HITSZ-ICRC: A Report for SMM4H Shared Task 2019-Automatic Classification and Extraction of Adverse Drug Reactions in Tweets

Shuai Chen[1], Yuanhang Huang[1], Xiaowei Huang[1], Haoming Qin[1], Jun Yan[2], Buzhou Tang[1*]
[1]Department of Computer Science, Harbin Institute of Technology, Shenzhen, China
[2]Yidu Cloud (Beijing) Technology Co., Ltd, Beijing, China
{chenshuai726, hyhang7, kitaharatomoyo, tangbuzhou}@gmail.com
qinhaoming1874@foxmail.com, junyan@yiducloud.cn
*corresponding author

Abstract

This is the system description of the Harbin Institute of Technology Shenzhen (HITSZ) team for the first and second subtasks of the fourth Social Media Mining for Health Applications (SMM4H) shared task in 2019. The two subtasks are automatic classification and extraction of adverse effect mentions in tweets. The systems for the two subtasks are based on bidirectional encoder representations from transformers (BERT), and achieves promising results. Among the systems we developed for subtask1, the best F1-score was 0.6457, for subtask2, the best relaxed F1-score and the best strict F1-score were 0.614 and 0.407 respectively. Our system ranks first among all systems on subtask1.

1 Introduction

Adverse drug reaction (ADR), namely adverse drug effect, is one of the leading causes of post-therapeutic deaths (Saha, Naskar, Dasgupta, & Dey, 2018). Nowadays, more and more people share information in social platform, including health information such as drugs and their ADRs. Twitter, as one of the most popular social platforms, has attracted a great deal of attention from researchers in the medical domain. Some methods, such as HTR_MSA (Wu et al., 2018) and Neural DrugNet (Nikhil & Mundra, 2018), have been proposed to detect tweets mentioning ADRs and medicine intake. In order to facilitate the use of social media for health monitoring and surveillance, the health language processing lab at University of Pennsylvania organized Social Media Mining for Health Applications (SMM4H) shared task four times. In 2019, the fourth SMM4H shared task was comprised of four subtasks: (1) Automatic classifications of adverse effect mentions in tweets, (2) Extraction of Adverse Effect mentions, (3) Normalization of adverse drug reaction mentions (ADR), and (4) Generalizable identification of personal health experience mentions (Weissenbacher et al., 2019).

We participated in subtask 1 and subtask2, and developed two systems based on bidirectional encoder representations from transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2018) for the two subtasks respectively. The system for subtask 1 achieved the best F1-score of 0.6457, ranking first. Among the systems we developed for subtask2, the best relaxed F1-score and the best strict F1-score were 0.614 and 0.407 respectively.

2 Task and Data Description

2.1 Task 1: Automatic Classifications of Adverse Effect Mentions in Tweets

Task 1 was formulated as follows: given a tweet, determine whether it mentions drug adverse effect mentions, denoted by 1 and 0, indicating a tweet mentions drug adverse effects and not, respectively. The organizers provided a train dataset consisting of 25,678 tweets for all participants to develop their system, and a test dataset consisting of 4,575 tweets to evaluate the performance of all systems. Table 1 shows the distribution of 0 and 1 labels over the training and test datasets, where #* denotes the number of tweets labeled with *, and NA denotes that the corresponding number is currently unknown.

Dataset	#1	#0	#all
Training set	2,377	23,301	25,678
Test set	NA	NA	4,575

Table 1: Distribution of labels over the training and test datasets of task1.

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 47–51
Florence, Italy, August 2, 2019.

2.2 Task 2: Extraction of Adverse Effect Mentions

Task 2 as a follow-step of Task 1 was formulated as follows: given a tweet, identify the text span of adverse effect mentions. The challenge of task 2 is to distinguish adverse effect mentions from similar non-ADR expressions. A training set of 3,225 tweets annotated with 1830 adverse effect mentions was provided for system development, and a test set of 1,573 tweets was provided for system evaluation. The statistics of the training and test datasets are listed in Table 2.

