# A BERT-based Model for Drug-Drug Interaction Extraction from Drug Labels

Lei Ding, Liang Liang, Yixuan Tong, Shanshan Jiang, Bin Dong Ricoh Software Research Center (Beijing) Co.,Ltd. {lei.ding, liang, yixuan.tong, shanshan.jiang, bin.dong}@srcb.ricoh.com

## Abstract

The SRCB team participated in entity recognition task, relation identification task and normalization task in TAC Drug-Drug Interaction Extraction (DDI) 2019. The entity recognition system and relation identification system are based on BioBERT architecture with task-specific improvement. The normalization system includes candidate generation and further re-rank algorithms to find the right answer.

# 1 Introduction

Drug-drug interactions can lead to a variety of adverse events, and it has been suggested that preventable adverse events are the eighth leading cause of death in the United States (Goldstein et al., 2004). The Drug-Drug Interaction Extraction from Drug Labels Track in the 2019 Text Analysis Conference (TAC) aims at the automatic extraction of drug-drug interaction (DDI) information from the content of Structured Product Labeling (SPL) documents for prescription drugs.

In DDI track this year, SRCB team focus on the three tasks including entity recognition, relation identification and normalization. In the entity recognition task (Task1), the submitted system is an ensemble model based on BioBERT (Lee et al., 2019) with other improvement like Universal Transformer (Dehghani et al., 2018) and data augmentation. In the relation identification task (Task2), the task was formulated as a sentence-pair classification task and the model was also finetuned on pre-trained BioBERT. In normalization task (Task3), candidate retrieval and further re-rank methods were included in the submitted system.

This paper is organized as follows: Section 2 describes the submitted system in the three tasks. Section 3 analyzes the experiments results. Section 4 describes the submitted results.

# 2 System Description

## 2.1 Task 1

Task1 is an entity recognition task, which aims to extract mentions of interacting Drugs/Substances and specific interactions at sentence level. This is similar to many NLP named entity recognition (NER) evaluations. We trained an ensemble model based on BioBERT architecture (Lee et al., 2019). In our initial experiments, we found that several base models with different input can result in higher recall with limited loss in precision. Thus, an ensemble model combined with several similar BioBERT models with different train dataset.

Universal Transformer (Dehghani et al., 2018), data augmentation and average checkpoint are other techniques that have been brought in our NER models. When several Universal Transformer layers are added after the original BioBERT model, significant improvement has been shown. Inspired by the idea of data augmentation, additional unlabeled data has pre-process automatically for the training process, which has been proved to be useful in experiment. Average checkpoint has shown its ability to recall more mentions and specific interactions as well, where the last 5 checkpoints of the model are averaged to get one trained model. The above techniques are combined in the state of practice to obtain best performance.

2.1.1 Basic Named Entity Recognition model BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining), which is a domain specific language representation model pre-trained on large-scale biomedical corpora. Based on the BERT architecture (Devlin et al., 2018), BioBERT effectively transfers the knowledge from a large amount of biomedical texts to biomedical text mining models with minimal task specific architecture modifications. Compared with BERT, BioBERT significantly outperforms BERT on the biomedical named entity recognition task. In this paper, we followed the BioBERT architecture for named entity recognition task. Based on that, we introduce some techniques for better performance.

Our model reads input sentence word by word which has been processed by WordPiece tokenization (Wu et al., 2016). Then, the initial NER model is constructed. Similar with BioBERT, the model outputs the probabilities of different BIO labels of each token. Where the B label indicates that the token or subword is the beginning of an entity. Label I represents that current token is in the span of the entity, while the O label reflects that the token is out of the entity. In this way, our labeling scheme can deal with multi-word entities and triggers (which has shown its benefit in model performance in the following tasks).

#### 2.1.2 Ensemble Models

Based on above basic NER model, we train several models in the same structure with different inputs. Then combine the outputs of those models to produce more reliable results.

