Comparing Encoder-Only and Encoder-Decoder Transformers for Relation Extraction from Biomedical Texts: An Empirical Study on Ten Benchmark Datasets

Mourad Sarrouti, Carson Tao, Yoann Mamy Randriamihaja
Sumitovant Biopharma, New York, USA
{mourad.sarrouti,carson.tao,yoann.randriamihaja}@sumitovant.com

Abstract

Biomedical relation extraction, aiming to automatically discover high-quality and semantic relations between the entities from free text, is becoming a vital step for automated knowledge discovery. Pretrained language models have achieved impressive performance on various natural language processing tasks, including relation extraction. In this paper, we perform extensive empirical comparisons of encoder-only transformers with the encoder-decoder transformer, specifically T5, on ten public biomedical relation extraction datasets. We study the relation extraction task from four major biomedical tasks, namely chemical-protein relation extraction, disease-protein relation extraction, drug-drug interaction, and protein-protein interaction. We also explore the use of multi-task fine-tuning to investigate the correlation among major biomedical relation extraction tasks. We report performance (micro F-score) using T5, BioBERT and PubMedBERT, demonstrating that T5 and multi-task learning can improve the performance of the biomedical relation extraction task.

1 Introduction

The scientific literature provides a rich source of biomedical knowledge (e.g., drug-drug interactions), and due to its rapid growth, it becomes increasingly difficult for scientists to keep up-to-date with the most recent discoveries hidden in literature (Zhang and Lu, 2019; Yadav et al., 2020). Moreover, manual curation of information from biomedical literature is time-consuming, costly, and insufficient to keep up with the rapid growth of the literature (Herrero-Zazo et al., 2013). Hence, there has been growing interest in using natural language processing (NLP) techniques for automatic relation extraction (RE) between biomedical entities from texts.

Recently, a variety of approaches based on pretrained language models such as BERT (Devlin et al., 2019) and other variants have shown promising results in various NLP tasks such as relation extraction (drissiya El-allaly et al., 2021b,a), question answering (Sarrouti et al., 2021c,a), text summarization (Goodwin et al., 2020; Yadav et al., 2021), and misinformation detection (Sarrouti et al., 2021b). In particular, RE with classification-based encoder-only pretrained transformers (BERT and variants) has been extensively studied (Lee et al., 2019; Peng et al., 2019a; Gu et al., 2022). In contrast, RE with pretrained language models based on encoder–decoder architecture, specifically Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020), has not been well-studied. Unlike encoder-only transformers, which are designed to predict a single prediction for an input sequence, T5 generates target tokens based on an encoder-decoder architecture.

In this paper, our goal is to compare pretrained sequence-to-sequence transformers with the encoder-only transformers for RE from biomedical texts. In order to satisfy this aim, we compare T5 with in-domain BERT-based models such as BioBERT and PubMedBERT on ten biomedical RE benchmark datasets. We also explore the use of multi-task fine-tuning (MTFT) on ten biomedical RE datasets (each with different entities and relation types) to investigate the correlation among four major biomedical RE tasks, namely chemical-protein relation extraction, disease-protein relation extraction, drug-drug interaction, and protein-protein interaction. Our experiments show that T5 performs better than in domain BERT-based models (encoder-only) such as BioBERT and PubMedBERT. The results also show that fine-tuning T5 with multi-task learning substantially improves the performance compared to single task fine-tuning.

2 Related Work

There has been a recent surge in interest from the NLP community to automatically extract re-

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 376–382
May 26, 2022. ©2022 Association for Computational Linguistics

lations between biomedical entities (proteins, gene, diseases, etc.) from the biomedical literature (Krallinger et al., 2008; Segura-Bedmar et al., 2013; Krallinger et al., 2017; Miranda et al., 2021). Recently, with the success of pretrained language models, several techniques based on transformers are widely utilized for extracting the relationships between entities from biomedical literature (Thillaisundaram and Togia, 2019; Wei et al., 2019; Hebbar and Xie, 2021; Hiai et al., 2021; Liu et al., 2021; Zhou et al., 2021; Su et al., 2021; Chang et al., 2021; Weber et al., 2021). The success of these systems has primarily been a result of encoder-only transformers such as BERT (Devlin et al., 2019) and its variants like SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2019), and PubMedBERT (Gu et al., 2022). Unlike RE with classification-based encoder-only transformers which have been widely studied, RE with encoder-decoder transformers has not been well-explored. Encoder–decoder-based transformer, specifically T5, (Raffel et al., 2020) has shown strong performance in various NLP tasks such as question answering and text summarization, etc.

