Pharma-InteractX Tool - LLM Application

Haneen Alfauri
Washington University in St. Louis

Lisa Liao
Washington University in St. Louis

Alice Wang
Washington University in St. Louis

Abstract

Physicians often prescribe several drugs together to combat different aspects of complex diseases. Understanding the effects of the interactions between these drugs is vital for ensuring drug safety and avoiding adverse effects due to unintended drug-drug interactions (DDIs). Databases like DrugBank (Knox et al., 2023) house valuable DDI information, but often require significant efforts from experts to update and maintain due to the rapidly growing literature volume. Recent advances in fine-tuning large language models (LLMs) for relation extraction (RE) have garnered attention in many fields, including DDI extraction from literature. Luo et al. (Luo et al., 2020) used a BiLSTM-CRF to jointly perform name entity recognition (NER) and RE on the DDI-2013 corpus (Herrero-Zazo et al., 2013). On the other hand, Weber et al. (Weber et al., 2022) have explored different ensembles of pre-trained LLMs for drug-protein RE using the DrugProt corpus (Miranda-Escalada et al., 2023), and incorporated knowledge graphs (KGs) to improve their extraction performance. In this report, we explore whether combining the sequence labeling framework from Luo et al. and knowledge graph from Weber et al. with 3 pre-trained LLMs (BERT-base, PubMedBERT, BioM-ELECTRA) to improve DDI extraction performance from the DDI-2013 corpus. We find that our fine-tuned models performed worse than the BiLSTM-CRF model from Luo et al. in the NER task, but has higher precision in the RE task (0.838, 0.729, and 0.828, respectively) than the 0.722 achieved by Luo et al. However, the recall for the RE task is moderately worse for all of our models (0.628, 0.577, and 0.599 versus 0.685). Incorporating information from a biomedical KG like PharmKG8k (Zheng et al., 2020) can assist in recovering false negative relations in the RE task, but the improvement is marginal due to the limited number of relation overlaps between PharmKG8k and the DDI-2013 corpus.

1 Introduction

Biomedical researches utilize information of different abstraction degrees, including words, disease codes, amino acids, and DNA. Recently, the volume of biomedical literature is growing continuously and it is challenging to extract information efficiently, especially for relations between biological entities (Miranda-Escalada et al., 2023). Efforts have been made to manually curate this information into a variety of databases such as ChEMBL (Gaulton et al., 2016), DrugProt (Miranda-Escalada et al., 2023) and DrugBank (Knox et al., 2023). However, these databases often require significant manual effort in maintenance and updates. This poses a challenge to centralize the information from fragmented literature.

To capture and mine the biomedical information and knowledge from texts, there is growing attention in the biomedical NLP community to adopt pre-trained, domain-specific large-language models (LLMs). Pre-trained LLMs could leverage these massive sequences without biomedical knowledge abstraction and human annotations, including but not limited to plain biomedical text, biomedical images, general text, protein sequences, and DNA sequences.

Relation extraction (RE) is an NLP task that concerns identifying and classifying relations/interactions between named entities extracted from texts. In order to perform accurate RE, correct name entity recognition (NER) is equally important. Some of the most common transformer-based pre-trained LLMs in the biomedical domain like BioBERT(Lee et al., 2019a), PubMedBERT(Gu et al., 2021) have been extensively tested for both NER and RE tasks, and achieved good performance.

Amongst the complex interactions between biological entities, drug-drug interactions (DDIs) is of special interest to researchers and physicians.

A comprehensive knowledge on these interactions is crucial for avoiding adverse DDIs and ensure safe drug administration. The availability of annotated datasets like the DDI-2013 corpus (Herrero-Zazo et al., 2013) provide a means to fine-tune pre-trained LLMs to improve DDI extraction from biomedical texts. Traditional approaches in DDI extraction primarily rely on classifying biomedical texts, which often lacks the structural and semantic richness needed to fully capture complex biomedical relationships. Methods for jointly Recognizing this gap, recent advancements have also incorporated knowledge graphs into language models to enhance predictive performance by providing a deeper contextual understanding (Weber et al., 2022).

2 Related Work

In the biomedical domain, both NER and RE are critical for tasks such as identifying drug-drug interactions, which are pivotal for clinical decision-making. Early work in this area, such as the BioCreative challenges, provided valuable benchmarks and data for NER and RE tasks in the biomedical literature. Traditional approaches initially used rule-based systems for detecting interactions, with Percha and Altman pioneering this work (Percha and Altman, 2013). Subsequently, machine learning methods like the convolutional neural networks used by Zhao et al. (Zhao et al., 2016) enhanced the ability to extract drug-drug interactions by leveraging sentence-level features, offering significant improvements over earlier systems.

