ROSE-NER: Robust Semi-supervised Named Entity Recognition on Insufficient Labeled Data

Haiyan Chen
School of Computer Science and Engineering
Southeast University
Nanjing, China
hy_chan@seu.edu.cn

Shuwei Yuan
College of Software Engineering
Southeast University
Nanjing, China
shuweiyuan@seu.edu.cn

Xiang Zhang*
School of Computer Science and Engineering
Southeast University
Nanjing, China
x.zhang@seu.edu.cn

ABSTRACT

With the wide application of deep learning technology in recent years, the effect of named entity recognition tasks has been significantly improved when there is sufficient labeled data. However, obtaining a large amount of labeled data is quite difficult in many domains, for example in the domain of medicine. Noisy data can also negatively affect the robustness of NER models. In this paper, we propose a robust semi-supervised NER approach ROSE-NER to tackle these challenges. A two-step semi-supervised model is introduced to expand a handful of labeled data with a large amount of predicted pseudo-labeled data. The combination of these data will be exploited in the model training. In addition, we introduce an adversarial training method to improve the robustness of the model by eliminating the impact of noise samples. Experiments on medical datasets show that our approach reduces dependency on massive labeled data, and it outperforms other State-of-the-Art approaches.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**.

KEYWORDS

Named Entity Recognition; Semi-Supervision; Adversarial Learning

ACM Reference Format:
Haiyan Chen, Shuwei Yuan, and Xiang Zhang. 2021. ROSE-NER: Robust Semi-supervised Named Entity Recognition on Insufficient Labeled Data. In *The 10th International Joint Conference on Knowledge Graphs (IJCKG'21), December 6–8, 2021, Virtual Event, Thailand*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3502223.3502228

1 INTRODUCTION

Named entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person, organizations, locations, drug, disease, gene, etc. in unstructured text. NER systems are usually a key component and play as the first step in various Natural Language Processing tasks like information retrieval, relation extraction and question answering.

Early NER systems were based on handcrafted rules, lexicons, features and ontologies. These systems were followed by NER systems based on feature-engineering and machine learning methods, such as Hidden Markov Model (HMM) [21], Conditional Random Field (CRF) [13], Maximum Entropy model [12], and other statistical models. Over the past few years, deep learning has been applied to the task of named entity recognition, and it has successively advanced the state-of-the-art performance with minimal feature engineering. Researchers have proposed models like Bi-LSTM-CRF [4], IDCNN-CRF [19], and many other neural architectures based on some form of recurrent neural networks (RNN) over characters, sub-words and/or word embeddings.

Although NER systems with deep models gained momentum these years, they are still facing challenges from insufficient and noisy data. Deep models usually have a strong requirement on training data. The performance of these models heavily rely on the quantity and quality of labeled data. But in many domains, the annotation budget for labeling is far less than the total number of unlabeled data. For example, in the domain of medicine, there are massive unlabeled medical records, clinical reports, and biomedical literature. Annotating these data is quite burdensome, because it requires a solid medical knowledge background, which can only be fulfilled by medical specialists. This situation of insufficient labeled data prevails in many real-world applications. Besides, noisy data can significantly affect the robustness of NER models. To make deep NER models more broadly useful, it is crucial to reduce its dependency on labeled data and meanwhile to subtract the negative impact of noisy data.

Based on above investigations, we propose a novel ROSE-NER (Robust Semi-supervised Named Entity Recognition) model, which aims at improving performance on insufficient and noisy data. Our model expands a small amount of training set with massive pseudo-labeled data, then the model is iteratively trained with expanded data. Adversarial training is also adopted to add a perturbation to the word embedding in the training process, so as to improve the robustness of the model and reduce the influence of noisy data.Our code and the supplementary materials are available at https://github.com/emask6/RoseNER. Overall, the contributions of this paper can be summarized as following:

