Improving Supervised Drug-Protein Relation Extraction with Distantly Supervised Models

Naoki Iinuma, Makoto Miwa and **Yutaka Sasaki**
Computational Intelligence Laboratory
Toyota Technological Institute
2-12-1 Hisakata, Tempaku-ku, Nagoya, Aichi, 468-8511, Japan
inaoki2628@gmail.com
{makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

Abstract

This paper proposes novel drug-protein relation extraction models that indirectly utilize distant supervision data. Concretely, instead of adding distant supervision data to the manually annotated training data, our models incorporate distantly supervised models that are relation extraction models trained with distant supervision data. Distantly supervised learning has been proposed to generate a large amount of pseudo-training data at low cost. However, there is still a problem of low prediction performance due to the inclusion of mislabeled data. Therefore, several methods have been proposed to suppress the effects of noisy cases by utilizing some manually annotated training data. However, their performance is lower than that of supervised learning on manually annotated data because mislabeled data that cannot be fully suppressed becomes noise when training the model. To overcome this issue, our methods indirectly utilize distant supervision data with manually annotated training data. The experimental results on the DrugProt corpus in the BioCreative VII Track 1 showed that our proposed model can consistently improve the supervised models in different settings.

1 Introduction

Drug-protein relations are important for drug discovery, metabolic, and drug response modeling, and their textual evidence is important in the development of evidence-based medicine. However, since drug-protein interactions are reported in the literature and the number of relevant articles is rapidly increasing (Coordinators, 2016), it is difficult for pharmacologists to read every single article to determine the interactions. Therefore, automatic interaction extraction from text has attracted much attention. The related shared tasks (Krallinger et al., 2021, 2017) are being conducted at BioCreative, an international workshop that aims to evaluate text mining and information extraction in the biological domain.

For drug-protein relation extraction, models using deep learning have achieved high performance. A typical deep learning model takes as input a sentence and the drug and protein mentions in the sentence, and predicts the relationship between the drug and the protein as expressed in the sentence. Gu et al. (2022) extracted the relationships using a large neural network model pretrained on a large biomedical literature (PubMedBERT). Deep learning models suffer from the problem of the huge cost of manually annotated training data.

A distantly supervised learning method has been proposed by Mints et al. (2009). The method enables the creation of a large amount of training data at low cost. However, this method still has the problem of producing data with incorrect labels, which become noise during training. Several methods have been proposed to mitigate the effects of such noisy examples. One of the most commonly-used methods is multi-instance learning (Riedel et al., 2010), where the distant supervision data is treated as a bag of instances corresponding to pairs in the database. Zeng et al. (2015) proposed a method to train instances with the representation with the highest prediction probability of the target label in the bag. Ji et al. (2017) proposed a method to weight instances in the bag so that correctly labeled instances will have large weights while noisy cases have small weights. Beltagy et al. (2019a) proposed a method of learning with distant supervision data by utilizing some of manually annotated training data to learn the weights. Although such methods show performance improvement in the distantly supervised training setting, the performance is still lower than that of the methods trained on manually annotated training data.

This study proposes a novel method of using distantly supervised relation extraction models for supervised drug-protein relation extraction. By using the model trained over the easy-to-create distant supervision data, we aim to improve the performance

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 161–170

of supervised drug-protein relation extraction while reducing the cost of building additional manually annotated data and the effect of noisy instances in the distant supervision data.

Our contributions are as follows:

1. We generate distant supervision data for drug-protein relation extraction from domain databases. By utilizing four databases, we create distant supervision data of the same scale as that of general domain distant supervision data.

2. We propose to utilizing representations obtained from a distantly supervised model for ordinary supervised training. The performance in extracting relations between drugs and proteins was consistently improved for two models (i.e., PubMedBERT and BioRoBERTa-large (Lewis et al., 2020)) with different parameter sizes.

3. The proposed method showed consistent performance improvement regardless of the data size of the manually annotated training data, indicating that it is effective for utilizing distantly supervised model to improve the extraction performance.

2 Methods

We propose a novel method for extracting drug-protein relations from manually annotated training data. The method uses a model trained on distant supervision data, which we call a *distantly supervised model*. By utilizing the distantly supervised model, we aim to improve the extraction performance while reducing the influence of noisy instances included in the distant supervision data.

