BIT.UA at Biocreative VIII track 1: A joint model for relation classification and novelty detection

Tiago Almeida[1,*], Richard A. A. Jonker[1], Dimitri da Silva[1], João Almeida[1], and Sérgio Matos[1]

[1]IEETA/DETI, LASI, University of Aveiro, Portugal

*Corresponding author: E-mail: tiagomeloalmeida@ua.pt

Abstract
The task of relation extraction has long posed a challenge within the Natural Language Processing (NLP) community, and its application in biomedical research is important for understanding scientific literature. The development of a tool capable of effectively addressing this task holds the potential to improve knowledge discovery by automating the extraction of relations from literature. The first track in the Biocreative VIII competition extended the scope of this challenge by introducing the detection of novel relations within literature. This paper presents the strategies used in this competition by our team, Biomedical Informatics and Technologies (BIT) at the University of Aveiro. We leveraged joint training to craft a singular, versatile model capable of not only classifying relations between two entities but also determining the novelty of the identified relation. Our experiments yielded promising results, with our submission outperforming the competition's average. This paper not only details our approach but also highlights the potential of joint training in relation extraction, paving the way for improved automated analysis of biomedical literature.

Introduction
In the domain of biomedical NLP, the task of relation extraction holds a central position, carrying significant implications for breakthroughs such as drug discovery. At its core, relation extraction seeks to discern and define semantic connections among two or more entities within a text. While most relation extraction datasets historically concentrated on single-relation, sentence-level extractions, the BioRED dataset (1) changes this. It provides a multiclass relation classification dataset as well as describing which of these are novel relations. In this study, we introduce an innovative joint training strategy, enabling the model to classify relations and determine their novelty. The notable advantage of this approach lies in its inherent efficiency, removing the need for multiple models and significantly reducing training and inference time.

Methodology
The data consists of 6 types of entities and 8 types of relations. Some inconsistencies in class naming, especially between the training and test sets (e.g., *'ChemicalEntity'* and *'Chemical'*), needed to be resolved to map them to the same class. Furthermore, an additional relation class was introduced to represent the negative class (9 total classes) that would indicate that the entities do not have a relation. Given a document D , containing E unique entities, the objective of this work is to identify every pair of entities (e_i, e_j) for which there exists a relation r_k between e_i and e_j. Additionally, the model must be capable of determining whether this triple(e_i, r_k, e_j) is novel.

Each document can potentially contain E^2 entity pairs per document as each entity can have a relation to itself, however only a certain number of those are valid relations. We use negative sampling to select the pairs belonging to the negative class. To balance the data more effectively, we randomly selected a subset of these potential pairs based on the number of valid relations in the document. To improve the quality of the negative pairs, we leveraged information provided by the event organisers to identify possible pair combinations, as only specific entity types had relations between them and we select negative samples from this pool. According to our interpretation of the annotation guidelines, an assumption was made that an entity relation triple is directional, $(e_i, r_k, e_j) \nRightarrow (e_j, r_k, e_i)$.

To accurately encode contextual entity information as input for the model, we introduce new tokens '[s1]', '[e1]', '[s2]', and '[e2]', which correspond to the start and end of the two entities in the text. These tokens are then directly inserted into the text. For example, in the sentence '... *high-grade* [s1]*glioma*[e1]...', *'glioma'* corresponds to entity 1. After tokenization, the positions of these tokens are provided to the model along with the tokens themselves. The model architecture we propose is based on the well-established Transformer architecture (2). We select contextualized embeddings corresponding to the special tokens for the entities. A multi-head attention layer is applied to these embeddings, which are then fed into two classifiers, one for relation classification and another for novelty classification. The model is trained end-to-end following a joint loss presented on Equation 1,

$$L = L_r + (y_r \neq 8)L_{n.} \qquad (1)$$

Here, we sum the cross-entropy loss for both the relation classifier and the novel classifier, but only considering the novel loss for samples with valid relations, which occurs when $(y_r \neq 8)$. To encode more domain knowledge in the mode, we construct a relation mask containing the possible relation pairs given two entity types. Using this relation mask, we apply it to relation classifier predictions, ensuring that only cases seen in the training data (valid combinations) are predictable.