*Corresponding Author

IJCKG'21, December 6–8, 2021, Virtual Event, Thailand

ACM ISBN 978-1-4503-9565-6/21/12…$15.00
https://doi.org/10.1145/3502223.3502228

1. We propose ROSE-NER, which is a span-based tagging model. Compared with existing NER models, our method achieves SOTA performance on a series of medical datasets.
2. We propose a two-step semi-supervision in ROSE-NER. In the first step, we enrich a handful of labeled data with massive model-predicted pseudo-labeled data. In the second step, the expanded data are exploited in model training to identify named entities in the text. We propose a revised loss function to control the weight of pseudo-labels on the golden-labels in model training.
3. We adopt an adversarial training method to alleviate the negative impact of noisy data. The robustness of the model is enhanced by a perturbation in the embedding layer through adversarial training.

2 RELATED WORK

Named Entity Recognition is a basic task of natural language processing and plays an important role in machine translation, sentiment analysis, intelligent question answering and other applications. Over the years, some deep learning models have been applied to NER with superior performance, but most of them rely on a large amount of labeled data. Obtaining labeled data often requires a price of manpower, and it is also time-consuming and prone to human errors. At the same time, it is difficult to deal with the problem of noisy data. In order to address these problems, researchers have attempted various unsupervised, semi-supervised or distant-supervised approaches in recent years.

Wang *et al.* [20] used an indirect supervision method to combine probabilistic logic and deep learning to perform joint reasoning on the entire data set, thereby getting rid of the dependence on labeled samples. Chen *et al.* [1] proposed a local additivity-based data augmentation (LADA) method for semi-supervised NER, which generates virtually-labeled-samples by interpolating sequences that are close to each other. Although manually labeled data is difficult to obtain, unlabeled domain data is easier to obtain. Our method uses a large amount of unlabeled domain data through a pseudo-label mechanism to significantly improve the results.

As a recent trend, distant supervision aims to reduce expensive manpower by exploring mention information in dictionaries or knowledge bases. Yang *et al.* [22] used dictionary-based matching to automatically generate training samples and solved the problem of incomplete annotations in NER. Sang *et al.* [18] used the method of distant supervision, using external resources such as knowledge bases to automatically annotate specific types of entities. However, the effectiveness of these distant-supervised methods usually relies on the availability and quality of domain dictionaries or knowledge bases, which might be unavailable in many domains. Besides, the problem of out-of-vocabulary (OOV) words and word sense disambiguation in mention linking also limit the performance of distant supervision in practical NER applications.

For the problem of noisy data in NER, Mishra *et al.* [14] proposed a semi-supervised NER model based on linear chain conditional random field (CRF), using the BIEOU coding scheme, and using random feature loss for the up-sampling of the training data. It can be applied to noisy, user-generated inaccurate data. Some researches based on distant supervision of NER directly regard distant tags as the basic facts of model training and rely on simple techniques. Ni *et al.* [16] used heuristic rules to filter out samples with low matching quality. This filtering strategy improved the accuracy at the cost of reducing the recall. Liang *et al.* [8] dealt with noise by using early stopping to prevent the model from over-fitting the incomplete annotation labels. Some methods [17] require additional manual annotation on a training set to build noise classification models.

Different with existing models, we propose a two-step semi-supervised training. First, a base model is trained by a handful of labeled training data, then this model is used to predict large-scale unlabeled corpus to obtain pseudo-labels. Our method can reduce the dependency on large-scale manual annotations without introducing external dictionaries or knowledge bases. In addition, to improve the anti-interference ability of our model, we adopt an adversarial training mechanism to fight against noisy labels. By perturbing the word embedding layer, the model can be trained with more generalized representations of words, which effectively enhance the model robustness.

