UIUC-BioNLP @ BioCreative VIII BioRED Track

M. Janina Sarol[1, *], Gibong Hong[2], and Halil Kilicoglu[2]

[1]Informatics Programs, University of Illinois at Urbana-Champaign
[2]School of Information Sciences, University of Illinois at Urbana-Champaign

*Corresponding author: E-mail: mjsarol@illinois.edu

Abstract

In this paper, we present a pipeline approach for the BioCreative VIII BioRED (Biomedical Relation Extraction Dataset) Track. Our approach combines fine-tuned PubMedBERT models for named entity recognition (NER), relation extraction (RE), and novelty detection (ND), with an entity linking (EL) approach based on PubTator and BERN2 models. Our end-to-end system achieved competitive results, ranking above the average and the median scores for all submissions.

Introduction

Biomedical publications contain vast amounts of scientific knowledge, often expressed as biomedical concepts (entities) and the semantic relationships between them. Automatically extracting these elements from biomedical publications can assist in constructing knowledge bases and tools that allow researchers to learn and sift through information more rapidly, and ultimately advance our understanding of biology and health.

While there has been considerable NLP research that focuses on extracting biomedical concepts (e.g., chemicals and diseases) and their relationships (e.g., chemical-induced diseases), these tasks remain challenging, particularly relation extraction (1). The BioRED Track in BioCreative VIII aims to advance state-of-the-art in these tasks, and provides a comprehensive dataset annotated for a variety of entity and relation types. In addition, entities are normalized to standard database identifiers, and relationships are labeled for novelty. In this paper, we present our end-to-end system for these tasks, which achieved competitive results in the BioRED Track.

Methods

Our end-to-end system employs a pipeline approach (see Figure 1). Our named entity recognition (NER) module is a fine-tuned PubMedBERT (2) model that identifies entity mentions and classifies them into one of six entity types: chemicals, gene/proteins, diseases, variants, species, and cell lines. The entity linking (EL) component combines information from two existing approaches, PubTator Central (3) and BERN2 (4), to map entity mentions to their unique database identifiers. In each abstract, each pair of normalized entities is then passed to our relation extraction (RE) component, which determines the type of relationship between the entities: positive correlation, negative correlation, association, binding, co-treatment, drug interaction, comparison, conversion, or no relation. Finally, for each related entity pair, the novelty detection (ND) model uses the abstract text to classify the relationship as novel (or not). Each of the NER, RE, and ND PubMedBERT models were separately fine-tuned on their respective task using the 500 abstracts in the BioRED training set.

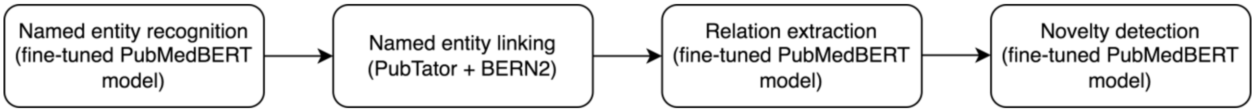
Named entity recognition (fine-tuned PubMedBERT model)
Named entity linking (PubTator + BERN2)
Relation extraction (fine-tuned PubMedBERT model)
Novelty detection (fine-tuned PubMedBERT model)

Figure 1: A Pipeline Approach for End-to-End Relation Extraction

Named Entity Recognition (NER)

We formulated the NER task as a token classification problem. Following BIO format, we labeled each token as either the beginning (B-type), part of (I-type), or outside of an entity (O), along with type information (e.g., *B-ChemicalEntity*, *I-CellLine*). We then fine-tuned a pretrained PubMedBERT model (*microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract*) with the following hyperparameters: epochs = 10, learning rate = 2e-05, batch size = 8. We used cross-entropy loss function. The model predictions are further refined by combining consecutive predicted entities based on the following set of rules: 1) there is no whitespace character between two entities of the same type (e.g., *A* and *(1)-adenosine receptor* vs. *A(1)-adenosine receptor*) and 2) there is a space in between two disease entities (e.g., *benign* and *tumor* vs. *benign tumor*). Our NER model obtained micro-F_1 score of 90.59 on the development set (100 abstracts).

Entity Linking (EL)

In the BioRED Track, the goal of EL is to map the chemical, gene/protein, disease, variant, species, and cell line entities to MESH (5), NCBI Gene (6), MEDIC (7), dbSNP (8), NCBI Taxonomy (9), and Cellosaurus (10) vocabularies, respectively. We leveraged existing tools for biomedical entity linking, namely, PubTator Central (PTC) and BERN2.

