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

Xin Jin
School of Information Science and Technology
Northwest University
Xi'an, Shaanxi, China
jin_xin@stumail.nwu.edu.cn

Xia Sun*
School of Information Science and Technology
Northwest University
Xi'an, Shaanxi, China
raindy@nwu.edu.cn

Jiacheng Chen
School of Information Science and Technology
Northwest University
Xi'an, Shaanxi, China
2015118051@stumail.nwu.edu.cn

Richard Sutcliffe*
School of Information Science and Technology
Northwest University
Xi'an, Shaanxi, China
School of Computer Science and Electronic Engineering
University of Essex
Wivenhoe Park, Colchester, UK
rsutcl@nwu.edu.cn,rsutcl@essex.ac.uk

ABSTRACT

The field of Drug-drug interaction (DDI) aims to detect descriptions of interactions between drugs from biomedical texts. Currently, researchers have extracted DDIs using pre-trained language models such as BERT, which often misclassify two kinds of DDI types, "Effect" and "Int", on the DDIExtraction 2013 corpus because of highly similar expressions. The use of knowledge graphs can alleviate this problem by incorporating different relationships for each, thus allowing them to be distinguished. Thus, we propose a novel framework to integrate the neural network with a knowledge graph, where the features from these components are complementary. Specifically, we take text features at different levels into account in the neural network part. This is done by firstly obtaining a word-level position feature using PubMedBERT together with a convolution neural network, secondly, getting a phrase-level key path feature using a dependency parsing tree, thirdly, using PubMedBERT with an attention mechanism to obtain a sentence-level language feature, and finally, fusing these three kinds of representation into a synthesized feature. We also extract a knowledge feature from a drug knowledge graph which takes just a few minutes to construct, then concatenate the synthesized feature with the knowledge feature, feed the result into a multi-layer perceptron and obtain the result by a softmax classifier. In order to achieve a good integration of the synthesized feature and the knowledge feature, we train the model using a novel multifocal loss function, KGE-MFL, which is based on a knowledge graph embedding. Finally we attain state-of-the-art results on the DDIExtraction 2013 dataset (micro F-score 86.24%) and on the ChemProt dataset (micro F-score 77.75%), which proves our framework to be effective for biomedical relation extraction tasks. In particular, we fill the performance gap (more than 5.57%) between methods that rely on and do not rely on knowledge graph embedding on the DDIExtraction 2013 corpus, when predicting the "Int" type. The implementation code is available at https://github.com/NWU-IPMI/DDIE-KGE-MFL.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Information extraction**; • **Applied computing** → **Bioinformatics**.

KEYWORDS

drug-drug interactions, knowledge graph, PubMedBERT, imbalance problem

ACM Reference Format:
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 and Knowledge Management (CIKM'22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557318

1 INTRODUCTION

There is always the possibility that unwanted side effects (called adverse drug reactions, or ADRs) can occur when a patient takes two or more drugs at the same time. These can endanger a patient and even result in their death [1]. It is therefore highly desirable to know whether there are drug-drug interactions (DDIs) between the drugs which a patient takes. Meanwhile, with the explosive growth of biomedical literature, there remain many latent drug interactions described in papers which have not yet been detected and added to DDI databases. For this reason, the DDI field has been a very active one in recent years.

*Corresponding author

CIKM'22, October 17–21, 2022, Atlanta, GA, USA.

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557318

There exist many neural network-based methods to extract DDIs in previous research, and these have undoubtedly achieved good performance. Specially, as the powerful Bidirectional Encoder Representations from Transformers (BERT) model [2] achieves the best result on many natural language processing tasks, many scholars are conducting the DDI extraction task by pre-trained biomedical language models such as BioBERT [3], SciBERT [4], BlueBERT [5] and PubMedBERT [6], all of which can improve the result compared with the Word2Vec or GloVe based methods.

However, as Zhu et al. [7] stated, because the "Int" type and "Effect" type are expressed in highly similar ways, where "Int" type just have an additional information to "Effect" type, this often leads to misclassification on the DDIExtraction 2013 dataset (henceforth DDIExtraction-13). Because a knowledge graph can express many potential relationships between two entities, its use can alleviate the misclassification problem.

Based on the above thinking, and following previous work, we aim to integrate the text information from a pre-trained language model-based neural network with the information from a knowledge graph, and propose a novel framework as shown in Figure 1. In the neural network part, most previous works partly use text information such as a position feature or a dependency tree feature. In order to take text features at different levels into account, we input the preprocessed texts with relative position information into PubMedBERT and a convolutional neural network (CNN) to get the word-level position feature, get the phrase-level key path feature alongside the position feature, input the preprocessed texts into PubMedBERT with an attention mechanism to obtain a sentence-level language feature, and finally integrate these three kinds of features into one synthetic feature. We construct a drug knowledge graph, called DrugKG, using the DrugBank database [8] and the DDIExtraction-13 [9]. We then obtain a knowledge feature by applying the knowledge graph embedding method to DrugKG, concatenate the synthetic feature and the knowledge feature, feed into a multi-layer perception (MLP) and softmax layer in order to obtain the classification result.

Meanwhile, we design a novel loss function called knowledge graph embedding multi focal loss (KGE-MFL) which not only incorporates external knowledge information, but also alleviates the imbalance problem. We evaluate our model on DDIExtraction-13 with a micro F-score of 86.24% and on the ChemProt dataset[10] with a micro F-score of 77.75%, which shows that our framework is effective for biomedical relation extraction tasks. To sum up, our contributions are as follows:

- We propose a novel framework to integrate a neural network with a knowledge graph, in which features from these are complementary.
- The approach benefits from the proposed KGE-MFL loss function, which not only alleviates the problem of imbalance, but also integrates external prior knowledge.
- We demonstrate that there mainly exist two kinds of patterns in the knowledge graph constructed from DDIExtraction-13, *antisymmetry* and *inversion*. This finding can assist future researchers in choosing a knowledge graph embedding method.