Joint models that simultaneously perform NER and RE have demonstrated the potential to outperform sequential models, where these tasks are handled separately. Xu et al. introduced a neural network that integrates these tasks, highlighting the efficiency of joint models in complex entity and relationship extraction scenarios (Xu et al., 2018). Luo et al. used a BiLSTM-CRF model with attention mechanism together with a novel sequence labeling framework to achieve better joint NER and RE performance over previous CNN based models (Luo et al., 2020).

The advent of transformer-based models, particularly those pre-trained on extensive text corpora, has been a game-changer in the field. Models such as BERT and its biomedical variant, BioBERT, have set new standards for NER and RE tasks (Devlin et al., 2018; Lee et al., 2019b). These models effectively capture complex linguistic nuances, drastically improving the extraction accuracy for biomedical applications. Despite these advancements, challenges remain in managing technical terminology and sparse data, indicating a need for innovative approaches in future research. The recent work by Fang et al. (Fang et al., 2024) is an example of integrating knowledge graphs with machine learning to enhance model performance, providing insights into survival prediction and biomarker discovery in cancer research. This method shows the growing trend of integrating structured external knowledge to improve the accuracy and reliability of biomedical information extraction. With RE as the main task, Weber et al. have successfully incorporated KG entity embeddings and textual side information from KGs with RoBERTa-large to effectively extract drug-protein relations (Weber et al., 2022). In our brief literature review, we have not seen joint NER and RE models used in conjunction with KG information. The rest of this report attempts to test one such combination using works from Luo et al. (Luo et al., 2020) and Weber et al. (Weber et al., 2022).

3 Approach

We combine the sequence labeling framework from Luo et al. (Luo et al., 2020) and the incorporation of knowledge graph information from Weber et al. (Weber et al., 2022) and Fang et al. (Fang et al., 2024) in an attempt to improve DDI RE. Based on Luo et al., we reframe the sentence classification task used in most RE models into a joint NER and RE problem.

Sequence labeling framework: Each input sentence is tokenized and each token receives a compound label according to the following rules: (1) "O" for non-entity tokens; (2) [drug type] for entity tokens not in a relation; (3) [drug type]-[relation type]-[1 or 2] for entity tokens in a relation. All labels are prefixed with the BIES tags to identify entity boundaries. The [1 or 2] suffix represent whether the entity is a head or a tail entity in the relation – head entity (suffix 1) can only be in a relation with a tail entity (suffix 2) during the relation extraction. With 4 drug types and 4 relation types, this labeling scheme produces a total of 126 distinct compound labels. The pre-trained LLMs try to classify each token into one of the compound label class.

Relation extraction: From the predicted labels, we extract all entities predicted to be in a relation. For each entity with suffix 1, we search to the right of the entity until we find the nearest entity with suffix 2 and the same relation type. If a multi-token entity is labeled with multiple relation types, we take the relation type of the first token. If the head entity relation type tag differs from that of the tail entity, we take the relation type from the head entity. During evaluation, predicted relations are considered a true positive if both entity strings and the relation type exactly match the ground truth annotations. An example input and output of the joint NER and RE task can be seen in Figure 1.

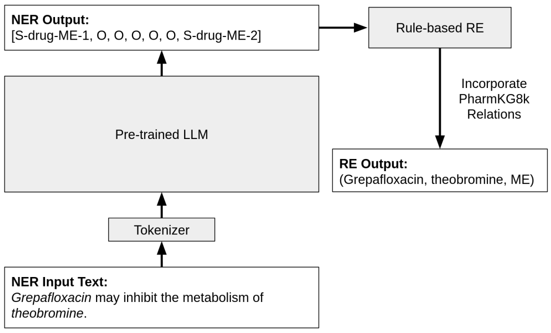
NER Output:
[S-drug-ME-1, O, O, O, O, O, S-drug-ME-2]
Rule-based RE
Incorporate PharmKG8k Relations
Pre-trained LLM
RE Output:
(Grepafloxacin, theobromine, ME)
Tokenizer
NER Input Text:
Grepafloxacin may inhibit the metabolism of *theobromine*.

Figure 1: General work flow for joint NER and RE.

