DATA PAPER

Semi-Supervised Noisy Label Learning for Chinese Clinical Named Entity Recognition

Zhucong Li[1,2]*, Zhen Gan[1,3]*, Baoli Zhang[1], Yubo Chen[1,2†], Jing Wan[3], Kang Liu[1,2], Jun Zhao[1,2†] & Shengping Liu[4]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

[3]Beijing University of Chemical Technology, Beijing 100029, China

[4]UNISOUND AI Technology Co., Ltd., Beijing 100096, China

Keywords: Named entity recognition; Electronic medical record; Noisy label learning; Semi-supervised; Adversarial training

Citation: Li, Z.C., et al.: Semi-supervised noisy label learning for Chinese clinical named entity recognition. Data Intelligence 3(3), 389-401 (2021). doi: 10.1162/dint_a_00099

Received: February 15, 2021; Revised: April 8, 2021; Accepted: April 30, 2021

ABSTRACT

This paper describes our approach for the Chinese clinical named entity recognition (CNER) task organized by the 2020 China Conference on Knowledge Graph and Semantic Computing (CCKS) competition. In this task, we need to identify the entity boundary and category labels of six entities from Chinese electronic medical record (EMR). We constructed a hybrid system composed of a semi-supervised noisy label learning model based on adversarial training and a rule post-processing module. The core idea of the hybrid system is to reduce the impact of data noise by optimizing the model results. Besides, we used post-processing rules to correct three cases of redundant labeling, missing labeling, and wrong labeling in the model prediction results. Our method proposed in this paper achieved strict criteria of 0.9156 and relax criteria of 0.9660 on the final test set, ranking first.

* These authors contributed equally to this work.

† Corresponding authors: Y.B. Chen (Email: yubo.chen@nlpr.ia.ac.cn; ORCID: 0000-0002-5485-9916) and J. Zhao (Email: jzhao@nlpr.ia.ac.cn; ORCID: 0000-0003-3370-2263).

1. INTRODUCTION

1.1 Evaluation Task

This task is a continuation of the series of evaluation carried out by China Conference on Knowledge Graph and Semantic Computing (CCKS) around the semantics of Chinese electronic medical records. It has been extended and expanded on the basis of the relevant evaluation tasks of CCKS2017, CCKS2018, and CCKS2019. For a given set of plain text documents of electronic medical records (EMRs), this Chinese medical record MER task in 2020 is to extract entity mentions and classify them into six predefined types of entities: disease and diagnosis, imaging examination, laboratory examination, operation, drug, and anatomy.

1.2 Data Set

The CCKS 2020 Medical Named Entity Recognition Competition provided 1,050 labeled data as a training set. The data included labels for six types of entities, including disease and diagnosis, imaging examination, laboratory examination, operation, drug, and anatomy. Besides, the evaluation task also provided 1,000 unlabeled data. The statistics of the number of entities in the training set are shown in Table 1.

Table 1. The statistics of the number of entities in the training set.

	Disease & Diagnosis	Imaging	Lab	Operation	Drug	Anatomy	Total
Deduplication	2,198	247	316	720	601	1,447	5,529
Duplication	4,345	1,002	1,297	923	1,935	8,811	18,313

1.3 Overview of Main Challenges and Solutions

Compared with named entity recognition (NER) in the general field [1], Chinese NER faces many new challenges. This paper introduces an algorithm modeling strategy towards the two significant challenges in this competition.

The first challenge is inconsistent entity labeling. Labelers from different medical departments may have various understandings of labeling standard, so labeling results of different standards are likely to appear. In the data set of this task, we did notice apparent inconsistencies in entity labeling. For example, this string "白细胞数" (white blood cell count) in some samples was labeled wholly as "白细胞数" (white blood cell count), while in other samples was labeled partly as "白细胞" (white blood cell). We did not know which standard was used in the test set. According to our estimation, about 13.69% of entities may be involved in inconsistent labeling, which seriously affected the model's final test performance. It was difficult to circumvent this problem with rules, nor could we directly correct the inconsistent entities in the training set.

