Using Knowledge Base to Refine Data Augmentation for Biomedical Relation Extraction

KU-AZ team at the BioCreative 7 DrugProt challenge

Wonjin Yoon [1], Sean Yi [1], Richard Jackson [2], Hyunjae Kim [1], Sunkyu Kim [1] and Jaewoo Kang [1,3]
1: College of Informatics, Korea University, Seoul, South Korea
2: AstraZeneca UK, Cambridge, United Kingdom
3: AIGEN Sciences, Seoul, South Korea

***Abstract*— This paper describes our participation in the BioCreative7 DrugProt challenge. We augmented the DrugProt dataset by predicting labels with transformer models and built a large-scale dataset to expose our model to diverse relation expression patterns. To alleviate the problem of noise inherited to the augmented dataset from the original dataset, we utilized a knowledge base to refine the augmented data points. Our experimental results on the development dataset and the result on the large track test dataset showed that models pre-trained on our augmented dataset produce slightly more accurate predictions. The effects of pretraining models on the augmented dataset varied between relationship types. Performances on rare types (i.e. relation types with smaller populations in the training dataset) benefitted more from the data augmentation method, and recall seemed to improve more than precision.**

Keywords— Biomedical Relation Extraction; Data Augmentation; Knowledge Base; Transformer; Language Models (*key words****)***

I. INTRODUCTION

Biomedical Relation extraction (BioRE) is a task of extracting and classifying relations between two biomedical entities in biomedical literature. The BioRE task can be used for real-world applications such as discovering Mechanism of Action for a drug, or building drug-drug interaction databases. Recent works have discovered effective methods to extract biomedical relations by harnessing large-scale biomedical language models (LMs) (1-3, 14). However, improving the model remains a challenge due to the relatively small sizes of datasets and limited selection for benchmark datasets.

Biomedical NLP (BioNLP) datasets should be annotated by domain experts, making it expensive and difficult to build a high-quality large-scale BioNLP dataset. To alleviate the problem of data scarcity in BioNLP datasets, researchers introduced strategies such as: utilizing large unlabeled data by training a model using self-supervised learning; creating distantly supervised datasets with domain-specific knowledge bases (KBs) (4, 5, 13); and utilizing model-generated labels to augment the existing dataset (6).

Distantly labeling biomedical relations using KBs is one alternative to make a RE dataset without manual annotations. Utilizing KBs enables the scalable automatic construction but has a drawback of producing considerably noisy labels. The noise of distantly labeled RE datasets mostly comes from the distant supervision assumption: if two entities listed as having a relation in KB exist in a sentence, then that sentence is labeled as having that relation. The assumption has the potential to produce erroneous labels: a sentence with relation can be labeled as negative example if the relation is not listed in KB (False Negative); and the sentence will be labeled as positive example if two entities exist in a same sentence, but the sentence does not have any clue of relation (False Positive).

Augmented datasets using model predictions as pseudo labels are created by training a model on small high-quality dataset and can generate large-scale dataset with extended relation patterns. However, labeling bias noises in the original dataset inherits to the augmented datasets and the limited coverage of knowledge in the original dataset adds noises to the augmented datasets (6). Jiang et al. (6) pointed the shortcomings of model-generated labels and proposed Noise-Aware Continual pre-training for biomedical NER.

To address these issues, we suggest a pipeline for BioRE that utilizes a KB to filter model-generated labels. Our pipeline consists of three phases: building a large-scale augmented dataset; pre-training a transformer model using the augmented dataset; and finally fine-tuning the model with the original human-annotated dataset (such as DrugProt (15) or ChemProt (16)). For the first phase of building an augmented dataset, we train a model with DrugProt dataset and predicted labels for the sentences in selected MEDLINE articles. Predicted labels are then compared with triples in the KB and are dropped if the labels from two sources disagree. For the subsequent phases, we trained a transformer-based sequence classification model on the augmented dataset and transfer the trained model weights to fine tune the model with original dataset, DrugProt.

This work was done while Wonjin Yoon worked under the Research Collaboration project at AstraZeneca.

This work is supported by National Research Foundation of Korea (NRF-2020R1A2C3010638, NRF-2014M3C9A3063541), ICT Creative Consilience program (IITP-2021-0-01819) and Korea Health Technology R&D Project (grant number: HR20C0021) through the KHIDI, funded by the Korea government

II. Methods and Experimental Settings

In this section, we describe elements in our pipeline for BioRE. We first describe methods for our first phase of our pipeline, building a large-scale augmented dataset (Figure 1). Then, we describe our sequence classification model that classifies the type of the relation for an entity pair in the given sentence. (Figure 2). Input sequences for the experiments throughout this paper are in the sentence level format with entities annotated in the sequence. Figure 3 illustrates the preprocessing steps.

