Reproducibility Project: Extracting Drug-drug Interactions from Biomedical Texts using Knowledge Graph Embeddings and Multi-focal Loss

Arun Joy and Alexei Romanov
{arunjoy2, romanov2}@illinois.edu

Group ID: 117, Paper ID: 26
Presentation link: https://www.youtube.com/watch?v=_0xI6iSqia4
Code link: https://colab.research.google.com/drive/1-9wox4J0ey0s16y3N3p9kchRxH30jgFR?usp=sharing

1 Introduction

Drug-Drug interaction (DDI) databases contain data on adverse reactions between pairs of drugs. Adverse drug reactions can cause side effects or even patient death when administered together. So, such databases are valuable for safely prescribing drugs for patients.

With the advent of Bidirectional Encoder Representations from Transformers(BERT) in 2018, efforts focused on using BERT, pre-trained on Biomedical text, to aid DDI extraction. To enhance performance further, external drug knowledge databases were incorporated into the architecture.

Extracting Drug-drug Interactions from Biomedical Texts using Knowledge Graph Embeddings and Multi-focal Loss (Jin et al., 2022), a CIKM '22 paper, created a novel architecture incorporating a PubMedBERT-based neural network and knowledge graph for DDI classification. Compared to previous such efforts like (Mondal, 2020), the paper also uses multi-level textual features from the text and a novel multifocal loss function, KGE-MFL, to handle imbalances in the training data set. This architecture was able to increase overall accuracy by reducing misclassifications.

Our work aims to reproduce the paper "Extracting Drug-drug Interactions from biomedical texts using Knowledge graph embeddings and Multi-focal Loss " (Jin et al., 2022). Specifically, we will rebuild the proposed framework for DDI classification, attempt to reproduce the evaluation results, and do the ablation experiments.

The DDI classification framework is comprised of a neural network component and a knowledge graph. BERT embeddings of the drug-interaction sentence is passed into multiple CNN networks to find the positional and grammatical features of the sentence. An attention mechanism over multiple hidden layers of BERT is used to capture additional language information in the sentence. These textual features along with KG embedding are then used for multi-label DDI classification. Learning additional textual features over BERT embeddings, reduced the misclassification of certain labels and increased the overall accuracy. Also, a new loss function, KGE-MLF, was used to handle the imbalances in training data. We found this novel framework, integrating neural networks with pertained PubMed BERT model and KG embeddings to be challenging to reproduce and a great learning experience for building neural network frameworks.

The aforementioned approach benefits from the proposed KGE-MFL loss function, which alleviates the problem of imbalance and integrates external prior knowledge into the framework.

Also, we will attempt to demonstrate that there exist two kinds of patterns in the knowledge graph constructed from DDIExtraction-13, antisymmetry, and inversion. This model improves upon the research of (Mondal, 2020), because the new framework improves on the former by taking different levels of text features into account, incorporating the knowledge graph and neural network by KGE-MFL loss, and achieves an overall improvement in accuracy.

2 Scope of reproducibility

In our reproduction study, we will focus on verifying papers' proposed claims: the viability of "incorporating a neural network and a knowledge graph in order to extract DDIs from biomedical texts", the efficacy of this method for biomedical relation extraction, and whether neural networks and knowledge graphs are mutually complementary. Specifically we will address the following claims.