Model architecture: We tested 3 pre-trained LLMs: BERT (Devlin et al., 2019), PubMedBERT (Gu et al., 2021) and BioM-ELECTRA (Alrowili and Shanker, 2021) as base models. A classification head is added to obtain token label predictions. We additionally tested an ensemble model where the token label predictions are determined by majority voting from the 3 base models. The training objective is to minimize the cross-entropy loss and predict a label for each input token. We will compare the performance of the 3 base models under the sequence labeling framework both before and after incorporating information from KG. To select the best hyperparameters, we also tested 3 weight decay rates (0.1,0.01,0.001) with the AdamW optimizer to prevent overfitting. The model with the lowest validation loss during training is selected as the final model.

Knowledge graph incorporation: We use known DDIs from PharmKG8k (Zheng et al., 2020) to recover missed relations resulted from mislabeled entities. Due to computational limitations, we use PharmKG8k (hereon referred to as PharmKG), a smaller subset of the full PharmKG. After obtaining predicted token labels, we extract all predicted entities and query PharmKG for all relations involving these entities (referred to as KG relations for the rest of this report). During RE, for each relation in the KG relations list, we check whether both entities exist in a sentence and whether at least one of the two has been labeled with a relation type. If so, these 2 entities are extracted into a relation, where the relation type is determined by the available relation type(s) in the predicted label. In the case of conflicting relation type between the 2 entities, we take the relation type of the head entity.

Code usage: Code for pre-processing the DDI-2013 corpus and sequence labeling was completely re-written from scratch due to unavailable code from Luo et al. Code logic for labeling the sentences for training was deduced from the Luo et al. (Luo et al., 2020). Code for fine-tuning the baseline models was adapted from a Github tutorial (here). Post-processing scripts is directly from the tutorial while evaluation scripts were written from scratch. Code for querying and incorporating PharmKG were also written from scratch.

In summary, our project makes 2 main contributions: (1) Evaluate the joint NER and RE sequence labeling framework on pre-trained LLMs instead of BiLSTM based models. (2) Utilize knowledge graphs to the pre-trained LLMs and evaluate whether side information enhance the NER and RE results.

4 Experiments

4.1 Data

DDI-2013 corpus: The DDI-2013 corpus contains annotated abstracts and full texts from DrugBank (Knox et al., 2023) and MedLine. The training set contains 571 DrugBank abstracts and 142 MedLine full texts, whereas the test set contains 158 DrugBank abstracts and 33 MedLine full texts. Each piece of text was annotated with (1) entities within the text, (2) types of the entities, (3) relations between the entities and (4) types of the relations. There are 4 types of entities (drug, brand, group, drug_n) and 4 types of relations between the entities (mechanism, effect, advise, int). A detailed breakdown of the entity and relation types by data split can be seen in Figures 2 and 3.

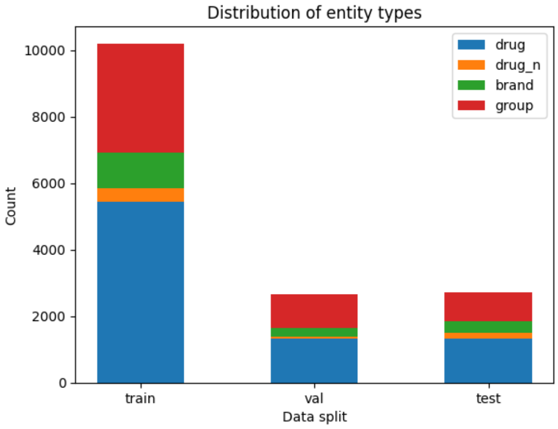

Figure 2: Distribution of entity types by data splits.

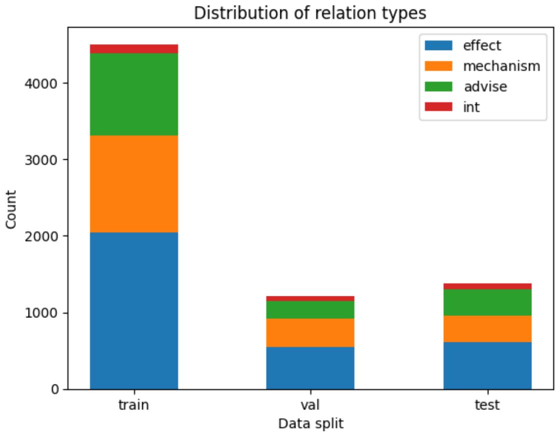

Figure 3: Distribution of relation types by data splits.

The entity type drug and drug_n both refers to generic drug names (e.g. alcohol, ascorbic acid etc.). The drug type is approved for human use while the drug_n type is not. The brand type refers to brand names of generic drugs, as different companies can produce the same drug with different names. In addition, generic drugs can be divided into groups with similar functions, and the group type is used to describe such drug groups.

