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

Authors: Zhucong, Li, Zhen, Gan, Baoli, Zhang, Yubo Chen, Jing, Wan, Kang, Liu, Jun, Zhao, Shengping, Liu, Yubo Chen, Jun Zhao

Date: 2022-11-27T00:00:00+00:00

Abstract

This paper describes our approach to 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 are required to identify the boundaries and category labels of six entity types from Chinese electronic medical records (EMR). We constructed a hybrid system comprising a semi-supervised noisy label learning model based on adversarial training and a rule-based post-processing module. The core idea of the hybrid system is to mitigate the impact of data noise by optimizing model performance. Additionally, we employed post-processing rules to correct three types of errors in model predictions: redundant labeling, missing labels, and incorrect labels. The method proposed in this paper achieved scores of 0.9156 under strict criteria and 0.9660 under relaxed criteria on the final test set, securing the first place.

Full Text

Preamble

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

*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)

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. The task requires identifying entity boundaries and category labels for six entity types from Chinese electronic medical records (EMRs). We constructed a hybrid system composed of a semi-supervised noisy label learning model based on adversarial training and a rule-based post-processing module. The core idea of our hybrid system is to mitigate the impact of data noise through model optimization, while using post-processing rules to correct three common error cases in model predictions: redundant labeling, missing labels, and incorrect labels. Our method achieved an F1 score of 0.9156 under strict criteria and 0.9660 under relaxed criteria on the final test set, ranking first in the competition.

1. Introduction

1.1 Evaluation Task

This task continues the series of evaluations conducted by the China Conference on Knowledge Graph and Semantic Computing (CCKS) focused on semantic analysis of Chinese electronic medical records. It extends and expands upon related evaluation tasks from CCKS 2017, 2018, and 2019. For a given collection of plain-text EMR documents, the 2020 Chinese medical record named entity recognition task requires extracting entity mentions and classifying them into six predefined types: 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 samples as the training set, with annotations for the six entity types mentioned above. Additionally, the task organizers provided 1,000 unlabeled samples. The statistics of entity counts in the training set are shown in Table 1 .

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

Entity Type	Disease & Diagnosis	Imaging	Operation	Anatomy	Total
Deduplication	2,198	4,345	1,002	1,297	8,811
Duplication	1,935	1,447	5,529	18,313	-

1.3 Overview of Main Challenges and Solutions

Compared with named entity recognition (NER) in the general domain [1], Chinese clinical NER faces several unique challenges. This paper introduces algorithmic strategies to address two significant challenges in this competition.

The first challenge is **inconsistent entity labeling**. Annotators from different medical departments may have varying interpretations of labeling standards, leading to inconsistent annotation patterns. In this task' s dataset, we observed clear labeling inconsistencies. For example, the string "白细胞数" (white blood cell count) was sometimes annotated in full as "白细胞数"(white blood cell count), while other times only partially as "白细胞" (white blood cell). It was unclear which standard would be used in the test set. We estimated that approximately 13.69% of entities might be affected by such inconsistencies, which seriously impacted model performance. This problem could not be easily resolved through rule-based approaches, nor could we directly correct the inconsistent annotations in the training set.

The second challenge is **insufficient training data leading to unstable model results**. Due to the social sensitivity of medical data, researchers often struggle to obtain sufficient labeled data. This scarcity typically leads to long-tail phenomena and poor model generalization. When training data is limited, model predictions can vary drastically with different parameter initializations. How can we maintain consistent results despite this data scarcity?

To address these challenges, we propose a hybrid system combining a semi-supervised noisy label learning model based on adversarial training with a rule-based post-processing module. The overall process is illustrated in Figure 1 [Figure 1: see original paper]. We introduced a five-fold cross-voting mechanism to handle annotation inconsistencies in the dataset. Model ensemble and semi-supervised training mechanisms helped stabilize results caused by insufficient training data. Additionally, adversarial training proved effective for both

challenges. Our method achieved the highest scores in the CCKS 2020 Chinese NER task: 0.9156 under strict criteria and 0.9660 under relaxed criteria on the official test set.

Figure 1. The overall process of our system.

