TTI-COIN at BioCreative VIII Track 1

Takuma Matsubara *, Taku Oi *, Ryuki Ida, Miwa Makoto, and Yutaka Sasaki

Toyota Technological Institute, Nagoya, Aichi, Japan

E-mail: {sd23439,sd23404,sd22401,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

*Equal contribution

Abstract

We built two neural-network based methods that use external data for NER and RE. For NER, We aimed to learn using multiple existing datasets. We propose Conditional VAE (CVAE) with conditions to create slightly different span representations for each dataset. For RE, we constructed a model that integrates the representations of the entities acquired from the neighborhood knowledge graphs, which are subgraphs around the entities, and the representations of the input document.

Introduction

Information extraction from biomedical documents is attracting attention for (semi-)automatic compilation of medical information from a large number of articles. This paper reports our participation in the BioCreative VIII Track 1, which focuses on Named Entity Recognition (NER), Entity Linking (EL), and Relation Extraction (RE) from biomedical articles. There are two subtasks in this track. Subtask-1 consists of RE, and Subtask-2 consists of NER, EL and RE. We took two different approaches that investigated using external data for NER and RE.

For the NER part, we focused on the use of publicly available labeled corpora. In the biomedical domain, there are many labeled corpora, partially thanks to many community-based activities such as BioCreative. However, each dataset has its own annotation target and definition, so learning by simply combining several labeled datasets can be problematic. Specifically, datasets have different types of labels and different terms labeled with different criteria, making it difficult to treat them as compatible data.

Therefore, we considered a way to include the differences among datasets in the span representation. Nguyen et al. (1) predicted the probability distribution of each label from the expression vectors of NE candidate spans and showed that reconstruction from these vectors and loss due to synonym generation improved classification performance. We aim at corpus-specific representation of spans by including the condition of which corpus is used for reconstruction in VAE.

For the RE part, we focused on the advantages of rich information in the databases, such as a wide range of relationships between entities that do not appear in the corpus.

Methods

Named Entity Recognition

We aimed to improve the NER performance by effectively using existing labeled datasets while alleviating differences in labels in multiple datasets. To alleviate label differences, we incorporated CVAE into NER, which compresses the span representation after concatenating a one-hot vector representing the gold label of the span. The overall picture of the proposed model is shown in Figure 1.

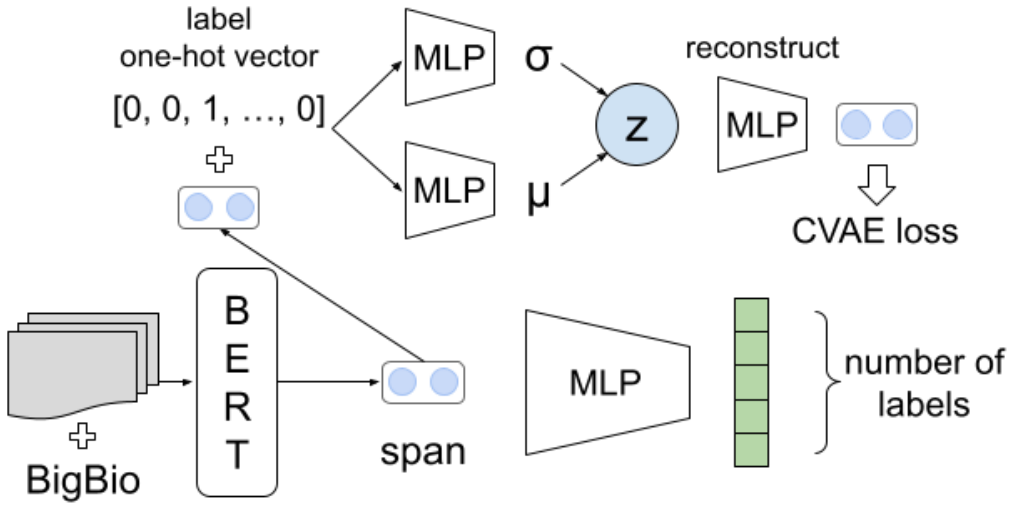

Figure1. Overview of NER model

a. Following Zhong et al. (2), we encoded the sentences to be predicted with a pre-trained BERT model and built the representation of a span of n words by combining the representations of the words at each end of the span and the length of the span as shown in the following equation.

$$h_{span} = Concat(x_1, x_n, \phi(n))$$

where Concat(-) is a vector concatenation and Φ(n) is the embedded representation for span length n.