In this work, we perform comprehensive comparisons of encoder-only transformers with the encoder-decoder transformer, specifically T5, on ten public biomedical relation extraction datasets. We also explore the use of multi-task learning to learn the shared complementary features across multiple biomedical relation extraction datasets.

3 Experiments

3.1 Problem statement

Given an input sentence S consisting of n tokens, i.e., $S = \{w_1, w_2, ..., w_n\}$ and a pair of entities (e_1, e_2) where $e_1 \in S$ and $e_2 \in S$, RE models are tasked with predicting the maximum probable label $\hat{y}$ from the set of labels in annotated data, y.

3.2 Datasets and processing

We explore ten benchmark datasets of RE between various entity types such as protein-protein, drug-drug, chemical-protein and disease-protein. Since the vast majority of relation instances are within single sentences in datasets of the aforementioned relation types, we model the RE task as sentence-level relation classification. The statistics of biomedical RE datasets are listed in Table 1.

Protein-protein interactions. We use five benchmark datasets, namely BioInfer, AIMed, IEPA, HPRD50, and LLL. These datasets are converted to a unified format by Pyysalo et al. (2008). Sentences that contain a pair of proteins are selected to generate positive and negative instances. All protein-protein pairs that occur in a sentence and do not have an explicit label in aforementioned datasets are considered as negative instances. Following previous work, we anonymized target named entities in a sentence using the pre-defined tag @PROTEIN$. For instance, a sentence with two protein names is represented as "*The POU domains of the @PROTEIN$ and Oct2 transcription factors mediate specific interaction with @PROTEIN$.*".

Drug-drug interactions. We use an existing preprocessed version of the Drug-Drug Interaction (DDI) 2013 corpus (Herrero-Zazo et al., 2013) and its corresponding train/dev/test split created by Peng et al. (2019b). Drug names were anonymized using the tag @DRUG$. For instance, a sentence with a pair of drug names is represented as "*Ketoconazole: @DRUG$ may inhibit both synthetic and catabolic enzymes of @DRUG$*". We evaluate four types of DDI relationships: "mechanism"', "effect", "advice", and "Int". The "mechanism" class defines the DDIs that are described by their pharmacokinetic mechanism. The "effect" type describes an effect or a pharmacodynamic mechanism in DDIs. The "advice" class describes DDIs that mention a recommendation or advice regarding a drug interaction. The "int" class is used when the text describes an interaction between drugs but without providing any additional information.

Disease-protein relationships. We use the existing preprocessed versions of the Genetic Association Database corpus (GAD) (Bravo et al., 2015) and EU-ADR datasets (van Mulligen et al., 2012). For both datasets, we use their corresponding train/dev/test splits created by Lee et al. (2019). Targeted entities were anonymized using the tags @DISEASE$ and @GENE$. For instance, a sentence with a pair of two entities (gene and disease in this case) is represented as "*In conclusion, @GENE$ 8092C > A polymorphism may modify the associations between cumulative cigarette smoking and @DISEASE$ risk.*"

Chemical-protein relationships. We use ChemProt (Krallinger et al., 2017) and DrugProt (Miranda et al., 2021) datasets that contain gene–chemical relations. For ChemProt, we use an existing preprocessed version and their corresponding train/dev/test split created by Peng et al.

Dataset	Train	Dev	Test	Metrics
AIMed	4938	-	549	micro F1
BioInfer	8544	-	950	micro F1
HPRD50	389	-	44	micro F1
IEPA	734	-	82	micro F1
LLL	300	-	34	micro F1
DDI	2937	1004	979	micro F1
ChemProt	4154	2416	3458	micro F1
DrugProt	17277	3765	-	micro F1
GAD	4796	-	534	micro F1
EU-ADR	318	-	37	micro F1

Table 1: Statistics of the biomedical relation extraction datasets. For DrugProt, we use the dev set as a test set.