2 RELATED WORK

There primarily exist two kinds of method for DDI extraction. The first one is traditional feature-based machine learning methods. At the beginning, Chowdhury et al. [11] used a kernel approach and Thomas et al. [12] used majority voting to extract DDIs. After this, many researchers improved the results by use more feature types such as relative position features, syntactic structure features, and phrase auxiliary features [13–15]. However, these methods usually rely heavily on feature engineering and feature selection, both carried out manually, which limits their effectiveness and leads to some possible patterns being overlooked.

The second group of methods are based on deep neural networks, mainly convolutional neural networks (CNNs) and recurrent neural networks (RNNs). In addition there are methods using pre-trained language models (often based on BERT), and knowledge graph methods fused with BERT. Liu et al. [16] were the first to apply CNNs to extract DDIs; they embedded word and position features by CNN to obtain the classification result. Subsequently, Liu et al. [17] fused syntactic information by dependency parsing tree (DPT), Sun et al. [18] increased the depth of the CNN and the convolution kernel size, and Asada et al. [19] combined an attention mechanism with CNN, all of which improved the classification result.

Meanwhile, there exist many methods based on Long Short-Term Memory (LSTM) for DDI extraction. Kavuluru et al. [20] used a word-based BiLSTM and a character-based BiLSTM to extract DDIs. Wang et al. [21] and Zheng et al. [22] separately constructed a BiLSTM model based on Dependency Parse Trees (DPTs) and an attention-based BiLSTM model to obtain different kinds of features. Sun et al. [23] and Xu et al. [24] fused external information to increase the prior knowledge.

Then due to the power of BERT, Zhu et al. [7] proposed a BioBERT-based model which fused external entity information and achieved a higher new baseline for DDI extraction. Asada et al. [25] proposed a SciBERT-based model which fused descriptions and molecular structures of drug entities. Huang et al. [26] combined BioBERT and Bio-GPT2 to construct the classification part and generation part of their model, which improved the DDI extraction performance. Mondal [27] proposed a model that integrates BERT and knowledge graphs. The model concatenates the output of BERT, drug mentions represented by BERT and drug embeddings from a knowledge graph constructed by them, and the result shows the effectiveness of knowledge graphs for this task.

There is no denying that the previous works have achieved good results, especially Mondal's model [27]. In comparison to Mondal, our method also integrates a neural network with a knowledge graph; however, because their method ignores text features in the neural network part and neglects the latent information of drug knowledge graph embeddings in a drug pair, the improvement of their approach is limited. By contrast, our method takes different levels of text features into account, incorporates the knowledge graph and neural network by KGE-MFL loss, and achieves a higher improvement.

3 METHODS

As Figure 1 shows, we firstly obtain the word-level position feature and phrase-level key path feature by PubMedBERT and CNN, get

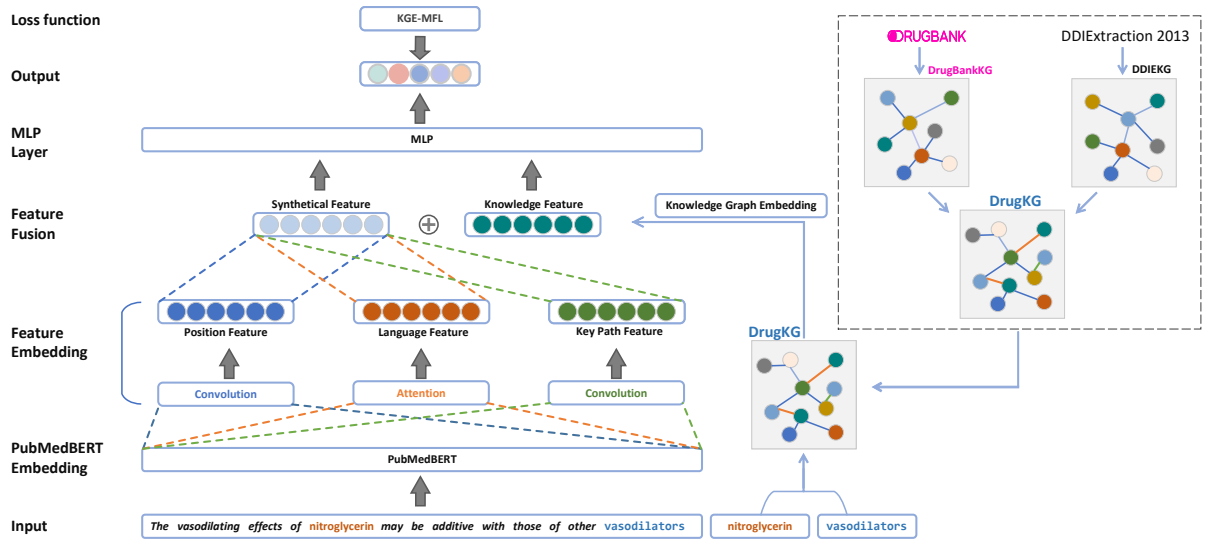

Figure 1: The architecture of our model. This figure uses the sentence "The vasodilating effects of nitroglycerin may be additive with those of other vasodilators", where "nitroglycerin" and "vasodilators" are drug entites.

Table 1: A preprocessing example for the sentence "Based on the chemical resemblance of itraconazole and ketoconazole, coadministration of astemizole with itraconazole is contraindicated."