First, we train 11 BioBERT NER models. The training data is divided into 11 folds. For each BioBERT NER models, 10 of them are used as training data, and the remaining fold data is used as validation data. Then voting strategy is adopted to combine the 11 BioBERT models: the mean of predicted probability of 11 models is obtained, which is used to determine the label of the current token in the sentence.

Average checkpoint is also another scheme of model ensemble, which could reduce the variance of our model's output. For one configured translation model, once the model finishes training, the last 5 checkpoints of the model are averaged to get one trained model. In the art of practice, average checkpoint could help improve the recall of our model while ensure high precision as well.

2.1.3 Universal Transformer Model

Universal Transformer is a parallel-in-time selfattentive recurrent sequence model which can be cast as a generalization of the Transformer model. In experiment, when several Universal Transformer layers are attached to the initial BioBERT model, we find out that Universal Transformer could yield significant model performance. Share the same parameters between different layers contributes to the performance either.

#### 2.1.4 Data augmentation

Data augmentation has been proved to be effective in many domains in artificial intelligence like neural machine translation (Sennrich et al., 2015). Promoted by the success application of data augmentation, we pre-process the unlabeled sentences collected as input, output the predicted entities of each sentences through our trained NER model. We then combine the result with other golden truth datasets for next training process. We continue above loop process until the performance meet our expectation.

# 2.2 Task 2

The relation identification in Task 2 aims to identify the interactions at sentence level, including: the interacting drugs, the specific interaction types: pharmacokinetic, Pharmacodynamic or unspecified, and the outcomes of Pharmacokinetic and Pharmacodynamic interactions.

More recently, the pre-trained language models, such as ELMo (Peters et al., 2018), OpenAI GPT (Radford et al., 2018), and BERT (Devlin et al., 2018), have shown their effectiveness to alleviate the effort of feature engineering. Especially, BERT has achieved excellent results in sentence-pair classification problem. In the relation identification system in Task 2, inspired by previous work on aspect-based sentiment analysis on BERT (Sun et al., 2019), the task is converted to sentence-pair classification task. Support sentence are constructed to represent the background information of the target interaction and a pre-trained BioBERT model is used as the input representation of the sentencepair.

#### 2.2.1 Pre-processing

The input of the model is the sentence text with some additional process to identify the *Precipitant* and *SpecificInteraction* in the target interaction. According to the data analysis and experiments, the words representing single drugs, substances, or drug/substance classes do not have special meaning in each interaction but the results of interactions do. So, the "Precipitant" in each sentence text is replaced with unified word and extra curly braces are used to surround the *SpecificInteraction* to remain the content of the *SpecificInteraction*.

2.2.2 Support sentence construction

The support sentence contains the information of the target interaction. For example, the support sentence of Pharmacodynamic interaction contains the content of *Precipitant* and *SpecificInteraction* in the interaction and the support sentence of Pharmacokinetic interaction is the joint of the Precipitant and some keywords of Pharmacokinetic interaction such as increase, decrease, reduce, half time.

#### 2.2.3 Relation extraction model

The basic relation extraction model is a sentencepair classification model based on BioBERT. The model is trained to judge whether the input sentence match the information in the support sentence or not. The input sample are formulated as multiple binary classification task to identify whether the sentence represent Pharmacodynamic interaction with the Precipitant and the SpecificInteraction or whether sentence represent kind of the some Pharmacokinetic interaction or unspecified interactions with the Precipitant.

2.2.4 Data augmentation

The NLM-180 data is added to the training data after some process steps. The process steps include the field mapping and the prediction of the subtypes in Pharmacokinetic interactions based on the model trained on the annotated data.

# 2.3 Task 3

Our method employed Apache Solr<sup>1</sup> to index the terminologies that mentions to be mapped to, including MED-RT, UNII, and SNOMED CT. Given a mention, top 30 candidates are retrieved, along with BM25 relevance scores. For those candidates, query-document (relevancy) and document-document (importance) features are calculated to re-rank them. To measure the similarity between mentions and terminologies, we employed Jaccard distance, Longest Common Subsequence (LCS), Levenshtein distance, and their combinations. For feature weighting, learning to rank and empirical values found in the experiments are used. In case that extracted mention

<sup>&</sup>lt;sup>1</sup> https://lucene.apache.org/solr/

is incomplete, multiple NER models in Task 1 are used to determine the maximum span of the possible mention. We also found it necessary to set non-result thresholds to replace low-score result to "*NO MAP*", especially for *SpecificInteractions*.