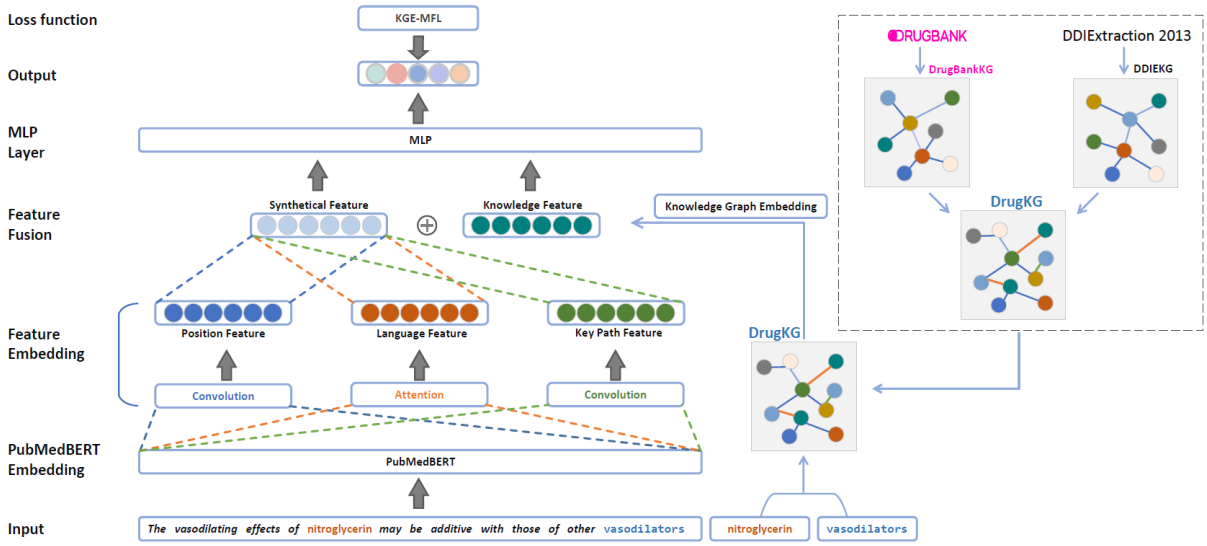


Figure 1: Neural network architecture

2.1 Addressed claims from the original paper

- Claim 1 : The classification model proposed by the paper, which incorporates a neural network and knowledge graph is effective in Drug-drug interaction classification from the text. It has better precision, accuracy, and f-score than previous neural models including ones that incorporated knowledge graphs like (Mondal, 2020)

- Claim 2 : Incorporating knowledge graph in the model has a complementary effect on neural network classification performance. Specifically, increased precision, accuracy, and f-score. We plan to verify this using an ablation experiment.

2.2 Ablations

Following ablations are planned to be included in the final report.

- 1 : Using PubMedBERT embeddings without additional textual features or knowledge graph for DDI classification. This will establish a baseline DDI classification performance without textual features using PubMedBERT.

- 2 : The effect of individual textual features like position, key path, and language and their combination on DDI classification performance will be studied.

- 3 : Effect of a KGE-MFL over Cross Entropy Loss function on the dataset will be studied. This will prove the author's claim the MFL gives better classification performance on an imbalanced dataset.

- 4 : In addition to the above, we also propose to do PubMedBert +Knowledge Graph (KG) to understand how this compares with the previous paper (Mondal, 2020). This will also help understand the contribution of KG to DDI classification performance.

3 Methodology

We will attempt to re-implement the framework from scratch and will reuse the data sets in the original study.

3.1 Model descriptions

The neural network architecture of the original paper is given in Figure 1. The core component of the model is the BERT (Bidirectional Encoder Representations from Transformers) encoder layer which is pre-trained on PubMed, a large medical corpus. 13 stacked encoder layers of the BERT learns a contextual embedding of the input text sequence. Positional, langage and key path feature representation of the sequence is created from the BERT output as follows:

The positional feature is incorporated into the model to capture the relative positions of the two drug entities within the text. The distance between the drug entities can provide useful information about their relationship and the likelihood of them interacting. In many cases, closer drug entities may

have a stronger relationship or more direct interaction, while those that are farther apart might be less related or have indirect interactions. By including the positional feature, the model can leverage the spatial information to better understand the context in which the drug entities appear. This additional feature helps the model to discriminate between different types of drug-drug interactions (DDIs) based on the relative positions of the drug entities.

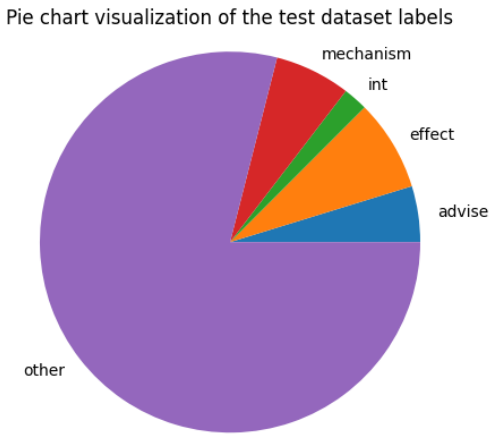
The language feature layers try to learn additional language properties of the sequence. BERT learns structure of language within different layers (e.g.: -semantics in higher layers). Instead of using just output of the final BERT layer, an attention based weighted average of the output of each of the 13 BERT encoder layers is used as the language feature. This will capture different language properties from BERT layers.