Label	Drug Pair	Preprocessed Sentence
negative	(itraconazole,ketoconazole)	Based on the chemical resemblance of DRUG1 and DRUG2, coadministration of DRUG0 with DRUG0 is contraindicated.
negative	(itraconazole,astemizole)	Based on the chemical resemblance of DRUG1 and DRUG0, coadministration of DRUG2 with DRUG0 is contraindicated.
negative	(itraconazole,itraconazole)	Based on the chemical resemblance of DRUG1 and DRUG0, coadministration of DRUG0 with DRUG2 is contraindicated.
negative	(ketoconazole,astemizole)	Based on the chemical resemblance of DRUG0 and DRUG1, coadministration of DRUG2 with DRUG0 is contraindicated.
negative	(ketoconazole,itraconazole)	Based on the chemical resemblance of DRUG0 and DRUG1, coadministration of DRUG0 with DRUG2 is contraindicated.
advice	(astemizole,itraconazole)	Based on the chemical resemblance of DRUG0 and DRUG0, coadministration of DRUG1 with DRUG2 is contraindicated.

the sentence-level language feature by PubMedBERT with an attention mechanism, and then incorporate these features into one synthetic feature. Meanwhile, we embed the knowledge feature by a knowledge graph embedding method from DrugKG. Finally, we concatenate the synthetic feature and the knowledge feature, feed them into a multi-layer perceptron and get the output through a softmax layer. We train the model using our proposed KGE-MFL loss. We now describe more details of the model.

3.1 Preprocessing

DDI extraction is a multi classification task to discern the drug pair interaction when given an input sentence. When a sentence has n entities, there will be C_n^2 drug pairs to be classified. Meanwhile, due to the negative samples being far more frequent than the positive, as is shown in Table 1, we adapt previous work [23] to filter some obvious negative samples in order to relieve the imbalance problem partly. Our negative filter strategy as follows:

- When two drug instances have the same name (like the example shown in the third row of Table 1) or one drug name is the abbreviation of the other, filter out the corresponding instances.
- When a drug pair is in a coordinate structure, then filter out the corresponding instances.
- When one name is a special case of the other in a drug pair, filter out the corresponding instances.

Meanwhile we apply a drug blinding method [16], to focus on the current interaction of drug pair entities and to improve the robustness of model. This masks a drug pair as "DRUG1" and "DRUG2" and the other drug entities as "DRUG0".

3.2 Feature representation

In order to consider the different levels of features, we will describe how to obtain the word-level position feature, the phrase-level key path feature and the sentence-level language feature. Next, we explain in more detail how to represent these features.

3.2.1 *Position feature representation.*

Given an input sentence $S = [w_1, w_2, ..., w_n]$ which includes drug pair entities w_{d1} and w_{d2} and has text length n, we first split the sentence using the WordPiece algorithm [28]. We then feed each wordpiece w_i into PubMedBERT to get the contextualized embedding $e_i^w \in \mathbb{R}^{d^w}$, where d^w is the dimension of the word embedding. Meanwhile, we calculate the d^p-dimensional relative position embeddings $e_i^{p1} \in \mathbb{R}^{d^p}$ and $e_i^{p2} \in \mathbb{R}^{d^p}$ for every wordpiece respectively. Finally, the wordpiece embedding and position embedding

Table 2: A misclassification example based on the preprocessed sentence "Barbiturates and glutethimide should not be administered to patients receiving coumarin drugs."

Preprocessed Sentence	True Label	Predict Label
DRUG1 and DRUG2 should not be administered to patients receiving DRUG0.	advice	advice
DRUG1 and DRUG0 should not be administered to patients receiving DRUG2.	advice	advice
DRUG0 and DRUG1 should not be administered to patients receiving DRUG2.	negative	advice

are concatenated as in Equation (1):

$$e_i^p = [e_i^w; e_i^{p1}; e_i^{p2}] \tag{1}$$

where ";" denotes the operation of concatenation. Then we capture the local and consecutive contextual features by the convolution operation as Equation (2):

$$Conv_i = ReLU(W_p \odot e^p_{i:i+k-1} + B_p) \tag{2}$$

where k is a window size, $W_p \in \mathbb{R}^{d^c \times (d^w + 2 \times d^p) \times k}$ denotes the application of the filter to all possible serial context windows of k words, and d^c is the filter size of the convolution, $\odot$ is an element-wise product, B_p is a bias term, and $ReLU(\cdot)$ represents the activation function, namely rectified linear units (ReLU). Then, we determine the output of the ReLU function as shown in Equation (3):

$$F_i^{pos} = max(Conv_i) \tag{3}$$

Finally, we obtain the position feature $F^{pos} \in \mathbb{R}^{d^w}$ by a max pooling layer. Because the position feature represents the relative position of all wordpieces in a sample, we consider this type property to be a word-level feature.

3.2.2 *Key path feature representation.*

Due word similarities, certain samples are always misclassified as illustrated in Table 2.

We observe that redundant words in a misclassification sample can lead to an incorrect result. Therefore, to address this problem, we construct the dependency parsing tree and cut the tree to obtain the path which contains drug entities comprising a drug pair. This focuses on the key phrase in the classfication sample and is named the *key path* by us. We construct the phrase-level key path as follows:

(1) We firstly build the dependency parse tree using the Stanza Python library [29] which is developed for biomedical NLP.
(2) Then we find the node which is the lowest common ancestor (LCA) of the drug pair.
(3) Finally, we figure out one route which includes the LCA node, the drug entity pair, and any other node whose distance to a dependent node is less than one hop. We call this the key path.

Take the sentence "**Barbiturates** and **glutethimide** should not be administered to patients receiving coumarin drugs" as an example; the current drug pair is (Barbiturates,glutethimide). We can calculate the key path as "Barbiturates and glutethimide", and delete the unrelated phrase "should not be administered...", as shown in Figure 2.

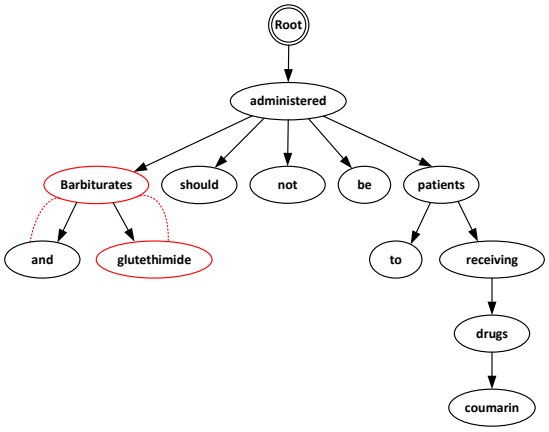

Figure 2: An example key path for the sentence "Barbiturates and glutethimide should not be administered to patients receiving coumarin drugs."

