Noisy Label Learning for Chinese Medical Named Entity Recognition Based on Uncertainty Strategy

Zhucong Li[1,2*], Zhen Gan[1,3*], Baoli Zhang[1], Yubo Chen[1,2], Jing Wan[3], and Jun Zhao[1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Beijing University of Chemical Technology
{zhucong.li,baoli.zhang,yubo.chen,jzhao}@nlpr.ia.ac.cn
{ganzhen,wanj}@mail.buct.edu.cn

Abstract. Medical named entity recognition(MER) is the basis of medical information extraction and a key technology for constructing medical knowledge graphs. This paper describes our approach for the Chinese MER 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, including disease, imaging examination, laboratory examination, drug, operation, and anatomy from Chinese electronic medical record(EMR). We construct a hybrid system composed of a semi-supervised noisy label learning model based on adversarial training and a rule post-processing module. The hybrid system's core idea is to reduce the aleatoric uncertainty caused by the inconsistent annotation standards of crowdsourced data and the epistemic uncertainty exacerbated by the lack of annotation data. Besides, we use 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.

Keywords: Medical Named Entity Recognition · Noisy Label Learning · Uncertainty.

1 Introduction

1.1 Task Definition

For a given set of plain text documents of EMR, this Chinese medical record MER task is to extract entity mentions and classify them into six predefined

* Equal contribution.

types of entities: disease & diagnosis, imaging examination, laboratory examination, operation, drug, and anatomy.

1.2 Overview of Main Challenges and Solutions

Compared with named entity recognition(NER) in the general field[9], MER faces many new challenges. This paper introduces uncertainty as 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 a various understanding of labeling standard, so labeling results of different standards are likely to appear. In the dataset of this task, we do notice apparent inconsistencies in entity labeling. For example, 白细胞数(white blood cell count), this string in some samples is labeled wholly as 白细胞数(white blood cell), while in other samples is labeled partly as 白细胞(white blood cell count). We do not know which standard is used in the test set. According to our estimation, about 13.69% of entities may be involved in inconsistent labeling, which seriously affects the model's final test performance. This phenomenon is difficult to circumvent with rules, nor can we directly correct the inconsistent entities in the training set.

The second challenge is that lacking training data leads 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 is insufficiency, 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?

In recent years, researchers have increasingly turned their attention to Bayesian deep learning methods that are more explanatory and mathematical. According to the Bayesian deep learning theory, labeling inconsistency can result in higher aleatoric uncertainty in training data, and lack of labeling data can lead to higher epistemic uncertainty in the model[3]. Therefore, designing algorithms to reduce aleatoric uncertainty and epistemic uncertainty will help alleviate the two major challenges' harmful effects.

This paper propose 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. To deal with annotation inconsistency in the dataset, we introduce a five-fold cross-voting mechanism to reduce aleatoric uncertainty. A model ensemble mechanism and a semi-supervised training mechanism help reduce epistemic uncertainty to cope with the unstable model results caused by lacking training data. Besides, an adversarial training mechanism is used to decrease aleatoric uncertainty and epistemic uncertainty simultaneously. The official test set results to 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 MER task.

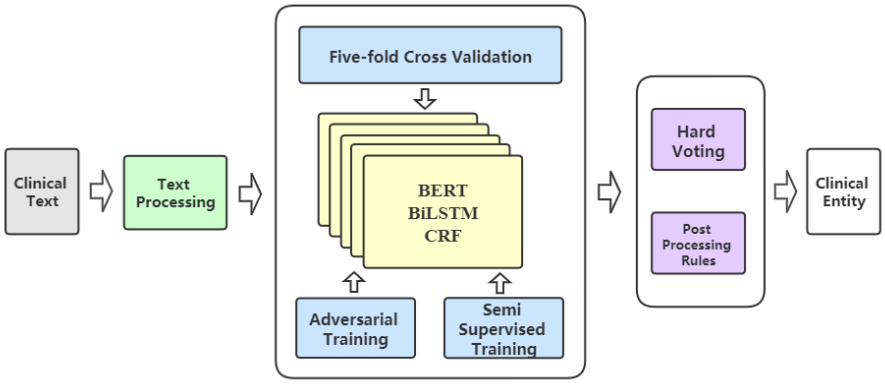

Fig. 1. The overall process of our system.