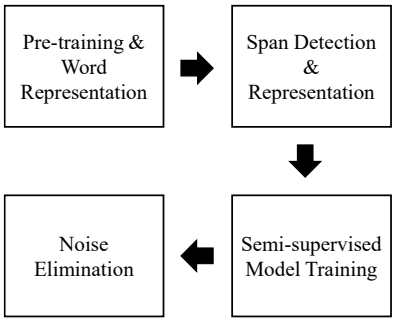


Figure 1: The Workflow of ROSE-NER

3 APPROACH

The workflow of ROSE-NER is shown in Figure 1. Our model is built in four main steps: To characterize the text features, we first pre-train the input text and generate a generalized word representations with perturbations. To identify and characterize text spans, the second step is to generate span-level representations. The third step is a two-step semi-supervision, in which pseudo labels are assigned to unlabeled data, and an enriched dataset is exploited for model training. In the final step, ROSE-NER reduces the impact of noisy data on the NER task by an adversarial training.

The training process of the ROSE-NER model is presented in Figure 2. First, the original BERT is fine-tuned using unlabeled domain data to obtain a BERT with specific domain knowledge, so that a semantic representation that is more suitable for downstream tasks can be obtained. Next, a ROSE-NER-base model is built with a handful of golden-labeled data with the help of features provided by fine-tuned BERT. In the next step, the base model acts as a predictor to assign pseudo labels to unlabeled data to generate an enriched training set. Since the pseudo-labeled data may contains a certain degree of noise comparing to high quality golden-labeled data, we propose a dynamic loss function during the process of

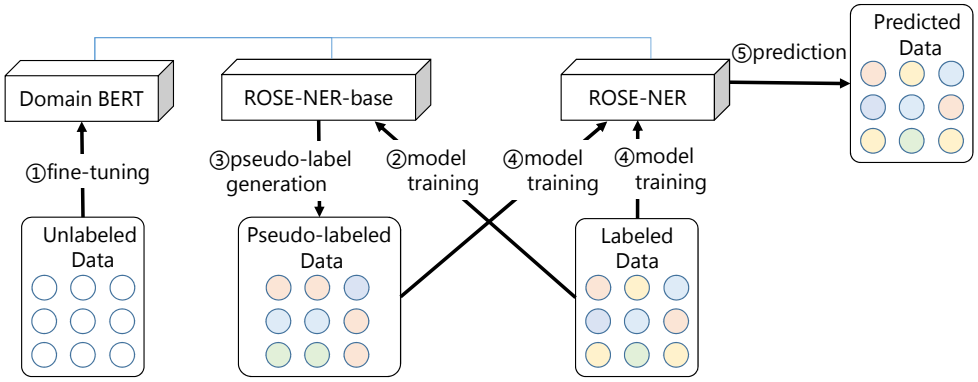


Figure 2: ROSE-NER model training steps. First, BERT is fine-tuned using unlabeled domain data to obtain the domain BERT, then combine labeled data with domain BERT training to obtain ROSE-NER base, and then use the ROSE-NER base model to predict pseudo-labeled data to obtain pseudo-labeled data. Finally, The pseudo-labeled data and labeled data are used as a new training set to train the final ROSE-NER model.