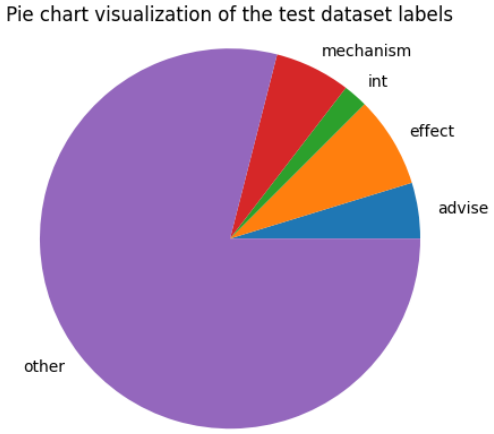
The key path feature is incorporated into the model to capture the essential syntactic and semantic information between the two drug entities in the text. By extracting and representing the key dependency paths between the drug entities, the model can better understand the context and relationships between them. Incorporating the key path feature allows the model to focus on the most relevant parts of the sentence, leading to more accurate predictions and improved performance.

The combination of the three features enables the model to learn a more comprehensive and context-rich representation of the input text. As a result, the model becomes better equipped to identify and classify DDIs with higher precision and recall, and thus a better final f-score.

Paper add drug knowledge graph embedding of the drugs in the input sequence in order to complement the BERT based embedding. Paper uses 'RotatE', a graph completion algorithm which can represent various types of drug graph relations (e.g.: symmetric, asymmetric etc.). Statistics of Knowledge graph is given in table 2.

Finally, the the synthetic BERT features and knowledge graph embeddings are concatenated and fed into a multi-label classifer to identify the DDI class type. Paper uses MFL (Multi-focal loss function) loss function since the input dataset is imbalanced and more weightage must be given to underrepresented DDI class types ('Mechanism', 'Effect', 'Advise', 'Int', 'Other').





3.2 Data descriptions

We are using the PubMedBERT data set, publicly released at Microsoft, DDI Corpus released as part of (Herrero-Zazo et al., 2013) at DDICorpus and DrugBank database at DrugBank.

Data set statistics is given in table 3 . A sample item from the dataset in given in table 1. Drug Knowledge base statistics is given in table 2.

3.3 Hyperparameters

We took the Hyperparameters proposed by the paper, and updated them to work with our model:

- Max sequence length n: 390
- Word embedding size d^w: 768
- Position embedding size d^p: 20
- Key path pos tag size d^k: 17
- Initial learning rate: 2e-5
- Linear warmup eps: 1e-8
- Number of fine-tuning epochs: 6
- Train batch size: 32
- Eval batch size: 16
- L2 weight decay: 0.01
- Dropout rate: 0.3

DDI Type	DDI Text	Drug1	Drug2
int	The authors report the case of an infant with confirmed congenital hypothyroidism on DRUG1 who experienced a possible drug interaction with DRUG2	levothyroxine	simeticone

Table 1: Sample Dataset Item (Preprocessed)

KG	Entity	Relation	Triplet(Total)	Triplet(Train)	Triplet (Test)
DBKG	3924	273	2681342	-	-
DDIEKG	3211	5	25222	16222	4500
DrugKG	7135	278	2706564	1706564	500000

Table 2: Knowledge graph statistics

DDI Type	Train	Test
Mechanism	1299	297
Effect	1649	358
Advice	802	219
Int	179	96
Negative	18781	3645
Total	22710	4615

Table 3: DDI Extraction 2013 dataset statistics

- Key-path LSTM layers: 1
- Key-path LSTM dropout: 0.5
- Knowledge embedding size d^{know}: 200

3.4 Implementation

Implementation for the reproduction study is here.

3.5 Computational requirements

We will attempt to use our own hardware to train the models. We have a combined 13696 Cuda cores available to us. Doing a preliminary run with 4651 tokens and running for 5 epochs took our local machine 1 hour and 15 minutes to train the model.