In the following sections, we will explain the baseline relation extraction model in Section 2.1, the construction of distant supervision data from databases in Section 2.2, and the methods for utilizing the distantly supervised model in Section 2.3.

2.1 Relation Extraction Model

We describe a supervised relation extraction model that is used as the baseline in this research. The model predicts the relation for a given entity pair from the input sentence.

First, the mentions of target drug and protein in the input sentence are masked with *"DRUG"* and *"PROTEIN"*, respectively. Table 1 shows an example of this preprocessing. The sentence contains three drug mentions (*androstenedione*, *oestrone*, *oestrone*) and one protein mention (*aromatase*), so three drug-protein pairs are created and their mentions are replaced.

Next, the input sentence with the target protein and drug entities is encoded with BERT (Devlin et al., 2019) to generate a feature representation vector $\boldsymbol{h}$ that represents the input sentence. For this vector, we use the representation vector of the [CLS] token since it contains the features of the whole sentence in BERT. Finally, based on the feature representation vector, the model then generates a prediction vector that represents the prediction probability for each relation by using one fully-connected layer and the softmax function. The model predicts the relation that has the maximum prediction probability. The optimizer is Adam (Kingma and Ba, 2015), and the model is trained to minimize the cross-entropy loss.

2.2 Building Distant Supervision Data

An overview of the process of building distant supervision data is shown in Figure 1. In this method, we use a medical literature database PubMed (Coordinators, 2016), a drug database DrugBank (DS et al., 2018), a protein database UniProt (Consortium, 2020), and a chemical substance database Comparative Toxicogenomics Database (CTD) (Davis et al., 2020). From these databases, we extract about 33 million articles, about 500 thousand drug entries, and about 570 thousand protein entries to create distant supervision data. In the following, we explain the process of building distant supervision data using these databases.

First, drug and protein entities are extracted from the medical literature in PubMed, as shown in Figure 1-(i). Sentence segmentation and entity extraction modules in SciSpacy (Neumann et al., 2019), a tool specialized for processing biomedical and scientific literature, are used to analyze the medical literature and extract drug entities and protein entities as named entities in the literature.

Next, we create relational triples as shown in Figure 1-(ii). ID relation triples are extracted from DrugBank. Here, an ID relation triple is a triple of drug ID, relation name, and protein ID. We create relation triples from the ID relation triples by mapping the IDs to their names using drug and protein name dictionaries. The drug name dictionary is created by mapping drug IDs to drug names and its synonyms on the information in DrugBank

Target drug	Target protein	Preprocessed input sentence
androstenedione	*aromatase*	*The **PROTEIN** enzyme, which converts **DRUG** to oestrone, regulates the availability of oestrogen so support the growth of hormone-dependent beast tumours.*
oestrone	*aromatase*	*The **PROTEIN** enzyme, which converts androstenedione to **DRUG**, regulates the availability of oestrogen so support the growth of hormone-dependent beast tumours.*
oestrogen	*aromatase*	*The **PROTEIN** enzyme, which converts androstenedione to oestrone, regulates the availability of **DRUG** so support the growth of hormone-dependent beast tumours.*

Table 1: Examples of preprocessing of drug-protein pairs in the sentence *The aromatase enzyme, which converts androstenedione to oestrone, regulates the availability of oestrogen so support the growth of hormone-dependent beast tumours.* (PMID:15341993)

DrugBank	DrugProt
ligand, binder, binding	DIRECT-REGULATOR
partial agonist	AGONIST-ACTIVATOR
inverse agonist	AGONIST-INHIBITOR
blocker, partial antagonist	ANTAGONIST
inducer, stimulator	INDIRECT -UPREGULATOR
product of	PRODUCT-OF
activator	ACTIVATOR
inhibitor	INHIBITOR
agonist	AGONIST
antagonist	ANTAGONIST
substrate	SUBSTRATE

Table 2: Mapping of relationships

and CTD. Similarly, a protein name dictionary is created from UniProt and CTD.

Then, as shown in Figure 1-(iii), the distant supervision data is created by strict matching the named entities extracted from the PubMed literature with drug and protein names in the relation triples after lowercasing the entities and names.

Finally, as shown in Figure 1-(iv), we map the relation types in DrugBank, which are the original labels of the distant supervision data, to the relation types in the DrugProt task (Krallinger et al., 2021) using a mapping dictionary as shown in Table 2. We manually build the mapping dictionary based on the relation annotation guideline (Rabal et al., 2021) in the DrugProt corpus.