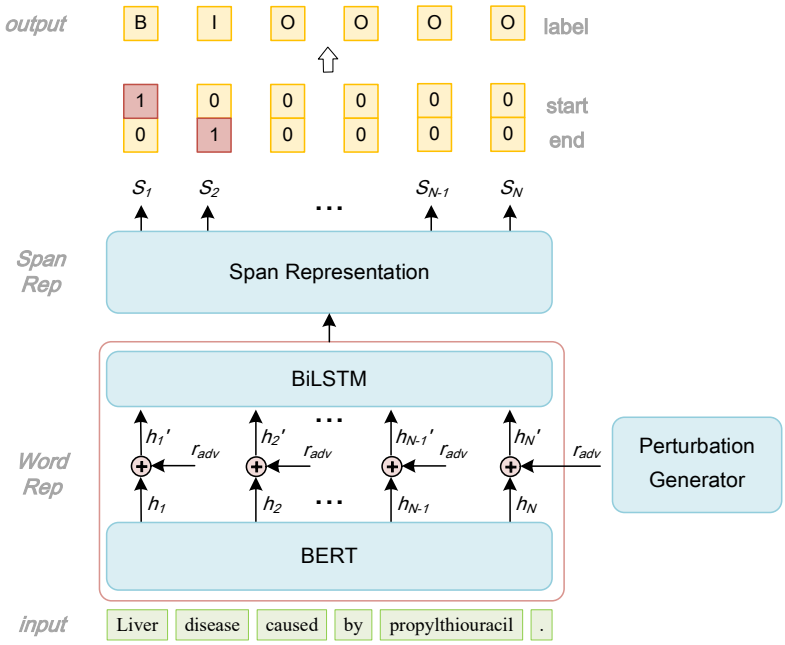


Figure 3: The structure of the ROSE-NER model. The input of the model is "Liver disease caused by propylthiouracil." h_i represents the original embedding of the word, $h_i^{'}$ represents the word representation after adding the perturbation r_{adv} to the original embedding. S_i is span representation. The output of the model is ['B', 'I', 'O', 'O', 'O', 'O'], and the entity "Liver disease" is recognized.

co-training on both golden and pseudo-labeled data, in order to adjust the impact of pseudo-labeled data on the performance of our model.

3.1 Word and Span Representation

The pre-training model BERT has an excellent performance in multiple tasks of NLP, so we adopt BERT as a module for obtaining word representations. Specifically, given an input sentence $x = \{x_1, x_2, ..., x_N\}$, where x_i represents the $i-th$ word in the sentence. x is sent to the BERT module, and BERT trains the embedding representation of words through the two pre-training tasks of Next Sentence Prediction (NSP in short) and Masked Language Model (MLM in short). Words in x are represented as tensors $H = \{h_1, h_2, ..., h_N\}$. After that, the parameters of the BERT is fine-tuned by exploiting a domain dataset we use in the experiment as a corpus.

On the span representation, we are inspired by the models in machine reading comprehension and question answering tasks, and apply the machine reading comprehension framework to NER tasks. In the specific method of extracting entities, a Pointer-tagging hybrid structure is adopted, which was first introduced in [2]. As shown in the output layer of Figure 3, in the two-layer labeling network, one layer marks the beginning position of the mention, and the other layer marks the end position of the mention. If there are a total of N types of entities, 2*N such labeling sequences are needed, and every two labeling sequences are a group, and there are N groups in total. The type of the mention is determined according to the label group where the "1" label is located. By using Softmax to process the logits of the model, the category with the maximum probability is taken as the final result, so as to ensure that each mention corresponds to a category. The label group in Figure 3 corresponds to "tumor".

In the design of the loss function of the model, we use the Focal Loss proposed by Tsung-Yi Lin [9]. The loss function was first proposed in the field of object detection. It can effectively solve the imbalance in the number of categories in the classification problem and the uneven distribution of samples. These two phenomena often appear in named entity recognition tasks, so the loss function is migrated from the CV domain to the NLP domain. Since the number of samples that are easy to classify is larger than the number of samples that are difficult to classify, simple samples will dominate the overall loss when using the cross-entropy loss function:

$$Focal(p_t) = -(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

where p_t is the prediction probability, γ ($\gamma \geq 0$) is the focus parameter so that the model can focus more on difficult-to-classify samples. It is found through experiments that it is also necessary to add a balance factor α on the basis of Equation(1), which helps to improve the accuracy. The modified loss function is shown as:

$$Focal(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t) \tag{2}$$

The two loss functions for the start position and end position of the mention are as following. The final loss function is the sum of the two losses:

$$L_{start} = Focal\left(P_{start}, Y_{start}\right) \tag{3}$$

$$L_{end} = Focal\left(P_{start}, Y_{start}\right) \tag{4}$$

$$L = \left(L_{start} + L_{end}\right) \tag{5}$$

3.2 Semi-supervised Learning

Semi-supervised learning uses a small amount of labeled data as a supervised signal and combines a large amount of unlabeled data to achieve data enhancement. This method has high application value and research value in fields where the cost of labeling data is relatively high, such as the field of medicine. We use a semi-supervised training mechanism to introduce unlabeled data into the training process, which reduces the cognitive uncertainty of the model to a certain extent.