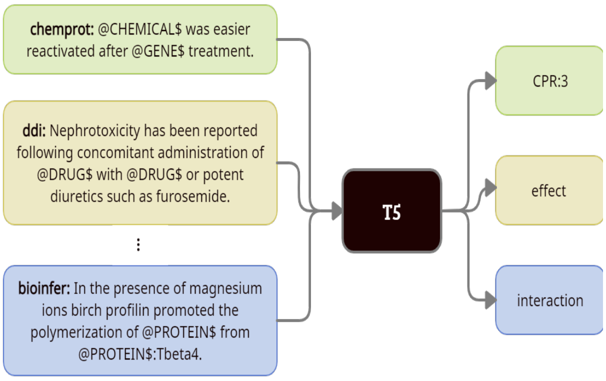
(2019b). We evaluate the same five classes: CPR:3, CPR:4, CPR:5, CPR:6 , CPR:9. The CPR:3 class describes upregulator, activator, and indirect upregulator. The CPR:4 class describes downregulator, inhibitor and indirect downregulator relation types. The CPR:5 category describes agonist, agonist activator and agonist inhibitor relation types. The CPR:6 type describes the antagonist relation. The CPR:9 class describes the following relation types: substrate, product of, and substrate product of. For DrugProt, we use the standard training and development sets in the DrugProt shared task and evaluate the same 13 classes: Activator, Agonist, Agonist-Inhibitor, Antagonist, Direct-Regulator, Indirect-Downregulator, Indirect-Upregulator, Inhibitor, Part-Of, Product-Of, Substrate, Substrate_Product-Of, Agonist-Activator. We first split abstracts into sentences using NLTK and then anonymized target entities in a sentence using the tags @CHEMICAL$ and @GENE$. For instance, a sentence with a pair of two entities (chemical and gene in this case) is represented as "*During differentiation, @CHEMICAL$ promoted early expression of osteoblast transcription factors, @GENE$ and osterix.*"

3.3 Models and setups

We compare in-domain BERT-based language models such as BioBERT (Lee et al., 2019) and PubMedBERT (Gu et al., 2022) with T5 (Raffel et al., 2020) and its variant SciFive (Phan et al., 2021), which is trained on biomedical texts (PubMed abstracts). For BERT-based models, we use a [CLS] token for the classification of relations. The [CLS] representation is fed into a softmax layer for a multi-way classification. For the T5-based models, the input sequence for the relation extraction task is "Processed sentence: [s] Relation: [r]". We fine-tuned T5 to generate tokens of relation types which are the ground truth labels in training datasets.

We also explore the use of MTFT on ten biomedical RE datasets. Figure 1 illustrates MTFT for RE tasks. We used the proportional and temperature-scaled task mixing as in (Raffel et al., 2020) for data mixture. During fine-tuning, a task-specific token (in our case, name of the dataset) is prepended to the input sequence.

In our experiments, we used the BioBERT (v1.1-base-PubMed), PubMedBERT, T5-base, and SciFive (SciFive-base-Pubmed) implementations provided in HuggingFace's Transformers package version 4.16.2 (Wolf et al., 2020). All models were trained with a batch size of 16 and maximum sequence length of 300 tokens for 10 epochs using single GPU (16 GB VRAM) on Amazon SageMaker. Adam optimiser with a learning rate of 1e-5 was used.



chemprot: @CHEMICAL$ was easier reactivated after @GENE$ treatment.

ddi: Nephrotoxicity has been reported following concomitant administration of @DRUG$ with @DRUG$ or potent diuretics such as furosemide.

⋮

bioinfer: In the presence of magnesium ions birch profilin promoted the polymerization of @PROTEIN$ from @PROTEIN$:Tbeta4.

Figure 1: Multi-task learning for biomedical RE

3.4 Results

In Table 2, we show the results of T5-based models compared to the in-domain and SOTA BERT-based models (pretrained on biomedical text) on ten benchmarking biomedical RE datasets, listed in Table 1. We compare the micro F1 scores obtained by T5 and its variant SciFive (pretrained on PubMed abstracts) to the BioBERT and PubMedBERT. On average (micro), T5 which was only pretrained on the general domain corpus, obtained a higher F1 score than BioBERT and PubMedBERT. T5 achieved the highest F1 scores on 5 out of 10 biomedical RE datasets. Models using biomedical text in pre-training generally perform better than models which pre-trained on general domain corpus. However, we observe that T5-scifive which was pre-trained on biomedical text (PubMed abstracts) did not perform well compared to T5.