The relation type mechanism and effect are closely related. The mechanism type describes DDIs by their pharmacokinetics (PK) mechanism, where PK mechanism refers to how a drug interacts with another at the level of drug absorption and metabolism (e.g. *Grepafloxacin* may inhibit the metabolism of *theobromine*) (Herrero-Zazo et al., 2013). The effect type describes the effect of DDIs (e.g. *clarithromycin* causes rash) or their pharmacodynamic (PD) mechanism, where PD mechanism refers to the direct interaction between 2 drugs (e.g. *Chlorthalidone* may potentiate the action of other *antihypertensive drugs*) (Herrero-Zazo et al., 2013). The advise type describes any comments or advise that has been given regarding a DDI (e.g. *UROXATRAL* should not be used in combination with other *alpha-blockers*). The int type describes any mention of DDIs without further detail (e.g. *omeprazole* and *ketoconazole* interact).

PharmKG: PharmKG (Zheng et al., 2020) is a comprehensive knowledge graph designed for the pharmaceutical domain, integrating a wealth of information related to drugs, their interactions, effects, and associations with diseases and genes. It serves as an invaluable resource for researchers and healthcare professionals who seek to understand the complex web of pharmacological data and apply this knowledge to clinical practice and drug development. PharmKG was curated from six renowned public databases, ensuring the utilization of high-quality, structured data. The curated interactions pertinent to DDIs encompass: Treats (T), indicating drugs that mitigate the effects of other drugs; Treated By (Te), detailing drugs that are mitigated by others; Inhibits (I), for drugs that restrain the effects of other drugs; Binds To (B), which denotes the binding action of drugs to others; Interacts With (Iw), which covers general drug interactions; Antagonizes (An), for drugs that counteract the effects of others; Disrupts (D), which involves drugs disrupting the function of other drugs; and Activates (A), highlighting drugs that enhance the effects of other drugs, (CC) which indicates a relationship between chemicals. Interaction types that are not directly implicated in DDIs have been excluded from our method.

4.2 Evaluation Methods

We use micro-averaged precision, recall and F1 scores across all relation types for evaluation. Only positive labels are included in calculating the metrics (true negative "O" labels are excluded).

4.3 Experimental Details

We fine-tuned 3 different base models: (1) PubMedBERT-base-uncased-abstract-fulltext , (2) BERT-base-uncased, and (3) BioM-ELECTRA-Large-Discriminator using the DDI-2013 training and validation set. We used a linear learning rate schedule with the AdamW optimizer. The initial learning rate is 1e-5 with 3,000 warm up steps. Batch size is fixed at 16 for all models as it is the largest batch we can use without exceeding Colab limits. For each base model, we choose the max sequence length based on the distribution of tokenized sentence lengths (Appendix Figure 1-3). We also tested 3 weight decay rates (0.1, 0.01, 0.001) for each base model to prevent overfitting. All models are trained for 30 epochs on 1 T4 GPU in a Google Colab notebook. A checkpoint is saved after each epoch, and the checkpoint with the loweset validation loss is chosen as the final model. Table 1 shows the final model configurations used to produce the results in this report.

Model	Batch size	Best weight decay
PubMedBERT	256	0.1
BERT	320	0.1
BioM-ELECTRA	256	0.1

Table 1: Final model hyperparameters

4.4 Results

Although we framed the task of our method to be joint NER and RE, we are mainly interested in the RE results and view NER as an auxiliary task. Table 2 and Table 3 shows the NER and RE performances from the 3 base models and the ensemble model on the test data. Rows with "+KG" show the results after recovering relations using PharmKG queries. The highest scores across the 4 models are in **bold**.

	Precision	Recall	F1
PubMedBERT	0.74	0.75	0.75
BERT	0.67	0.69	0.68
BioM-ELECTRA	**0.78**	**0.77**	**0.77**
Ensemble	0.76	**0.77**	0.76

Table 2: Base models' NER performance on test set.

	Precision	Recall	F1
PubMedBERT	0.838	0.628	0.718
PubMedBERT+KG	**0.838**	0.629	**0.718**
BERT	0.729	0.577	0.644
BERT+KG	0.731	0.581	0.647
BioM-ELECTRA	0.828	0.599	0.695
BioM-ELECTRA+KG	0.828	0.600	0.696
Ensemble	0.831	0.607	0.701
Ensemble+KG	0.831	0.608	0.702
Luo et al.	0.722	**0.685**	0.702

Table 3: Base models' RE performance on test set.