---

2. Related Work

2.1 Adversarial Training

Adversarial samples [2] are created by adding small perturbations to input data that are imperceptible to humans but can severely disrupt neural network predictions. Adversarial training aims to build more robust and generalized models by continuously defending against such samples [3]. Madry et al. [3] defined adversarial training from an optimization perspective with Equation (1):

$$L_{adv} = \max_{\|r\| \leq \epsilon} \mathcal{L}(\theta, x + r, y)$$

The adversarial training process involves finding small perturbations that maximize training loss, then optimizing model parameters θ to minimize loss under these attacks, iterating until convergence.

2.2 Semi-supervised Learning

Semi-supervised learning leverages a small amount of labeled data as supervised signals combined with large amounts of unlabeled data for augmentation. This approach holds significant practical and research value in domains where labeled data is expensive to obtain, such as medicine. We employed a semi-supervised training mechanism to incorporate the 1,000 unlabeled samples provided by the CCKS organizers into the training process, thereby alleviating the shortage of annotated data.

---

3. Methodology

3.1 Basic Model Structure

Our basic model architecture is shown in Figure 2 [Figure 2: see original paper]. Input sequences obtain embedding representations through a pre-trained model [4]. A BiLSTM layer [5, 6, 7] then encodes these embeddings with contextual information, followed by a Conditional Random Field (CRF) [8, 9] for decoding to produce final annotations. We experimented with five different pre-trained models, which provide richer semantic representations and incorporate extensive world knowledge and commonsense.

Figure 2. Our basic model structure.

3.2 Five-fold Cross-voting

We applied five-fold cross-validation to split the training set into five different subsets, each exhibiting varying degrees of entity labeling inconsistency. Using a fixed model architecture, we trained five separate models on these subsets and integrated their predictions on the test set through hard voting.

3.3 Model Ensemble

To further reduce the impact of parameter randomness on predictions, we ensembled multiple models through voting to mitigate performance fluctuations caused by individual model variations. Figure 3 [Figure 3: see original paper] illustrates the model ensemble process combined with five-fold cross-voting. Two voting sequences were considered: (1) first fusing five models trained on the same fold (red box), then fusing across five folds (totaling 25 models); or (2) first aggregating results across five folds for each model architecture, then fusing five different architectures (green box, also totaling 25 models). Since both sequences yielded similar results, we adopted the green box sequence 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 consists of two stages. In the first stage, we train on all 1,050 labeled samples plus the 1,000 unlabeled samples. In the second stage, pseudo-labeled data obtained from the first stage is added to the training set to produce the final model.

3.5 Adversarial Training

Following the FGM [10] adversarial training mechanism, we directly apply small perturbations to the model' s embedding representations. Treating the embedding representation of an input text sequence $[v_1, v_2, \ldots, v_T]$ as x, the perturbation r_{adv} is computed using Equations (2) and (3):

$$r_{adv} = \epsilon \cdot \frac{g}{\|g\|} = \epsilon \cdot \frac{\nabla_x \mathcal{L}(\theta, x, y)}{\|\nabla_x \mathcal{L}(\theta, x, y)\|}$$

These equations move the input one step in the direction of increasing loss, creating an adversarial attack. The model must then find more robust parameters to defend against such attacks. Applying perturbations to embeddings simulates natural labeling errors in the dataset, encouraging the model to learn more robust parameters. Adversarial training makes the model more tolerant to parameter fluctuations, thereby reducing the impact of data noise.

3.6 Post-processing Rules

If multiple labeling standards exist for an entity, the distribution of these standards in the test set should match that in the training set. Based on this assumption, we can filter out entities in predictions whose distribution deviates from the training set. For these filtered entities, we further categorized errors into three types—redundant labeling, missing labels, and wrong labels—and constructed correction dictionaries for each case.

---

4. Experiments

4.1 Evaluation Metrics

This task uses two F1 criteria. **Strict F1** requires exact match of both entity boundary and type with the gold standard. **Relaxed F1** considers a prediction correct if either the entity type matches or the boundary overlaps with the gold standard. For local evaluation, we used only strict F1 to better reflect model performance.

4.2 Pre-processing

We applied the following pre-processing steps to all data.

4.2.1 Sentence Segmentation Since BERT models have a maximum input length of 512 tokens, we segmented EMR texts while preserving semantic coherence to ensure each input segment remained below this limit.

4.2.2 Text Normalization This step unified text and symbols, converted English cases, and removed invisible characters from input medical records.

4.3 Implementation Details

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

Table 2. Implementation details of our five basic models.