The second challenge is that lacking training data led to inconsistent model results. Due to data's social sensitivity in the medical field, it is often difficult for researchers to obtain sufficient labeled data. The lack of annotated data is generally considered to lead to long-tail phenomena and poor model generalization. When training data are insufficient, the model prediction results may change drastically with different model parameters. How should we maintain the consistency of model results with the absence of training data?

This paper proposed a hybrid system composed of a semi-supervised noisy label learning model based on adversarial training and a rule post-processing module. The overall process of the system is shown in Figure 1. We introduced a five-fold cross-voting mechanism to deal with annotation inconsistency in the data set. A model ensemble mechanism and a semi-supervised training mechanism helped to cope with the unstable model results caused by lacking training data. Besides, an adversarial training mechanism is effective for the above two challenges. The official test set results show that our method achieved the highest score of 0.9156 on the strict criteria and 0.9660 on the relax criteria in the CCKS 2020 Chinese NER task.

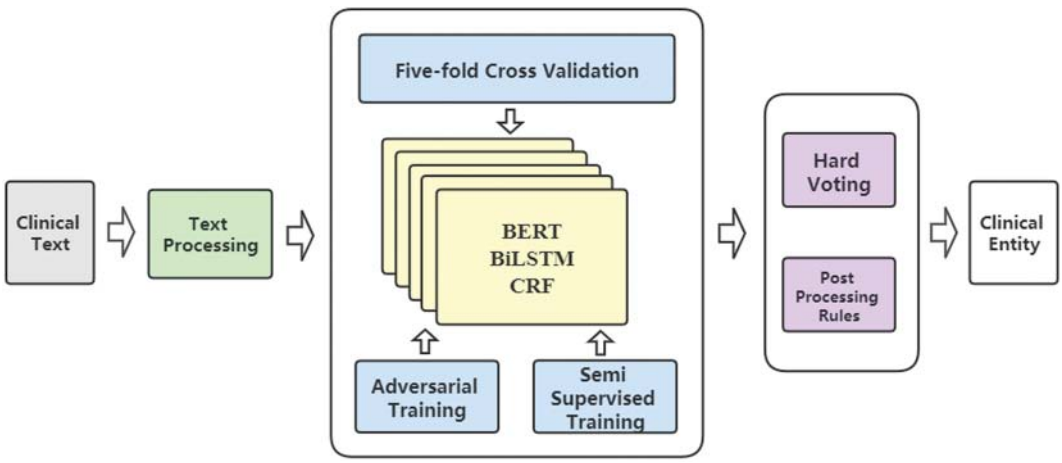


Figure 1. The overall process of our system.

2. RELATED WORK

2.1 Adversarial Training

The adversarial sample [2] is that adding small disturbances to the input samples that are difficult for humans to detect. Such attacks will seriously interfere with the prediction results of the neural network. The adversarial training is to train a more robust and generalized model by continuously defending against adversarial samples [3]. Madry et al. [3] defined adversarial training from an optimization perspective with Equation (1):

$$\min_{\theta} E_{(x,y)\sim\mathbb{D}}\left[\max_{r_{adv}\in\mathcal{S}} L\left(\theta, x + r_{adv}, y\right)\right] \quad (1)$$

The process of adversarial training is to find a small disturbance that can maximize the training loss and then optimize the model parameters θ to make the model loss smaller and continue to iterate to resist the current attack until it converges.

2.2 Semi-supervised Learning

Semi-supervised learning employs a small amount of labeled data as a supervised signal and combines numerous unlabeled data to achieve data augmentation. It has high application value and research value in fields where labeled data acquisition is expensive, such as medicine.

We used a semi-supervised training mechanism to incorporate the unlabeled data provided by the CCKS organizer[①] into the training process, which reduced the lack of annotated data to a certain extent.

