Data Augmentation for BERT in the Medication Extraction Task of BioCreative VII

You-Qian Lee[1], Chen-Kai Wang[2,3*], Chung-Hong Lee[1], Vincent S Tseng[3], Hong-Jie Dai[1,4,5]
[1] Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan, R.O.C.
[2] Big Data Laboratory, Chunghwa Telecom Laboratories, Taoyuan, Taiwan, R.O.C
[3] Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, ROC
[4] Department of Post-Baccalaureate Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan, R.O.C.
[5] National Institute of Cancer Research, National Health Research Institutes, Tainan, Taiwan, R.O.C.

***Abstract*—Identifying medical entities such as disease and medications mentioned in short, informal, and noisy in social media text is challenging. We participated the track 3 of the BioCreative VII challenge with the goal to extract the mentions of medications or dietary supplements in tweets. We use different solutions based on BERT and BiLSTM (bidirectional long short-term memory) to develop our system under highly unbalanced data distribution. Four systems were developed for the task including the original BERT fine-tuning on the official training set, BERT with data augmentation (BERT-DA), BiLSTM, and BiLSTM with the focal loss. Owning to the limit of time for producing the predictions of the testing set, we only submitted two results (BERT and BERT-DA) for evaluation. The best performed model we submitted is the BERT-DA, which obtained an F1-score of 70.4%. From the evaluation results, we confirmed the effectiveness of the proposed data augmentation method, which can greatly improve the recall of the developed system.**

Keywords—Social media; named entity recognition; data imbalance

I. INTRODUCTION

Twitter has been utilized as an important source for monitoring public health-related messages for drug abuse and pharmacovigilance. To reach the goal, named entity recognition (NER) plays a very important role [1]. Machine learning (ML)-based methods is the main stream for addressing the NER task, which have evolved over recent years, with a noticeable shift from support vector machine (SVM) [2] and conditional random field (CRF) [3] trained on carefully engineered features to deep neural networks (DNN) that automatically discover relevant features from word embeddings [4]. In particular, the transformer-based language representation networks, such as Bidirectional Encoder Representations from Transformers (BERT) [5], is now be applied in many downstream natural language processing (NLP) tasks.

In this study, we present our system developed for the BioCreative VII track 3 [4]. The organizers provided datasets of English tweets at the message level with text spans of medication names, and their normalized IDs. The goal of the track is to extract the spans mentioning medications or dietary supplements in tweets. This track is especially challenging due to the reason that user generated texts from social media contain the natural and highly imbalanced distribution of medication mentions, with only approximately 0.2% of the tweets mentioning a medication. Motivated by the recent success of transformer-based architectures, we develop our systems based on the BERT and compare the performance with that of the bidirectional long short-term memory (BiLSTM) [6] systems.

II. METHODS

A. Dataset

The organizers provided a training set and a validation set consisting of 89,004 and 38,194 tweets respectively for all participants to develop their systems. In addition, the organizer also provides a dataset consisting of 9,622 tweets of social media mining for health applications (SMM4H) shared task in 2018 as an additional data. Table I shows the distribution of the medication mentions on the training and validation sets. We can find that the dataset is highly imbalanced.

TABLE I. DISTRIBUTIONS OF MEDICATION MENTIONS OVER THE DATASETS USED IN THIS STUDY.

Dataset	w/ Mentions	w/o Mentions	Total
Training Dataset	234 (0.003%)	88,770 (0.997%)	89,004
SMM4H Dataset	4,975 (0.517%)	4,647 (0.483%)	9,622
Validation Dataset	105 (0.003%)	38,044 (0.997%)	38,149
Test Dataset	N/A	N/A	54,482
Unlabel Dataset	10,531 (0.002%)	6,328,926 (0.998%)	6,339,457

B. IOB2 Tagging Scheme

The IOB2 tagging scheme is used to encode the span information. B (Beginning) indicates that the word is the start of a medication entity, I (Inside) indicates that the word is inside a medication entity, and O (Outside) indicates that the word is outside a medication entity. To tokenize the tweets used for training, we use the Twitter tokenizer provided by the

