Enhanced Protein-Protein Interactions Extraction from the Literature using Entity Type- and Position-aware Representation

Anonymous NAACL submission

Abstract

Since protein-protein interactions (PPIs) are crucial to understanding living systems, harvesting these data is essential to probe the development of diseases and to understand gene/protein functions and biological processes. Some curated datasets exist containing PPI data derived from the literature and other sources (e.g., IntAct, BioGrid, DIP and HPRD), but these are far from exhaustive and their maintenance is a labor intensive process. On the other hand, machine learning (ML) methods to automate PPI knowledge extraction from the scientific literature have been limited by a shortage of appropriate annotated data. In this work, we create a unified multi-source PPI corpora with vetted interaction definitions, and augmented by binary interaction type labels. We also present a Transformer-based deep learning method, exploiting entity type and positional information for relation representation to improve relation classification performance. We evaluated our model's performance on three widely studied relation extraction datasets from biology and computer science domains as well as our work's target PPI datasets to observe the effectiveness of the representation to relation extraction tasks in various domains, and found it to outperform prior state-of-the-art (SOTA) models.

1 Introduction

Much effort in modern molecular biology either involves or is entirely focused on learning and understanding the functions and interactions of the millions of proteins which are the basic building blocks of life. The functions of most of these are currently unknown, and are only definitively established for a small fraction on which extensive and labor-intensive labwork has been performed. These gold-standard protein function assignments have been extended computationally via DNA & amino acid sequence homology throughout the ever-expanding collection of protein sequences determined from genome sequencing. However, inference from homology is often inaccurate. Helpfully, clues about function can come from other sources, including interactions with proteins for which the function is known. While experiments that definitively determine interactions can be labor-intensive, several relatively high-throughput methods are in use, such as Two-hybrid screening (Brückner et al., 2009) and affinity purification followed by mass spectrometry (Dunham et al., 2012). Several databases such as IntAct[1], STRING[2], DIP[3], BioGrid[4], HPRD[5], and MINT[6] are now dedicated to collecting and curating PPI results obtained with these techniques and others, and from the scientific literature. However, mining the literature requires manual effort and is slow. We aim to develop a ML model that effectively identifies statements of PPIs in scientific text.

Efforts to fully automate text knowledge extraction are widespread and ongoing, with supervised learning approaches currently being the most favored. A key challenge in applying these methods to PPI extraction is a shortage of training data specifically annotated for this purpose. Several publicly available PPI training datasets suffer from biases of restricted biological focus (i.e., human, medical, or microbial only) and also from differences in the concept of what defines an interaction. For this work we combine all these training sets, vet them for uniformity in interaction definition, and also add interaction type labels. We propose Transformer architecture-based models (Vaswani et al., 2017), which leverage entity type and positional information to build a relation representation to improve relation classification performances.

[1] https://www.ebi.ac.uk/intact
[2] https://string-db.org
[3] https://dip.doe-mbi.ucla.edu/dip
[4] https://thebiogrid.org
[5] https://www.hprd.org
[6] https://mint.bio.uniroma2.it

In this paper, our contribution is two-fold:

1. We augment public PPI corpora with labels for protein types (*enzyme* and *structural*), which further delineate the functional role of proteins and consequently provides a helpful protein classification for the biology community. We provide the interaction-typed PPI corpora for the community.

2. We present a Transformer-based relation prediction method that exploits types of entities and positional information to build an improved relation representation. Our study showed the effectiveness of the proposed approach on not only the PPI datasets, but also three relation extraction datasets from biology and computer science domains.

2 Related Work

There have been ongoing efforts to consolidate biological knowledge pertinent to PPIs from literature by creating machine-processible data and designing protein relation extraction methods.