2.3 Relation Extraction Using Distantly Supervised Models

We propose two alternatives to utilize the distantly supervised model. One is the initialization approach that initializes the supervised model with the distantly supervised model (`Initialization`), and the other is the mixture approach that combines representations obtained from a fixed distantly supervised model and representations obtained from a supervised model in training the supervised model (`Mixture`).

2.3.1 Initialization

In the task of natural language processing, pretraining on datasets close to the domain sometimes improves the performance of the model on the target dataset. (Beltagy et al., 2019b) Following this line, for `Initialization`, we perform pretraining using distant supervision data to initialize the model for supervised learning. Specifically, we first train the relation extraction model described in Section 2.1 using the distant supervision data, use the model parameters to initialize another relation extraction model for supervised learning, and then train the relation extraction model using manually annotated training data.

2.3.2 Mixture

For `Mixture`, we pretrain a relation extraction model explained in Section 2.1 using distant supervision data to extract additional features from the input. Similarly, another relation extraction model is pretrained with manually annotated training data[1]. The two pretrained feature extraction models, i.e., BERT, are used to mix the feature representations. In training, the feature extraction model pretrained on the distant supervision data is fixed, while the feature extraction model trained on the manually annotated training data is not fixed and further fine-tuned[2].

Predictions are made by mixing representations obtained from the model pretrained with distant supervision data and representations obtained from the model that is specific to supervised training with manually annotated training data as shown in Figure 2. We propose two mixing methods that use the importance weights of the representations, which mix the representations obtained from dis-

[1] We find this pretraining can improve the performance in our preliminary experiments.

[2] In our preliminary experiments, we tried to fine-tune the feature extraction model pretrained on the distant supervision data, but the performance with the model was lower than one with fixed parameters.

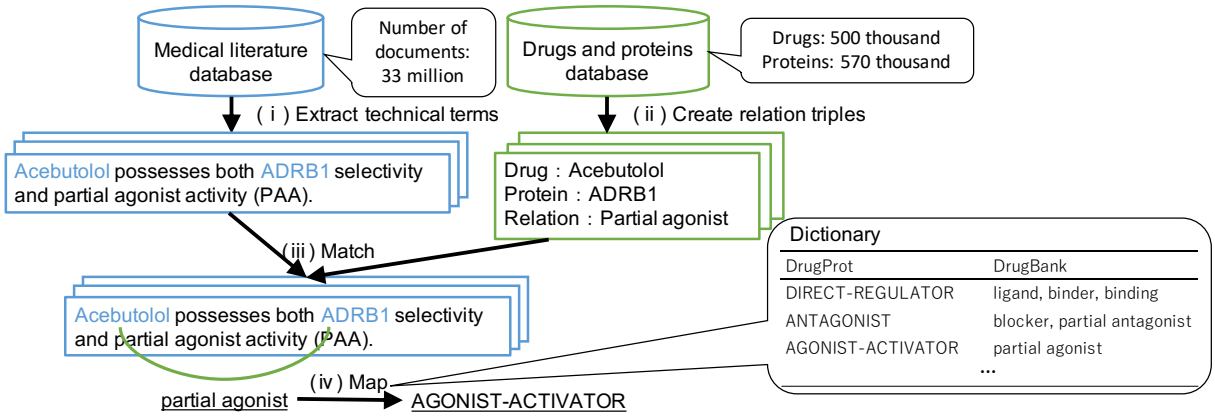

Figure 1: Overview of the creation of distant supervision data

tant supervision data with those obtained from manually annotated training data.

First, as shown in Figure 2-(i), the representations $\boldsymbol{h_{ds}}$ obtained from the fixed BERT model in the fixed pre-trained distantly supervised model are mixed with the representations $\boldsymbol{h_{sv}}$ from the BERT model in another relation extraction model that is pre-trained on the manually annotated training data. Next, as shown in Figure 2-(ii), we mix the representations $\boldsymbol{h_{ds}}$, $\boldsymbol{h_{sv}}$. In mixing the representations, we propose two mixing methods, `Add` and `Concat`, which are defined as follows:

$$\boldsymbol{h}_{\texttt{Add}} = \alpha \boldsymbol{h_{ds}} + \beta \boldsymbol{h_{sv}} \quad (1)$$

$$\boldsymbol{h}_{\texttt{Concat}} = [\alpha \boldsymbol{h_{ds}}; \beta \boldsymbol{h_{sv}}] \quad (2)$$

$[\cdot;\cdot]$ denotes the concatenation of vectors. α and β are the importance weights of each feature, which are scalar-valued parameters that are trained during training. Here, `Add`, as shown in Eq. (1), sums $\boldsymbol{h_{ds}}$ and $\boldsymbol{h_{sv}}$ after multipying the corresponding weight, which indicate the importance, to each representation. `Concat` is mixed by concatenating $\boldsymbol{h_{ds}}$ and $\boldsymbol{h_{sv}}$ after multiplying weights to the parameters, as shown in Eq. (2).