PubMedBERT and BioM-ELECTRA both performed better than BERT in the NER task, with BioM-ELECTRA being the top performer. The less accurate predictions from BERT is expected, as it is the only model out of the 3 that has not been pre-trained on biological texts. Surprisingly, the ensemble model is marginally worse than BioM-ELECTRA alone. However, all 3 models are far from the current state of the art (micro-F1 > 0.9). Luo et al. reported a F1-score of 0.911 for their BiLSTM-CRF model, which also exceeds the NER performance we obtained.

For the RE task, PubMedBERT is the best performing model with or without KG queries. PharmKG queries consistently recovered a very small number of false negative relations for all 3 models, marginally improving the recall metric for PubMedBERT, BERT and the ensemble model. Similar to the NER task, the ensemble model did not produce better results than PubMedBERT+KG model alone. When compared to the BiLSTM-CRF from Luo et al., all 3 of our models reached better precision, with PubMedBERT+KG performing better in the F1 score as well due to much higher precision. However, the recall of all models are at least 0.06 lower than that of Luo et al., which requires further investigation into our prediction results.

Overall, the increased precision of our models shows the value of using pre-trained LLMs as the base models for extracting true positive relations under this sequence labeling framework. The use of known DDIs from PharmKG contributes to the recovery of some missed relations due to mis-labeled entities, but the improvements are extremely limited. This can be mainly attributed to the small overlap between entities and DDIs in PharmKG8k and DDI-2013, where the largest number of relations found from test set entities is 648 – only 21.4% of the 3,031 annotated relations in the DDI-2013 test set. Using the full PharmKG that includes 3 fold more entities and relations may provide more improvements in the recovery of false negatives. However, the amount of relations we can recover using the KG queries will always be upper-bounded by what is available in the KG itself. On the other hand, the higher RE false negative rate suggests either larger error propagation from the NER task, or limitations in the rule-based extraction. Due to code unavailability from Luo et al., we do not have a ground truth to compare our implementation to. In the next section, we conduct an error analysis to study both the NER task error propagation and behavior of our RE implementation in comparison to the descriptions provided by Luo et al.

5 Analysis

We investigate 3 main components of our method: (1) whether the large number of token label classes negatively impacted NER and downstream RE performance, (2) the main cause of erroneous predictions in both the NER and RE tasks and (3) how much errors in the NER task propagated to the RE task. We use outputs from PubMedBERT+KG for all analysis in this section since it produced the best RE results.

NER error analysis: To analyze the main sources of errors, we will refer to labels with relation types as "relation labels" (e.g. S-drug-ME-1) and those without relation types as "non-relation labels" (e.g. S-drug).

A false positive entity means an "O" token was predicted with a non "O" label (vice versa for false negatives). We refer to these errors as FP or FN entity. A false positive relation means a non-relation entity was predicted with a relation label (vice versa for false negatives). We refer to these errors as FP or FN relation. A mismatched entity means a non-relation label was correctly predicted, but with the wrong entity type and/or BIES tag. A mismatched relation means that there is a mismatch in one of the 4 tags in the label. We refer to both of these errors as mismatches. An example of these error types can be seen in Table 4.

	True label	Pred label
FP entity	O	S-drug
FN entity	S-drug	O
FP relation	S-drug	S-drug-ME-1
FN relation	S-drug-ME-1	S-drug
Mismatch entity	S-drug	S-group
Mismatch relation	S-drug-ME-1	S-drug-AD-1

Table 4: Error type examples.

There are 433 sentences with at least 1 mis-labeled token, and a total of 1,046 errors. The most common type is mismatched relation, accounting for 26.1% of the total errors. The errors for non-relation entities (mismatched entity, FP entity, FN entity) are lower in frequency, but still represents 15.1% , 14.3% and 9.7% of total errors in the model. Through manual examination, the most common erroneous label for all error types involves mis-identifying drug_n entities into drug entities. This is likely due to drug_n being the rarest entity type in all data splits (Figure 2), therefore the model does not have enough instances to learn from. The drug entity type is the most abundant, but the drug entities and drug_n entities are also the most semantically similar. Hence these 2 entity types may be particularly hard to distinguish.

In addition to the mis-identification of entity types, we further looked into whether the FP/FN relation errors resulted from incorrect entity boundaries (i.e. wrong BIES tag) or incorrect entity type. For mismatched relation errors, an incorrect relation type is the main cause and comprises 59.7% of all mismatched relation errors. The most common mismatch is mis-classifying an "effect" relation type into a "mechanism" relation type. Despite being the 2 most abundant relation types in the DDI-2013 corpus (Figure 2), these two relation types also have semantically similar definitions, which potentially have led to confusion in learning.