Dataset	#tweets	#ADRs
Training set	3,225	1,830
Test set	1,573	NA

Table 2: Statistics of the training and test datasets of task 2

3 Methods

Our systems for both task 1 and task 2 were based on BERT, an unsupervised language representation method to obtain deep bidirectional representations of sentences by jointly conditioning on both left and right context in all layers from free text. Below we described in detail the methods for the two tasks: task 1 and task 2, respectively.

3.1 Task 1: BERT and BERT+Knowledge Base

In this task, we designed two methods, BERT and BERT +Knowledge Base. The model architecture is shown in Fig. 1.

BERT: Like what BERT did, we took the final hidden state of the first input token [CLS] as the representation of a tweet. Then we applied a softmax layer over the output to classify a tweet. We denote the representation vector as H, then the predicted label $\hat{y}$ is computed as:

$$\hat{y} = softmax(WH + b) \quad (1)$$

where W, b is the parameters of the fully connected layer.

BERT+Knowledge Base: Inspired by Li et al. (2018), we tried to combine the BERT output with features from knowledge bases to improve the performance of systems. We firstly extracted drugs which appear in the SIDER 4.1 (a side effect resource which contains information on marketed medicines and their recorded adverse drug reactions) from the train dataset, and obtained a drug lexicon of 538 drugs. Then we extracted corresponding adverse effects in SIDER according to the drug lexicon, and obtained 4,411 <drug, ADR> pairs. For each tweet, according to the presence of <drug ADR> pairs, we could build a binary feature. We incorporated the binary feature into representation vectors of a tweet. The final representation of a tweet is a concatenation of its BERT output and lexicon feature. Then we used a fully connected layer to fuse information from different feature spaces, and applied a softmax layer on it to classify tweets. We denote the output of BERT as H_1, the lexicon feature as H_2, then the predicted label $\hat{y}$ of a tweet is computed as :

$$\hat{y} = softmax(W[H_1, H_2] + b) \quad (2)$$

where W, b is the parameters of the fully connected layer. The loss function for two models training is crossentropy:

$$L = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}.log(\hat{y}_{ij}) \quad (3)$$

Where y_{ij} and $\hat{y}_{ij}$ are gold label and predicted label for the i_{th} sample in the j_{th} label category. N is the number of samples in a batch, C is the number of label categories.

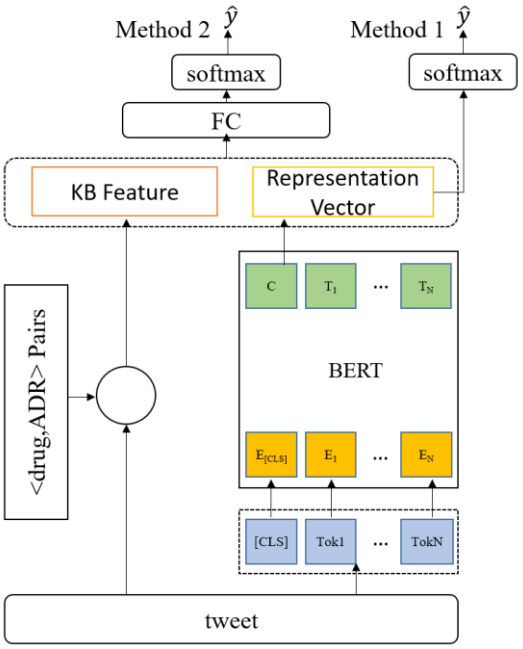

Figure 1: The model architecture in Task 1

3.2 Task 2: BERT and BERT+CRF

In task2, we still took BERT as the basic architecture, and designed two methods. The model architecture is shown in Fig. 2.

BERT: This method is very similar to the first method in Task 1. The difference is that we feed the

final hidden representation for to each token into a classification layer over the NER tags set, because we need to obtain predicted tag of each input token.