Finally, as shown in Figure 2-(iii), the obtained representations, i.e., $\boldsymbol{h}_{\texttt{Add}}$ or $\boldsymbol{h}_{\texttt{Concat}}$, are used to predict the relation between the drug and the protein with one fully connected layer (FC) and the softmax function. The model is trained on the manually annotated training data to minimize the cross-entropy loss.

3 Experimental Settings

In this section, we explain the settings for the data sets, tasks and hyper-parameter tuning.

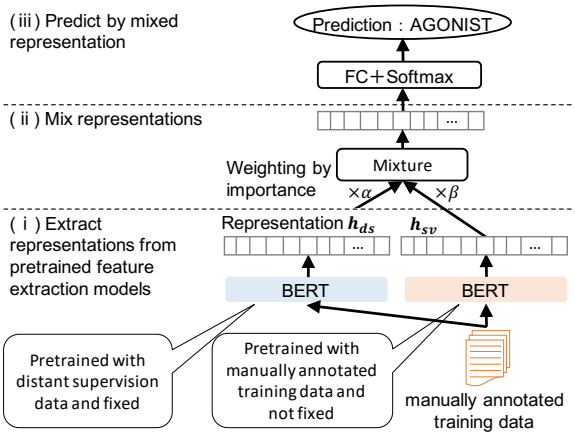

Figure 2: Overview of the `Mixture` of the representations

We used the data set from the BioCreative VII Track 1 - Text mining drug and chemical-protein interactions (DrugProt) (Krallinger et al., 2021) for the evaluation. This data set is composed of documents annotated with drug mentions, protein mentions, and their relations. The DrugProt corpus consists of train, develop, and test. Since the annotations for the test data are not publicly available, this study evaluates the model on the development data. In addition, the distant supervision data built by the method in Section 2.2 were used to train the model. The number of instances per relation in the DrugProt corpus and the distant supervision data are shown in Table 3. We followed the task setting of DrugProt. The task is to classify a given pair of a drug and a protein into 13 relation types or no relation. We evaluated the performance with the F-score on each relation type and the micro-averaged F-score on all relation types. Micro-averaged F-score is also shown for reference. We used the

	DrugProt		Distant
	train	develop	supervision data
ANTAGONIST	972	218	69,234
AGONIST	659	131	89,704
AGONIST -ACTIVATOR	29	10	875
AGONIST -INHIBITOR	13	2	1,107
DIRECT -REGULATOR	2,250	458	18,945
ACTIVATOR	1,429	246	31,745
INHIBITOR	5,392	1,152	173,400
INDIRECT-DOWN-REGULATOR	1,330	332	0
INDIRECT-UP-REGULATOR	1,379	302	11,981
PART-OF	88	625	0
PRODUCT-OF	921	158	1,565
SUBSTRATE	2,003	495	2,311
SUBSTRATE _PRODUCT-OF	25	3	0
Total	17,288	3,765	400,867

Table 3: The number of instances per relation in the DrugProt corpus and the distant supervision data

official evaluation script[3] provided by the task organizers.

We used the Successive Halving Algorithm from the open-source hyper-parameter auto-optimization framework Optuna (Akiba et al., 2019) for hyper-parameter tuning. We chose the dropout rate from the region of [0.0, 0.5], the learning rate of Adam from the region of [1e-6, 1e-4], the weight decay of Adam from the region of [1e-10, 1e-3]. Hyper-parameters are determined by a parameter search to maximize the micro-averaged F-score on the development data of the DrugProt corpus[4].

4 Results

To evaluate the proposed method, we conducted three experiments: evaluation of the performance of extracting drug-protein relations, analysis of prediction results, and comparison of extraction performance on small-scale manually annotated training data. In this section, we describe these three experiments.