2.1 PPI corpora

BioCreative VI (Islamaj Doğan et al., 2019) proposed a PPI relation extraction challenge task related to genetic mutations to foster the development of mining PPI information from biomedical literature. Bunescu et al. (2005) annotated 1000 titles and abstracts from the MEDLINE repository that discuss human genes/proteins, so-called AIMed corpus, which includes roughly 5000 protein names and 1000 protein interactions. Pyysalo et al. (2007) created BioInfer (Bio Information Extraction Resource), containing 1100 sentences with named entities and their relationships tagged from abstracts of biomedical research articles. Fundel et al. (2007) tagged the sentences of 50 abstracts referenced by the Human Protein Reference Database (HPRD) with direct physical interactions, regulatory relations, and modifications between genes/proteins. The IEPA (Information Extraction Processing Assessment) corpus (Ding et al., 2001) was created to conduct a comparative study on the merits of different text processing units for interactions between biochemical entities. And the Learning Language in Logic Workshop (LLL05) designed the genic interaction extraction challenge task, the purpose of which is to promote protein/gene interactions information extraction from biology abstracts from the MEDLINE bibliography database (Nédellec, 2005). The challenge focused on gene interactions in Bacillus subtilis which is a model bacterium and many papers have been published on direct gene interactions involved in sporulation.

Although the number of corpora and methods for PPI information extraction from biomedical text has increased as the interest to automatic mining system is growing, the lack of consensus with respect to PPI annotation has hindered consolidation of heterogeneous datasets, and thereby making it difficult for researchers to properly evaluate their methods on a standardized dataset for PPI extraction. Pyysalo et al. (2008) conducted a comparative analysis of the five PPI datasets AIMed, BioInfer, HPRD50, IEPA and LLL and unified the PPI annotations to be shared with the community for clear and comparative method evaluation. To merge the diverse datasets, Pyysalo et al. (2008) found common categories across the five corpora and generated a unified PPI corpora composed of sentences tagged with undirected and untyped binary interactions (i.e., positive and negative). These unified versions of PPI datasets – hereafter called the five benchmark PPI corpora – has been widely used to evaluate various approaches on PPI extraction tasks (Tikk et al., 2010; Bui et al., 2011; Warikoo et al., 2021). In the biological literature, single sentences often discuss more than two proteins, and not all such statements are declarations of interactions between the proteins mentioned. These datasets include all identified protein/gene entity names found within each training sentence, and also a pairwise evaluation of positive/negative interaction between each possible pairing.

However, some issues remain regarding the content and annotations in these benchmark PPI datasets, as we detail in section 3.1. In this paper, we present an augmented, refined version of the five benchmark PPI corpora plus the BioCreative VI corpus that further specify positive interactions into two types of interactions: *enzyme, structural*. These interaction types are desirable to construct networks of protein interactions.

2.2 PPI extraction methods

In the early stages of adopting ML approaches for the PPI extraction task, feature-based and kernel-based approaches were commonly used (Baumgartner et al., 2008; Bui et al., 2011). Murugesan et al. (2017) developed a Distributed Smoothed Tree ker-

nel (DSTK) composed of distributed lexical parse trees and semantic feature vectors in an attempt to capture syntactic and semantic information of sentences and demonstrated that the shallow linguistic information aided to enhance the PPI extraction capability with the model evaluation on the five benchmark PPI corpora.

With the recent success of deep learning in a number of applications, deep neural network models have emerged to tackle the PPI extraction task. Peng and Lu (2017) demonstrated that their multichannel dependency-based convolutional neural network model (McDepCNN) effectively captured syntactic features of sentences by adding a separate channel for the dependency information of the sentence syntactic structure on the PPI task using AIMed and Bioinfer corpora. Recently, attention mechanisms in natural language processing (NLP) have shed some light on solving long dependency issues between tokens in sequential data. The self-attention based Transformer architecture (Vaswani et al., 2017) has proven to well preserve long term dependencies and establish effective contextual representations. NLP models built upon Transformer architecture such as BERT (Devlin et al., 2018) have achieved SOTA results in various NLP tasks, including in biology domains (Vig et al., 2020). Warikoo et al. (2021) proposed a Lexically aware Transformer-based Bidirectional Encoder Representation model (LBERT) that generated syntactic contexts emphasized representations for sentence-level bio-entity relation extraction tasks. LBERT is a modified version of BERT that takes *n*-gram parts-of-speech frames as an additional input embedding in order to deliver latent lexical properties, and the model outperformed the prior SOTA models on a PPI task with the five benchmark PPI corpora.

3 Additional PPI curation

This section details our further curation and enhancement of the aforementioned data sets.