We also explored the impact of MTFT on four

Relation	Datasets	BioBERT	PubMedBERT	T5	T5-SciFive	T5-MTFT
Protein-protein	AIMed	92.36	93.31	**94.35**	94.17	93.62
	BioInfer	**95.97**	94.59	95.36	95.89	95.16
	HPRD50	85.45	90.56	84.09	90.90	**95.95**
	IEPA	86.58	86.46	87.80	87.80	**90.24**
	LLL	88.24	**100.0**	97.05	94.11	97.05
Drug-drug	DDI	89.67	90.69	91.01	90.60	**91.83**
Chemical-protein	ChemProt	90.11	91.64	90.45	92.39	**96.56**
	DrugProt	88.69	89.40	88.71	**89.56**	89.37
Disease-protein	GAD	79.91	80.87	**81.46**	81.27	80.71
	EU-ADR	57.42	64.63	78.38	75.67	**83.78**
Average score		85.44	88.22	**89.47**	89.23	**91.42**

Table 2: Biomedical relation extraction test results. In T5-MTFT, we fine-tuned T5 with multi-task learning on ten datasets and then evaluate on the test set for each dataset.

benchmark biomedical RE tasks, i.e., drug-drug interaction, protein-protein interaction, chemical-protein relation extraction, and disease-protein relation extraction. On average, the results clearly show that the performance improves when using MTFT (an improvement of 1.95 F-score over the best single performing model). For instance, on the ChemProt dataset, T5-MTFT was able to achieve significant performance improvement of 6.11 and 6.45 F-score points over T5 and BioBERT respectively. While overall results indicate that MTFT provides improved RE performance on the four biomedical RE tasks (tasks with clear knowledge transfer), we observe a slight drop in the performance on some datasets such as AIMed, BioInfer, and GAD. In MTFT, we believe that in addition to the sample size of each task, the difficulty of the task/dataset can have an impact on the overall performance (the model underfits or overfits a dataset). More efforts and ablation studies are needed to study the impact of different biomedical RE tasks/datasets on downstream performance.

3.5 Error analysis

We performed a manual analysis of the test sets where the best performing model (T5-MTFT) predicted an incorrect label. Table 3 presents some examples.

Protein-protein interaction. The error analysis has shown that sentences are mostly classified incorrectly when they contain repetitive protein mentions (examples #1 and #3). Multiple protein mentions tend to add noise, which can prevent the model to extract the relevant contextual information. In addition, numerical or statistical findings might be a cause of error (example #1). We also observed that when the protein interacting words (e.g., bind, interact, localization) are mentioned in a sentence, the model predicts the class label “true” (i.e, interacting) (examples #2, #3 and #4).

Drug-drug interaction. The model tends to classify “Int” class as “Effect” type (examples #5 and #6). “Int” type is used whenever there exists an interaction between two drugs (i.e., a coarse-grained relation type). Having coarse-grained and fine-grained categories can be a cause of error. We also observed that when the input sentence contains some class-specific words (e.g., effect, interact, interaction, advise) that are not associated with the target entities, the model fails to predict the correct label (examples #7 and #8).

Chemical-protein relation extraction. Being a common source of mis-classification, the CPR:3 type was often predicted as CPR:4 and vice versa (examples #9 and #10). The CPR:3 class usually describes up-regulation, and its instances usually include up-regulation words such as “promote”, “increase”, and “activate”. The CPR:4 class is usually related to down-regulation and contains down-regulation words such as “decrease”, “inhibitor”, and “deposition”. Having both up-regulation and down-regulation words in the same sentence creates confusion, which can lead to mis-classification. The model also misclassified some instances due to the presence of multiple entities in a single sentence (example #11). Multiple entities can also create noise and make it difficult for the model to identify if there is a relation between the two target entities.

Disease-protein relation extraction. We found that our model fails to predict the correct label for instances (examples #12, #13, #14 and #15) that contain association words (e.g., associated)