4.1 Drug-Protein Relation Extraction

We conducted experiments to compare the extraction performance of the proposed method with a baseline trained only on manually annotated training data. As the baselines, we trained relation extraction models based on PubMedBERT and BioRoBERTa-large, both of which were pretrained in a domain close to the dataset, with manually annotated training data. BioRoBERTa-large is a large-scale pretrained model with a parameter size approximately three times larger than PubMedBERT. The baseline model with BioRoBERTa-large is the same as the model by Yoon et al. (2021) that achieved the high performance of 77.46% on the development data without external knowledge. [5]

The results are shown in Table 4. First, we focus on the performance of the proposed methods when they are applied to the PubMedBERT baseline model. For all the proposed methods, the prediction performance for AGONISTand PRODUCT-OF, which have less manually annotated training data, is greatly improved. This is because the representations obtained from the distantly supervised model can compensate for the lack of manually annotated data. Besides, the performance of AGONIST-ACTIVATORand AGONIST-INHIBITOR, which have particularly less manually annotated training data, was significantly improved by `Initialization`, but not by `Mixture`. This shows that the representations obtained from the distantly supervised model with `Initialization`more directly influenced the performance than those with `Mixture`. In addition, `Add` and `Concat`, which mixed the representations from the distantly supervised model data with the representations specific to the supervised model, improved the micro-averaged F scores by 0.6 and 0.8 points, respectively. This indicates that `Mixture` is a more effective way to use distantly supervised model than `Initialization`.

Next, we discuss the performance of the proposed method for the BioRoBERTa baseline. Overall, the proposed method improves the micro-averaged F-score by 0.5 points. Furthermore, when we compare the F-score of each relation, the performance of all relations except ACTIVATOR, ANTAGONIST, and SUBSTRATE is improved or maintained. From these results with two different BERT models, we show that the proposed

[3] https://github.com/tonifuc3m/drugprot-evaluation-library

[4] This setting can cause overfitting to the development data sets, but since this is an official development set, we decided to report the best score to make the scores comparable to other methods in the shared task.

[5] Weger et al. (Weber et al., 2021) showed a slightly better performance with 78.3% on the development data by adding input start and end markers for target entities in the sentences, instead of masking the target entities like us. Since our main focus is not investigating a better baseline model, we leave investigating the representation of target entities for future work.

method can improve the performance regardless of the parameter size of the model.

4.2 Analysis of Prediction Results

We show the confusion matrices between gold labels and predicted labels by the baseline and the proposed method to analyze the prediction tendency of the two methods, and visually check the prediction cases. The confusion matrix is a table that visualizes the differences in two different sets of labels for instances. It has gold labels in the row direction and predicted labels in the column direction, and each element has the number of cases for the pair of gold and predicted labels. For the proposed method, we used a model that employs `Mixture` with `Concat`, which showed the best performance improvement from the baseline in the approach to utilize the distantly supervised model as shown in Section 4.1, based on PubMedBERT. The confusion matrices of the baseline and the proposed method are shown in Figure 3. The left and right confusion matrices are for the baseline and the proposed method, respectively.

First, we focus on the cases of different predictions in relation types. We can see that the number of cases that the proposed method mistakenly predicts INHIBITOR for DIRECT-REGULATOR is reduced from 14 to 2. Some example cases, where the predictions are improved by the proposed method, are shown in Table 5. The reason for the incorrect prediction by the baseline model is that the sentence contains "inhibit", "inhibited", and "inhibition", which are important for predicting the INHIBITOR type. For these cases, the baseline may predict the relations as INHIBITOR even though the sentence indicated DIRECT-REGULATOR between DRUG and PROTEIN entities. The reason why the proposed method was able to correctly predict such cases may be that the proposed method uses representation obtained from distantly supervised models that are trained on large-scale distant supervision data, and thus places more emphasis on the context than on word-level expressions.

Conversely, the number of cases in which the proposed method predicted INHIBITOR for the instances with the gold INDIRECT-DOWNREGULATOR type has increased from 19 to 25. The cases where the baseline made a correct prediction and the proposed method made a wrong prediction are shown in Table 6. The reason why the proposed method made such incorrect predictions in these cases may also be due to the existence of inhibit, inhibited, and inhibition in the sentences, which are important for predicting INHIBITOR, similarly to the baseline's wrong predictions for the cases in Table 5. This is because the sentence contains "inhibit", "inhibited", and "inhibition", which are important for predicting both INHIBITOR and INDIRECT-DOWNREGULATOR. Furthermore, the context of the cases is similar because these types are both related to the cases that drugs inhibit proteins. Therefore, the proposed method is likely to make INHIBITOR predictions based on such keywords for cases that the prediction is difficult with the context, without much consideration on the differences in the actions of drugs on proteins.