2 Related Work

2.1 Uncertainty in Deep Learning

Bayesian deep learning theory believes that there are two main types of uncertainty in deep learning: aleatoric uncertainty and epistemic uncertainty. They can cause fluctuations in model results, hinder the model generalization, and damage the model performance. The aleatoric uncertainty comes from the error of the data annotation itself. The more disordered the annotation noise in the dataset, the greater the aleatoric uncertainty. The epistemic uncertainty comes from the observation error on results caused by the model parameter sensitivity. It is worth noting that lacking training data will aggravate the negative impact of epistemic uncertainty on the model[3, 14].

2.2 Adversarial Training

The adversarial sample[13] 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[8].

Madry et al[8]. defined adversarial training from an optimization perspective:

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

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.

2.3 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 use a semi-supervised training mechanism to incorporate the unlabeled data provided by the CCKS organizer into the training process, which reduces the model's epistemic uncertainty to a certain extent.

3 Our Method

3.1 Basic Model Structure

Our basic model structure is shown in Figure 2. The sequence samples get their embedding representation through the pre-training model[2]. Then BiLSTM[15, 7, 5] is connected to the embedding representation for context encoding, and CRF[4, 12] is used to decode the context representation. Finally the annotation result is obtained.

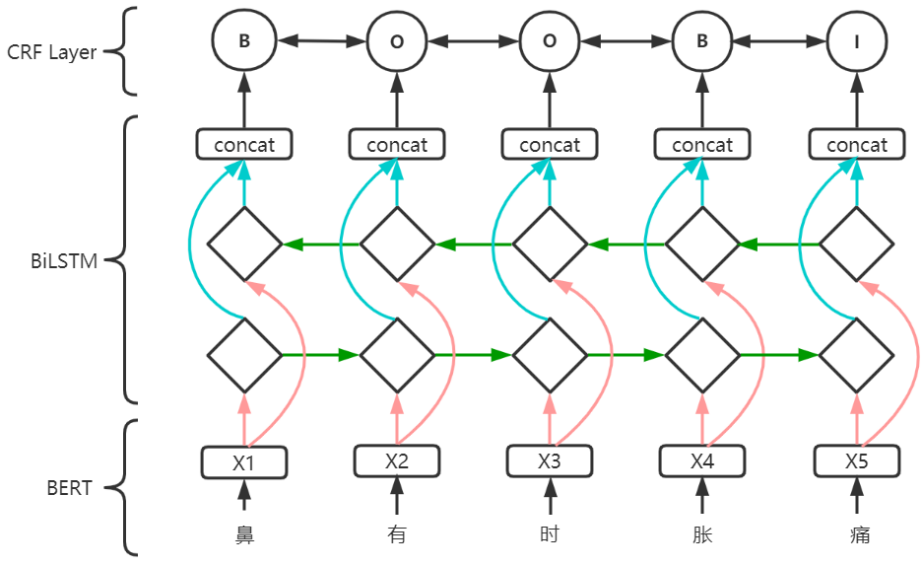

Fig. 2. Our basic model structure.

We tried five different pre-training models. The pre-training model can bring richer semantic representation, a large amount of world knowledge, common sense knowledge, and grammatical knowledge contained therein can play a similar role in data expansion.

3.2 Reducing Aleatoric Uncertainty

Five-fold Cross-voting We use five-fold cross-validation to divide the training set into five different datasets, and the inconsistencies of entity labeling in each dataset are various. We fix the same model structure, train five models on five training sets, and integrate their prediction results on the same test set by hard voting.

3.3 Reducing Aleatoric Uncertainty

Model Ensemble To further reduce the impact of the randomness of the model parameters on the prediction results, we ensemble 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 are first fused, and then the five fusion models obtained on the five-fold data set are 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 are first obtained, and then the five models obtained from the five model results are continued to be merged, for a total of 25 models. Because the two sequences' results are similar, we follow the sequence represented by the green box by default.

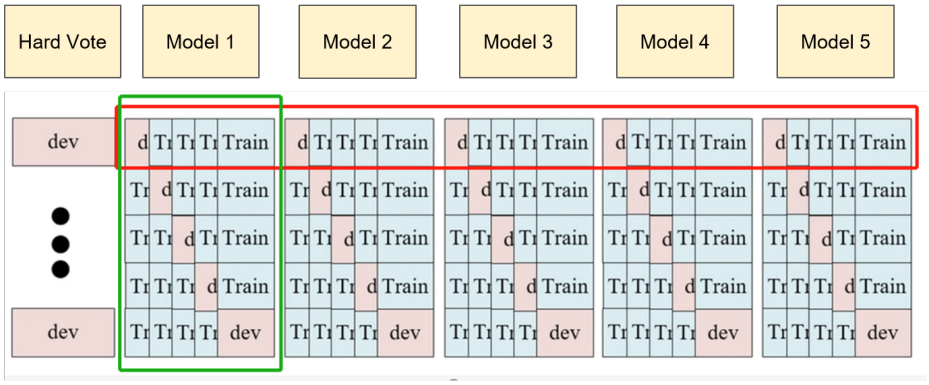

Fig. 3. The process of model ensemble combined with five-fold cross-voting.