Relation labels are the largest and most complex compound labels in the sequence labeling framework, and is therefore the hardest to distinguish between. The models will have to learn to classify tokens to the correct entity type in conjunction with the relation type, which is a harder task than classifying tokens into entity types or relation types alone. Further, the number of training instances for each of the compound relation label is small, which makes classification more challenging. The top 3 most common error types are all relevant to relation entities (mismatcted relation, 26.1%; FP relation, 17.4% ; FN relation, 9.7%), further demonstrating the difficulty in predicting the compound relation labels. To understand whether the large number of compound token labels hindered the model's ability to accurately classify each token, and whether the joint entity type plus relation type labels improve RE results, we tag all entities with the same entity type (drug type) in the token labels (e.g. S-group-ME-1 becomes S-drug-ME-1) and re-ran all models for NER and RE. This simplification reduced the number of possible compound labels from 126 to 41, increasing all NER performance metrics by 0.07. However, the RE performance decreased by 0.03-0.04, suggesting there is some value in learning both entity types and relation types together in this labeling framework.

RE error analysis and error propagation: To investigate errors in the RE task, we manually examined 30 test set sentences with at least 1 false positive or false negative relations, totaling 67 errors. Below, we provide some examples of common errors observed. Erroneous relations are in red . Relations are displayed in the form of triplets: (head entity, tail entity, relation type).

The first type of error occurs when a head entity is in a relation with multiple tail entities of the same type. In the example below,

- Text: Other HDAC Inhibitors: Severe thrombocytopenia and gastrointestinal bleeding have been reported with concomitant use of ZOLINZA and other HDAC inhibitors (e.g. valproic acid).
- True relations: (HDAC Inhibitors, ZOLINZA, effect), (HDAC Inhibitors, valproic acid, effect)
- Extracted relation: (HDAC Inhibitors, ZOLINZA, effect)

There are 2 true relations in this sentence, but only 1 relation is picked up by all 4 models. NER label predictions are correct for all tokens in this sentence: and the false negative relation is due to the extraction rule only allowing a head entity to match with the nearest tail entity of the same type. As valproic acid is the second tail entity to the right of HDAC Inhibitors, it becomes a false negative relation. This type of error is the most common source of false negatives, comprising 67.2% (45) of the total errors we examined.

The second type of error occurs when there is a erroneous label prediction from the NER task.

- Text: Concurrent and/or sequential systemic or topical use of other potentially neurotoxic and/or nephrotoxic drugs, such as amphotericin B, aminoglycosides, bacitracin, polymyxin B, colistin, viomycin, or cisplatin, when indicated, requires careful monitoring.
- True relations: None
- Extracted relation: (amphotericin B, cisplatin, advise)

There are multiple entities in this sentence, all have the correct NER label except amphotericin B and cisplatin. The predicted NER labels for both are [B-drug-AD-1, E-drug] and [S-drug-AD-2], whereas the true NER labels are [B-drug, E-drug] and [S-drug]. The NER task correctly labeled the entity type and boundary for both, and correctly labeled the second token of amphotericin B, but falsely added a relation to both the first token of amphotericin B and cisplatin, leading to a false positive RE result. RE errors due to false positive or false negative relation labels from NER account for 27.4% (19) of the total errors we examined.

The third type of error is due to entity boundary mis-labeling resulted from overlapping entities:

- Text: Other eye drops or medications such as acetylcholine chloride (Miochol) and carbachol (Carboptic, Isopto, Carbachol) may decrease the effects of suprofen ophthalmic.
- True relations: (acetylcholine chloride, suprofen, effect), (miochol, suprofen, effect), (carbachol, suprofen, effect), (carboptic, suprofen, effect), (isopto carbachol, suprofen, effect)
- Extracted relations: (acetylcholine chloride, suprofen, effect), (miochol, suprofen, effect), (carbachol, suprofen, effect), (carboptic, suprofen, effect), (isopto , suprofen, effect)

There are 5 true relations in this sentence, and we were able to correctly extract 4 of them. The 5th relation we extracted is correct in both the tail entity and the relation type, but mismatched in the head entity. The predicted labels for isopto carbachol is [S-brand-EF-1, S-brand-EF-1], but the true label is [B-brand-EF-1, E-brand-EF-1]. The two-token entity was mis-identified as 2 single-token entities due to incorrect BIES tags. Hence when we encounter the the "isopto" token in isopto carbachol, we immediately start searching for the closest tail entity with relation type effect (EF) and suffix 2 to the right, ignoring "suprofen". The entity boundary error is likely because carbachol is also a single-token entity in the same sentence. In addition, carbachol is also in an effect type relation with suprofen just like isopto carbachol is. This type of error accounts for 5.97% (4) of the 67 errors we examined.