Then, we focus on the cases where the miss prediction is made between a relation type and a negative type. We can see that the proposed method reduces the number of cases in which the negative examples are mistakenly predicted as the INHIBITOR type from 204 to 178, the number of cases in which the negative examples are mistakenly predicted as PRODUCT-OF from 70 to 44, and the number of cases in which the negative examples are mistakenly predicted as SUBSTRATE from 142 to 86. The cases, where the baseline incorrectly predicted the negative cases as PRODUCT-OF while the proposed method correctly predicted them, are shown in Table 7. The numbers of improved cases and example cases suggest that the proposed method is more context-sensitive in its prediction than the baseline model.

These results suggest that the proposed method places more emphasis on contextual expression than on word expressions in making predictions compared to the baseline models. However, for cases where it is difficult to make predictions based on context, we found that the proposed method made incorrect predictions.

4.3 Performance Comparison with Small-Scale Manually Annotated Training Data

This section examines the effectiveness of the proposed method in training with small-scale manually annotated training data. We aim to improve the performance of drug-protein interaction extraction while reducing the cost of creating additional manually annotated training data by utilizing distant supervision data that have low creation costs. In

	PubMedBERT					BioRoBERTa		#Manually-annotated instances
	Manual data	Distant data	+ Init	+ Mix (Add)	+ Mix (Concat)	Manual data	+ Mix (Concat)	
INDIRECT -DOWNREGULATOR	76.7	0.0	74.6	77.7	78.7	79.3	79.9	1,330
INDIRECT-UPREGULATOR	73.3	1.9	75.1	73.7	73.6	75.6	76.2	1,379
DIRECT-REGULATOR	65.9	6.1	62.1	66.9	67.7	66.9	69.4	2,250
ACTIVATOR	77.3	5.2	70.6	77.5	76.7	75.7	73.8	1,429
INHIBITOR	84.2	29.4	84.7	84.6	84.3	85.1	86.1	5,392
AGONIST	75.5	6.7	79.7	78.2	77.0	76.1	77.2	659
AGONIST-ACTIVATOR	0.0	0.0	46.2	0.0	0.0	0.0	0.0	29
AGONIST-INHIBITOR	0.0	0.0	80.0	0.0	0.0	0.0	0.0	13
ANTAGONIST	90.6	26.0	89.6	92.2	91.8	91.7	90.2	972
PRODUCT-OF	59.0	10.6	63.7	62.9	62.5	61.2	62.0	921
SUBSTRATE	69.5	13.1	69.1	68.4	69.9	72.7	71.8	2,003
SUBSTRATE_PRODUCT-OF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25
PART-OF	71.7	0.0	70.6	72.2	71.7	72.8	74.4	886
Macro-averaged F-score	57.2	7.6	66.6	58.0	58.0	58.2	58.5	—
Micro-averaged F-score	76.2	16.6	75.6	76.8	77.0	77.5	78.0	—

Table 4: Relation extraction performance on the development data set. +Init, +Mix (Add), and +Mix (Concat) denote `Initialization`, `Add` of `Mixture`, and `Concat` of `Mixture`, respectively

DRUG inhibit (125) i - PROTEIN binding to recombinant rat eta receptors.
N - (diphenylmethyl) - 2 - phenyl - 4 - quinazolinamine (DRUG), n - (2, 2 - diphenylethyl) - 2 - phenyl - 4 - quinazolinamine (sori - 20040), and n - (3, 3 - diphenylpropyl) - 2 - phenyl - 4 - quinazolinamine (sori - 20041) partially inhibited [(125) i] 3beta - (4'- iodophenyl) tropan - 2beta - carboxylic acid methyl ester (rti - 55) binding, slowed the dissociation rate of [(125) i] rti - 55 from the PROTEIN, and partially inhibited [(3) h] dopamine uptake.
DRUG (parent compound), has moderate affinity for the PROTEIN (competitive inhibition).