3.1 Problems Discovered during Curation

In vetting the five benchmark PPI training corpora, we identified the following problems:

3.1.1 Bias due to restricted biological focus for each set

In particular, the AIMed and IEPA corpora are focused on human medical biochemistry and phenomena, including viral pathogens, whereas the set LLL is limited to a single bacterial species, *B. subtilis*. These differences manifest in skew and distribution of protein/gene name frequency counts between the five sets, as well as other domain-specific terminology. In fact, the most frequently occurring protein in IEPA, *insulin*, accounts for 14% of the protein mentions in all the IEPA positives, whereas it doesn't occur in the AIMed positives set, for which the most common protein, *p53*, accounts for only 1.75% of the protein names. These sets all sampled very different populations in the literature. Combining all sets together helps towards countering this bias, but in the future, we plan to collect more training data to better address this issue.

3.1.2 Differences in notion of the definition of an *interaction*

The five sets largely restrict PPI positive cases to clear statements of direct interaction between the two subjects. LLL further restricted positive PPI declarations to cases wherein a protein binds to DNA and causes or inhibits the transcription of the gene of another protein, thus a statement of gene regulation – a very particular type of interaction.

We intentionally broaden our acceptance of a positive PPI indication. Our goal is to provide biologists with a tool which will identify possible interactive connections between proteins directly from the scientific literature text. Because of the likelihood that claims of direct PPI will end up in databases in the future, if not there already, a less restrictive interpretation will allow a text mining system to report results of value that won't necessarily be found in a PPI database.

Along these lines, we did not distinguish between gene or protein for this work. In addition to direct binding between two proteins or a protein and itself (i.e. *dimers* and *multimers*), we also considered as interacting cases where two proteins bound to a larger complex of other proteins without necessarily contacting each other directly.

The following is an example (from BioCreative corpus) where a direct connection between proteins *PVA12* and *ORP3a* is made, but is not declared an actual interaction.

> *The targeting of the oxysterol-binding protein **ORP3a** to the endoplasmic reticulum relies on the plant VAP33 homolog **PVA12**.*

On the other hand, we were mindful of the possibility of being too broad, which would result in too many PPI calls to be meaningful.

3.1.3 Confusion over PPI-negative annotations

This expanded threshold for PPI-positive impacted the public negative annotations. Here are two example cases (from AIMed corpus) where we disagreed with the given negative labels.

> *In addition to this unique pathway,* ***FGFR3*** *also links to* ***GRB2****.*

A negative interaction between proteins *FGFR3* and *GRB2* was declared in the public set.

> *After a brief historical incursion regarding RAS of renal origin, we present the main extrarenal* ***angiotensin****-forming enzymes, starting with* ***isorenin****,* ***tonin****, D and* ***G cathepsin*** *and ending with the conversion enzyme and* ***chymase****.*

In this case negative interactions were annotated between *angiotensin* and each of *isorenin*, *tonin*, *G cathepsin* and *chymase* respectively, even though they are declared as forming angiotensin.

Here is an example of a negative PPI sentence where we agree and which we included in our curated set (from AIMed corpus).

> *The molar ratio of serum* ***retinol-binding protein (RBP)*** *to* ***transthyretin (TTR)*** *is not useful to assess vitamin A status during infection in hospitalised children.*

To reduce confusion in our initial models regarding updated positive and negative re-labels, we considered only those negatively labeled sentences where no positive pairs were declared in a sentence, and we then manually examined each case to make sure we agree, disregarding for now those where we differ. For the same reason, for this work, we also disregarded negative pair cases in sentences with both positive and negative annotations.

3.2 Interaction Type Annotation

PPIs aid with biological engineering: i) structure and protein subunit complex knowledge is critical to protein engineering ii) transient interactions (e.g., chaperone to client protein) knowledge needed for engineering at a broader scale. To make the public PPI corpora more useful for this purpose, we added interaction type labels for the positively defined pairs in the unified datasets as well as for the BioCreative set. In determining the interaction type labels, we first considered top-level protein function categories from IntAct's molecular interaction ontology but discovered that we did not have enough training examples to provide sufficient statistics in each of the 28 categories to properly train a model (not all interaction types occur with equal frequency). We then attempted to reduce the number of categories by making them coarser, first lowering to roughly ten and then three types, but we found that making assignments in this manner proved too complicated and provided questionable scientific value.