The semi-supervised training process is divided into three stages, as presented in Figure 2. In the first stage, the gold-labeled data is used as the training set to train the benchmark model. In the second stage, the benchmark model is used to predict the unlabeled data to obtain pseudo-labeled data. In the third stage, the pseudo-labeled data is merged with the training set of the first stage to obtain a new training set.

Due to the unbalanced distribution of mention categories in the data set, the confidence of using ROSE-NER-base to predict entities with fewer categories is lower. If the pseudo-label data is filtered by setting a threshold, the mentions of fewer categories will be filtered out, which aggravates the imbalance of the category distribution and violates our original intention of generating pseudo-labels. Regarding the setting of the loss function, Since the confidence of the pseudo-labeled data set is not high, it will introduce noise to the model. Therefore, the loss of the gold-labeled data and the pseudo-labeled data cannot be directly added, but the loss function is adjusted to the combination of golden-labeled data and the pseudo-labeled data. From the perspective of the model, we introduce a learnable parameter β during the training process to adjust the impact of the pseudo-labeled data set on loss:

$$L_{semi} = (1-\beta)L(X_{label}) + \beta L(X_{pseudo}) \tag{6}$$

Experiments prove that this paper has a positive effect on the result by adjusting the weight of the pseudo-label loss function.

3.3 Adversarial Training

Referring to the adversarial training mechanism of FGM [15], the model directly imposes a small perturbation to the word embedding and assumes that the input text sequence $x = \{x_1, x_2, ..., x_N\}$ is denoted as x , then the small disturbance r_{adv} is defined as:

$$r_{adv} = \epsilon \cdot g/\|g\|_2 \tag{7}$$

$$g = \nabla_x L(\theta, x, y) \tag{8}$$

where g represents the gradient of the loss function L with respect to the input x, r_{adv} walks along the gradient direction, and such a perturbation is added to the input x, that is, $x + r_{adv}$, so the input moves in the direction where the loss rises fastest, thus forming an attack. In the case of fixed perturbations, the model needs to find more robust parameters in the optimization process to deal with the perturbation on the sample.

The small perturbation to the embedding simulates the natural error of the data set in the annotation to a certain extent. The model can be guided to find more robust parameters in the training process to weaken the influence of random uncertainty. Then the embedding representation of the model will be optimized together with the model, and the adversarial training will make the model more tolerant of sample noise, thereby reducing the impact of cognitive uncertainty.

Because token embedding is a vector representation of the semantic space, we can intuitively understand the perturbation generated in adversarial training as allowing the model to recognize not only the words in the corpus, but also their synonyms. From the results, it has a positive impact on the model and can improve the robustness of the model.

4 EXPERIMENTS

4.1 Data Sets

We will compare our model with existing methods on two medical benchmark datasets. NCBI-Disease [3] is a named entity recognition dataset in the field of diseases, which contains only one type of mention. While BC5CDR [7] is a dataset formed by selecting 1,500 abstracts from Pubmed, which contains two types of entities, chemical, and disease. The statistics of the data sets are described in Table 1.

Table 1: Statistics of data sets NCBI and BC5CDR

	NCBI-Disease	BC5CDR
Train set	5,424	4,560
Dev set	923	4,581
Test set	940	4,797
entities	7,025	28,545

4.2 Experimental Setup

In the training process, in order to obtain the domain pre-training model, we first use the training set to finetune the pre-training model, thereby obtaining word has embedded domain knowledge. We use the pre-trained BERT-base model as the backbone model and using the last layer output as token embedding, and then use the training set to train ROSE-NER. Finally, test on the validation set to obtain the best parameters.