3. METHODOLOGY

3.1 Basic Model Structure

Our basic model structure is shown in Figure 2. The sequence samples obtained their embedding representation through the pre-training model [4]. Then BiLSTM [5, 6, 7] is connected to the embedding representation for context encoding, and the Conditional Random Field (CRF) [8, 9] was used to decode the context representation. Finally the annotation result was obtained.

We tried five different pre-training models. The pre-training model can bring richer semantic representation, a large amount of world knowledge, and common sense knowledge.

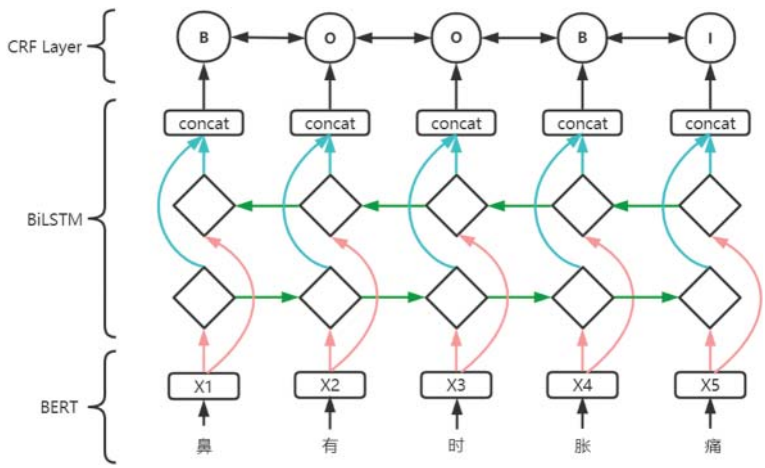


Figure 2. Our basic model structure.

① http://openkg.cn/dataset/yidu-s4k

3.2 Five-fold Cross-voting

We used five-fold cross-validation to divide the training set into five different data sets, and the inconsistencies of entity labeling in each data set were various. We fixed the same model structure, trained five models on five training sets, and integrated their prediction results on the same test set by hard voting.

3.3 Model Ensemble

To further reduce the impact of the randomness of the model parameters on the prediction results, we ensembled a variety of models through voting to weaken the impact of performance fluctuations caused by a single model parameter change on the prediction results.

Figure 3 shows the process of model ensemble combined with five-fold cross-voting. There are two voting sequences. The red box indicates that the five models trained on the same training set were first fused, and then the five fusion models obtained on the five-fold data set were continued to be fused, for a total of 25 models. The green box indicates that the five models obtained from the five-fold data set for each model structure were first obtained, and then the five models obtained from the five model results were continued to be merged, for a total of 25 models. Because the two sequences' results were similar, we followed the sequence represented by the green box by default.

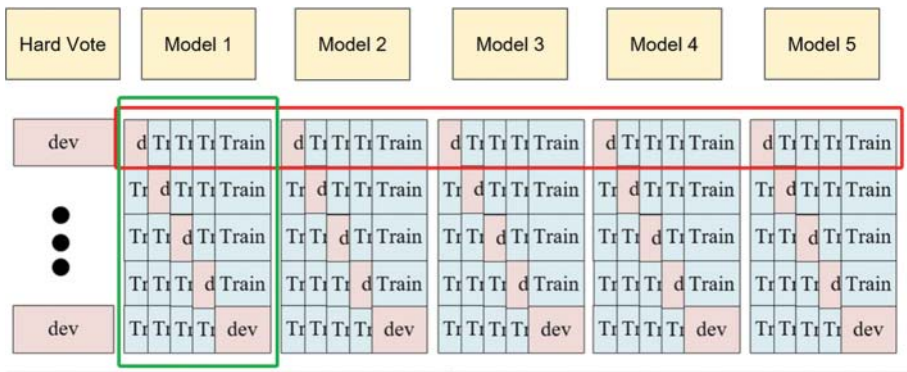


Figure 3. The process of model ensemble combined with five-fold cross-voting.