We finally decided on a simple binary classification, with interactions being declared either *enzyme* or *structural* for our first pass in that *enzyme* or *structural* accurately delineates the functional role of almost all proteins and consequently provides a concise but meaningful protein classification. The *structural* label was applied to protein assemblages of large, permanent cellular components such as cell walls, histones, golgi apparatus, microtubules, membranes, inter-cellular structures and the like. All other interactions were classified as *enzyme*. Type was determined by examining the given function for each protein/gene, where it could be obtained from any of several online protein databases such as Uniprot, NCBI, and GeneCards, and from the sentence context itself. For the five sentence-based data sets, interaction type labels were applied for positively identified protein pairs. An example of a structural interaction label for the proteins *alpha-syntrophin* and *utrophin* (from BioInfer corpus) follows:

> *Absence of* ***alpha-syntrophin*** *leads to structurally aberrant neuromuscular synapses deficient in* ***utrophin****.*

The remaining non-structural interactions were considered *enzymatic*, a label we applied to nominal enzyme activity (proteins which catalyze chemical reactions of metabolites in reaction pathways) as well as proteins which activate other proteins (*kinases*), but which we also applied to all proteins which activated, inhibited, signaled, and formed temporary complexes with other proteins, and also those which bind to DNA to regulate gene expression, chaperones which help proteins fold, and

those which destroy proteins (proteases), An example of an enzyme-labeled PPI between *JAK2* and *Ref-1* (from AIMed corpus):

> *The cytokine-activated tyrosine kinase **JAK2** activates **Raf-1** in a p21ras-dependent manner.*

This process of adding type labels was the most difficult and labor-intensive aspect of the training data curation. There were thousands of gene names and symbols which required external look-ups in addition to an equally large host of specialized biological jargon and acronyms (chemical names, cell lines, experimental conditions, *etc.*) which also needed to be researched in order to differentiate from proteins and to established context necessary for understanding each sentence. Importantly, because this annotation effort is informed by resources and knowledge external to the text in question, it encodes specialized domain knowledge that makes the PPI type classification task more challenging, and which puts further pressure on ML models to adequately capture sufficiently informative context to make a class determination.

Two domain experts performed the PPI annotation. With respect to inter-annotator agreement, one of them annotated, the second was able to veto, with disagreements resolved by mutual agreement. The definition of an interaction and the annotation rules were carefully determined ahead of time, according to domain expertise. The rules will be also released along with the dataset.

4 Methodology

We adopted a Transformer-based approach for the PPI classification task. In particular, we improved a relation representation exploiting entity type and positional information.

4.1 Entity Type- and Position-aware Relation Representation

To generate embeddings for relation representations, we applied a marker-free representation approach used in earlier models (Zhang et al., 2021; Eberts and Ulges, 2020). Basically, the representation consists of a pair of two max-pooled entity embeddings and a local context that is a max-pooled embedding from a series of tokens between the two entities. We have further enhanced the representation embodying entities' types and positional information. This rationale has been adapted from the earlier findings that adding entity type indicators to representations improved a relation extraction task (Qin et al., 2021; Zhou and Chen, 2021) and that entity positional information was important for a model to focus on the relation pairs of interest (Qin et al., 2021; Zhang et al., 2017).

Unlike the previous relation representations, our approach takes advantage of entity types and positional information without using additional markers. Considering the resource efficiency, a marker-free method can be preferable to a marker-used method where an input is pre-tagged with extra marker tokens (e.g., `[TAG1]`entity 1`[/TAG1]` is involved ...`[TAG2]`entity 2`[/TAG2]`...). It is commonly found in relation extraction datasets that a sentence can have more than two entities and multiple relations. In such cases, models using a marker-free representation can handle multiple relations in a sentence at once whereas marker-used approaches might need to generate a separate input for each pair. This leads to lower resource use by shorter input length and less training cost. Also, a marker-free method is more flexible in case entities are not defined in the first place (e.g., hierarchical joint-learning (Eberts and Ulges, 2020; Takanobu et al., 2019)).