For the NCBI-Disease and BC5CDR data sets, the maximum sequence length(max_len) is set to 120. The dropout rate and batch size are 0.1 and 12 respectively; the adversarial training method is FGM [15]; we use AdamW [10] as the optimization. In terms of the experimental environment, all experiments were run on one NVIDIA GeForce GTX 2070 GPU.

4.3 Comparison with State-of-the-Art Methods

In order to demonstrate the effectiveness of this method, we will compare it with some existing methods, including traditional machine learning methods and some representative deep learning models.

Among the machine learning methods, Dnorm [5] is a pipeline model applied to medical NER(Named Entity Recognition) and NEN(Named Entity Normalization), which uses TF-IDF features to learn the bilinear mapping matrix for standardized tasks. TaggerOne [6] is a joint modeling medical NER model based on Semi-Markov. Transition-based Model [11] designed a state transition function for the NER task.

In order to reduce manual feature engineering, researchers began to apply deep learning to NER tasks. IDCNN [19] proposed an improved CNN model for the NER task. MCNN [25] is composed of multi-label CNN modules, which can achieve better results on NER tasks. CollaboNet [23] uses multi-source data sets to train multi-task models and achieves better results on benchmark data sets. MTL-MERN [24] contains two task frameworks, NER and NEN, and uses feedback strategies to improve the results of the two tasks.

Table 2 shows the evaluation results of the two datasets. The first three models represent traditional machine learning methods, and the rest are deep learning methods. Among them, the combined model TaggerOne [6] and Transition-based Model [11] are better than the pipeline model Dnorm. When the deep learning

Table 2: F1 scores of each model on the NCBI-Disease and BC5CDR datasets

	NCBI-Disease	BC5CDR
Dnorm [10]	0.798	-
TaggerOne [5]	0.829	0.826
Transition-based Model [6]	0.8205	0.8302
IDCNN [19]	0.7983	0.8011
MCNN [11]	0.8517	0.8783
CollaboNet [25]	0.8636	0.8818
MTL-MERN [23]	0.8743	0.8763
ROSE-NER	**0.8886**	**0.9095**

Table 3: ROSE-NER frame ablation experiment results

	NCBI-Diease			BC5CDR		
Model	Precision	Recall	F1	Precision	Recall	F1
Rose-NER	**0.8956**	**0.8817**	**0.8886**	**0.9187**	**0.9005**	**0.9095**
w/o adversarial training	0.8914	0.8757	0.8835	0.9116	0.891	0.9012
w/o focal loss	0.8908	0.8807	0.8857	0.9058	0.8917	0.8987
w/o dynamic loss parameters	0.8851	0.8773	0.8812	0.9034	0.8897	0.8965

"w/o adversarial training" means not using adversarial training, "w/o focal loss" and "w/o dynamic loss parameters" are the same.

Table 4: Case study of NER results on NCBI

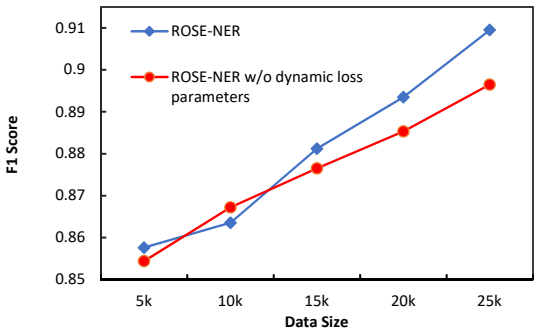
Method	NER results
BERT	Skin fragility in most cases is due to mutations in the gene encoding type XVII collagen (COL17A1) .
In-Domain BERT	Skin fragility disease in most cases is due to mutations in the gene encoding type XVII collagen (COL17A1) .
BERT	Complement component C6 deficiency (C6D) was diagnosed in a 16 - year - old African - American male with meningococcal meningitis disease .
In-Domain BERT	Complement component C6 deficiency disease (C6D) was diagnosed in a 16 - year - old African - American male with meningococcal meningitis disease .
BERT	In SCA3 , gaze - evoked nystagmus was often present as was saccade hypometria and smooth pursuit gain was markedly decreased .
In-Domain BERT	In SCA3 , gaze - evoked nystagmus disease was often present as was saccade hypometria and smooth pursuit gain was markedly decreased .