Table 5: Improved cases with wrong predictions by the baseline model. The baseline model mistakenly predicted INHIBITOR for DIRECT-REGULATOR for the DRUG and PROTEIN pairs.

Section 4.1, we trained models using all manually annotated training data and confirmed that the proposed method can improve the performance of the baseline models. To verify the effectiveness of the proposed method in training with a small amount of manually annotated training data, we trained with only a small portion of the manually annotated training data and compared the performance of relation extraction between the PubMedBERT baseline model and the model with the proposed method. We checked the performance of the proposed method on the development data when the model was trained with the small number of cases, we chose the number from [3, 5, 7, 10, 20, 50, 100, 200, 500, 1,000], for each relation in the manually annotated training data. For the proposed method, we used a model that mixes feature representations with `Concat`, which showed the best performance improvement from the baseline with Section 4.1.

The results are shown in Figure 4. As in the case of Section 4.1, we did not obtain a significant performance improvement over the baseline as we saw when training with all manually annotated training data, but the performance consistently improved for all the cases. This indicates that the proposed method can improve performance by using representations obtained from the distantly supervised model, regardless of the number of cases of manually annotated training data.

5 Conclusions

We aimed to improve the performance of drug-protein relation extraction by creating distant supervision data at low cost and utilizing the model pre-trained on the data while reducing the noise contained in the distant supervision data. We proposed two methods of utilizing distant supervision data. Both methods improved the prediction performance from the baseline for relation types with less manually annotated training data. In addition, the method that mixes representations also improved the F-scores for many relation types, some of them have a large amount of manually annotated training data, as well as the micro-averaged F-score, demonstrating the effectiveness of the proposed method. In addition, we showed that the performance im-

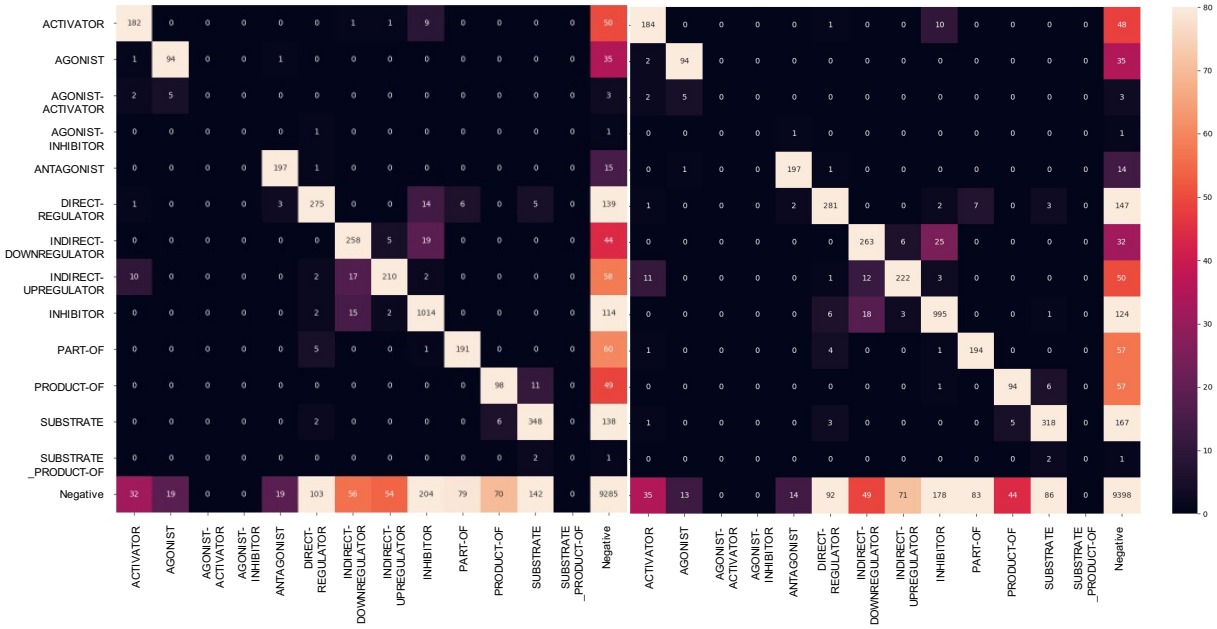

Figure 3: The confusion matrices (left: baseline, right: proposed method)

The upregulation of calpain, PROTEIN and caspase - 3 activity were further inhibited by treatment with DRUG in the presence of ald.
The mechanism of action of DRUG was related to the inhibition of the cleavage of pro - caspase - 1, PROTEIN and pro - il - 18 which in turn suppressed the activation of nlrp3 inflammasome.

