 0.3248 in entity pair, 0.1381 in Entity Pair + Relation Type, 0.129 in Entity Pair + Novelty and 0.0552 at Entity Pair + Relation Type + Novelty. As for Run 2, the F-scores were 0.3248 in entity pair, 0.134 in Entity Pair + Relation Type, 0.129 in Entity Pair + Novelty and 0.0518 in Entity Pair + Relation Type + Novelty. Last, in run 3, the F-scores were 0.3248 in entity pair, 0.0727 in Entity Pair + Relation Type, 0.129 in Entity Pair + Novelty and 0.0296 at Entity Pair + Relation Type + Novelty.

We took advantage of these extra runs to fully submit what was previously partially submitted at the official run of Sub-task 2 and correct the offset error in the NER phase, this was the unofficial run 1 submission. This time, the NER task results were properly evaluated since it was

in the right format and the results were 0.7703, 0.6411 and 0.6998 for precision, recall and F-score respectively. However, these results were below the average and median reported for the task. Additionally, we applied the K-RET-E20-Clean model at run 2 which led to a slight decline in the scores regarding relation type. The F-score obtained were 0.051 in Entity Pair, 0.01 in Entity Pair + Relation Type, 0.0247 in Entity Pair + Novelty, and 0.0069 in Entity Pair + Relation Type + Novelty.

Conclusion and Future Work

This manuscript presented the lasigeBioTM team approach to BioCreative VIII BioRED track, which focused primarily on the RE task employing the K-RET system in combination with five distinct ontologies that cover the majority of the BioRED entities, GO, ChEBI, HPO, DO and NCBITaxon Ontology.

Substantially our results were below the average and median performance reported for the task. The results clearly demonstrate the impact of a cleaner dataset to get better performance in our model and that it is dependable on a larger number of example instances because it was unable to detect low-represented labels. Moreover, we had complications in sub-task 2 regarding the lack of time and computational resources. Future similar tasks should take this into account and provide better accommodations for smaller teams with fewer resources, such as more time, or less noise masking papers. Lastly, it was not possible to evaluate the impact of the external knowledge in RE, we would like to perform ablation studies to investigate this point in the future.

For future work, we would like to explore different combinations of ontologies and hyperparameters. For novelty, it would be interesting to explore trigger words and external gold standard datasets for distant supervision.

Funding

This work was supported by Fundação para a Ciência e a Tecnologia (FCT) through funding of LASIGE Computer Science and Engineering Research Centre (LASIGE) Research Unit [UIDB/00408/2020 and UIDP/00408/2020]; and FCT and FSE through funding of PhD Scholarship [SFRH/BD/145221/2019] attributed to DFS; FCT through funding of PhD Scholarship [UI/BD/153730/2022] attributed to SIRC.

References

(1) Li, X., Yang, J., Liu, H., *et al.* (2021) Overview of Distant Supervised Relation Extraction. In 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), pp. 1287-1292.

(2) Luo, L., Lai, P. T., Wei, C. H., *et al.* (2022). BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, **23**(5), bbac282.

(3) Sousa, D. F., and Couto, F. M. (2023). K-RET: knowledgeable biomedical relation extraction system. *Bioinformatics*, **39**(4), btad174.

(4) Lai, T., Ji, H., Zhai, C., *et al.* (2021). Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Vol.1: Long Papers, pp. 6248–6260

(5) Lamurias, A., Sousa, D., Clarke, L. A., *et al.* (2019). BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies. *BMC bioinformatics*, **20**(1), 1-12.

(6) Sousa, D., and Couto, F. M. (2020). BiOnt: deep learning using multiple biomedical ontologies for relation extraction. In European Conference on Information Retrieval. *Cham: Springer International Publishing,* 367-374

(7) Sousa, D. and Couto, F. M. (2022). Biomedical relation extraction with knowledge graph-based recommendations. *IEEE Journal of Biomedical and Health Informatics*, **26**(8), 4207-4217.

(8) 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

(9) Ashburner, M., Ball, C. A., Blake, J. A., *et al.* (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25(1), 25-29.

(10) Gene Ontology Consortium. (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, 47(D1), D330-D338.

(11) Degtyarenko, K., De Matos, P., Ennis, M., *et al.* (2007). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, 36(suppl_1), D344-D350.

(12) Köhler, S., Gargano, M., Matentzoglu, N., *et al.* (2021). The human phenotype ontology in 2021. *Nucleic Acids Res.*, 49(D1), D1207-D1217.

(13) Schriml, L. M., Munro, J. B., Schor, M., *et al.* (2022). The human disease ontology 2022 update. *Nucleic Acids Res.*, 50(D1), D1255-D1261.

(14) Federhen, S. (2012). The NCBI taxonomy database. *Nucleic Acids Res.*, 40(D1), D136-D143.

(15) Weber, L., Sänger, M., Münchmeyer, J., *et al.* (2021). HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. Bioinformatics, 37(17), 2792-2794.

(16) Davis, A. P., Wiegers, T. C., Wiegers, J., *et al.* (2023). CTD tetramers: a new online tool that computationally links curated chemicals, genes, phenotypes, and diseases to inform molecular mechanisms for environmental health. Toxicological Sciences, kfad069.

(17) Bairoch, A. (2018). The cellosaurus, a cell-line knowledge resource. Journal of biomolecular techniques: JBT, 29(2), 25.

(18) Brown, G. R., Hem, V., Katz, K. S., *et al.* (2015). Gene: a gene-centered information resource at NCBI. Nucleic Acids Res., 43(D1), D36-D42.