Model	Learning Rate	Epochs	Dropout [11]	Optimizer
BERT-base+BiLSTM+CRF	2e-5	20	0.2	AdamW [12]
BERT-wwm-ext+BiLSTM+CRF[13]	2e-5	20	0.2	AdamW
RoBERTa-wwm-ext+BiLSTM+CRF[13]	2e-5	20	0.2	AdamW
RoBERTa-wwm-ext-large+BiLSTM+CRF[13]	1e-5	10	0.3	AdamW
RoBERTa-wwm-ext-large+CRF[13]	1e-5	10	0.3	AdamW

4.4 Results

We divided the 1,050 training samples into five folds, with 840 samples for training and 210 for development in each fold. Table 3 shows results on the local development sets, reported as the average F1 across all five folds. In all tables, we abbreviate Semi-supervised Training as **ST**, Adversarial Training as **AT**, and Post-processing Rules as **PR**.

Table 3. Results on the local development set.

Model	F1 Score
BERT-base+BiLSTM+CRF	0.8390
BERT-wwm-ext+BiLSTM+CRF	0.8473
RoBERTa-wwm-ext+BiLSTM+CRF	0.8524
RoBERTa-wwm-ext-large+BiLSTM+CRF	0.8563
RoBERTa-wwm-ext-large+CRF	0.8514
BERT-base+BiLSTM+CRF+ST	0.8472
BERT-base+BiLSTM+CRF+AT	0.8491
Model Ensemble	0.8641
Model Ensemble+ST	0.8705
Model Ensemble+AT	0.8723
Model Ensemble+ST+AT+PR	**0.8768**

Table 3 demonstrates that model ensemble, semi-supervised training, and adversarial training each brought significant improvements over the basic model, with the best performance achieved by combining all three mechanisms.

Official test set results are shown in Table 4 . We refer to BERT-base+BiLSTM+CRF as the **Single Model**. Notably, the Single Model scored 0.0384 higher on the official test set than locally, suggesting that annotation inconsistencies in the official test set may be substantially less than in the training set. In our final model, we applied five-fold cross-voting to each fused model to mitigate the impact of limited training data.

Interestingly, while post-processing rules showed modest overall improvement on the local development set, they yielded significant gains of 0.0242 and 0.0414 on the two entity types with fewer samples (imaging examination and laboratory examination).

Table 4. Results on the official test set.

Model	Disease & Diagnosis	Imaging	Operation	Anatomy	Total
Single Model	0.9156	0.9660	-	-	-
+ST+AT	-	-	-	-	0.9156
Our Method	-	-	-	-	0.9660

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

Criteria	Disease & Diagnosis	Imaging	Operation	Anatomy	Total
Relaxed	-	-	-	-	0.9660
Strict	-	-	-	-	0.9156

5. Conclusion and Future Work

To address the two core challenges in this task' s dataset—inconsistent entity annotations and insufficient labeled data—we introduced semi-supervised data augmentation and adversarial training methods, achieving state-of-the-art performance. Future research will focus on precisely quantifying annotation inconsistency in NER datasets and developing models that can better overcome such noise.

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. Zhang (baoli.zhang@nlpr.ia.ac.cn) performed the experiments. 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 work. All authors reviewed the manuscript.

Data Availability Statement

The datasets generated and/or analyzed during this study are not publicly available due to their production by medical expert consultants of Yidu Cloud based on their professional experience. Public release requires consent from all expert consultants, but the data are available from the corresponding authors upon 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 the School of Artificial Intelligence, University of Chinese Academy of Sciences. His research interests include information extraction, knowledge graphs, 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 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 at the Institute of Automation, Chinese Academy of Sciences. His research interests include information extraction, knowledge graphs, and natural language processing. ORCID: 0000-0002-5815-7292

Yubo Chen is an Associate Professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences. His research interests include information extraction, knowledge graphs, 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 at Beijing University of Chemical Technology. Her research interests include knowledge graphs and cultural heritage digital protection. She has published over 40 papers. ORCID: 0000-0002-4232-7883

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

Jun Zhao is a Professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, and the School of Artificial Intelligence, University of Chinese Academy of Sciences. Prof. Zhao has published over 90 peer-reviewed papers in prestigious conferences and journals, including ACL and AAAI. He received the 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. He is now a senior technical expert at UNISOUND AI Technology Co., Ltd.

Note: Figure translations are in progress. See original paper for figures.