PTC (3), available through an application programming interface, performs NER and EL using four different systems (TaggerOne (11) for diseases, chemicals, and cell lines, tmVar (12) for variants, GNormPlus (13) for genes, and SR4GN (14) for species) and addresses ambiguity problems using a convolutional neural network. We used PTC as a lookup table, that is, for each entity mention found by our NER model, we queried the PTC results to find an exact string match. Using this approach, we obtained a micro-F_1 score of 70.71 on the development set with gold standard entities. BERN2 (4) combines NER and EL, employing a hybrid approach which uses rule-based normalization and BioSyn (15), a neural network-based biomedical EL model. Since our NER model already identifies entity mentions, we only used the EL component of BERN2. We obtained a micro-F_1 score of 65.36 using BERN2 on the development set.

While PTC has better overall results, experiments on the BioRED development set showed that PubTator only obtained better results for gene, proteins and diseases, while BERN2 performed better for the rest of the entities. Following these results, we used PTC to link gene, protein and disease mentions, and BERN2 to link other entity mentions. Entities that could not be mapped with PTC or BERN2 were discarded. This approach produced a micro-F_1 score of 72.25.

Relation Extraction (RE)

We adapted the PURE (16) model for entity and relation extraction. This model is composed of two BERT models separately fine-tuned for NER and RE. We utilized the RE model, which uses the generated entity representations for classifying entity pairs with a relation type (or no relation). For this purpose, all tokens belonging to an entity mention are enclosed with marker tokens denoting the entity type and whether the entity is the subject or object of a relation. For each entity, the embedding of its corresponding marker token (from the last hidden state of the BERT model) is taken as its representation. The embeddings of each possible entity pair are concatenated and passed to the classification layer, which predicts the pair's relation type. PURE performs sentence level extraction, assumes a single mention for each entity in each instance, and is designed for unidirectional relations. In contrast, the BioRED dataset contains full abstracts, multiple mentions of the same entity are common, and most relation types are bidirectional (e.g., *Y is associated with X* is equally valid as *X is associated with Y*). Therefore, we made several key updates to the model:

- We remove the directionality of relations. For a given entity pair [ENTITY1, ENTITY2], we generated two embeddings: [ENTITY1, ENTITY2, ENTITY1 x ENTITY2], and [ENTITY2, ENTITY1, ENTITY2 x ENTITY1], each corresponding to the concatenation of two entity representations and their element-wise product. These concatenated embeddings are individually passed to the relation classifier. The loss is the sum of the cross-entropy losses of both relation representations. To address the bidirectionality during prediction, the logits of both representations are summed up.
- We tag multiple mentions of the same entity. Each entity mention has its own corresponding marker token. However, for prediction, we select the pair of mentions (one for entity1 mentions and one for entity2 mentions) that best helps with classification. Our intuition is that not all mentions are important in identifying the relation, and may introduce unnecessary noise for the model. We take the dot product of each mention pair, which represents the importance of each mention pair to classifying the relation. We take the mention pair with the highest dot product and use it as the final relation representation for a given entity pair.
- We also remove the distinction between different entity types for our marker tokens; that is, instead of using [ENTITY-GENE] as a marker token, we only used [ENTITY]. Our initial experiments showed that including the entity type information in the marker token was not helpful for relation prediction.
- Finally, to improve model robustness, we use projected gradient descent attacks (17) during training. After the model's weights are updated using the combined loss, we perturb the token embeddings three times, adding noise, and train the model to correctly classify relations using the perturbed input.

For model training, we included all pairs of non-related entities in the same abstract as negative examples. We trained the model with the following hyperparameters: 10 epochs, 2e-5 learning rate, 32 batch size, and Adam optimizer. With this model, we obtained a micro-F_1 score of 56.92 on the development set using ground truth entities.

Novelty Detection (ND)

We used a similar approach for ND task with two notable changes: 1) we did not include negative examples for training (as the input entity pairs already have an identified relation) and 2) we used a different entity representation. Instead of picking the best pair of entity mentions, we weigh all mentions based on their importance for the ND task by computing the log sum of exponentials for all mentions of the entities in the entity pair. This generates a single vector for each entity, which we concatenate to obtain the final relation representation. Our intuition is that some mentions are more important than others, only in this case, we still consider all mentions as possibly contributing to the novelty prediction task. We trained the model with the following hyperparameters: epochs (15), learning rate (1e-5), batch size (8), and Adam optimizer. We obtained a micro-F_1 score of 75.58 on the development set, using ground truth entities and relations.