* Corresponding author

Python NLTK library [7]. An example is given as follows to illustrate the results of tokenization and encoding:

- **Tweet:** @TakinHearts_ 😂😂😂 I hope that don't happen to me I'm running until I get back on **birth control**
- **Tokens:** TakinHearts_ 😂😂😂 I hope that do n't happen to me I 'm running until I get back on **birth control**
- **IOB2:** O O O O O O O O O O O O O O O O O O O **B-DRUG I-DRUG**

C. BERT-based Model for Medication Recognition

BERT is an unsupervised language representation method for providing deep bidirectional representations of given sentences by jointly conditioning on both left and right context in all layers. We experimented with different combinations based on BERT and observed the performance of each NER model on the validation set. In our implementation, we fine-tuned the BERT-based pre-trained models on the released training set to develop our first system.

In addition to the training and the SMM4H datasets, we used the classifier based on our previous work in SMM4H 2020 task 1 [8] to classify tweets mentioned medications in a large unlabeled twitter corpus collected by our team. As shown in Table 1 "Unlabeled Dataset", the corpus consisted of 6,339,457 tweets on Twitter collected from January to April 2019, according to 183,593 drug names recorded in RxNorm [9] and 13,699 ADRs released by Nikfarjam, et al. [10]. After classification, a total of 10,531 tweets were regarded as containing medication mentions. Then we applied the aforementioned BERT-based system on these 10,531 tweets to recognize medication mentions, and used the results to augment our training set for training our data augmented model —BERT-DA.

For both models, the same parameters were used for fine-tuning BERT-base-cased; each model was trained by using the Adam optimizer with a learning rate of 3e-5 for 20 epochs with a batch size of 16.

D. BiLSTM-based Models

BiLSTM is built based on LSTM, which contains two LSTM layers, one is the forward LSTM and the other is the backward LSTM. Because BiLSTM performs forward and backward passes on the given sequential data to model dependencies in both directions, so that it can better capture temporal nuances.

Word embeddings have become the norm in neural NLP, which provide an effective mechanism for encoding the semantic and temporal context information [11]. It has been shown that using the pre-trained language model as the word embedding layer can improve the models' performance in sequence tagging [12]. In this work, the output of the last encoder layer of BERT was extracted as pre-trained word embeddings to cover contextualized embedding. We mapped each word in a sentence to the the 768-dimensional pre-trained word embedding and then used the BiLSTM to develop our third system.

As shown in Table I, the dataset is highly imbalance. Inspired by the work of focal loss [13] in the image recognition field, we implemented the focal loss for training our BiLSTM network to avoid the issues of imbalance and over-fitting, the loss function generalizes binary and multiclass cross-entropy loss and penalizes hard-to-classify examples. The developed BiLSTM-FL system is the fourth configuration in our experiments. For both BiLSTM models, embedding layer was initialized with the BERT-base-cased with the dimension of 768; the following hyper-parameters were set for training; each model was trained by using the Adam optimizer with a learning rate of 3e-5 for 20 epochs with a batch size of 16. For the focal loss, we experimented with two different combinations of gamma 0 and 5, respectively.

III. RESULTS

Table II shows the performance of the developed systems in terms of precision (P), recall (R), and F1-score (F1) on the validation and test set. For the validation set, the performance was evaluated by using the seqeval library [14]. From the result, we can observe that the BERT-based systems performed slightly better than the BiLSTM-based systems. In addition, we can find that for the BiLSTM-based systems, the result of using focal loss is better than the result of using cross-entropy loss. This result confirms that the use of the focal loss for imbalance handling can improve the generalization ability of models trained on highly imbalanced datasets.

Owing to the unexpected long time for processing the test set of the developed BiLSTM systems, we only submitted the predictions generated by the BERT-based systems for the official evaluation. The relevant results based on BiLSTM in Table II are not official results, these results are a full evaluation of the work done. In the results of the test set, the BERT-DA model significantly outperformed the BERT model in terms of recall and resulted in a better F-score. The phenomenon demonstrated the effectiveness of the proposed data augmentation method. Compared with the average scores of 16 teams, the two submitted systems have better precision. However, the recall is lower even with the proposed data augmentation method, handcrafted features and features with domain-specific knowledge [15] should be considered to be incorporated to boost the performance

TABLE II. PERFORMANCE ON VALIDATION AND TEST DATA FOR TRACK 3.