We take $e_i^k \in \mathbb{R}^{d^k}$ as the key path embedding, where d^k is the dimension of key path feature. Then we concatenate the wordpiece embedding and key path embedding as shown in Equation (4):

$$e_i^{key\_path} = [e_i^w; e_i^k] \tag{4}$$

Then, we obtain the key path feature in the same way as the position feature by a CNN model as was shown in Equations (2) and (3). Finally, we represent the phrase-level key path feature as $F^{key\_path} \in \mathbb{R}^{d^w}$.

3.2.3 *Language feature representation.*

Although BERT learns about the structure of language within several different layers [30], previous researchers usually just embed the final output in their model, an approach which ignores some features. Therefore, we obtain a sentence-level language feature to catch different language properties by the hidden states of PubMedBERT with an attention mechanism. We obtain the language feature as in Equation (5):

$$F^{lang} = \sum_{i=1}^{m} Softmax(W_i * PubMedBERT(S) + b_i) \tag{5}$$

where $PubMedBERT(S)$ is the hidden state output of PubMedBERT when fed the sentence S, W_i is the i-th layer of the hidden states of PubMedBERT, and b_i is a bias term, $Softmax(\cdot)$ is used for normalization, m is the number of hidden layers which we need in order to calculate the attention weight. Finally, we can obtain the language feature as $F^{lang} \in \mathbb{R}^{d^w}$. Then we concatenate different kinds of features as in Equation (6):

$$F^{synth\_feature} = \frac{1}{M} \sum_{i=1}^{M} \alpha_i F_i \tag{6}$$

where M is the feature number, and α_i is the corresponding feature weight. Here, we consider that these three kinds of features are equally important for DDI extraction. Thus we set $M = 3$ and $\alpha_i = 1$,

then average these features to combine their properties. We take the embedding of $F^{synth\_feature}$ as the final feature from the neural network.

3.3 DrugKG construction

Generally, a knowledge graph is described by a triplet $G = (E, R, S)$. $E = \{e_1, e_2, ..., e_m\}$ is the collection of entities, where m is the number of entities; $R = \{r_1, r_2, ..., r_n\}$ is the collection of relations, where n is the number of relations; Finally, $S \subseteq E \times R \times E$ is the collection of triplets in the knowledge base.

We usually describe a specific triplet as Equation (7):

$$(head, relation, tail) \in S, head, tail \in E, relation \in R \quad (7)$$

which means the head entity and tail entity have this relation. This is a simple but expressive form to store the information, and is often abbreviated to (h, r, t). Then we build the drug knowledge graph using the datasets DDIExtraction-13 and DrugBank by the following steps:

(1) Firstly, we extract the drug pair as a head entity and tail entity respectively, and determine the label as a relation from DDIExtraction-13, which can be explicitly expressed as $(DRUG1, label, DRUG2) \rightarrow (h, r, t)$.
(2) Then, we list the result collection of (h, r, t) as a triplet, and construct a drug knowledge graph from DDIExtraction-13 named *DDIEKG*.
(3) In the same way, we extract the drug entity and relation expression from DrugBank, expressed as $(DrugA, relation, DrugB) \rightarrow (h, r, t)$, and then build the drug knowledge graph named *DBKG*.
(4) Finally, we merge *DDIEKG* and *DBKG*, filter the repetitive drug entities and relations, and hence get the final knowledge graph called *DrugKG*.

3.4 Knowledge feature representation

Because knowledge graphs store information in a direct and clear way, there exist many methods to model the embedding of entities and relations using them, such as TransE [31], DistMult [32], ComplEx [33] and RotatE [34]. The most important step for evaluating such knowledge graph embeddings (KGEs) is a score function, which usually takes the form $f_r(h, t)$. Meanwhile, there are different kinds of patterns in a knowledge graph; some relations are symmetric (e.g., marriage) while others are antisymmetric (e.g., filiation), some relations are the inverse of others (e.g., hypernym and hyponym), while some types of relation may be composed of other relations (e.g., father's father is grandfather). Therefore, a knowledge graph can contains four forms: symmetry, antisymmetry, inversion, and composition. Diverse methods of KGE could model the different patterns as is shown in Table 3. We adapt RotatE as our KGE method because it can represent all patterns, and its score function is shown in Equation (8):

$$d_r(h, t) = \|h \circ r - t\| \quad (8)$$

where $\circ$ is the element-wise product, and $h, r, t \in \mathbb{C}^k$, where C^k is a k-dimensional complex. We obtain the entity embedding by the loss function proposed by RotatE as Equation (9):

$$L = -log\sigma(\gamma - d_r(h, t)) - \sum_{i=1}^{n} p(h'_i, r, t'_i) log\sigma(d_r(h'_i, t'_i) - \gamma) \quad (9)$$

where σ is the sigmoid function, γ is a fixed margin, (h'_i, r, t'_i) is the i-th negative triplet, $d_r(\cdot)$ is the score function. We minimize the loss function to get the embedding of entity and relation, then we obtain the entity embedding $e^{know} \in \mathbb{C}^k$ as a knowledge feature. Finally, we match the entities in drug pairs to get the knowledge feature, and concatenate it with the synthetic feature as in Equation (10):

$$F = [F^{synth\_feature}; e^{know}_{d1}; e^{know}_{d2}] \quad (10)$$

where e^{know}_{d1} and e^{know}_{d2} are the entity embedding of drug1 and drug2 in a drug pair. Then, input the vector to an MLP layer to get the prediction probability of all DDI types as Equation (11):

$$p = [p_1, p_2, ...p_n], p_j = Softmax(MLP(M^{pred}F)) \quad (11)$$

where p is the prediction probability of each DDI type in n types, $M^{pred} \in \mathbb{R}^{n \times d^w}$ is a weight matrix, and the prediction result is obtained by a $Softmax(\cdot)$ layer.