Table 6: Deteriorated cases with wrong predictions by the proposed model. The model wrongly predicted INHIBITOR, instead of INDIRECT-DOWNREGULATOR, for the DRUG and PROTEIN pairs.

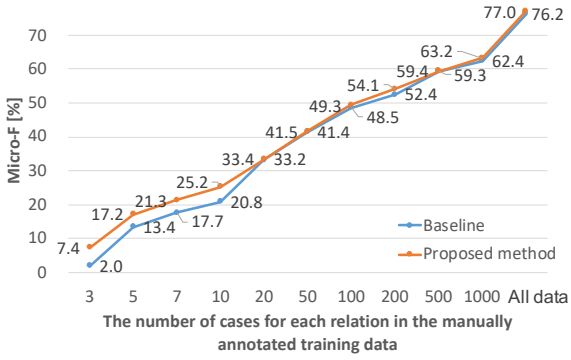

Figure 4: Micro-averaged F-scores for the number of manually annotated training instances for each relation type

provement was independent of the parameter size of the model and the number of cases of manually annotated training data.

To improve the extraction performance, we plan to investigate the `Mixture` method for its way of mixing representations and pretraing.

Acknowledgements

This work was supported by JSPS Grant-in-Aid for Scientific Research JP20K11962.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Iz Beltagy, Kyle Lo, and Waleed Ammar. 2019a. Combining distant and direct supervision for neural relation extraction. 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)*, pages 1858–1867, Minneapolis, Minnesota. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019b. Scibert: Pretrained language model for scientific text. In *EMNLP*.

The structures solved after the diffusion of oligosaccharides (either maltotetraose, g4 or maltopentaose, g5) into PROTEIN / glc1p crystals show the formation of DRUG and elongation of the oligosaccharide chain
DRUG biosynthesis in plants : molecular and functional characterization of PROTEIN and three isoforms of folylpolyglutamate synthetase in arabidopsis thaliana.
Knockdown of nadph oxidase, nox5 - s, a variant lacking calcium - binding domains, by nox5 sirna significantly inhibited acid - induced increase in PROTEIN expression, thymidine incorporation, and DRUG production.

Table 7: Improved cases with wrong predictions by the baseline model. The baseline model mistakenly predicted PRODUCT-PRODUCT-OF for the negative DRUG and PROTEIN pairs.

The UniProt Consortium. 2020. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489.

NCBI Resource Coordinators. 2016. Database resources of the national center for biotechnology information. *Nucleic Acids Res.*

Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. 2020. Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research*, 49(D1):D1138–D1143.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, and Wilson M. 2018. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.*

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Martin Krallinger, Obdulia Rabal, Saber Ahmad Akhondi, Martín Pérez Pérez, Jesus Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio Baso López, Umesh K. Nandal, Erin M. van Buel, Anjana Chandrasekhar, Marleen Rodenburg, Astrid Lægreid, Marius A. Doornenbal, Julen Oyarzábal, Anália Lourenço, and Alfonso Valencia. 2017. Overview of the biocreative vi chemical-protein interaction track.

Martin Krallinger, Obdulia Rabal, Antonio Miranda-Escalada, and Alfonso Valencia. 2021. DrugProt corpus: Biocreative VII Track 1 - Text mining drug and chemical-protein interactions.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 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*, pages 146–157, Online. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Obdulia Rabal, Jose Antonio López, Astrid Lagreid, and Martin Krallinger. 2021. DrugProt corpus relation annotation guidelines [ChemProt - Biocreative VI].

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, Heidelberg. Springer Berlin Heidelberg.

Leon Weber, Mario Sänger, Samuele Garda, Fabio Barth, Christoph Alt, and Ulf Leser. 2021. Humboldt@ drugprot: Chemical-protein relation extraction with pretrained transformers and entity descriptions. In *Proceedings of the BioCreative VII Challenge Evaluation Workshop*.

Wonjin Yoon, Sean Yi, Richard Jackson, Hyunjae Kim, Sunkyu Kim, and Jaewoo Kang. 2021. Using knowledge base to refine data augmentation for biomedical relation extraction. In *Proceedings of the BioCreative VII Challenge Evaluation Workshop*.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.