# **3** Experiments

# 3.1 Task 1

#### 3.1.1 Setup

For development, we use 22 training labels and two labeled evaluation datasets supplied in DDI 2018, with additional 66 training labels provided this year. About 20,000 unlabeled xml files from the website have also been utilized for data augmentation.

Most parameters are initialized by randomly sampling from the default settings in BioBERT with grid search algorithm to obtain better results. All our models were trained on a NVIDIA GeForce GTX 1080X GPU. The training stage of each model took about 2.0 hours.

3.1.2 Results and Analysis

SRCB submitted three runs to the Task1 evaluation this year (called srcb1, srcb2 and srcb3). We used the ensemble model in all three runs for named entity recognition. We tried different combination of parameters for different run.

The performance of our model with different technical points on the 2018 DDI evaluation data for Task1 are listed in Table 1. All the scores are computed using the official scorer in 2018. In the results, we conclude that the data augmentation and Universal Transformer shows significant improvement over our final NER system.

# 3.2 Task 2

#### 3.2.1 Setup

In the experiments of Task2, the 22 SPLs used for training in 2018 (training 2018) and the set of 8,000 sentences from 180 SPLs re-annotated according to the 2018 guidelines (additional 66) are used as training data and the two test set in TAC DDI 2018 are used as evaluation data.

The parameters in the model are initialized by randomly sampling from the default settings in BioBERT and the hyper-parameters like the epoch number are determined with grid search on the evaluation data.

3.2.2 Results and Analysis

Results on two test sets in TAC DDI 2018 with are presented in Table 2. According to the results, the support sentence method brings significant improvement compared with the base method because it contains the meaning of the interactions. The NLM-180 data improves the result because it expands the training data and brings the kind of *Precipitant* without interactions in the sentence which is not included in the annotated dataset.

	Test_1 dataset		Test_2 dataset			
Technical point	Prec	Rec	F1	Prec	Rec	F1
BioBERT+Training22(Base)	30.65	41.56	35.28	32.45	40.39	35.99
+Data Augmentation (DA)	45.98	40.52	43.08	47.07	42.31	44.56
+Additional66	61.75	64.16	62.93	68.01	65.02	66.48
+Average	60.98	66.30	63.53	67.30	67.66	67.48
+Universial	66.93	68.27	67.59	70.95	68.39	69.65
+Ensemble	70.77	65.07	67.80	77.53	68.88	72.95

Table 1: Performance of Task1 on the test set in DDI 2018

	Test_1 dataset		Test_2 dataset			
	(Task1 F1=67.80)			(Task1 F1=72.95)		
Technical point	Prec	Rec	F1	Prec	Rec	F1
BioBERT (Base)	53.89	45.57	49.38	55.21	46.69	50.59
+ Support sentence	57.11	50.67	53.70	57.80	51.37	54.39
+NLM-180	58.40	51.74	54.87	58.50	52.02	55.07

Table 2: Performance of Task2 on the test set in DDI 2018

#### 3.3 Task 3

## 3.3.1 Setup

According to task description, drug classes should be normalized to MED-RT this time instead of NDF-RT in TAC 2018 DDI. Therefore, although XML-22 (a set of 22 gold-standard SPLs annotated with drug-drug interaction, from the 2018 training set), ADD-66 (A set of 8,000 sentences from 180 SPLs re-annotated according to the 2018 guidelines), and testset2018 (the test set of 128 SPLs from 2018) are all available for training, we chose to use ADD-66 as development data because it is supposed to be the most consistent dataset with testset2019. As for XML-22 and testset2018, we extracted the golden mappings from mentions strings to terminologies in UNII and SNOMED CT as a dictionary (gold\_dict).