System	Validation			Test		
	P	R	F1	P	R	F1
BERT	0.68	0.64	0.66	0.810	0.115	0.201
BERT-DA	0.7	0.65	0.67	0.840	0.605	0.704
BiLSTM	0.67	0.55	0.61	0.827	0.581	0.683
BiLSTM-FL (gamma = 0)	0.67	0.55	0.61	0.842	0.574	0.683
BiLSTM-FL (gamma = 5)	0.71	0.58	0.64	0.752	0.578	0.654
Average scores				0.811	0.709	0.749

* Corresponding authors

IV. Error Analysis

In our error analysis, we observed that the developed systems can accurately recognize words that clearly refer to drugs, such as medications containing pills and zofran. However, they failed to recognize drugs expressed by their effect, such as sleeping medicine. We also observed that our systems tend to recognize words that frequently co-occur with drugs but are not drug names. For example: "Serotonin" and "Dopamine" were recognized as medications in the tweet “The answer is usually insufficient serotonin and dopamine”. Finally, we did not apply case normalization on the dataset, but many drug names may be described in upper- or lower-cases, such as "vitamin d" and "Vitamin D" or "birth control" and "BIRTH CONTROL". In the future, we will adopt the uncased BERT models with case normalization to improve the recall of the developed models.

V. Conclusion

In this study, we presented a data augmentation for the BERT models and applied the focal loss for BiLSTM to develop our systems to address the issue of the highly imbalanced data distribution. Our experimental results provide evidence to support the effectiveness of the proposed data augmentation method, which can greatly improve the recall of our systems. In the future, we will apply more sophisticated data preprocessing methods and continue to investigate the application of more advanced architecture and their combination with different data augmentation methods to improve the performance of our systems.

References

1. S. Sekine and C. Nobata, "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy," in LREC, 2004: Lisbon, Portugal, pp. 1977-1980.

2. W. S. Noble, "What is a support vector machine?," Nature biotechnology, vol. 24, no. 12, pp. 1565-1567, 2006.

3. H. Tseng, P.-C. Chang, G. Andrew, D. Jurafsky, and C. D. Manning, "A conditional random field word segmenter for sighan bakeoff 2005," in Proceedings of the fourth SIGHAN workshop on Chinese language Processing, 2005.

4. D. Weissenbacher, A. Sarker, A. Klein, K. O’Connor, A. Magge, and G. Gonzalez-Hernandez, "Deep neural networks ensemble for detecting medication mentions in tweets," Journal of the American Medical Informatics Association, vol. 26, no. 12, pp. 1618-1626, 2019.

5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

6. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

7. E. Loper and S. Bird, "Nltk: The natural language toolkit," arXiv preprint cs/0205028, 2002.

8. C.-K. Wang et al., "ISLab System for SMM4H Shared Task 2020," in Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, 2020, pp. 42-45.

9. S. Liu, W. Ma, R. Moore, V. Ganesan, and S. Nelson, "RxNorm: prescription for electronic drug information exchange," IT professional, vol. 7, no. 5, pp. 17-23, 2005.

10. A. Nikfarjam, A. Sarker, K. O’connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," Journal of the American Medical Informatics Association, vol. 22, no. 3, pp. 671-681, 2015.

11. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, 2013, pp. 3111-3119.

12. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," Journal of machine learning research, vol. 12, no. ARTICLE, p. 2493−2537, 2011.

13. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980-2988.

14. H. Nakayama, "seqeval: A python framework for sequence labeling evaluation," Software available from https://github. com/chakki-works/seqeval, 2018.

15. H.-J. Dai, "Family member information extraction via neural sequence labeling models with different tag schemes," BMC medical informatics and decision making, vol. 19, no. 10, pp. 1-12, 2019.