To incorporate entity type and position information without explicit markers, we used a dedicated embedding table for entity types (e.g., protein, gene, chemical, drug) which the model looks up to find the type embedding of an entity. The entity type table is filled by a pre-defined entity type list, and the table contains a pair of start and end embedding for each entity type (e.g., `[protein]`, `[/protein]`). For the entity's positional information, we leveraged Transformer's inherent positional embeddings of a preceding/succeeding token of an entity span as the entity's start/end positional information. This can be seen as utilizing entity start/end markers' positional information. The model performs an element-wise addition of the entity type embeddings and position embeddings before appending them to the final representation. Figure 1 illustrates the construction of relation representation in a PPI sample. We evaluated the proposed method on three relation extraction datasets from biology and computer science as well as our work's target PPI datasets to validate the effectiveness of the representation to relation extraction tasks in various domains.

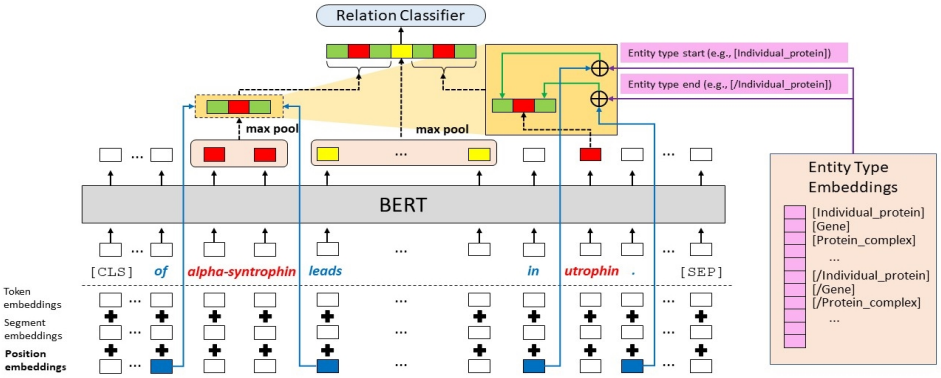

Figure 1: The representation consists of the max-pooled of two entity outputs, the max-pooled of local context (a series of tokens between the two entities), two pairs of entity span's start/end positional embeddings, and two pairs of entity type embeddings. $\oplus$ denotes element-wise addition. The example sentence is *"Absence of alpha-syntrophin leads to structurally aberrant neuromuscular synapses deficient in utrophin."* (source from: BioInfer corpus)

4.2 Model Architecture

Our Transformer-based relation extraction model performs a token-level classification task using a logistic regression with softmax to determine the probability of token class (e.g., $c \in \{\texttt{enzyme}, \texttt{structural}, \texttt{negative}\}$) as follows.

$$P(c|X) = softmax(Wh_0), \quad (1)$$

where h_0 is the output of the model. The model parameters are optimized using a categorical cross entropy as follows.

$$-\sum_{c} \delta(X, c) \log P(c|X), \quad (2)$$

where $\delta(X, c)$ indicates whether the class of X is correctly predicted ($\delta(X, c) = 1$) or not ($= 0$).

5 Experimental Setup

We demonstrate the effectiveness of the proposed approach on the three widely studied relation extraction datasets in biomedical and computer science domains, the five PPI benchmark corpora, and our PPI corpus with interaction types by comparing the performance with SOTA models.

5.1 Datasets

The relation extraction datasets from biomedical and computer science domains are ChemProt (Krallinger et al., 2017), DDI (Herrero-Zazo et al., 2013), and SciERC (Luan et al., 2018). The descriptions of the datasets can be found in Appendix A. Table 1 displays the number of relations of the datasets.

The five PPI benchmark corpora include AIMed (Bunescu et al., 2005), BioInfer (Pyysalo et al., 2007), HPRD50 (Fundel et al., 2007), IEPA (Ding et al., 2001), LLL (Nédellec, 2005). We adopted the unified version of PPI benchmark datasets provided by Pyysalo et al. (2008) that has been used in the SOTA models. In the datasets, the PPI relations are tagged with either *positive* or *negative*. The statistics of the corpus is described in Table 2. Our PPI annotations with interaction types (*enzyme*, *structural*, or *negative*) are the expanded version of the five benchmark corpora plus the BioCreative VI protein interaction dataset (Islamaj Doğan et al., 2019), and Table 3 displays the statistics of the corpora. The annotation work in all corpora has been carried out in a sentence boundary as in the five PPI benchmark corpora.

	Train	Dev	Test
ChemProt	4,154	2,416	3,458
DDI	25,296	2,496	5,716
SciERC	1,861	275	551

Table 1: Statistics of relation extraction datasets from biology and computer science domains.