b. The resulting span representation is classified into its corresponding entity type through a two-layer fully-connected layer with a ReLU activation function and a softmax layer. When an instance of a corpus is classified, the values corresponding to the types of the other corpora are masked before the softmax layer so that only the labels of the corpus are predicted. Cross Entropy Loss (L_{CE}) is employed for the classification loss. In addition to the classification, CVAE compresses the representation of the span after concatenating a one-hot vector representing the gold label of the span. Specifically, the mean (μ) and variance (σ) for the span are calculated from the concatenated representation of the span and the one-hot vector of the label through two corresponding Multi-Layer Perceptrons (MLPs). A vector z sampled from the distribution of the mean and variance is concatenated with the one-hot vector, and it is used to reconstruct the original span representation through a two-layer fully-connected layer. The mean squared error between the original and reconstructed span representations and the Kullback Leibler (KL) divergence, which

makes the predicted distribution closer to a Gaussian distribution, are added to the classification loss with a weight as the total loss. AdamW (5) is used as the optimization method.

$$L = \alpha L_{CE} + \left(h_{span} - z\right)^2 + KL\left[q\left(z|h_{span}\right)||p(z)\right]$$

Linking

We linked only entities recognized as chemical, disease, and gene in the named entity recognition. First, we performed an exact match-based linking with a database. Next, we employed an off-the-shelf linking model (3) trained on the NLM-Chem dataset for the unlinked chemical and disease entities. The model maps an entity mention to the corresponding concept ID in the MeSH thesaurus.

Relation Extraction

Our model is based on (4). The model integrates the representations of the entities acquired from the neighborhood knowledge graphs, which are subgraphs around the entities, and the representations of the input document. We perform relation extraction by integrating entity information of the neighborhood knowledge graphs calculated with GCN into the text. Unlike (4), we employ the BERT-based relation extraction model in ATLOP (5) for the base relation extraction model. Specifically, we obtain the representation of each entity through Localized Context Pooling (LOP), which is part of ATLOP. We classified entity pairs into relation types as well as novel or not.

GCN and BERT are trained in an end-to-end manner using Adam (7) as the optimization method and the cross entropy loss.

Experimental Settings

BERT was initialized with PubMedBERT (8) for named entity recognition and BioLM (9) for relation extraction. Hyperparameter tuning was performed using Optuna.

Results and Discussion

Named Entity Recognition

For NER, the results are shown in Table 1. We prepared 41 additional datasets that were available from the BigBio (10) dataset, which includes 126 biomedical NLP datasets. Available datasets are those datasets for English NER, where training data are currently available to download from huggingface. We trained a model on the BioRED (11) training dataset by adding each dataset from BigBio. Experiments were conducted five times with different seed values for adding each dataset. BioRED in Table 1 refers to the model trained only with BioRED. The top 1 shows the results of the model that was best on the development data among all the trained models. Furthermore, for the top k models on the development set, voting was employed to obtain the final prediction. The voting threshold was chosen based on the results submitted to CodaLab. For the top 10, spans predicted by 60% of the models were chosen as the final prediction, while for the top 20 and 40, spans predicted by 70% of the models were chosen.

Table 1: Subtask2 NER results on the test set. Top k shows the result of an ensemble with k models trained by adding top k datasets to BioRED.

datasets	All (P/R/F [%])	Gene (F [%])	Disease (F [%])	Chemical (F [%])	Species (F [%])	Cellline (F [%])	Variant (F [%])
BioRED	85.89/87.47/86.67	88.20	86.18	84.11	89.51	66.12	85.66
Top 1	87.14/87.41/87.28	88.14	87.48	84.29	90.68	69.32	86.92
Top 10	87.36/87.33/87.34	88.18	87.46	84.32	90.90	70.63	87.08
Top 20	87.42/87.22/87.32	88.17	87.34	84.29	90.96	70.63	87.10
Top 40	87.42/87.20/87.31	88.18	87.33	84.30	90.89	70.63	87.10
Average	80.38/74.48/76.87	79.69	72.85	73.19	83.80	67.77	74.21
Median	83.35/74.33/78.58	83.55	74.02	74.97	86.35	67.19	87.42

Linking

For linking, the results are shown in Table 2.

Table 2: Subtask2 Linking results on the test set. The first line indicates the type of dataset used in NER, and all linking methods are the same.

datasets	BioRED	Top1	Top10	Top20	Top40
P/R/F [%]	46.03/39.79/42.68	45.77/39.97/42.67	45.80/39.93/42.66	45.83/39.91/42.66	45.83/39.89/42.65

Relation Extraction

Table 3, 4, and 5 show the performances of our RE model.

Table 3 shows the RE results on the BioRED test set, not the test set for this task. The proposed model (+NKG) improved the micro-averaged F-score by 1.9 percentage points compared to the baseline (BioLM). This result indicates that the information of neighborhood knowledge graphs can improve prediction performance and that the relationship extraction can take the knowledge graph information into account.

We also incorporated our neighborhood KG into the BioLM+LOP model, and it improved the performance.