3.3.2 Results and Analysis

Firstly, a very small-sized dictionary was introduced in candidate retrieval, to normalize synonyms such as {drug, agent, medication}, {toxicity, poison}, Arabic numerals and corresponding Roman numerals, Greek alphabet, and etc. Search parameters were tuned to achieve a relatively higher recall rather than a higher precision. Secondly, we experimented on mentions similarities between strings and terminologies, including feature selection and feature weighting. Mention-candidate similarity and candidate-candidate similarity both worked. We employed a combination of Jaccard distance, Levenshtein distance, and LCS to calculate the similarity. Experimental results showed that, empirical weights outperformed those trained by learning to rank methods. Thirdly, we extend the mention span by voting from NER systems when the mentions are possibly incomplete. Finally, nonresult thresholds are selected for Precipitant and SpecificInteration respectively. Experiments on ADD-66 using golden mentions are showed in Table 3, while using extracted mentions by a Task 1 system (Primary F1=58.80) are showed in Table 4.

	Macro-F
Retrieval (top1)	67.58
Retrieval (top30) + features (empirical)	80.87
Retrieval (top30) + features (learning to rank)	76.44
Retrieval (top30) + features (empirical) + non-result threshold	81.90
Retrieval $(top30)$ + features $(empirical)$ + non-result threshold + gold dict	83.12

Table 3: Relaxed Match results on ADD-66 (golden mentions)

Task1 F1=58.80	Macro-F1
Retrieval (top1)	53.78
Retrieval (top30) + features (empirical)	59.13
Retrieval (top30) + features (learning to rank)	55.17
Retrieval (top30) + features (empirical) + span extension	
Retrieval (top30) + features (empirical) + span extension + non-result threshold	
Retrieval (top30) + features (empirical) + span extension + non-result threshold + gold_dict	

Table 4: Relaxed Match results on ADD-66 (extracted mentions)

Entity Recognition (Task1)						
Run	Precision	Recall	F1-score			
srcb_1	70.9276	56.5161	62.9070			
srcb_2	71.3284	55.8109	62.6227			
srcb_3	72.4608	55.5236	62.8715			
Relation Identification (Task2)						
Run	Precision	Recall	F1-score			
srcb_1	53.8435	41.3241	46.7603			
srcb_2	54.7015	40.8436	46.7675			
srcb_3	53.8435	41.3241	46.7603			
Normalization (Task3)						
Run	Precision	Recall	F1-score			
srcb_1	67.5508	59.3683	61.4320			
srcb_2	65.7809	56.4855	59.4293			
srcb_3	70.8757	58.4930	62.3889			

Table 5: Result for test set 2019

# 3.4 Submissions

Table 5 shows the result of the three submitted runs in test set 2019. The three runs come from different combination of hyper-parameters and some postprocess steps like whether remove the mentions that do not have interaction in Task2. The result shows that there is significant reduction of the performance in different dataset which will be an importance direction for future research.

## 4 References

Goldstein, J., Jaradeh, I., Jhawar, P., Stair, T. ED Drug-Drug Interactions: Fre-quency & Type, Potential & Actual, Triage & Discharge. The Internet Journal of Emergency and Intensive Care Medicine 8(2), (2004)

- Lee J, Yoon W, Kim S, et al. BioBERT: pre-trained biomedical language representation model for biomedical text mining[J]. arXiv preprint arXiv:1901.08746, 2019.
- Dehghani M, Gouws S, Vinyals O, et al. Universal transformers[J]. arXiv preprint arXiv:1807.03819, 2018.
- Devlin J, Chang M W, Lee K, et al. Bert: Pretraining of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- Wu Y, Schuster M, Chen Z, et al. Google's neural

machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint arXiv:1609.08144, 2016.

Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual

data[J]. arXiv preprint arXiv:1511.06709, 2015.

Sun C, Huang L, Qiu X. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence[J]. 2019.