	Positive	Negative
AIMed	1,000	4,834
BioInfer	2,534	7,132
HPRD50	163	270
IEPA	335	482
LLL	164	166
TOTAL	4,196	12,884

Table 2: The number of samples in the five PPI benchmark corpora for *positive* and *negative* classes.

	Enzyme	Structural	Negative
BioC†	378	83	0
AIMed	548	182	1,371
BioInfer	604	1,465	2,148
HPRD50	103	34	87
IEPA	271	2	224
LLL	163	0	0
TOTAL	2,067	1,766	3,830

Table 3: The number of samples in the interaction typed PPI corpora for *enzyme*, *structural*, and *negative* classes. † Annotations using the PPI data from BioCreative VI Track 4: Mining protein interactions and mutations for precision medicine (PM)

5.2 Implementation details

We deployed the pre-trained BERT models in the experiments including BioBERT (Lee et al., 2020). BioBERT is a pre-trained BERT model on biomedical literature, which has demonstrated excellent performance in biomedical NLP applications. We compared the BioBERT models with the default BERT models to observe the effectiveness of biomedical knowledge transfer. We built the model using PyTorch (version 1.9.0) and using the HuggingFace's Transformers package (version 4.12.0) (Wolf et al., 2019). The model architecture and weight initialization follow the used pre-trained model, and the hyper-parameters were set as follows: *training epochs = 10, learning rate = 5e-05 with Adam, batch size = 16*. We used a dense layer with linear activation as post transformer layer. We chose a micro F-score as the performance measure that was adopted in the SOTA works.

6 Results

6.1 Evaluation on Bio/CS datasets

Table 4 shows the performance comparison of our approach with some SOTA methods on the Bio/CS datasets. We compared our model with not only SOTA scored works, but also the methods using the identical pre-trained model to observe the effectiveness of our proposed relation presentation. We used the BioBERT base-cased model for the ChemProt and DDI datasets and the SciBERT cased model (Beltagy et al., 2019) for the SciERC dataset. Our model achieved SOTA performances, and although some discrepancy might exist between model training environments such as specific hyper-parameter settings and used machines, our representation produced the higher results for all datasets and for ChemProt and SciERC outperformed the earlier works that used the same pre-trained model with different representations for relation classification such as the output embedding of the token `[CLS]`, the combinations of two entity outputs.

Corpus	Method	F1
ChemProt	Peng et al. (2020)†	73.5
	Phan et al. (2021)	89.0
	Our Method†	**93.2**
DDI	Zhu et al. (2020)†	80.9
	Su et al. (2021)†	80.6
	Our Method†	**81.1**
SciERC	Beltagy et al. (2019)‡	80.0
	Zhou et al. (2021)	82.3
	Our Method‡	**89.2**

Table 4: The F-scores of model evaluation on test data. †trained on BioBERT. ‡trained on SciBERT.

6.2 Evaluation on PPI datasets

To compare the performance of the proposed approach with SOTA works, we evaluated the models via 10 fold cross-validation (CV), as adopted in the SOTA models. Table 5 displays the F-scores of 10 fold CV on the PPI classification task along with the SOTA results. Our models produced the best performances and outperformed SOTA models on the overall classification as described in the average F-scores. The BioBERT models achieved greater improvements than the BERT models, which is in line with expectations and empirical studies (Peng et al., 2019) claiming that domain-specific pre-training benefits solving domain problems.

We further examined the model's ability on our PPI corpora with interaction types. In this experiment, we combined the six corpora where some datasets contain only single class samples or highly skewed samples so that the model can be trained on more balanced data. The model evaluation was also carried out in a 10 fold CV manner, and Table 7 shows the evaluation results. The results demonstrate that the models yield consistent predictions with over 80 F-scores as compared to the previous experiments, and the BioBERT model constantly outperformed the BERT model.