Table 3: Subtask1 RE results on the original dataset, not the test set for this task

Model	Entity Pair (P/R/F [%])	Entity Pair + Relation Type (P/R/F [%])	Entity Pair + Novelty (P/R/F [%])	Entity Pair + Relation Type + Novelty (P/R/F [%])
BioLM	72.20/70.35/71.26	57.65/49.20/53.09	55.29/49.03/51.97	36.11/44.02/39.67
BioLM+NKG	73.70/71.38/72.52	60.92/52.55/56.43	56.39/52.87/54.57	37.04/46.03/41.05
BioLM+NKG+LOP	74.45/71.78/73.09	63.41/56.02/59.49	56.42/53.09/54.70	38.87/47.92/42.92

Table 4: Subtask1 RE results on the task dataset. The value in the parentheses on the "Model" column indicates the seed value.

Model	Entity Pair (P/R/F [%])	Entity Pair + Relation Type (P/R/F [%])	Entity Pair + Novelty (P/R/F [%])	Entity Pair + Relation Type + Novelty (P/R/F [%])
BioLM+NKG+LOP (42)	64.28/44.61/52.67	45.48/31.56/37.26	48.16/33.42/39.46	34.52/23.96/28.29
BioLM+NKG+LOP (43)	68.91/40.69/51.17	51.19/30.23/38.02	50.86/30.03/37.77	38.18/22.55/28.35
BioLM+NKG+LOP (44)	67.38/42.49/52.11	48.71/30.72/37.67	49.45/31.18/38.24	36.23/22.85/28.02
Average	69.22/68.60/67.03	49.01/48.39/47.74	50.92/50.02/49.23	36.15/35.73/35.22
Median	77.93/69.65/73.56	51.64/54.79/53.17	52.97/60.42/56.45	41.61/39.88/40.73

Table 5: Subtask2 RE results on the task dataset. The RE model is the same for all the settings.

Model	Entity Pair (P/R/F [%])	Entity Pair + Relation Type (P/R/F [%])	Entity Pair + Novelty (P/R/F [%])	Entity Pair + Relation Type + Novelty (P/R/F [%])
NER (BioRED) + NKG	15.95/3.02/5.08	11.66/2.21/3.71	12.27/2.32/3.91	8.94/1.69/2.85
NER (top1) + NKG	16.64/3.32/5.54	11.31/2.26/3.76	12.90/2.57/4.29	8.82/1.76/2.93
NER (top10) + NKG	16.19/3.22/5.37	11.10/2.21/3.68	12.77/2.54/4.24	8.68/1.73/2.88
NER (top20) + NKG	16.53/3.30/5.51	11.21/2.24//3.74	12.87/2.57/4.29	8.80/1.76/2.93
NER (top40) + NKG	16.51/3.30/5.51	11.20/2.24/3.74	12.86/2.57/4.29	8.80/1.76/2.93
Average	34.14/26.48/28.62	25.13/19.87/21.39	25.83/20.23/21.82	19.00/15.12/16.25
Median	30.14/40.26/34.47	22.15/29.75/25.40	23.29/31.50/26.78	17.18/23.33/19.79

References

1. Nguyen NTH, Miwa M, Ananiadou S. Span-based Named Entity Recognition by Generating and Compressing Information. In: Vlachos A, Augenstein I, eds. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. May 2023.

2. Zhong, Z., and Chen, D. (2021) A Frustratingly Easy Approach for Entity and Relation Extraction. In Association for Computational Linguistics. pp. 50-61

3. Tsujimura, T., Miwa, M., and Sasaki, Y. (2023). Large-scale neural biomedical entity linking with layer overwriting. Journal of Biomedical Informatics, 143, 104433.

4. Matsubara, T., Miwa, M., and Sasaki, Y. (2023). Distantly Supervised Document-Level Biomedical Relation Extraction with Neighborhood Knowledge Graphs. In *the 22nd Workshop on* Biomedical Natural Language Processing and BioNLP Shared Tasks (pp. 363-368). Association for Computational Linguistics.

5. Zhou, W., Huang, K., Ma, T., & Huang, J. (2021). Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling. In Proceedings of the AAAI Conference on Artificial Intelligence.

6. Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. In International Conference on Learning Representations.

7. Kingma, D., and Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference for Learning Representations.

8. Lewis, P., Ott, M., Du, J., & Stoyanov, V. (2020). Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In Proceedings of the 3rd Clinical Natural Language Processing Workshop (pp. 146-157). Association for Computational Linguistics.

9. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans. Comput. Healthcare, 3(1), Article 2.

10. Fries, J. A., et al. (2022) BigBio: A Framework for Data-Centric Biomedical Natural Language Processing. Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

11. Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, Zhiyong Lu. (2022). BioRED: a rich biomedical relation extraction dataset, Briefings in Bioinformatics, Volume 23, Issue 5, bbac282,