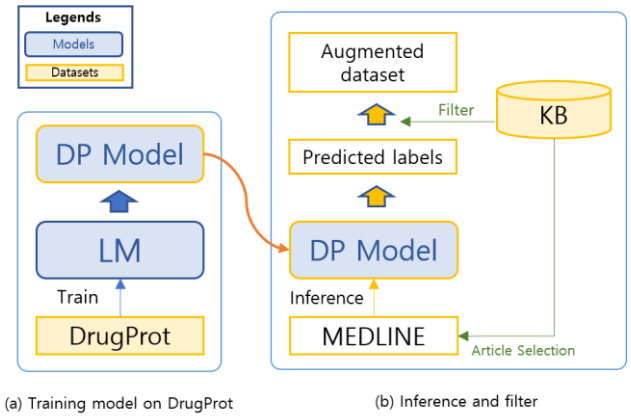

Figure 1 Pipeline for building a large-scale augmented dataset

A. *Building a large-scale augmented dataset*

Training a model that can generalize to diverse relation expression patterns is a challenging task and a key factor to get a robust model. In order to expose our model to diverse patterns, we augmented the dataset using the outputs of the model and refined the generated data points (i.e. input sequence and a label) with KB.

We selected Chemical–gene interactions dump from Comparative Toxicogenomics Database (CTD) (7) as our KB source. The steps for preparing source (unlabeled) dataset for the augmented dataset is analogous to the pre-processing steps for the DrugProt dataset (Figure 3). First, we selected articles that appear as reference articles in the database. We collected abstracts of the articles from MEDLINE using PMID and split the abstracts into sentences using the Stanza library (8, 9). Gene and Chemical type entities within the sentences are automatically recognized with BERN, a BioBERT based online NER tool (10). We have also utilized the dictionaries of the KB and external source (HGNC gene set (11)) for additional steps to detect entities. Finally, we attached entity markers at the beginning and end of the entity.

We first trained a sequence classification model with the DrugProt dataset and used the trained model to predict the potential label for the unlabeled dataset. A KB was used to check the validity of predicted labels. Since the labeling schema for the dataset and KB is different, we used KB to only check the presence of the relation between two entities. For example if a prediction for a sequence indicates that there is a relation between the entity pair in the sequence, the predicted label will only be used for the augmented dataset if the entity pair is listed as having a relation in the KB. Sequences without agreement are dropped. Although this filtering may drop potentially useful samples, we believe that reducing noise is much more important.

B. *Sequence Classification Model*

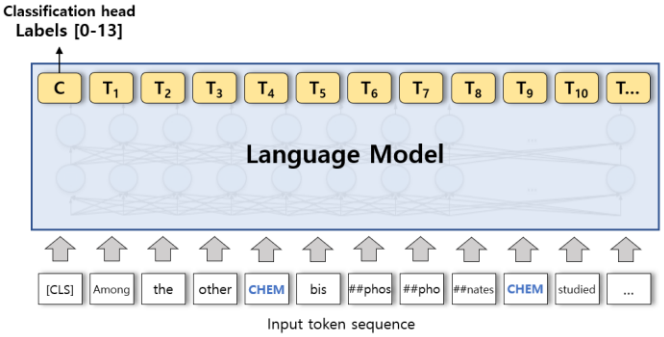

Figure 2 Sequence classification model. Output of [CLS] token is used for the classification head.

Our sequence classification model applied a straightforward method. Following BERT (12) and BioBERT (1), we used the output of special token, [CLS], as the sentence representation for the classification head. A linear classification layer was used. Different from the experiments of BioBERT, we did not anonymize entities and used different entity markers for entity types and for the start and end of the entity. We registered 4 entity markers to the vocabulary file. We discovered through ablation experiments that using independent entity markers, registering markers and non-anonymized entities showed the best performance, showing about 0.5~1 percent performance gain for each element.