Semi-supervised Training The semi-supervised training process is divided into two stages: the first stage uses all 1050 labeled data for training and 1000 unlabeled data; the second stage adds the obtained pseudo-labeled data to the training set to get the final model.

3.4 Reducing Aleatoric Uncertainty and Epistemic Uncertainty Simultaneously

Adversarial Training Referring to the FGM[10] adversarial training mechanism, we directly impose a small disturbance on the embedding representation of the model and assume 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:

$$r_{adv} = \epsilon \cdot g / \|g\|_2 \tag{2}$$

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

The meaning of the formulas 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 dataset in the labeling to a certain extent. It encourages the model to find more robust parameters during the training process to weaken the impact of aleatoric uncertainty. 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 epistemic uncertainty.

3.5 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 continue to subdivide entities based on the three cases of redundant labeling, miss labeling, and wrong labeling and establish a redundant labeling dictionary, a missing labeling dictionary, and a wrong labeling dictionary for correction.

4 Experiments

4.1 Dataset

The CCKS 2020 Medical Named Entity Recognition Competition provides 1,050 labeled data as a training set. The data includes labels for six types of entities, including disease & diagnosis, imaging examination, laboratory examination, operation, drug, and anatomy. Besides, the evaluation task also provided 1,000 unlabeled corpora. 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	2198	247	316	720	601	1447	5529
Duplication	4345	1002	1297	923	1935	8811	18313

4.2 Evaluation

There are two F1 criteria for this task. The strict F1 criteria are right only when the entity boundary and entity type are consistent with the gold answer. The other relax F1 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 use strict F1 criteria in the local evaluation.

4.3 Pre-processing

We perform the following pre-processing for each piece of data:

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

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.

4.4 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[6]
BERT-wwm-ext+BiLSTM+CRF[1]	3e-5	50	0.3	AdamW
RoBERTa-wwm-ext+BiLSTM+CRF[1]	3e-5	50	0.3	AdamW
RoBERTa-wwm-ext-large+BiLSTM+CRF[1]	3e-5	20	0.3	AdamW
RoBERTa-wwm-ext-large+CRF[1]	3e-5	20	0.3	AdamW

4.5 Results

We divided the 1050 training data into five data according to the five-fold cross method, and each data 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 F1 on five local development sets. In all tables of this paper, we abbreviate 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 is achieved.

Table 3. Results on the local development set.

Model	F1
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**

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

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 accidental uncertainty in the data.

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

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.

5 Conclusions

To solve the two core challenges in the dataset of this task: inconsistent entity annotation and lack of annotated data, we innovatively introduced the concept of uncertainty in deep Bayesian theory to guide the design of corresponding algorithms, thus achieving the best Good performance.

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

Acknowledgement

This work is supported by the Natural Key R&D Program of China (No.2017 YFB1002101), the National Natural Science Foundation of China (No.61533018, No.61976211, No.61806201) and the Key Research Program of the Chinese Academy of Sciences (Grant NO. ZDBS-SSW-JSC006). This work is also supported by the CCF-Tencent Open Research Fund, a grant from Ant Group and independent research project of National Laboratory of Pattern Recognition.

References

1. Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Revisiting pre-trained models for chinese natural language processing. In: Findings of EMNLP. Association for Computational Linguistics (2020)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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)
3. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in neural information processing systems. pp. 5574–5584 (2017)
4. Lafferty, J.D., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289 (2001)

5. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: 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)
6. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)
7. 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)
8. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
9. Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., Gómez-Berbís, J.M.: Named entity recognition: fallacies, challenges and opportunities. Computer Standards & Interfaces **35**(5), 482–489 (2013)
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., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
12. Sutton, C., McCallum, A.: An introduction to conditional random fields for relational learning. Introduction to statistical relational learning **2**, 93–128 (2006)
13. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
14. Xiao, Y., Wang, W.Y.: Quantifying uncertainties in natural language processing tasks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 7322–7329 (2019)
15. Xu, K., Zhou, Z., Hao, T., Liu, W.: A bidirectional lstm and conditional random fields approach to medical named entity recognition. In: International Conference on Advanced Intelligent Systems and Informatics. pp. 355–365. Springer (2017)