The results were as follows for epoch 5:

- Precision = 0.8636363636363636
- Recall = 0.822680412371134
- microF = 0.8426610348468849

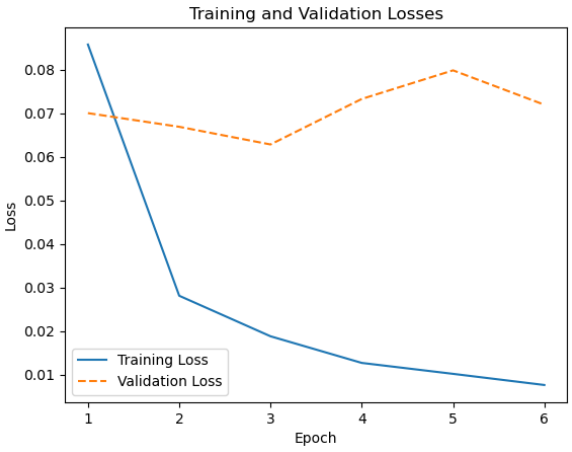
For local testing, we will used a 3090ti which has 10752 Cuda cores. We have determined that our local machines will suffice as we can use smaller data sets that will allow us to measure compute time on the scale of minutes.

For the final model, the creation of the data sets takes 1 hour 30 mins, but since we cache the results, this only has to be ran once. Training and evaluation takes about 45 mins for 6 epochs, with an average of 2.25 iterations per second on the 3090ti. A note has to be made that upwards of 20GB of v-cache is required to parse and store the entire dataset.

Loss Fn	Precision	Recall	MicroF
Cross Entropy	0.807	0.750	0.779
Multi Focal Loss	0.851	0.831	0.841

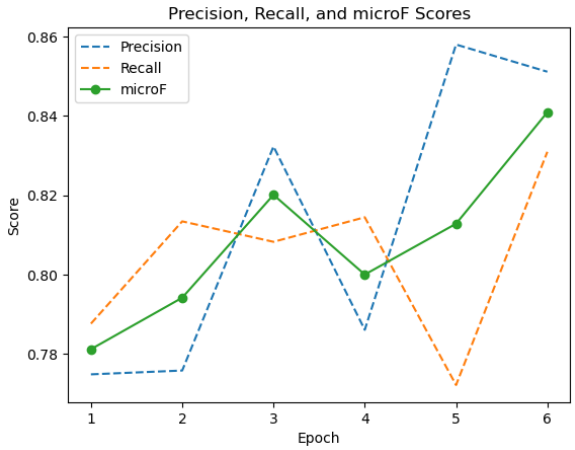
Table 4: Loss Function's impact on performance

4 Results



4.1 Overall performance

Our experiment results show the a micro F-score of 0.84 for DDI classification which is approximately 0.0224 less than what is claimed by the paper (86.24).Since a previous paper (Mondal, 2020) achieved 0.84, we are unable to conclude that novel framework proposed in the paper provide state of the art results on DDIExtraction 2013 dataset. Also



we observed that framework successfully complemented language feature with Drug Knowledge Graph . In Table 5, After adding drug knowledge embedding, recall increased by 0.03 and microF increased by 0.01.

4.2 Performance impact of KGE-MFL loss function

Table 4 shows the performance improvement of MFL over cross entropy. As claimed by the paper, MFL was able perform better than cross entropy loss function for the imbalanced DDI extraction dataset.

4.3 Ablations

We conducted additional ablations and the results are listed in table 5. It has a combination of every possible feature with and without Knowledge Graph Embedding.

5 Discussion

5.1 Implications of the experimental results

Our experiment results show a micro-F-score of 0.84 which is approximately 0.0224 less than what is claimed by the paper (86.24).Since a previous paper (Mondal, 2020) achieved 0.84, we are unable to conclude that the novel framework proposed in the paper provide state of the art results. Framework successfully complemented language feature with drug Knowledge Graph . After adding drug knowledge embedding, recall increased by 0.03 and f-score increased by 0.01 MFL (F-score= 0.84) was able perform better than cross entropy loss function (F-score= 0.78) for the imbalanced DDI extraction dataset.