6.3 Ablation study

We conducted a detailed ablation study to examine the effect of entity types and entity positional

Method	AIMed	BioInfer	HPRD50	IEPA	LLL	*Avg.*
DeepResCNN (Zhao et al., 2016)	77.6	86.9	77.7	75.5	83.2	80.2
DSTK (Murugesan et al., 2017)	71.0	76.2	80.0	80.2	89.2	79.3
LBERT (Warikoo et al., 2021)	74.0	72.8	85.5	83.7	86.0	80.4
Our Method (w/ BERT)	89.2	86.9	86.0	81.7	85.3	85.8
Our Method (w/ BioBERT)	**91.4**	**88.2**	**86.7**	**85.1**	**90.6**	**88.4**

Table 5: The F-scores of 10 fold CV on the PPI classification with the five benchmark PPI corpora.

Rep.	AIMed	BioInfer	HPRD50	IEPA	LLL	Typed PPI	ChemProt	DDI	SciERC
baseline	90.4	86.6	86.6	82.9	87.1	84.1	93.1	79.0	87.6
+ET	90.7	87.6	**87.7**	84.2	90.2	**85.1**	93.1	**83.1**	88.1
+EP	90.9	86.8	87.1	84.7	89.9	84.2	**93.4**	80.8	88.7
+ET+EP	**91.4**	**88.2**	86.7	**85.1**	**90.6**	**85.1**	93.2	81.1	**89.2**

Table 6: Ablation test results on relation representations. The baseline representation is a concatenated vector of two max-pooled vectors of two entity tokens and a max-pooled vector of a local context. ET and EP stand for entity type embedding and entity position embedding, respectively.

Method	Typed PPI
Our Method (w/ BERT)	82.4
Our Method (w/ BioBERT)	**85.1**

Table 7: The F-scores of 10 fold CV on the PPI corpora with interaction types.

information as representation features for relation classification performance. We set a concatenated embedding of two max-pooled embeddings of entity tokens and a max-pooled embedding of context vectors as a baseline representation which was used in the earlier relation extraction studies. We tested the model's performance by adding the entity type feature, the entity position feature, or both features to the baseline representation. As in the previous experiments, the test was performed using SciBERT cased for SciERC and BioBERT base-cased for the other datasets. The summary of the results are shown in Table 6. Both entity type embeddings and entity span's start/end positional embeddings increased the models' performance from the baselines on all of the datasets. In some cases, their contributions to the results were marginal, but the two features did not degrade the models' performance in any case. Although individual features had more positive effects on some datasets (the entity type for HPRD50 and DDI, the entity position for ChemProt), the combination of the two features produced the best prediction in a majority of cases.

7 Conclusion

In this paper, we augmented existing PPI corpora annotated with interaction types, which is expected to be beneficial to extracting further PPI information from scientific publications. We also presented a Transformer architecture-based model for relation extraction. In particular, we improved a relation representation leveraging entity type and positional information without using additional markers. Our models outperformed prior SOTA models and also proved the effectiveness of entity type and positional information for the classification on three relation extraction datasets from biomedical and computer science and the PPI datasets.

We will continue to improve our PPI annotations by resolving identified problems, including de-biasing the training data: more examples are needed from across biological subject areas (plants, environmental, microbiomes etc). Our goal is a tool which works across all subfields of biology. Granularity in type classifications also needs to be increased, which will require more training data and manual annotation. Finally, statements of interaction that span two (or more) sentences also require further attention in the future.

References

William A Baumgartner, Zhiyong Lu, Helen L Johnson, J Gregory Caporaso, Jesse Paquette, Anna Lindemann, Elizabeth K White, Olga Medvedeva, K Bretonnel Cohen, and Lawrence Hunter. 2008. Concept recognition for extracting protein interaction relations from biomedical text. *Genome biology*, 9(2):1–15.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Anna Brückner, Cécile Polge, Nicolas Lentze, Nicolas Auerbach, and Uwe Schlattner. 2009. Yeast two-hybrid, a powerful tool for systems biology. *International Journal of Molecular Sciences*, (10):2763–2788.

Quoc-Chinh Bui, Sophia Katrenko, and Peter MA Sloot. 2011. A hybrid approach to extract protein–protein interactions. *Bioinformatics*, 27(2):259–265.

Razvan Bunescu, Ruifang Ge, Rohit J Kate, Edward M Marcotte, Raymond J Mooney, Arun K Ramani, and Yuk Wah Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial intelligence in medicine*, 33(2):139–155.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jing Ding, Daniel Berleant, Dan Nettleton, and Eve Wurtele. 2001. Mining medline: abstracts, sentences, or phrases? In *Biocomputing 2002*, pages 326–337. World Scientific.