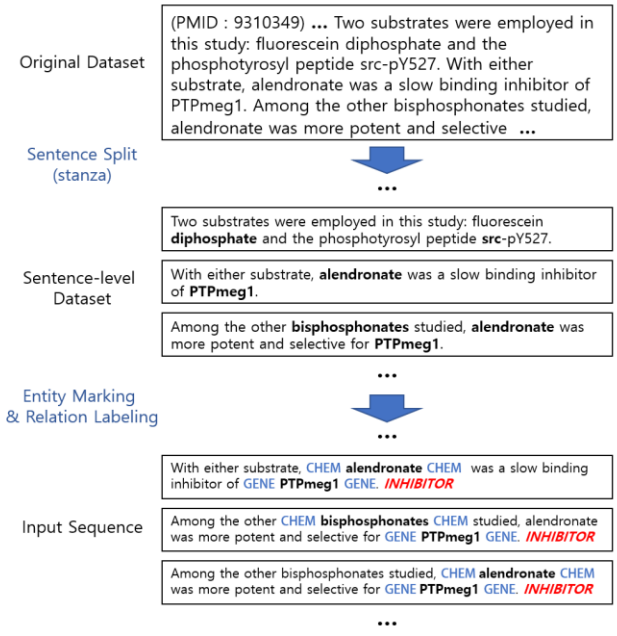

Figure 3 Pre-processing steps. Entity markers wrap entities in input sequences

C. Datasets

Table 1 shows the statistics for the DrugProt dataset (15) and the augmented dataset. We were able to build an augmented dataset about 13 times larger than the original DrugProt dataset.

Among the 1,262,404 sentences we predicted, 387,054 sentences (about 30% of the prediction) were dropped due to the disagreement between the model prediction and the KB. Pre-processing steps and the prediction steps took less than 12 hours using one CPU (16 cores) and one GPU (TITAN RTX 24GB).

TABLE I. STATISTICS OF THE DATASETS

Dataset	The number of input sequences	
	Train	Development
Original (DrugProt)	64,779	13,480
Augmented dataset	875,350	-

TABLE II. STATISTICS OF THE RELATION TYPES

Type	Train	Development
INHIBITOR	5392 (31.3%)	1152 (30.6%)
DIRECT-REGULATOR	2250 (13.0%)	458 (12.2%)
SUBSTRATE	2003 (11.6%)	495 (13.2%)
ACTIVATOR	1429 (8.3%)	246 (6.5%)
INDIRECT-UPREGULATOR	1379 (8.0%)	302 (8.0%)
INDIRECT-DOWNREGULATOR	1330 (7.7%)	332 (8.8%)
ANTAGONIST	972 (5.6%)	218 (5.8%)
PRODUCT-OF	921 (5.3%)	158 (4.2%)
PART-OF	886 (5.1%)	258 (6.9%)
AGONIST	659 (3.8%)	131 (3.5%)
AGONIST-ACTIVATOR	29 (0.2%)	10 (0.3%)
SUBSTRATE_PRODUCT-OF	25 (0.1%)	3 (0.1%)
AGONIST-INHIBITOR	13 (0.1%)	2 (0.1%)

Fig. 1. Statistics of the datasets Table 1 shows the number of dat points for the original dataset and augmented dataset. Table 2 shows the number of data points and its proportion for each class.

III. RESULTS AND DISCUSSION

In this section, we discuss our experimental results for both development dataset and the test data submission for the challenge. For the test dataset, we report scores received from the challenge organizers.

A. Results on the development dataset

Table 3 shows the performance of our models on the DrugProt development dataset. Since the performance variance for the model was high, we report statistics of 10 independent runs using different random seeds. Scores in Table 2 are the average of the performance of 10 runs. We saved the checkpoints with 2k, 4k or 10k intervals and evaluated the saved checkpoints to find the best models.

Models initiated from BioLM (14) large weights (RoBERTa structure) showed better performance than PubMedBERT (2) weights. For *3-RoBERTaLarge CTD* model, the model is transferred from the BioLM model trained on the augmented data. We pre-trained the model on the augmented data for 70,000 steps (with mini batch size of 32). Note that we need fewer training steps for the transferred model as the models are expected to learn common knowledges about RE the task from the augmented dataset and the final fine tuning steps are for fine-grading the models with high-quality dataset.

A mini batch size of 16 was used for training the models on DrugProt dataset. In our experiment with the ensemble method on develop set, we observed an increase of 1.5 percent (F1), showing 0.789 and 0.795 for *1-RoBERTaLarge* model and *3-RoBERTaLarge_CTD* model respectively (10 models ensembled).