method is applied to the pipeline model, IDCNN [19] achieves better results than the pipeline model using traditional machine learning methods, such as Dnorm. Compared with MCNN, CollaboNET uses multi-source datasets as input and uses multi-task learning to improve the effect of NER tasks. MTL-MERN [24] makes full use of multi-task learning and deep semantic representation, and the effect is better than the above methods.

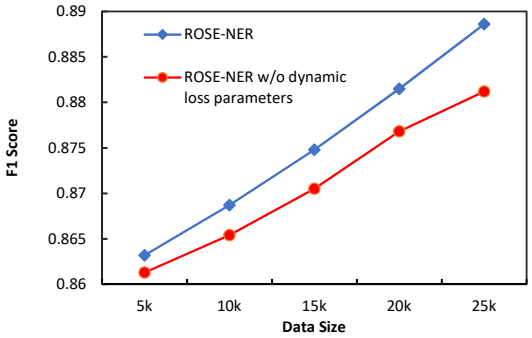
Compared with the baselines, ROSE-NER achieved the best results on both data sets, which is 1.43% higher than MTL-MERN [24] on the NCBI-Disease dataset and 3.32% better on the BC5CDR dataset. Because the domain pre-training model obtained by our framework through finetune can better represent the medical text. At the same time, the semi-supervised method makes full use of unlabeled data, and the robustness of the model is improved through adversarial training. In addition, the Span method used in the annotation framework can better identify the mention boundary.

4.4 Ablation Study of ROSE-NER Model

In order to deeply study the influence of each component in the ROSE-NER framework on the results, we conducted ablation experiments. The experimental results are shown in Table 3. From the results, it can be found that various components in our framework have a positive effect on the results. Where adversarial training perturbs the word embeddings generated by the pre-training model, which can make the model more robust and improve the results. Since the non-mention part of the NER task is much larger than the mention part, the same is true on the NCBI-Disease and BC5CDR datasets in our experiments. From the experimental results, the focal loss has a positive effect on solving the problem of category imbalance. The dynamic loss parameters designed for pseudo-label data have the greatest impact on the results, so it can be proved that this parameter can effectively suppress the noise caused by the pseudo-label data set to the model.



(a) Results on BC5CDR



(b) Results on NCBI-Disease

Figure 4: Experiments were performed on different numbers of unlabeled field data using ROSE-NER and ROSE-NER without dynamic loss parameter. (a) is the experiments on the BC5CDR data set, (b) is the experiments on the NCBI-Disease data set.

We also studied the influence of original BERT and In-Domain BERT on named entity recognition. As shown in Table 4, it shows the mention recognition results of three samples selected from the NCBI data set using the original BERT and the domain BERT respectively. In contrast to original BERT, In-Domain BERT is able to detect new mentions. As can be seen from the previous section, it is an important step in ROSE-NER to finetune the original BERT to obtain the domain BERT through the domain corpus. It can be seen from the results that the recall rate of using the domain BERT is better than using the original BERT. Although these entities do not appear in the training set, ROSE-NER can recognize these entities after the domain corpus fine-tuning.