BERT +CRF: This method is a follow step of the first method. For BERT method, the predictions are not conditioned on the surrounding predictions. A CRF layer has a state transition matrix as parameters (Huang, Xu, & Yu, 2015). With such a layer, the system can efficiently use past and future tags to predict the current tag. Therefore, we applied a CRF layer on the classification layer. We denote the output sequence after softmax layer as $H = [h_1, h_2, \dots h_n]$, then the predicted tag sequence $Z = [z_1, z_2, \dots z_n]$ is as follows:

$$Z = \underset{y}{\operatorname{argmax}} \frac{exp(score(H,y))}{\sum_{y'} exp(score(H,y'))} \quad (4)$$

where $score(H,y) = \sum_{t=1}^{n} E_{t,y_t} + \sum_{t=0}^{n-1} T_{y_t y_{t+1}}$, $E_{t,y_t} = w_{y_t}^{\mathrm{T}} h_t$ is the score of predicting tag y_t at the t_{th} time, and $T_{y_t y_{t+1}}$ is the score of transitioning from y_t to y_{t+1}.

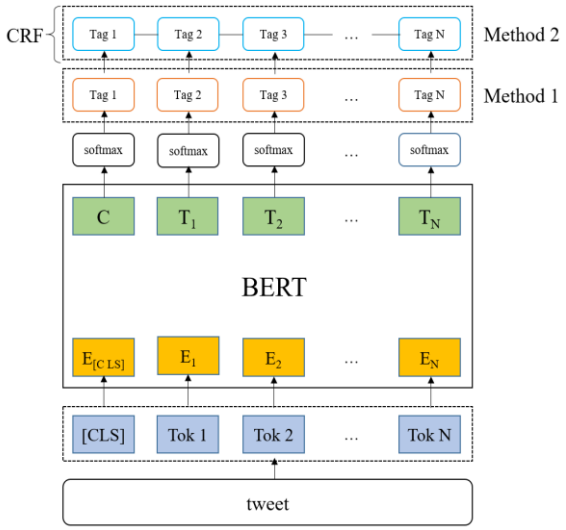

Figure 2: The model architecture in Task 2

3.3 Experiments

For task1, we compared BERT and BERT+knowledge base with two classic deep learning methods, TextCNN (Kim, 2014) and LSTM (Hochreiter & Schmidhuber, 1997), and also investigated the effect of different BERT models, including the BERT model (Devlin et al., 2018) publicly released by (https://github.com/google-research/bert) (denoted by BERT_noRetrained) and the BERT model retrained on a large-scale tweet unlabeled corpus based on the previous BERT model (denoted by BERT_Retrained). The unlabeled corpus consisted of 1,500,000 tweets crawled from Twitter according to 150 drug names collected from the training set. For task2, we only used the retrained BERT model.

In our experiments, we set batch size to 32, learning rate to 5e-5 when training all models. The epoch number was set to 8 for BERT retraining, and 20 for other models. The dimension of word embeddings used in TextCNN and LSTM was set to 200. We split out about 10% from the training set as a validation set for parameter optimization. The performance of all methods for the two tasks were measured by precision, recall and F1-score, which can be calculated by the official tools provided by the organizers. For task2, there were two criteria for system performance evaluation: relaxed and strict.

4 Results

Table 3 and Table 4 show the performance of our systems for task 1 and task 2 on the test set, respectively.

For task 1, among the systems we developed, "BERT_Retrained" achieved the best F1-score of 0.6457 and recall of 0.6885 on the test set, "BERT_Retrained+Knowledge Base" achieved the best precision of 0.6916 on the test set. Compared with TextCNN and LSTM on the validation set, methods based on BERT showed much better performance. As officially reported, "BERT_Retrained" ranked first among all systems.

For task 2, among the systems we developed, "BERT_Retrained+CRF" achieved the best relaxed F1-score of 0.614 and the best strict F1-score of 0.407, outperforming "BERT_Retrained" by 0.024 in relaxed F1-score and 0.060 in strict F1-score.