	Example
(1)	**AIMed_sentence:** Chemokines that could compete with high affinity for MIP-1beta binding could also compete for monomeric gp120 binding, although with variable potencies; maximal @PROTEIN$ binding inhibition was 80% for MCP-2, but only 30% for @PROTEIN$. **Gold label:** True **Predicted label:** False
(2)	**AIMed_sentence:** We investigated whether @PROTEIN$, which binds to tyrosine-phosphorylated ITAM, interacts with @PROTEIN$ following T cell activation. **Gold label:** False **Predicted label:** True
(3)	**AIMed_sentence:** We further demonstrated that @PROTEIN$ and E3 but not @PROTEIN$ can decrease the fusogenic activity of Abeta(29-42) via a direct interaction. **Gold label:** False **Predicted label:** True
(4)	**BioInfer_sentence:** In localization studies with mammalian cells, all fusion proteins showed the localization expected for @PROTEIN$ in areas of high @PROTEIN$ dynamics, such as leading lamellae and ruffles induced by epidermal growth factor. **Gold label:** False **Predicted label:** True
(5)	**DDI_sentence:** Other drugs which may enhance the neuromuscular blocking action of @DRUG$ such as MIVACRON include certain antibiotics (e.g., aminoglycosides, tetracyclines, bacitracin, @DRUG$, lincomycin, clindamycin, colistin, and sodium colistimethate), magnesium salts, lithium, local anesthetics, procainamide, and quinidine. **Gold label:** Int **Predicted label:** Effect
(6)	**DDI_sentence:** @DRUG$ may decrease the effectiveness of oral contraceptives, certain antibiotics, @DRUG$, theophylline, corticosteroids, anticoagulants, and beta blockers. **Gold label:** Int **Predicted label:** Effect
(7)	**DDI_sentence:** Drugs Eliminated by Active Tubular Secretion: Although studies to assess drug-drug interactions with Sanctura have not been conducted, @DRUG$ has the potential for pharmacokinetic interactions with other drugs that are eliminated by active tubular secretion (e.g. digoxin, procainamide, pancuronium, morphine, @DRUG$, metformin and tenofovir). **Gold label:** Mechanism **Predicted label:** Int
(8)	**DDI_sentence:** Since Celontin (@DRUG$) may interact with concurrently administered @DRUG$, periodic serum level determinations of these drugs may be necessary (eg methsuximide may increase the plasma concentrations of phenytoin and phenobarbital). **Gold label:** Advise **Predicted label:** Int
(9)	**ChemProt_sentence:** EVn-50 possessed a broad spectrum of in vitro anticancer activity for those tested cancer cells, especially sensitive to MDA-MB-435, SKOV-3, BXPC-3, SMMC-7721, MCF-7, HO-8910, SGC-7901, BEL-7402, HCT-116, and 786-O, with the respective IC50 below 10mg/ml. Treatment with @CHEMICAL$ or VB1 resulted in arresting the MDA-MB-435 and SMMC-7721 cells at G2/M phase, which was further supported by observations of increased phosphorylation of Histone 3 at Ser10, phosphorylation of @GENE$ at Tyr15, expression of cyclin B1, and decreased expression of Cdc25c. **Gold label:** CPR:3 **Predicted label:** CPR:4
(10)	**ChemProt_sentence:** @CHEMICAL$ also increases Amyloid b (@GENE$) deposition and tau pathology. **Gold label:** CPR:4 **Predicted label:** CPR:3
(11)	**ChemProt_sentence:** Agonist and antagonist actions of yohimbine as compared to @CHEMICAL$ at alpha(2)-adrenergic receptors @GENE$, serotonin (5-HT)(1A), 5-HT(1B), 5-HT(1D). **Gold label:** CPR:5 **Predicted label:** CPR:6
(12)	**GAD_sentence:** Our results possibly indicate an association of @DISEASE$ with @GENE$ homozygosity (P=0.056). **Gold label:** False **Predicted label:** True
(13)	**GAD_sentence:** Our results suggest that the @GENE$ 168His variant is associated with reduced susceptibility to @DISEASE$. **Gold label:** False **Predicted label:** True
(14)	**GAD_sentence:** Our results indicate that the intron 2 CYP46 @GENE$ genotype may predispose to @DISEASE$, and this association is independent of the apolipoprotein E genotype. **Gold label:** False **Predicted label:** True
(15)	**GAD_sentence:** Although there remains a possibility that the @GENE$ TaqI A polymorphism plays some role in modifying the phenotype of the @DISEASE$, these results suggest that neither the A1 allele nor the homozygous A1 genotype is associated with alcoholism. **Gold label:** False **Predicted label:** True

Table 3: Examples of sentences that were incorrectly classified by the MTFT model.

with non-conclusive evidence (“possibly indicate”, “suggest”, “may predispose”, “possibility”).