WH Dunham, M Mullin, and AC Gingras. 2012. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics*, 12(10):1576–90.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020*, pages 2006–2013. IOS Press.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2007. Relex—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Rezarta Islamaj Doğan, Sun Kim, Andrew Chatr-Aryamontri, Chih-Hsuan Wei, Donald C Comeau, Rui Antunes, Sérgio Matos, Qingyu Chen, Aparna Elangovan, Nagesh C Panyam, et al. 2019. Overview of the biocreative vi precision medicine track: mining protein interactions and mutations for precision medicine. *Database*, 2019.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, and Ander Intxaurrondo. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *EMNLP*.

Gurusamy Murugesan, Sabenabanu Abdulkadhar, and Jeyakumar Natarajan. 2017. Distributed smoothed tree kernel for protein-protein interaction extraction from the biomedical literature. *PLoS One*, 12(11):e0187379.

Claire Nédellec. 2005. Learning language in logic-genic interaction extraction challenge. In *4. Learning language in logic workshop (LLL05)*. ACM-Association for Computing Machinery.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on bert for biomedical text mining. *BioNLP 2020*, page 205.

Yifan Peng and Zhiyong Lu. 2017. Deep learning for extracting protein-protein interactions from biomedical literature. *BioNLP 2017*, page 29.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, and Tapio Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. In *BMC bioinformatics*, volume 9, pages 1–11. BioMed Central.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):1–24.

Yongbin Qin, Weizhe Yang, Kai Wang, Ruizhang Huang, Feng Tian, Shaolin Ao, and Yanping Chen. 2021. Entity relation extraction based on entity indicators. *Symmetry*, 13(4):539.

Peng Su, Yifan Peng, and K Vijay-Shanker. 2021. Improving bert model using contrastive learning for biomedical relation extraction. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 1–10.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7072–7079.

Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. 2010. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS Comput Biol*, 6(7):e1000837.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Nazneen Rajani, et al. 2020. Bertology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations*.

Neha Warikoo, Yung-Chun Chang, and Wen-Lian Hsu. 2021. Lbert: Lexically aware transformer-based bidirectional encoder representation model for learning universal bio-entity relations. *Bioinformatics*, 37(3):404–412.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Haiyang Zhang, Guanqun Zhang, and Yue Ma. 2021. Syntax-informed self-attention network for span-based joint entity and relation extraction. *Applied Sciences*, 11(4):1480.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.

Zhehuan Zhao, Zhihao Yang, Ling Luo, Hongfei Lin, and Jian Wang. 2016. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics*, 32(22):3444–3453.

Meng Zhou, Zechen Li, and Pengtao Xie. 2021. Self-supervised regularization for text classification. *arXiv preprint arXiv:2103.05231*.

Wenxuan Zhou and Muhao Chen. 2021. An improved baseline for sentence-level relation extraction. *arXiv preprint arXiv:2102.01373*.

Yu Zhu, Lishuang Li, Hongbin Lu, Anqiao Zhou, and Xueyang Qin. 2020. Extracting drug-drug interactions from texts with biobert and multiple entity-aware attentions. *Journal of biomedical informatics*, 106:103451.

A Datasets from Bio/CS

ChemProt contains chemical-protein interactions extracted from 1,820 PubMed abstracts, and the task is evaluated using three entity types (CHEMICAL, GENE-Y, GENE-N) and five relation classes (CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9). We used the ChemProt dataset in Biomedical Language Understanding Evaluation (BLUE) benchmark (Peng et al., 2019). DDI consists of drug-drug relations with four entity types (DRUG, BRAND, GROUP, DRUG-N) and four relation classes (Advice, Effect, Mechanism, Int) based on 792 texts from DrugBank and 233 Medline abstracts. SciERC is a collection of relations from 500 AI papers that was initially designed for knowledge graph construction. The sentences in SciERC includes six entity types (Task, Method, Material, OtherScientificTerm, Metric, Generic) and seven relation types (Compare, Conjunction, Evaluate-For, Used-For, FeatureOf, Part-Of, Hyponym-Of). We leveraged the preprocessed ChemProt and DDI datasets by Phan et al. (2021) and the preprocessed SciERC dataset by Eberts and Ulges (2020)