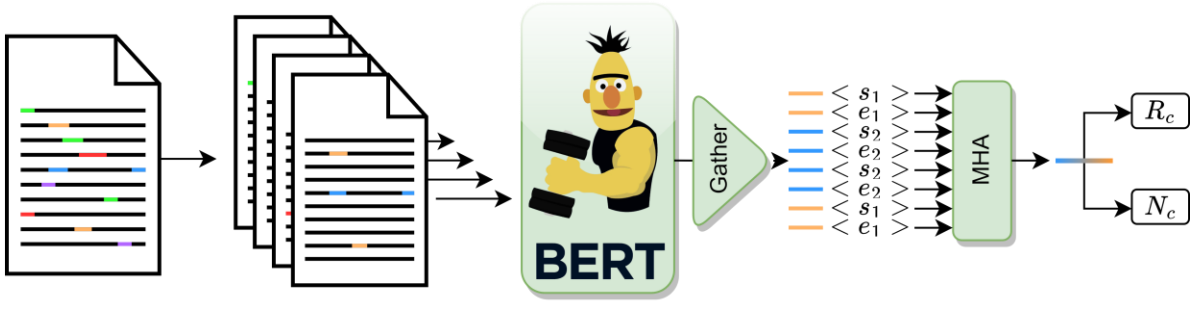

Figure 1 Overview of our joint model for relation classification and novelty detection.

To enhance the performance of this model, we employed ensemble methods to fuse results from various distinct runs. The ensemble process takes place after eliminating the negative pairs. Initially, we employ a majority voting mechanism to decide whether a relation candidate should be regarded as a valid relation. If it surpasses the majority voting threshold, it is assigned the class with the highest level of support. The novelty score is allocated based on the majority class from the novelty predictions for this relation class. In a tie, a class is selected randomly.

Results

To prepare the best models for submission to the competition, we made use of an online validation system provided by the event organisers. The original training data comprised 500 documents, the validation data contained 100 documents, and the test data included 400 documents. There are three evaluation metrics used in the competition. The main evaluation metric is the F1-score, which is evaluated on the entity pair (e_i, e_j), relation type (e_i, r_k, e_j), and the novel score for each relation (e_i, r_k, e_j, n). The novelty score is the evaluation used for determining competition winners.

Validation results

Initial tests were conducted using PubMedBERT (3) as a baseline, and more refined models were chosen after fine-tuning. This is a BERT-based model pre-trained on the PubMed corpus. We performed tests to evaluate the efficacy of the special tokens introduced ('[s1]', '[e1]', '[s2]', and '[e2]'). We experimented with 's', 'e', and 'both' types of tokens. The results of these initial tests revealed that using both tokens yielded the best results, however, in some later tests, it was found that using only 's' yielded slightly better results, with only a marginal difference of about one percentage point. Further tests will be necessary to draw concrete conclusions regarding these methods.

In testing various model architectures, the primary observation was the use of BioLinkBERT (4), which was pre-trained using document links, enabling better inter-document dependencies. BioLinkBERT improved upon the base PubMedBERT models by approximately 2 percentage points. Moreover, the larger versions of these models further improved performance over the base models by an additional percentage point.

Subsequent tests focused on negative sampling of entity pairs. Initially, we used all possible combinations of negative samples, but this skewed the model heavily toward the negative relation class, negatively impacting performance. We then balanced the data to ensure an equal number of negative samples compared to positive samples and conducted further tests with both double and triple the number of negative samples in comparison to positive samples. Both double and triple negative samples led to improved performance, with the double negative sample showing a roughly 5 percentage points advantage and the triple sample showing less of an improvement. The limitation to this approach of sampling is that the negative samples are randomly selected, suggesting that the random seed will impact the model's performance. After testing several seeds, one seed was found to significantly outperform the others, improving performance by 4 percentage points.