TABLE III. PERFORMANCE OF THE MODELS ON DEVELOPMENT DATASET

Submission name (Main track)	Settings	Performance (Develop)		
	Model	Training Steps	F1 (Avg)	F1 (Std)
-	PubMedBERT (2)	40K (~10epoch)	0.7721	0.005
1-RoBERTaLarge	BioLM Large (14)	110k (~28epoch)	0.7746	0.003
3-RoBERTaLarge CTD	BioLM Large (pretrained on augmented dataset)	4k (~0.1epoch)	0.7864	0.003

[a.] Best performance across the training steps

Fig. 2. Performance of the models on the development dataset. We report statistics of 10 independent runs using different random seeds.

B. Results on the test data submission

For the main track submission, we used various ensembles of models. For runID 1 and 2, we trained 10 models without the augmented dataset. For runID 3 and 4, we trained 5 models, which were also pretrained on the augmented dataset.

We only used 5 models for the submissions using augmented data due to lack of time to train the models. This may have lead to a relatively lower performance of our augmented dataset models. The fact that augmented dataset models perform better than the plain model for the Large track, where we submitted predictions of equal condition (i.e. single models), supports our assumption.

Another method we used for the main track was to utilize the development dataset for training our model. We first merged the training dataset and development dataset and split the merged dataset into 10 partitions. We created 10 reorganized train-develop dataset pairs, each has one partition as a new dev set and 9 partitions as the new training set. We trained 10 models with the pairs. Since we were using ensemble method, we assume that the ensembled model is trained on knowledge of both train and develop datasets. Submission names under *2-RoBERTaLarge-10-traindev* (runID 2) and *4-RoBERTaLarge_CTD-5-traindev* (runID 4) were

trained using this strategy, and showed superior performance than models only trained on the training dataset. This would suggest that increasing the size of the training set further still would yield additional gains.

5-Best-CTD (runID 5) was an ensemble of 10 checkpoints across all settings and steps that showed the best performance for the develop dataset.

For the large track submission, we were not able to use ensemble methods due to the computational cost. We submitted the predictions of single models and the pre-processing and the prediction took 9 hours with 16 distributed GPUs (each with at least 24 GB GPU memory). *1-BioLM-CTD-lr1e5-filter* and *2-BioLM-CTD-lr5e6-filter* are models pre-trained with augmented data and *3-BioLM-lr2e5-filter* are trained without augmented data. The difference between runID 1 and 2 is the learning rate during training.

TABLE IV. PERFORMANCE OF THE MODELS ON MAIN TRACK TEST DATA

runID	Settings			Performance (Test)			
	Aug.	Data	#Mod.	F1 % (All)	F1 % (INHI)	F1 % (DIR)	F1 % (SUBS)
1	X	T	10	78.53	87.45	69.75	67.94
2	X	**T+D**	10	78.93	87.78	66.74	68.27
3	**O**	T	5	78.38	87.44	67.85	68.78
4	**O**	**T+D**	5	78.61	87.51	69.07	69.25
5	**O**	T	10	78.36	86.96	67.72	70.25

TABLE V. RECALL PERFORMANCE OF THE MODELS ON MAIN TRACK

runID	Performance (Develop)				Performance (Test)			
	Rec % (All)	Rec % (INHI)	Rec % (DIR)	Rec % (SUBS)	Rec % (All)	Rec % (INHI)	Rec % (DIR)	Rec % (SUBS)
1	77.50	86.08	61.21	76.14	78.13	86.86	66.66	63.48
2	-	-	-	-	78.16	88.20	62.47	64.20
3	78.70	88.50	63.58	77.66	79.08	87.82	66.66	67.06
4	-	-	-	-	77.93	88.01	64.56	66.10
5	-	-	-	-	78.08	86.96	64.56	67.06

TABLE VI. PERFORMANCE OF THE MODELS ON LARGE TRACK TEST DATA

runID	Settings		Performance (Test)			
	Aug.	Data	F1 % (All)	F1 % (INHI)	F1 % (DIR)	F1 % (SUBS)
1	**O**	T	75.76	85.22	66.98	65.42
2	**O**	T	75.82	84.66	66.91	64.07
3	X	T	75.18	84.76	65.74	62.98

Fig. 3. Performance of the model on test data (evaluated by the organizers). Base model (Language model) for all submissions was BioLM Large. Aug. in the column denotes that the model is pretrained on augmented dataset. #Mod. denotes the number of models used for the ensemble method. T and T+D in the Data column denotes Training data and Training + Develop data respectively. F1(INHI), F1(DIR), F1(SUBS) denotes F1 score for INHIBITOR, DIRECT-REGULATOR and SUBSTRATE, which are the most abundant types of relation.