From the method in Section 3 and Figure 2, we can see that our ROSE-NER framework introduces pseudo-label data. Pseudo-labeled data is obtained through ROSE-NER base's prediction of unlabeled field data. In order to explore the influence of different amounts of unlabeled field data on the experimental results, we conducted comparative experiments on different data volumes. The experimental results are shown in Figure. 4. From the results, we can see that regardless of whether it is in the BC5CDR dataset or the NCBI-Disease dataset, the F1-score is positively correlated with the scale of the pseudo-label data, which proves that the introduction of the pseudo-label data set can significantly improve the experimental results. Since the pseudo-label data is obtained through the low-confidence ROSE-NER base model test, if you do not do any processing directly to mix the pseudo-label and artificially labeled data for training, it will introduce noise to the model, which may reduce the precision of the model. Therefore, ROSE-NER proposed a dynamic loss parameter to reduce the impact of low-confidence tags on the results. It can be seen from Figure 4 that the F1-score in the BC5CDR and NCBI-Disease data sets has decreased without using the dynamic loss parameter. It proves that the dynamic loss parameter we proposed is effective in reducing the noise in the pseudo-labeled data.

5 CONCLUSIONS

For the task of named entity recognition, we propose an efficient and robust semi-supervised learning approach named ROSE-NER.. On one hand, we use adversarial training to solve the noise problem in the data. On the other hand, we use a two-step semi-supervision. A base model is first trained with a handful of labeled data to automatically annotate massive unlabeled data, then the base model expands the pseudo-labeled data to the training set. We introduce dynamic loss parameters during the training process to reduce the effect of data noise brought by pseudo labels. The results show that ROSE-NER achieved a state-of-the-art performance on both NCBI-Disease and BC5CDR comparing to existing methods.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2019YFB2101802).

REFERENCES

[1] Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020. Local additivity based data augmentation for semi-supervised NER. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 1241–1251.

[2] Kalpit Dixit and Yaser Al-Onaizan. 2019. Span-level model for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5308–5314.

[3] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics* 47 (2014), 1–10.

[4] Bin Ji, Rui Liu, Shasha Li, JinTao Tang, Jie Yu, Qian Li, and WeiSang Xu. 2018. A BiLSTM-CRF Method to Chinese Electronic Medical Record Named Entity Recognition. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*. 1–6.

[5] Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* 29, 22 (2013), 2909–2917.

[6] Robert Leaman and Zhiyong Lu. 2016. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics* 32, 18 (2016), 2839–2846.

[7] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *The Journal of Biological Databases and Curation. 2016* (2016).

[8] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1054–1064.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[10] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

[11] Yinxia Lou, Yue Zhang, Tao Qian, Fei Li, Shufeng Xiong, and Donghong Ji. 2017. A transition-based joint model for disease named entity recognition and normalization. *Bioinformatics* 33, 15 (2017), 2363–2371.

[12] Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. 2000. Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of 17th International Conference on Machine Learning*, Vol. 17. 591–598.

[13] Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of 7th Conference on Computational Natural Language Learning*. 188–191.

[14] Shubhanshu Mishra and Jana Diesner. 2016. Semi-supervised Named Entity Recognition in noisy-text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. 203–212.

[15] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. In *Proceedings of the 5th International Conference on Learning Representations*.

[16] Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 1470–1480.

[17] Yasumasa Onoe and Greg Durrett. 2019. Learning to denoise distantly-labeled data for entity typing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2407–2417.

[18] Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2054–2064.

[19] Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2670–2680.

[20] Hai Wang and Hoifung Poon. 2018. Deep probabilistic logic: A unifying framework for indirect supervision. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1891–1902.

[21] Qingyao Wu, Michael K Ng, and Yunming Ye. 2013. Markov-miml: A markov chain-based multi-instance multi-label learning algorithm. *Knowledge and information systems* 37, 1 (2013), 83–104.

[22] Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised NER with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*. 2159–2169.

[23] Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. 2019. Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC bioinformatics* 20, 10 (2019), 55–65.

[24] Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. 2019. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 817–824.

[25] Zhehuan Zhao, Zhihao Yang, Ling Luo, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. 2017. Disease named entity recognition from biomedical literature using a novel convolutional neural network. *BMC medical genomics* 10, 5 (2017), 75–83.