Table 3: Patterns which can be recognised by different methods.

Method	TransE	DistMult	ComplEx	RotatE
ScoreFunction	$-\|\|h + r - t\|\|$	$< h, r, t >$	$Re(< h, r, \bar{t} >)$	$-\|\|h \circ r - t\|\|$
Symmetry	✗	✓	✓	✓
Antisymmetry	✓	✗	✓	✓
Inversion	✓	✗	✓	✓
Composition	✓	✗	✗	✓

3.5 Training

In order to alleviate the imbalance problem and incorporate external prior knowledge, we propose a novel loss function called KGE-MFL – knowledge graph embedding multi-focal loss. Multi-focal loss is based on the focal loss function [35] which considers different DDI types as dissimilar weights and is used as follows:

$$MFL = -\frac{\sum_{i=1}^{n} \alpha_i (1 - p_t)^\gamma log(p_t)}{n} \quad (12)$$

where n is the number of DDI types, α_i is the weight of each type as defined in Equation (13), and p_t is the previous result of Equation (11), where $\gamma = 2$.

$$\alpha_i = \frac{Count_i}{\sum_{i=1}^{n} Count_i} \quad (13)$$

where $Count_i$ represents the number of the $i - th$ type, n is the number of different types. Then we define KGE-MFL loss as follows to train our model:

$$Loss = \beta MFL(p) + (1 - \beta) F(KGE_{drug2} - KGE_{drug1}) \quad (14)$$

where β is a weight parameter to balance the output of MFL and the classfication of KGE, $MFL(p)$ is the output of the MFL loss function as defined in Equation (12), $F(\cdot)$ is a classification network which contains a category for each DDI type, KGE_{drug2} and KGE_{drug1} are the knowledge graph embeddings obtained previously. Due to the result of knowledge embedding being complex and high-dimensional, we consider that the difference between KGE_{drug2}

Table 4: Statistics of the DDIExtraction 2013 dataset

	Train		Test	
DDI Type	Original	Filtered	Original	Filtered
Mechanism	1319	1299	302	297
Effect	1687	1649	360	358
Advice	826	802	221	219
Int	188	179	96	96
Negative	23665	18781	4712	3645
Total	27685	22710	5691	4615

and KGE_{drug1} contains the information needed to find the correct label. Hence we design the classification network to use the relation information.

4 RESULTS

4.1 Dataset

As stated earlier, we evaluate our model on the DDIExtraction-13 dataset, which is the baseline corpus for DDI extraction, and contains data mainly originating in DrugBank and MEDLINE. It contains four positive types and one negative type as follows:

- Mechanism: This type is assigned when a pharmacokinetic mechanism happened in a drug pair, e.g., *In a study in normal volunteers, concomitant administration of buspirone HCl and haloperidol resulted in increased serum haloperidol concentrations.*
- Effect: This type is assigned when an effect or a pharmacodynamic appeared in a drug pair, e.g., *Digitalis: Thyroid preparations may potentiate the toxic effects of digitalis.*
- Advice: This type is assigned when the text gives a suggestion or advice relating to a drug pair interaction, e.g., *Avoid the concomitant use of chlorprothixene and tramadol.*
- Int: This type is assigned when additional information about a drug interaction is being given, e.g., *Bromocriptine mesylate may interact with dopamine antagonists, butyrophenones, and certain other agents.*
- Negative: This type is assigned when there is no drug interaction, e.g., *It is unknown whether the concomitant administration of proton pump inhibitors affects duloxetine absorption.*

The statistics of the original and filtered datasets are shown in Table 4. As mentioned earlier, the negative quantity is almost six times the positive, meaning that the dataset is extremely imbalanced.

4.2 Drug Knowledge Graph

We divide the KG dataset into three parts: Train dataset, development dataset, and test dataset. Information about the three KGs is shown in Table 5. We use the train dataset to train our model and validate the hyperparameters on development dataset. Finally, we evaluate on the test dataset.

Table 5: Statistics of different knowledge graphs. DBKG is an abbreviation for DrugBankKG

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

Table 6: Hyperparameters of our model

Parameter	Value
Max sequence length n	390
Word embedding size d^w	768
Position embedding size d^p	20
Key path embedding size d^k	20
Initial learning rate	2e-5
Number of fine-tuning epochs	5
Batch size	16
L2 weight decay	0.01
Dropout rate	0.1
Attention number m	13
Convolution window size k	3
Convolution filter size d^c	768
Knowledge embedding size d^{know}	200
Fixed margin γ	24
Loss function weight β	0.6

4.3 Evaluation Metric

The original DDIExtraction 2013 shared task used P_{micro}, R_{micro} and F_{micro} to evaluate the performance of models used by participating research teams. These metrics are calculated as follows:

$$P_{micro} = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}, R_{micro} = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \tag{15}$$

$$F_{micro} = \frac{2 \times P_{micro} \times R_{micro}}{P_{micro} + R_{micro}} \tag{16}$$

where $\overline{TP}$, $\overline{FP}$, and $\overline{FN}$ represent the average of true positives, false positives and false negatives respectively.

4.4 Hyper-parameter settings

We apply the Pytorch[1] library and Python3.7 to code our model, and run it on the Ubuntu[2] operating system. We optimize the hyperparameters by NNI[3] using the development set which consists of 4542 samples selected from the train dataset as described earlier. Finally, we obtain the hyperparameter settings as listed in Table 6.