Submission results

In our competition submission, we included our top two performing models from the validation set, as well as three ensemble models. The highest-performing model in the validation set achieved a novelty F1 score of 56.24 (run0), ranking as the fourth-best model among all participating teams (validation). This model was trained on BiolinkBERT-base, with double negative sampling, utilizing both 's' and 'e' special tokens, and without a relation mask. The second-best model achieved a validation F1 score of 53.36 (run1) and was trained using BiolinkBERT-large, single negative sampling, both 's' and 'e' special tokens, and no relation mask. We also submitted various ensemble combinations, including our top eight performing

models (run2: mean and standard deviation of the models in the ensemble: 51.19 ± 2.70), top five performing models (run3: 52.80 ± 1.95), and top three performing models (run4: 53.92 ± 1.70). A summary of results can be seen in Table 1.

Table 1: Table of results submitted, precisions, recall and F1 of our 5 submitted runs, as well as the median and average of all participants runs.

Config.	Entity Pair (P/R/F %)			Entity Pair+Relation Type (P/R/F %)			Entity Pair+Relation Type+Novelty (P/R/F %)		
run0	**66.06**	78.33	71.67	46.82	57.05	51.43	36.19	44.71	40.00
run1	63.91	85.72	73.22	47.23	65.98	**55.05**	**36.88**	53.00	**43.50**
run2	59.75	**88.96**	71.48	43.67	**68.79**	53.42	33.68	**54.77**	41.71
run3	64.52	86.19	73.79	47.28	65.40	54.88	36.68	51.87	42.97
run4	66.18	84.63	**74.27**	**48.26**	63.27	54.76	37.76	50.37	43.16
Median	**77.93**	69.65	73.56	**51.64**	54.79	**53.17**	**41.61**	39.88	40.73
Average	69.22	68.6	67.03	49.01	48.39	47.74	36.15	35.73	35.22

Upon evaluating the test set results, we observed that our best performing model on the validation set (run0) did not perform as well in comparison to other models. This suggests that the model's performance is closely tied to the random seed used during training, highlighting the need for a system that is less dependent on random assignment. Interestingly, our best-performing model in the test dataset was the second-best model in the validation dataset, which can be attributed to the use of a larger, more robust model. Regarding the ensemble models, it was evident that larger ensembles tended to perform better, largely due to the subpar performance of the best validation model. The best-performing ensemble was only slightly behind the best model, trailing by a mere 0.34 percentage points.

Comparing our system's performance to that of the competition, we found that almost all of our systems ranked above the median and mean, which were in line with our 4th place ranking on the validation data. Our models exhibited significantly higher recall when compared to the median recall. We also note that our models have a higher rate of novel score, given the entity pair score, when compared to the median scores.

Conclusion

In this work, we present the results for our submissions to the BioCreative-VIII BioRED task. We investigated a joint training approach to train relation class and novel score for each pair of entities given in a document. Our results performed above average in the competition. We have made some conclusions about the models to be used, however, the approach we used for negative sampling led to some results dependent on random seeding. In future work, a more robust way of selecting negative sampling should be used. We further note that our work has a better conversion rate of novel score given entity pair, when compared to the median results. We also note that some changes in the architecture could lead to better performance of the model.

Funding
This work was partially supported by national funds through the Foundation for Science and Technology (FCT) in the context of the projects DSAIPA/AI/0088/2020 and UIDB/00127/2020. Tiago Almeida is funded by FCT under the grant 2020.05784.BD.

References

1. Luo L, Lai P-T, Wei C-H, Arighi CN, Lu Z. 2022. BioRED: a rich biomedical relation extraction dataset. Briefings in Bioinformatics. 23(5). doi:10.1093/bib/bbac282

2. Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) [Internet]. Minneapolis, Minnesota: Association for Computational Linguistics; p. 4171–4186. https://aclanthology.org/N19-1423

3. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. 2022. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare. 3(1):1–23. doi:10.1145/3458754

4. Yasunaga M, Leskovec J, Liang P. 2022. LinkBERT: Pretraining Language Models with Document Links. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) [Internet]. Dublin, Ireland: Association for Computational Linguistics; p. 8003–8016. https://aclanthology.org/2022.acl-long.551