Table 4 and Table 5 show the performance of our models for the main track of the challenge. Overall performances (F1 % (All)) of models pre-trained on augmented datasets (runID 2, 4) are slightly lower than models without augmented dataset (runID 1, 3). This suggests that our augmentation strategy was not wholly effective. However, as illustrated in Table 5, pre-training models on the augmented dataset showed constant performance gain in recall for development dataset, and test dataset. This may be due to exposing the model to more relationship patterns than are represented in the training data alone. Please note that for runID 2 and 4, we could not report the performance for the development dataset, as it was used for training.

Augmented models performs better than the unaugmented equivalent models for most relationship types, except the most common INHIBITOR type. This INHIBITOR type constitutes a large portion of the total dataset (about 30%), perhaps suggesting that it is well enough represented already, and augmentation is a hindrance.

We assume that pre-training the model on augmented dataset is beneficial for predicting relations with underrepresented training examples. However, our model did not show better performance for the very poorly represented relation types (i.e. less than 5% of the total dataset). All of our approaches struggled to learn these relationships – suggesting that a minimum number of examples of these types needs to be present in the training data for a machine learning strategy to be viable. Discovering this minimum example number remains a topic of further work.

Similar results are observed for the large track submission (Table 6). Our augmented models (runID 1, 2) were more robust than the plain model (runID 3) on scarce relation types. The results for the large track were lower than the expectation based on our experiments on the development dataset. We suspect a technical issues on our pre-processing pipeline as a main cause of the suboptimal performance for predicting at a scale. Further work will investigate this performance degradation.

REFERENCES

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

2. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. arXiv preprint arXiv:2007.15779.

3. Shin, H. C., Zhang, Y., Bakhturina, E., Puri, R., Patwary, M., Shoeybi, M., & Mani, R. (2020, November). Bio-Megatron: Larger Biomedical Domain Language Model. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4700-4706)..

4. Peng, Y., Wei, C. H., & Lu, Z. (2016). Improving chemical disease relation extraction with rich features and weakly labeled data. Journal of cheminformatics, 8(1), 1-12.

5. Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinformatics. 2015;16:55. Published 2015 Feb 21. doi:10.1186/s12859-015-0472-9

6. Jiang, H., Zhang, D., Cao, T., Yin, B., & Zhao, T. (2021). Named Entity Recognition with Small Strongly Labeled and Large Weakly Labeled Data. arXiv preprint arXiv:2106.08977.

7. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, Mattingly CJ The Comparative Toxicogenomics Database: update 2021. Nucleic Acids Res. 2020 Oct 17.

8. Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020.

9. Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D. Manning, Curtis P. Langlotz. Biomedical and Clinical English Model Packages in the Stanza Python NLP Library, Journal of the American Medical Informatics Association. 2021.

10. Kim, D., Lee, J., So, C. H., Jeon, H., Jeong, M., Choi, Y., ... & Kang, J. (2019). A neural named entity recognition and multi-type normalization tool for biomedical text mining. IEEE Access, 7, 73729-73740.

11. Tweedie S, Braschi B, Gray KA, Jones TEM, Seal RL, Yates B, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res. PMID: 33152070 PMCID: PMC7779007 DOI: 10.1093/nar/gkaa980

12. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.. In J. Burstein, C. Doran & T. Solorio (eds.), NAACL-HLT (1) (p./pp. 4171-4186), : Association for Computational Linguistics. ISBN: 978-1-950737-13-0

13. Verga, P., Strubell, E., & McCallum, A. (2018, Junie). Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 872–884. doi:10.18653/v1/N18-1080

14. Lewis, P., Ott, M., Du, J., & Stoyanov, V. (2020, November). Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. Proceedings of the 3rd Clinical Natural Language Processing Workshop, 146–157. doi:10.18653/v1/2020.clinicalnlp-1.17

15. Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia and Martin Krallinger. 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. 2021.

16. Krallinger, M., Rabal, O., Akhondi, S. A., Pérez, M. P., Santamaría, J., Rodríguez, G. P., ... & Intxaurrondo, A. (2017, October). Overview of the BioCreative VI chemical-protein interaction Track. In Proceedings of the sixth BioCreative challenge evaluation workshop (Vol. 1, pp. 141-146).