4.5 Experiment results

4.5.1 *Comparison with prior methods.*

We evaluate on DDIExtraction-13 and compare with other established methods as shown in Table 7. We can see that our model

[1]https://pytorch.org/
[2]https://ubuntu.com/
[3]https://nni.readthedocs.io/

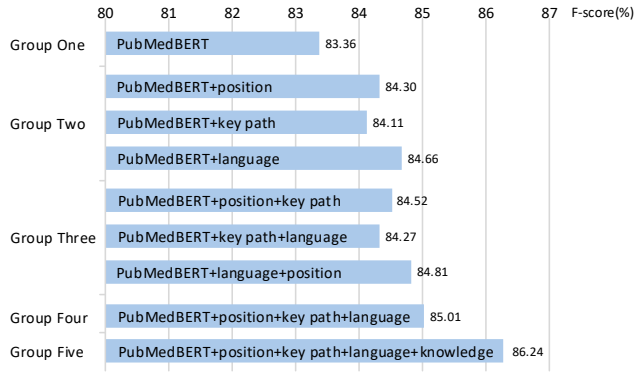

Figure 3: Ablation study to show the effect of different feature combinations on the performance of the proposed model. We train all models using MFL. The Group One model is the baseline and uses no features. Group Two models use one feature, Group Three models use two features, the Group Four model uses three features and the Group Five model uses all four features.

achieves the highest performance, in particular far higher than existing methods in respect of "Mechanism", "Effect", and "Int" types. The values of Precision, Recall and F-score all are higher than the previous methods. In particular, as discussed earlier, Mondal's method [27] does not use the text feature and the latent information between drug knowledge graph embeddings. This could account for the fact that our method exceeds theirs on all evaluation metrics.

4.5.2 Ablation experiment.
To investigate the effect of different features, we add features to our model step-by-step. The model configurations are divided into five groups. Group One consists only of PubMedBERT. Group Two consists of PubMedBERT plus one feature (with three choices). Group Three allows two features (with three choices). Group Four allows three features (one choice) and Group Five allows four features (one choice) and is the complete proposed model. The results are shown in Figure 3.

Comparing Group One with Group Two, we can see that the performance of Group Two improved, whichever feature we integrate. Thus position, key path, and language features all have a positive boosting effect for DDI extraction.

Relative to Group Two, models with two features in Group Three either show a small increase or a decrease in performance. This may imply that the additional information is similar to that already contained in the Group Two models.

However, from Group Three to Group Four, we can see an improvement in all cases, indicating that the three features in Group Four are complementary. Finally, between Groups Four and Five, the performance improved by 1.23% when fusing the external knowledge, which proves that this knowledge, obtained by the proposed KGE-MFL loss, makes up the shortage of features obtained by the neural network.

5 DISCUSSION

5.1 The effectiveness of different feature combinations in BERT-based models

We evaluate our method on four different pre-trained language models: BERT [36], SciBERT [4], BioBERT [3], and BlueBERT [5]. Each model is evaluated by F-score when used with various combinations of features as shown in Table 8.

The results show that the performance all the models improved when we incorporated the various features into the base pre-trained language models. Moreover, for all models, the highest score was seen when all the features were added. Specifically, we can see that when just one feature is added, the language feature gives the largest improvement, proving its usefulness for DDI extraction. On the other hand, the enhancement of BioBERT and BlueBERT is the smallest because these BERTs are trained for use in the biomedical domain and therefore already contain the appropriate domain knowledge.

5.2 The results of various knowledge graph and knowledge graph embedding methods

Due to DBKG not containing all the drug entities in DDIExtraction-13, we evaluate the model on DDIEKG and DrugKG, and the results are shown in Table 9. The table shows that DrugKG, the larger KG, can achieve a better performance because it provides more information and embeds more accurate representations. Meanwhile, we can see that the ComplEx and TransE methods obtain a similar performance on *DDIEKG*, while ComplEx and RotatE produce almost equal results on *DrugKG*. Different knowledge embedding methods can model different patterns. DistMult can only express symmetry patterns, and it obtains poor results on both KGs. This suggests that no matter which KG we construct, we mainly have two kinds of pattern, antisymmetry and inversion, so a method which cannot handle both of these may not perform well. This is a useful point for future researchers to keep in mind when selecting KGE methods.

5.3 The influence of different loss functions

To determine which DDI types the proposed KGE-MFL loss function improved, we change the loss function of our model and the results are shown in Figure 4.

We can observe that the KGE-MFL improves F-score when measured on all DDI types together (bottom right hand box in the figure) when compared to CE, with an improvement of 2.63%. There are two aspects which together improve the result. The first is that the KGE-based losses include external knowledge which achieves a better performance. We can see this from comparing CE with KGE-CE, and comparing MFL with KGE-MFL, for all the DDI types shown individually in Figure 4 (the first four boxes). The second point is that MFL can alleviate the imbalance problem which we can see from comparing CE with MFL. The results show improvements of 2.85%, 0.64%, 3.19% and 1.4% respectively on "Effect", "Mechanism", and "Int" types, and on "All DDI types".

Table 7: Comparison of proposed method with previous approaches, all evaluated on DDIExtraction-13. The highest value in each column is shown in bold.

	Methods	F-score on each DDI type				Overall		
		Mechanism	Effect	Advice	Int	Precision	Recall	F-score
Feature based methods	Chowdhury et al. [11]	67.90	62.80	69.20	54.70	64.60	65.60	65.10
	Kim et al. [14]	69.30	66.20	72.50	48.30	-	-	67.00
	Raihani et al. [15]	73.60	69.60	77.40	52.40	73.70	68.70	71.10
CNN and RNN based methods	Liu et al. [16]	70.23	69.32	77.72	46.37	75.70	64.66	69.75
	Asada et al. [19]	73.83	71.03	81.62	45.83	73.31	71.81	72.55
	Sun et al. [23]	78.25	73.49	80.54	58.90	77.30	73.75	75.48
	Zheng et al. [22]	77.50	76.60	85.10	57.70	78.40	76.20	77.30
Bert based methods	Zhu et al. [7]	84.60	80.10	86.00	56.60	81.00	80.90	80.90
	Asada et al. [25]	87.61	82.05	**90.79**	58.74	85.36	82.83	84.08
	Huang et al. [26]	87.20	84.80	87.70	57.90	83.40	85.00	84.20
Knowledge graph based method	Mondal [27]	87.00	81.00	88.00	59.00	-	-	84.00
	Ours	**89.00**	**86.47**	89.64	**64.47**	**86.20**	**86.29**	**86.24**

Table 8: The effects of various feature combinations when using different BERT-based pre-trained language models. "pos", "kpath", "lang" and "know" are the abbreviations of "position feature", "key path feature", "language feature" and "knowledge feature". Each number in brackets is the improvement of F-score.