3.4 Semi-supervised Training

The semi-supervised training process is divided into two stages. The first stage used all 1,050 labeled data for training and 1,000 unlabeled data; the second stage added the obtained pseudo-labeled data to the training set to obtain the final mode.

3.5 Adversarial Training

Referring to the FGM [10] adversarial training mechanism, we directly imposed a small disturbance on the embedding representation of the model and assumed the embedding representation of the input text sequence $[v_1, v_2, \ldots, v_T]$ as x. Then the small disturbance r_{adv} applied each time is computed with Equations (2) and (3):

$$\mathrm{r}_{\mathrm{adv}} = \epsilon \cdot \mathrm{g} / |\mathrm{g}|_2 \tag{2}$$

$$g = \nabla_x L(\theta, x, y) \tag{3}$$

The meaning of the equations is to move the input one step further in the direction of rising loss, which will make the model loss rise in the fastest direction, thus forming an attack. In contrast, the model needs to find more robust parameters in the optimization process to deal with attacks against samples.

Among them, applying a small disturbance to the embedding characterization simulates the natural error of the data set in the labeling to a certain extent. It encourages the model to find more robust parameters during the training process. Then the model's embedding representation will be optimized together with the model. Adversarial training will make the model more tolerant of changes brought about by model parameter fluctuations, thereby decreasing the impact of data noise.

3.6 Post-processing Rules

If an entity has multiple labeling standards, then the number ratio between each labeling standard of the test set should be consistent with the training set. Based on this assumption, entities in the prediction results inconsistent with the distribution in the training set can be directly screened out. For the selected entities, we continued to subdivide entities based on the three cases of redundant labeling, miss labeling, and wrong labeling and established a redundant labeling dictionary, a missing labeling dictionary, and a wrong labeling dictionary for correction.

4. EVALUATION

4.1 Evaluation Metrics

There are two *F*1 criteria for this task. The strict *F*1 criteria are right only when the entity boundary and entity type are consistent with the gold answer. The other relax *F*1 criteria are right when the entity type is consistent with the gold answer or the entity boundary overlaps with the gold answer boundary. To reflect model performance more accurately, we only used strict *F*1 criteria in the local evaluation.

4.2 Pre-processing

We performed the following pre-processing for each piece of data.

4.2.1 Sentence Segmentation

Since the maximum input sequence of the data BERT model was only 512, the input medical record text was segmented under the premise of ensuring the relatively complete semantic information in the office to ensure that each input's text length was less than 512.

4.2.2 Text Normalization

This part mainly realizes the unification of the text and symbols in the input medical record, the conversion of English cases, and the processing of invisible characters.

4.3 Implementation Details

Implementation details of our five basic models are shown in Table 2.

Table 2. Implementation details of our five basic models.

Model	Learning Rate	Epoch	Dropout [11]	Optimizer
BERT-base+BiLSTM+CRF	5e-5	50	0.3	AdamW [12]
BERT-wwm-ext+BiLSTM+CRF[13]	3e-5	50	0.3	AdamW
RoBERTa-wwm-ext+BiLSTM+CRF[13]	3e-5	50	0.3	AdamW
RoBERTa-wwm-ext-large+BiLSTM+CRF[13]	3e-5	20	0.3	AdamW
RoBERTa-wwm-ext-large+CRF[13]	3e-5	20	0.3	AdamW

4.4 Results

We divided the 1,050 training data into five data groups according to the five-fold cross method, and each data group contains 840 training set and 210 development set. Table 3 shows the results of the local development set. The results in Table 3 are the average of *F*1 on five local development sets. In all tables of this paper, we abbreviated Semi-supervised Training as ST, Adversarial Training as AT, and Post-processing Rules as PR.

It can be noticed from Table 3 that the model ensemble mechanism and semi-supervised training mechanism, and the adversarial training mechanism have brought significant improvements to the basic model. Furthermore, after combining the three mechanisms, the best model result was achieved.