5 Discussion

For Task 1, the distribution of 0 and 1 is highly imbalanced, 90% of samples are negative, 10% of samples are positive. When we used CNN and LSTM, if we did not deal with the data imbalance problem, the performance of them was quite poor, most tweets were classified to 0. In order to balance the number of positive and negative samples, we randomly divided into the negative

System	Validation			Test		
	F1	P	R	F1	P	R
TextCNN	0.491	0.464	0.522	\		
LSTM	0.483	0.516	0.453	\		
BERT_noRretrained	0.618	0.646	0.593	\		
BERT_Retrained	**0.665**	0.611	**0.728**	**0.6457**	0.6079	**0.6885**
BERT_Retrained+Knowledge Base	0.642	**0.720**	0.579	0.6289	**0.6916**	0.5767
Average of participants' systems	\	\	\	0.5019	0.5351	0.5054

Table 3: Results on validation and test data for Task 1

System	Relaxed			Strict		
	F1	P	R	F1	P	R
BERT_Retrained+CRF	**0.614**	0.538	0.716	**0.407**	0.357	0.474
BERT_Retrained	0.59	0.529	0.666	0.347	0.311	0.392
Average of participants' systems	0.5383	0.5129	0.6174	0.3169	0.3026	0.3581

Table 4: Results on test data for Task 2

samples into five equal parts, and combined each part with the positive samples to form a new training dataset. After this operation, we obtained five new balanced training datasets. Then we trained five models on them, and ensembled the five models. The ensembled model brought an increase of about 8% in F1-score. However, when applying this operation to BERT and "BERT+Retrained", we obtained little increase on F1-score.

By analyzing results of "BERT_Retrained", we found that the main errors are:

- ADR mentions cannot be compeletely distinguished from the reason mentions of taking drugs. For example, in "oxycodone just took my headache away so fast", "headache" is the reason of taking oxycodone, not an adverse effect mention of oxycodone. The tweet was wrongly classified to 1.
- Implicit adverse effect mentions are difficult to identified. For example, "pristiq and im livin in a cold world" and "uhh my gabapentin does went up today and I don't even know what planet i'm on. i hope i adjust to this quickly ... #endometriosis".

For task 2, because the CRF layer takes full advantages of relations between neighbor labels, "BERT_Retrained+CRF" could avoid some terrible tag sequences such as "I-B-B-O-O". The main errors appearing in task 2 are the same as task 1.

For further improvement, a possible direction is dealing with task 1 and task 2 at the same time using joint learning methods.

6 Conclusion

In this paper, we developed systems for task 1 and task 2 of the SMM4H shared task in 2019. Our systems were based on BERT and achieved promising results, especially ranking first on task 1.

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *ArXiv Preprint ArXiv:1508.01991.*

Kim, Y. (2014). Convolutional neural networks for sentence classification. *ArXiv Preprint ArXiv:1408.5882*.

Li, H., Yang, M., Chen, Q., Tang, B., Wang, X., & Yan, J. (2018). Chemical-induced disease extraction via recurrent piecewise convolutional neural networks. *BMC Medical Informatics and Decision Making*, *18*(2), 60.

Nikhil, N., & Mundra, S. (2018, October). Neural DrugNet. *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task.*

Saha, R., Naskar, A., Dasgupta, T., & Dey, L. (2018). Leveraging Web Based Evidence Gathering for Drug Information Identification from Tweets. *Proceedings of* the *2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*.

Weissenbacher, D., Sarker, A., Magge, A., Daughton, A., O'Connor, K., Paul, M., & Graciela Gonzalez-Hernandez. (2019). Overview *of* the Fourth Social Media Mining for Health (SMM4H) Shared Task at ACL 2019. *Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task*.

Wu, C., Wu, F., Liu, J., *Wu*, S., Huang, Y., & Xie, X. (2018, October). Detecting Tweets Mentioning Drug Name and Adverse Drug Reaction with Hierarchical Tweet Representation and Multi-Head Self-Attention. *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task.*