Method	BERT(Δ)	SciBERT(Δ)	BioBERT(Δ)	BlueBERT(Δ)
-	79.71(-)	81.11(-)	81.89(-)	82.64(-)
+pos	80.98(1.27)	83.31(2.2)	82.23(0.34)	82.94(0.3)
+kpath	81.29(1.58)	83.17(2.06)	82.89(1)	82.96(0.32)
+lang	81.77(2.06)	83.28(2.17)	82.91(1.02)	83.01(0.37)
+pos+kpath+lang	82.14(2.43)	83.71(2.6)	83.40(1.51)	83.53(0.89)
+pos+kpath+lang+know	**82.73**(3.02)	**84.85**(3.74)	**84.03**(2.14)	**84.31**(1.67)

Table 9: The F-scores resulting from using different KGE methods with the DDIEKG and DrugKG knowledge graphs.

	TransE	DistMult	ComplEx	RotatE
DDIEKG	85.01	83.54	85.22	84.30
DrugKG	85.55	83.87	86.37	86.24

5.4 The generalization of our method

To examine the generalization of our model, we evaluate it on the ChemProt dataset [10]. Like the DDIExtraction 2013 task, this is also a multi-classification task on biomedical texts, but in this case the aim is to extract the relations between chemicals and proteins with F-score as the evaluation metric. The result of running the proposed model on ChemProt is shown in Table 10.

Even though we apply our method on ChemProt without changing hyperparameter settings, it still has a good performance and the results surpass most previous models.

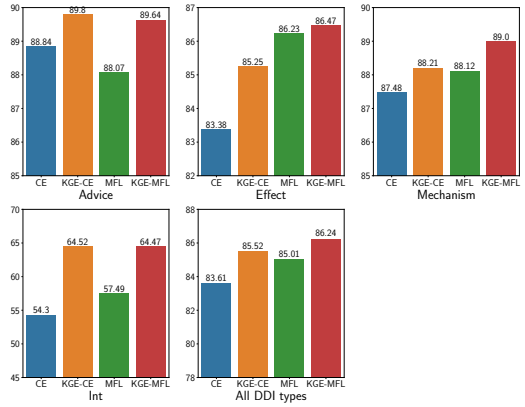

Figure 4: The effect of different loss functions on our model, where CE is cross-entropy loss, MFL is multi-focal loss, and KGE-CE is the loss function based on KGE and CE.

Table 10: Comparison of the proposed model with previous methods on the test dataset of ChemProt.

Methods	Peng et al. [5]	Lee et al. [3]	Gu et al. [6]	Su et al. [37]	Ours
Precision	71.74	77.02	78.80	80.60	78.81
Recall	74.13	75.90	75.90	76.90	76.68
F-score	72.93	76.46	77.24	78.70	77.75

5.5 Error Analysis

We draw the confusion matrix in Figure 5 to study the types of errors made by the model. The confusion matrix reveals that although our method has alleviated the problem of misclassifying "Int" into "Effect", the problem still exists on DDIExtraction-13. We

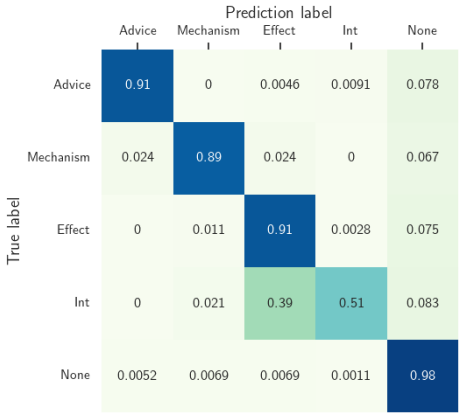

Figure 5: Performance of the proposed model on each DDI type.

believe there are three main reasons for this. First, as Table 4 shows, the frequency of "Int" is the smallest and "Effect" is the largest in positive samples. This results in the model not being able to learn sufficiently the difference between "Int" and "Effect" types. The second reason is that there are a few mislabeled samples in the gold standard dataset such as the sentence "A potential interaction between oral miconazole and oral hypoglycemic agents leading to severe hypoglycemia has been reported." which is labeled as "Effect" in the training dataset of "Glibenclamide_ddi.xml", but labeled as "Int" in the training dataset of "Glipizide_ddi.xml". However, if we filter these samples, the imbalance problem could be further increased. The last reason is that these types all describe the interaction of two drugs with little diversity. As a result, they have a similar semantic information and that is the reason why the problem of mislabeling has happened in the gold standard.

6 CONCLUSION

In this paper, we proposed a novel model incorporating a neural network and a knowledge graph in order to extract DDIs from biomedical texts. The results indicate that our method is effective for biomedical relation extraction, which suggests that the features from neural networks and knowledge graphs are mutually complementary.

In future, we will try to construct a larger drug knowledge graph to obtain more abundant drug information. Meanwhile, in order to handle the misclassification of "Int" types is as "Effect", we will attempt to increase the number of "Effect" and "Int" type training examples by data augmentation methods.

ACKNOWLEDGMENTS

This work is supported by Program for International Science and Technology Cooperation Projects of Shaanxi Province-general project 2021KW-63, and supported by National Natural Science Foundation of China(61877050).