5.2 What was easy

The author's code was very helpful reference for the reproduction study, the pre-processed dataset and drug knowledge base helped us to recreate the experiment faster.

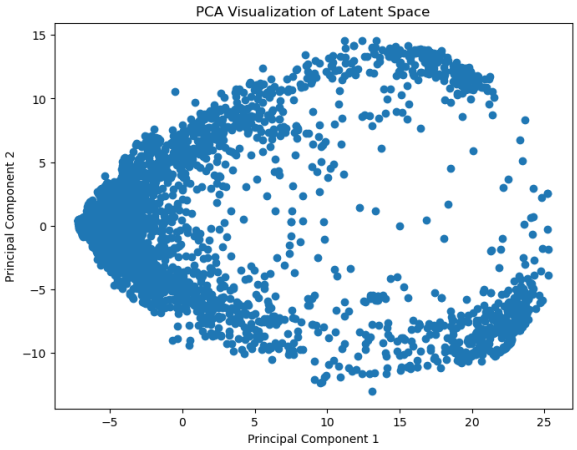
5.3 What was difficult

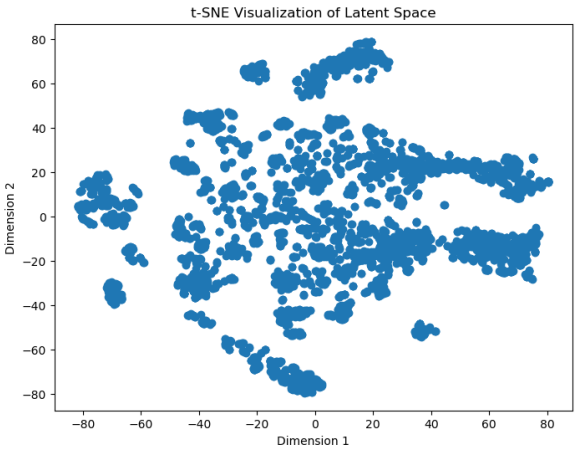
We ran paper's original code with the hyperparameter as set in paper , but was unable to achieve author's claim of F-score of 0.86 . Instead we were able to achieve a F-score of 0.84 only.

5.4 Recommendations for reproducibility

The reproduction of this code proved to us that BERT alone was more than capable of achieving good classification scores. I think training BERT on more relevant text would help accuracy more than adding features on top of BERT.

Additional research should be done to identify multiple drug combinations, not just two.





Experiment (without Knowledge Graph)	Precision	Recall	MicroF
Baseline PubMedBERT	0.867	0.718	0.785
PubMedBERT + Language Feature	0.811	0.765	0.787
PubMedBERT + KeyPath Feature	0.839	0.767	0.801
PubMedBERT + Language + KeyPath Feature	0.802	0.773	0.787
PubMedBERT + Position + KeyPath Feature	0.824	0.804	0.814
PubMedBERT + Position + Language Feature	0.807	0.808	0.807
PubMedBERT + Position + Language + KeyPath Feature	0.857	0.805	0.830
Experiment (with Knowledge Graph)	Precision	Recall	MicroF
Baseline PubMedBERT	0.805	0.765	0.785
PubMedBERT + Language Feature	0.816	0.794	0.805
PubMedBERT + KeyPath Feature	0.820	0.810	0.815
PubMedBERT + Language + KeyPath Feature	0.805	0.819	0.812
PubMedBERT + Position + KeyPath Feature	0.832	0.789	0.810
PubMedBERT + Position + Language Feature	0.810	0.729	0.767
PubMedBERT + Position + Language + KeyPath Feature	0.851	0.831	0.841

Table 5: Ablation experiment results

References

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.

Xin Jin, Xia Sun, Jiacheng Chen, and Richard Sutcliffe. 2022. Extracting drug-drug interactions from biomedical texts using knowledge graph embeddings and multi-focal loss. In *Proceedings of the 31st ACM International Conference on Information amp; Knowledge Management*, CIKM '22, page 884–893, New York, NY, USA. Association for Computing Machinery.

Ishani Mondal. 2020. Towards incorporating entity-specific knowledge graph information in predicting drug-drug interactions. *CoRR*, abs/2012.11142.