The error analysis suggests that only extracting the nearest tail entity is likely the main cause for false positives in the RE task. Further, RE results are heavily dependent on the accuracy of the NER predictions. To evaluate the performance of the rule-based RE in the absence of NER errors, we replace all predicted token labels with their true labels, then re-extracted the relations. False positive relations are significantly reduced (only 8 left), which resulted in a significant increase in precision from 0.838 to 0.987, recall increased from 0.629 to 0.670 and F1-score increased from 0.718 of 0.798. This indicates that the error propagation from NER mainly increased false positive rate, whereas the "nearest rule" for extracting relations is the primary cause for false negatives. For each head entity, if we continue searching beyond the nearest tail entities of the same relation type, the recall metric can be improved from 0.629 to 0.891. However, the precision also suffers significantly (0.838 to 0.638), likely because in the case where the sentence tokens are ordered as [head1, head2, O, O, tail1, tail2], the tail entities beyond the nearest one can be in a relation with other head entities to the right of the current head.

6 Conclusion

In this project, we fine-tuned 3 pre-trained LLMs for joint NER and RE for DDI extraction based on the sequence labeling framework introduced in Luo et al. Although we did not outperform Luo et al. in the NER task, we were able to achieve higher precision in the RE task, affirming the value of using pre-trained LLMs under this framework. In addition, we supplemented the rule-based RE process with PharmKG relations and consistently made marginal improvements on the recall metric. Due to computational resource constraints, PharmKG8k was used instead of the full PharmKG dataset. This resulted in limited entity and relation overlaps between PharmKG and the DDI-2013 corpus. Being able to utilize the full PharmKG dataset could potentially improve our recall metric. The lower NER performance from our models led to a drop in RE precision due to error propagation. Swapping the base models we tested here for a model designed specifically for biomedical NER can increase both NER and RE performance. Another significant limitation lies in the design of the rule-based RE process, where extracting tail entities based on proximity led to significantly higher false negative rates. In the future, a better tagging scheme could be designed to aid the LLMs in identifying head and tail entities in a specific relation. Further, KG embeddings and textual side information can also be evaluated in the future to see if they improve DDI extraction performance.

One of the larger limitations for the method we employed (and many other RE models) lies within the size of the benchmarking datasets and an overall lack of generalizability assessment. Although there have been multiple gold standard datasets that language models frequently use for benchmarking, these manually curated and annotated datasets may not be representative of the larger biomedical corpus in general. Granted, the absence of generalizability evaluation in current models can be attributed to the lack of "ground truths" in the biological domain. In recent years, technological advancements have allowed more accurate experimental measurements both *in vitro* and *in vivo* on larger scales. As high-throughput biological measurements continues to be generated, there are potentials in curating a large enough dataset to allow a more comprehensive evaluation of the generalizability of existing models.

7 Appendix

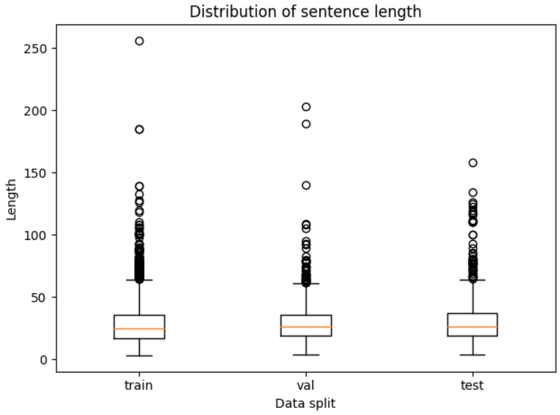

Figure 1: PubMedBERT tokenized input lengths distribution.

References

Sultan Alrowili and Vijay Shanker. 2021. BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Online, pages 221–227. https://www.aclweb.org/anthology/2021.bionlp-1.24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep

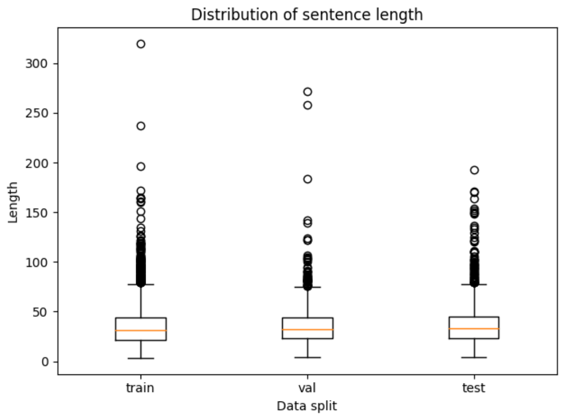

Figure 2: BERT tokenized input lengths distribution.