REFERENCES

[1] Vanessa Miranda, Angelo Fede, Melissa Nobuo, Veronica Ayres, Auro Giglio, Michele Miranda, and Rachel P Riechelmann. Adverse drug reactions and drug interactions as causes of hospital admission in oncology. *Journal of pain and symptom management*, 42(3):342–353, Sep 2011.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, Jun 2019.

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

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics, Nov 2019.

[5] Yifan Peng, Qingyu Chen, and Zhiyong Lu. An empirical study of multi-task learning on BERT for biomedical text mining. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, BioNLP 2020, Online, July 9, 2020*, pages 205–214. Association for Computational Linguistics, Jul 2020.

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

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

[8] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, Jan 2018.

[9] Isabel Segura-Bedmar, Paloma Martínez Fernández, and María Herrero Zazo. 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: Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 341-350*, Atlanta, Georgia, USA, Jun 2013. Association for Computational Linguistics, ACL.

[10] Sangrak Lim and Jaewoo Kang. Chemical–gene relation extraction using recursive neural network. *Database*, 2018, Jun 2018.

[11] Md Faisal Mahbub Chowdhury and Alberto Lavelli. Fbk-irst: A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 351–355, Atlanta, Georgia, USA, June 2013.

[12] Philippe Thomas, Mariana Neves, Tim Rocktäschel, and Ulf Leser. Wbi-ddi: drug-drug interaction extraction using majority voting. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 628–635, Atlanta, Georgia, USA, June 2013.

[13] Behrouz Bokharaeian and Alberto Díaz. Nil_ucm: Extracting drug-drug interactions from text through combination of sequence and tree kernels. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 644–650, Atlanta, Georgia, USA, June 2013.

[14] Sun Kim, Haibin Liu, Lana Yeganova, and W John Wilbur. Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. *Journal of biomedical informatics*, 55:23–30, 2015.

[15] Anass Raihani and Nabil Laachfoubi. Extracting drug-drug interactions from biomedical text using a feature-based kernel approach. *Journal of Theoretical and Applied Information Technology*, 92(1):109, 2016.

[16] Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. Drug-drug interaction extraction via convolutional neural networks. *Computational and mathematical methods in medicine*, 2016, 2016.

[17] Shengyu Liu, Kai Chen, Qingcai Chen, and Buzhou Tang. Dependency-based convolutional neural network for drug-drug interaction extraction. In *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 1074–1080, Shenzhen, China, Dec 2016. IEEE.

[18] Xia Sun, Long Ma, Xiaodong Du, Jun Feng, and Ke Dong. Deep convolution neural networks for drug-drug interaction extraction. In *2018 ieee international conference on bioinformatics and biomedicine (bibm)*, pages 1662–1668, Madrid, Spain, Dec 2018. IEEE, IEEE.

[19] Masaki Asada, Makoto Miwa, and Yutaka Sasaki. Extracting drug-drug interactions with attention cnns. In *BioNLP 2017*, pages 9–18, Vancouver, Canada, Aug 2017. ACL, Association for Computational Linguistics.

[20] Ramakanth Kavuluru, Anthony Rios, and Tung Tran. Extracting drug-drug interactions with word and character-level recurrent neural networks. In *2017*

IEEE International Conference on Healthcare Informatics (ICHI), pages 5–12, Park City, UT, USA, Sep 2017. IEEE, IEEE.

[21] Wei Wang, Xi Yang, Canqun Yang, Xiaowei Guo, Xiang Zhang, and Chengkun Wu. Dependency-based long short term memory network for drug-drug interaction extraction. *BMC bioinformatics*, 18(16):99–109, Dec 2017.

[22] Wei Zheng, Hongfei Lin, Ling Luo, Zhehuan Zhao, Zhengguang Li, Yijia Zhang, Zhihao Yang, and Jian Wang. An attention-based effective neural model for drug-drug interactions extraction. *BMC bioinformatics*, 18(1):1–11, Oct 2017.

[23] Xia Sun, Ke Dong, Long Ma, Richard Sutcliffe, Feijuan He, Sushing Chen, and Jun Feng. Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. *Entropy*, 21(1):37, Jan 2019.

[24] Bo Xu, Xiufeng Shi, Zhehuan Zhao, Wei Zheng, Hongfei Lin, Zhihao Yang, Jian Wang, and Feng Xia. Full-attention based drug drug interaction extraction exploiting user-generated content. In *2018 ieee international conference on bioinformatics and biomedicine (bibm)*, pages 560–565, Madrid, Spain, Jan 2018. IEEE, IEEE.

[25] Masaki Asada, Makoto Miwa, and Yutaka Sasaki. Using drug descriptions and molecular structures for drug–drug interaction extraction from literature. *Bioinformatics*, 37(12):1739–1746, Jun 2021.

[26] Lei Huang, Jiecong Lin, Xiangtao Li, Linqi Song, Zetian Zheng, and Ka-Chun Wong. Egfi: drug–drug interaction extraction and generation with fusion of enriched entity and sentence information. *Briefings in Bioinformatics*, 23(1):bbab451, Jan 2022.

[27] Ishani Mondal. Bertkg-ddi: Towards incorporating entity-specific knowledge graph information in predicting drug-drug interactions. In *SDU@ AAAI*, Feb 2021.

[28] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics, 2018.

[29] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Apr 2020.

[30] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jul 2019. ACL.

[31] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

[32] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, May 2015.

[33] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, volume 48, pages 2071–2080. PMLR, PMLR, Jun 2016.

[34] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, May 2019.

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988. ICCV, IEEE, Oct 2017.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[37] Peng Su, Yifan Peng, and K Vijay-Shanker. Improving bert model using contrastive learning for biomedical relation extraction. In *Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP@NAACL-HLT 2021, Online, June 11, 2021*, pages 1–10. Association for Computational Linguistics, Jun 2021.