Results and Discussion

The BioCreative VIII BioRED Track consists of two tasks: Subtask 1 focuses on relation extraction and novelty detection only. The ground truth named entities and identifiers are provided. In Subtask 2, the systems are expected to identify named entity mentions and map them to standard database identifiers as well. We submitted three official runs: one for subtask 1 and two for subtask 2. For subtask 2, our first submission only used BERN2 for EL, while our second submission used the combined BERN2 and PTC approach described above.

Subtask 1: Relation Extraction and Novelty Detection

Table 1 shows our results for Subtask 1. Our results are higher than the average scores for all official submissions, but lower than the median scores. Our experiments on the development set showed that our model had difficulty distinguishing between association and positive correlation relationships, and it is likely that this issue persisted in the test set. This is especially notable given that these two relations make up a majority of the training and development sets (78.52% and 82.55%, respectively). Analyzing the results of the development set, we find that expressions of some of the positive correlations are somewhat less direct; an example is shown below, where a positive correlation between *V1763M* and *arrhythmias* is annotated.

"These findings suggest that the Na(v)1.5/V1763M channel dysfunction and possible neighboring mutants contribute to a persistent inward current due to altered inactivation kinetics and clinically congenital LQTS with perinatal onset of arrhythmias that responded to lidocaine and mexiletine."

Table 1: Our official subtask 1 results.

	Precision	Recall	F_1	Average F_1	Median F_1
RE	55.90	49.96	52.76	47.74	53.17
RE+ND	42.07	37.61	39.71	35.22	40.73

Subtask 2: End-to-End System

Table 2 shows our results for Subtask 2. Our results are higher than the average and median scores for all official submissions. The lower NER F_1 score in submission 2 compared to submission 1 is due to the higher number of unmapped entities in the latter submission, which were discarded. The increase in EL F_1 score is mainly due to the improvement in mapping gene/protein entities. Using PTC instead of BERN2 for this entity type produced a substantial increase in the test set results (precision: 56.34 to 88.90, recall: 60.72 to 73.47, F1: 58.45 to 80.45). It is unknown how much of the relation extraction scores are affected by missing entities, i.e., entities that were not found by our NER model or mapped by our EL system.

Table 2: Our official subtask 2 results.

	Submission 1			Submission 2				
	Precision	Recall	F_1	Precision	Recall	F_1	Avg. F_1	Median F_1
NER	90.13	77.72	**83.47**	95.33	68.01	79.39	76.87	78.58
NER+NEL	69.01	64.75	66.81	85.53	67.33	**75.35**	63.36	66.81
NER+NEL+RE	27.63	19.46	22.84	39.94	21.43	**27.90**	21.39	25.40
NER+NEL+RE+ND	20.75	14.61	17.15	29.90	16.05	**20.89**	16.25	19.79

References

1. Luo L, Lai PT, Wei CH, Arighi CN, Lu Z. (2022) BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*. 2022;23(5):bbac282.

2. Gu Y., Tinn R., Cheng H., et al. (2021) Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3, 1-23.

3. Wei, C. H., Allot, A., Leaman, R., Lu, Z. (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.*, 47, W587-W593.

4. Sung, M., Jeong, M., Choi, Y., et al. (2022) BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38, 4837-4839.

5. Lipscomb C.E. (2000) Medical subject headings (MeSH). *Bull. Med. Library Assoc* ., 88 , 265.

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

7. Davis, A. P., Wiegers, T. C., Rosenstein, M. C., Mattingly, C. J. (2012) MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database*, *2012*, bar065.

8. Sherry, S. T., Ward, M. H., Kholodov, M., et al. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29, 308-311.

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

10. Bairoch A. (2018) The Cellosaurus, a cell-line knowledge resource. *J. Biomol. Tech*., 29, 25-38.

11. Leaman, R., Lu, Z. (2016) TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, 32, 2839-2846.

12. Wei, C. H., Harris, B. R., Kao, H. Y., Lu, Z. (2013) tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, 29, 1433-1439.

13. Wei, C. H., Kao, H. Y., Lu, Z. (2015) GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *Biomed. Res. Int*., 2015.

14. Wei, C. H., Kao, H. Y., Lu, Z. (2012) SR4GN: a species recognition software tool for gene normalization. *PloS One*, 7, e38460.

15. Sung, M., Jeon, H., Lee, J., Kang, J. (2020) Biomedical entity representations with synonym marginalization. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, 3641–3650.

16. Zhong, Z., Chen, D., (2021) A Frustratingly Easy Approach for Entity and Relation Extraction. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics. 2021, 50-61.

17. Madry, A., Makelov, A., Schmidt, L., et al. (2018) Towards deep learning models resistant to adversarial attacks. In: *Proceedings of the 6th International Conference on Learning Representations*. Vancouver, Canada: ICLR. 2018.