Table 3. Results on the local development set.

Model	*F*1
BERT-base+BiLSTM+CRF	0.8398
BERT-wwm-ext+BiLSTM+CRF	0.8415
RoBERTa-wwm-ext+BiLSTM+CRF	0.8412
RoBERTa-wwm-ext-large+BiLSTM+CRF	0.8463
RoBERTa-wwm-ext-large+CRF	0.8445
BERT-base+BiLSTM+CRF+Semi-supervised Training	0.8530
BERT-base+BiLSTM+CRF+Adversarial Training	0.8473
Model Ensemble	0.8717
Model Ensemble+Semi-supervised Training	0.8731
Model Ensemble+Adversarial Training	0.8735
Model Ensemble+Semi-supervised Training+Adversarial Training	0.8741
+Model Post-processing Rules	**0.8849**

The results of the official test set are shown in Table 4. We call BERT-base+BiLSTM+CRF the Single Model. The Single Model score is 0.0384 higher than that of the local, indicating that the inconsistency of entity annotations on the official test set may be much less than that in the training set. In the final model, we used a five-fold cross-voting mechanism for each model used for fusion to reduce the impact of lack of training data.

It is worth noting that although the overall improvement brought by the post-processing rule is not apparent in the local development set, it has brought significant improvements of 0.0242 and 0.0414 in the inspection and verification of the two classes with fewer entities.

Table 5 shows final performance obtained on the official test set.

Table 4. Results on the official test set.

Model	Disease & Diagnosis	Imaging	Lab	Operation	Drug	Anatomy	Total
Single Model	0.8591	0.8586	0.8141	0.9193	0.9213	0.8778	0.8782
+ST+AT	0.8902	0.8567	0.8240	0.9279	0.9266	0.9042	0.8992
Our Method	0.9093	0.8996	0.8594	0.9485	0.9356	0.9162	0.9156
- PR	0.9056	0.8754	0.8180	0.9441	0.9330	0.9088	0.9088

Table 5. Final performance obtained on the official test set.

Criteria	Disease & Diagnosis	Imaging	Lab	Operation	Drug	Anatomy	Total
Relax	0.9712	0.9239	0.9258	0.9754	0.9778	0.9667	0.9660
Strict	0.9093	0.8996	0.8594	0.9485	0.9356	0.9162	0.9156

5. CONCLUSION AND FUTURE WORK

To solve the two core challenges in the data set of this task, inconsistent entity annotations and lack of annotated data, we introduced the semi-supervised data augmentation and the adversarial training methods, which achieved the best good performance.

The task of MER, precisely quantifying the inconsistency of entity annotations in data, and letting the model better overcome this noise, is our future research goal.

ACKNOWLEDGEMENTS

This work is supported by the National Key R&D Program of China (2020AAA0106400), the National Natural Science Foundation of China (No. 61831022, No. 61806201) and the Key Research Program of the Chinese Academy of Sciences (Grant No. ZDBS-SSW-JSC006). This work is also supported by Beijing Academy of Artificial Intelligence (BAAI).

AUTHOR CONTRIBUTIONS

Z.C. Li (zhucong.li@nlpr.ia.ac.cn) and Z. Gan (ganzhen@mail.buct.edu.cn) planned and contributed equally to the overall work. Z.C. Gan, Z. Li and B.L. Z (baoli.zhang@nlpr.ia.ac.cn) performed the approach. Z.C. Li and Z. Gan wrote the manuscript. Y.B. Chen (yubo.chen@nlpr.ia.ac.cn), J. Wan (wanj@mail.buct.edu.cn), K. Liu (kliu@nlpr.ia.ac.cn), J. Zhao (jzhao@nlpr.ia.ac.cn), and S.P. Liu (liushengping@unisound.com) supervised the whole work. All the authors reviewed the manuscript.