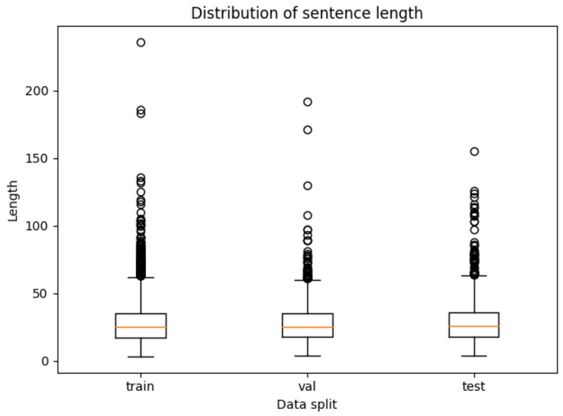

Figure 3: BioM-ELECTRA tokenized input lengths distribution.

bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

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

Chao Fang, Gustavo Alonso Arango Argoty, Ioannis Kagiampakis, Mohammad Hassan Khalid, Etai Jacob, Krishna Bulusu, and Natasha Markuzon. 2024. Integrating knowledge graphs into machine learning models for survival prediction and biomarker discovery in patients with non–small-cell lung cancer. *bioRxiv* https://doi.org/10.1101/2024.02.29.582842.

Anna Gaulton, Anne Hersey, Michał Nowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J. Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María Paula Magariños, John P. Overington, George Papadatos, Ines Smit, and Andrew R. Leach. 2016. The ChEMBL database in 2017. *Nucleic Acids Research* 45(D1):D945–D954. https://doi.org/10.1093/nar/gkw1074.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare* 3(1):1–23. https://doi.org/10.1145/3458754.

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. https://doi.org/https://doi.org/10.1016/j.jbi.2013.07.011.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun (Lucy) Chin, Seth A Strawbridge, Marysol Garcia-Patino, Ray Kruger, Aadhavya Sivakumaran, Selena Sanford, Rahil Doshi, Nitya Khetarpal, Omolola Fatokun, Daphnee Doucet, Ashley Zubkowski, Dorsa Yahya Rayat, Hayley Jackson, Karxena Harford, Afia Anjum, Mahi Zakir, Fei Wang, Siyang Tian, Brian Lee, Jaanus Liigand, Harrison Peters, Ruo Qi (Rachel) Wang, Tue Nguyen, Denise So, Matthew Sharp, Rodolfo da Silva, Cyrella Gabriel, Joshua Scantlebury, Marissa Jasinski, David Ackerman, Timothy Jewison, Tanvir Sajed, Vasuk Gautam, and David S Wishart. 2023. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Research* 52(D1):D1265–D1275. https://doi.org/10.1093/nar/gkad976.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019a. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4):1234–1240.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019b. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* .

Ling Luo, Zhihao Yang, Mingyu Cao, Lei Wang, Yin Zhang, and Hongfei Lin. 2020. A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of Biomedical Informatics* 103:103384. https://doi.org/https://doi.org/10.1016/j.jbi.2020.103384.

Antonio Miranda-Escalada, Farrokh Mehryary, Jouni Luoma, Darryl Estrada-Zavala, Luis Gasco, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2023. Overview of DrugProt task at BioCreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical–protein relations. *Database* 2023:baad080. https://doi.org/10.1093/database/baad080.

Bethany Percha and Russ B. Altman. 2013. Informatics confronts drug-drug interactions. *Trends in Pharmacological Sciences* .

Leon Weber, Mario Sänger, Samuele Garda, Fabio Barth, Christoph Alt, and Ulf Leser.

2022. Chemical–protein relation extraction with ensembles of carefully tuned pretrained language models. *Database* 2022:baac098. https://doi.org/10.1093/database/baac098.

Kehai Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2018. A joint model of entity recognition and relation extraction based on a hybrid neural network. *Neurocomputing* .

Zhihao Zhao, Zhi Yang, Lin Luo, Hongfei Lin, and Jian Wang. 2016. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics* .

Shuangjia Zheng, Jiahua Rao, Ying Song, Jixian Zhang, Xianglu Xiao, Evandro Fei Fang, Yuedong Yang, and Zhangming Niu. 2020. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in Bioinformatics* 22(4):bbaa344. https://doi.org/10.1093/bib/bbaa344.