4 Conclusion

In this paper, we present a comprehensive evaluation of encoder-only and encoder-decoder transformers on four benchmark biomedical RE tasks. We also explored the use of MTFT to investigate the correlation among these biomedical RE tasks. For that, we used ten popular datasets, namely AIMed, BioInfer, HPRD50, IEPA, LLL, DDI, ChemProt, DrugProt, GAD, and EU-ADR. The experiments showed that T5 and MTFT achieved better performance than BERT-based models (BioBERT and PubMedBERT) in extracting relations between bio-entities from texts. In the future, we plan to study the impact of each RE task/dataset on the downstream performance.

References

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, Hong Kong, China. Association for Computational Linguistics.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16(1).

Ting-Wei Chang, Tzu-Yi Li, Yu-Wen Chiu, Sheng-Jie Lin, Panchanit Boonyarat, Wen-Chao Yeh, Neha Warikoo, and Yung-Chun Chang. 2021. Identifying Drug/chemical-protein Interactions in Biomedical Literature using the BERT-based Ensemble Learning Approach for the BioCreative 2021 DrugProt Track. *Development*, 750(18858):199620.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ed drissiya El-allaly, Mourad Sarrouti, Noureddine En-Nahnahi, and Said Ouatik El Alaoui. 2021a. Deep-CADRME: A deep neural model for complex adverse drug reaction mentions extraction. 143:27–35.

Ed drissiya El-allaly, Mourad Sarrouti, Noureddine En-Nahnahi, and Said Ouatik El Alaoui. 2021b. MTT-LADE: A multi-task transfer learning-based method for adverse drug events extraction. *Information Processing & Management*, 58(3):102473.

Travis Goodwin, Max Savery, and Dina Demner-Fushman. 2020. Towards Zero-Shot Conditional Summarization with Adaptive Multi-Task Fine-Tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3215–3226, Online. Association for Computational Linguistics.

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.

Shashank Hebbar and Ying Xie. 2021. Covidbert-biomedical relation extraction for covid-19. In *The International FLAIRS Conference Proceedings*, volume 34.

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.

Satoshi Hiai, Kazutaka Shimada, Taiki Watanabe, Akiva Miura, and Tomoya Iwakura. 2021. Relation extraction using multiple pre-training models in biomedical domain. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 530–537, Held Online. INCOMA Ltd.

Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(Suppl 2):S4.

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. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Xiaofeng Liu, Kaiwen Tan, and Shoubin Dong. 2021. Multi-granularity sequential neural network for document-level biomedical relation extraction. *Information Processing & Management*, 58(6):102718.

Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2021. Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. In *Proceedings of the seventh BioCreative challenge evaluation workshop*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019a. 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, Florence, Italy. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019b. 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*. Association for Computational Linguistics.

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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Mourad Sarrouti, Asma Ben Abacha, and Dina Demner-Fushman. 2021a. Multi-task transfer learning with data augmentation for recognizing question entailment in the medical domain. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 339–346. IEEE.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021b. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mourad Sarrouti, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2021c. NLM at BioASQ Synergy 2021: Deep Learning-Based Methods for Biomedical Semantic Question Answering about COVID-19. In *2021 Working Notes of CLEF-Conference and Labs of the Evaluation Forum, CLEF-WN 2021*, pages 335–350.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Ashok Thillaisundaram and Theodosia Togia. 2019. Biomedical relation extraction with pre-trained language representations and minimal task-specific architecture. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 84–89, Hong Kong, China. Association for Computational Linguistics.

Erik M. van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A. Kors, and Laura I. Furlong. 2012. The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, 45(5):879–884.

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.

Qiang Wei, Zongcheng Ji, Yuqi Si, Jingcheng Du, Jingqi Wang, Firat Tiryaki, Stephen Wu, Cui Tao, Kirk Roberts, and Hua Xu. 2019. Relation extraction from clinical narratives using pre-trained language models. In *AMIA annual symposium proceedings*, volume 2019, page 1236. American Medical Informatics Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Shweta Yadav, Srivastsa Ramesh, Sriparna Saha, and Asif Ekbal. 2020. Relation extraction from biomedical and clinical text: Unified multitask learning framework. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Shweta Yadav, Mourad Sarrouti, and Deepak Gupta. 2021. NLM at MEDIQA 2021: Transfer learning-based approaches for consumer question and multi-answer summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 291–301, Online. Association for Computational Linguistics.

Yijia Zhang and Zhiyong Lu. 2019. Exploring semi-supervised variational autoencoders for biomedical relation extraction. *Methods*, 166:112–119.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14612–14620.