DATA AVAILABILITY STATEMENT

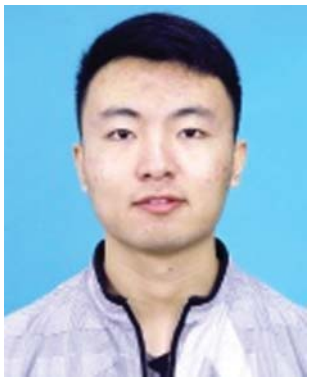
The data sets generated and/or analyzed during the current study are not publicly available due to the fact that the data sets are produced by medical expert consultants of Yidu Cloud based on their own experience. The publicly released version of the data sets needs the consent of all expert consultants, and they are available from the corresponding authors on reasonable request.

REFERENCES

[1] Marrero, M., et al.: Named entity recognition: Fallacies, challenges and opportunities. Computer Standards & Interfaces 35(5), 482–489 (2013)
[2] Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
[3] Madry, A., et al.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
[4] Devlin, J., et al.: 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, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019)

[5] Xu, K., et al.: A bidirectional LSTM and conditional random fields approach to medical name identity recognition. In: International Conference on Advanced Intelligent Systems and Informatics, pp. 355–365 (2017)
[6] Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1064–1074 (2016)
[7] Lample, G., et al.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260–270 (2016)
[8] Lafferty, J.D., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, pp. 282–289 (2001)
[9] Sutton, C., McCallum, A.: An introduction to conditional random fields for relational learning. Introduction to Statistical Relational Learning 2, 93–128 (2006)
[10] Miyato, T., Dai, A.M., Goodfellow, I.: Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725 (2016)
[11] Srivastava, N., et al.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1), 1929–1958 (2014)
[12] Loshchilov, I., Hutter, F.: Fixing weight decay regularization in Adam. arXiv preprint arXiv:1711.05101 (2018)
[13] Cui, Y., et al.: Revisiting pre-trained models for Chinese natural language-processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 657–668 (2020)

AUTHOR BIOGRAPHY



Zhucong Li is currently a graduate student at School of Artificial Intelligence, University of Chinese Academy of Sciences. His research interests include information extraction, knowledge graph and natural language processing.
ORCID: 0000-0002-6057-5784



Zhen Gan is currently a graduate student at Beijing University of Chemical Technology. His research interests mainly focus on natural language processing and information extraction.
ORCID: 0000-0002-5128-3501



Baoli Zhang received his Master's degree in Computer Science from Beijing University of Posts and Telecommunications. He is now an engineer of Institute of Automation, Chinese Academy of Sciences. His research interests mainly focus on information extraction, knowledge graph and natural language processing.
ORCID: 0000-0002-5815-7292



Yubo Chen is an Associate Professor of the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences. His research interests include information extraction, knowledge graph and natural language processing. He has published over 30 papers in prestigious conferences such as ACL, EMNLP, COLING, AAAI and IJCAI.
ORCID: 0000-0002-5485-9916



Jing Wan is an Associate Professor of Beijing University of Chemical Technology. Her research interests include knowledge graph and cultural heritage digital protection. She has published over 40 papers.
ORCID: 0000-0002-4232-7883



Kang Liu is currently a Professor of the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences. His research interests include natural language processing, knowledge graph, and question answering. He has published over 90 papers in journals like *IEEE Transactions on Knowledge and Data Engineering* (TKDE) and conferences like ACL, IJCAI, CIKM, EMNLP, and COLING. He has won COLING 2014 Best Paper Award.
ORCID: 0000-0002-6083-8433



Jun Zhao is a Professor of the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, and School of Artificial Intelligence, University of Chinese Academy of Sciences. Prof. Zhao has published over 90 peer-reviewed papers in the prestigious conferences and journals, including ACL and AAAI. He has won COLING 2014 Best Paper Award.
ORCID: 0000-0003-3370-2263



Shengping Liu received his PhD degree from the Department of Information Science, School of Mathematics, Peking University. Now he is a senior technical expert of UNISOUND AI Technology Co., Ltd.