Seq-GAN-BERT：Sequence Generative Adversarial Learning for Low-resource Name Entity Recognition

Anonymous ACL submission

Abstract

Named entity recognition (NER), as an important basic task of natural language processing, has been widely studied. In the case of relatively sufficient labeled data, traditional NER methods have achieved remarkable results. However, due to the lack of labeled data in many fields and the difficulty of manual annotation, the task of low-resource NER has become a research hotspot. To effectively improve the recognition accuracy of low-resource NER, this paper proposes the semi-supervised learning model Seq-GAN-BERT，which integrates the adversarial generative network based on the pre-trained language model BERT, and uses the domain unlabeled corpus to train the adversarial generative network to learn the important general semantic information of the data. The proposed Seq-GAN-BERT method can further optimize BERT-based supervised training and improve the ability of entity recognition. The experimental results show that our model greatly reduces the dependence on labeled samples and effectively improves the performance of low-resource NER task.

1 Introduction

Since the pre-trained language model BERT(Jacob Devlin et al., 2018) was proposed, it has been widely used in natural language processing tasks. Tremendous improvements have been achieved by fine-tuning downstream tasks such as NER, text classification, machine reading comprehension, etc. However, fine-tuning for downstream tasks often requires high-quality labeled samples, and the performance drops dramatically when there are few labeled samples. NER, as one of the most basic natural language processing tasks, is different from text classification and reading comprehension tasks, and has the characteristics of large data annotation workload and data enhancement difficulties. The types of entities in this task are often diverse, resulting in complex annotation. Each entity in the annotation must specify its beginning, middle, end and entity type. Data augmentation is among the research methods of low-resource NER, which has been continuously studied as a mainstream method (Jason Wei and Kai Zou, 2019; Xiang Zhang et al., 2019; Xiang Dai and Heike Adel, 2020). However, since entities are fine-grained information in the sentences, it is difficult to effectively enhance the data. Mixing the enhanced data with the original label data for training will inevitably introduce harmful noise. Motivated by the work of (Christian Szegedy et al., 2014; Joey Tianyi Zhou et al., 2019; Ting Chen et al., 2019), this paper will use an adversarial learning approach to improve the small sample learning ability of BERT for the low-resource NER task.

At present, there are three methods related to adversarial learning: adversarial training, GAN-based transfer learning, and GAN-based semi-supervised learning. Adversarial training (Christian Szegedy et al., 2014) is an important method to improve the robustness of the model. The main idea is to add a small perturbation to the sample in the direction of the negative gradient of the model training, and such perturbation is likely to cause the model to misclassify. The model gradually adapts to this perturbation during training to enhance the robustness of the model. This method can alleviate the model's dependence on annotated data, but the improvement is limited. GAN-based transfer learning can solve the low-resource problem to a certain extent. Through adversarial learning, (Joey Tianyi Zhou et al., 2019) enables the constructed shared network layer to effectively extract the data features shared by the source domain and the target domain, which can transfer the feature extraction capability of the network in the source domain to the target domain. Although these works can improve the ability of target domain data features extraction to a certain extent, thereby alleviating the problem of insufficient labeled data. When the source domain data is relatively small, the improvement effect of GAN-based transfer learning methods is limited. We can know the above-mentioned adversarial training and GAN-based transfer learning methods do not effectively utilize unlabeled corpora in the field. The GAN-based semi-supervised learning method uses unlabeled data to learn general information representation while performing supervised learning. Which effectively improve the small sample learning ability of the model. SS-GANs (Tim Salimans et al., 2016) represents this idea. The discriminator of SS-GANs is not like the discriminator of traditional GANs, it needs to identify the generated samples and the k categories to which the real examples belong. Through the adversarial training of the internal generator and the discriminator, the model produces the better representations of data, improving classification performance. At present, SS-GANs can achieve relatively good results in image recognition and text intent recognition tasks using only unlabeled corpora and a few dozens of labeled samples (Jason Weston et al., 2008; Thomas N.Kipf and Max Welling et al.. 2017; Zhilin Yang et al., 2016). In addition, SS-GANs have also been applied to recently proposed methods such as Kernel-based GAN (Danilo Croce et al., 2019), etc. Although SS-GANs have achieved relative success, they are limited to image and text classification tasks, and there is no relevant research progress on low-resource NER task.

Therefore, to address the low-resource NER task, we propose the Seq-GAN-BERT[1] semi-supervised learning model. First, the input labeled data and real unlabeled data are encoded by BERT to generate a semantic vector representation, which is fed into the discriminator together with the fake sample representation generated by the generator of the SS-GANs. The discriminator has two training objectives: 1). Identify whether unlabeled samples are real or generated; 2). Perform entity k classification on labeled data. The former is unsupervised training and the latter is supervised training. The general semantic representation ability learned by unsupervised training in real unlabeled samples has a positive impact on the supervised training of the model. By alternately training the generator and the discriminator, the BERT parameters are further optimized to improve the entity recognition effect.

In summary, the main contributions of this paper are as follows:

(1) Our proposed Seq-GAN-BERT model, using adversarial learning, provides a new solution to the low-resource NER.

(2) Seq-GAN-BERT can also be used to solve the low-resource problem of sequence tagging tasks such as word segmentation and part-of-speech tagging.

(3) Our experiment shows that Seq-GAN-BERT model greatly reduces the dependence on manually annotated samples in the sequence labeling task, can effectively utilize unlabeled corpus, and has excellent small-sample learning ability.

2 Related work

At present, many methods have been studied and proposed in scenarios of low-resource task. One of the most direct methods is data augmentation, which augments training samples by removing words, replacing words, and back-translating. (Rico Sennrich et al., 2016; Li Dong et al., 2017) performed data augmentation through back-translation, which has been improved on text classification tasks, but this method is not suitable for NER. To effectively expand the training data for NER, (Bosheng Ding et al., 2020) proposed a generative model to generate annotated samples.

In addition to data augmentation, some semi-supervised learning methods are improved. (Samuli Laine et al., 2017) proposed the semi-supervised learning method Π-Model, which uses dropout and other ways to perform data augmentation on all data, including unlabeled data, by encouraging the original and augmented data to have a consistent output probability distribution optimize the model. (Rui Wang and Ricardo Henao, 2021) used a consistent training semi-supervised method to improve the model, which augmented unlabeled data with back-translation to encourage model predictions of

[1] We release the code at https://github.com/camel2000/seq_GAN_BERT

original and back-translated samples containing the same entity type. This method avoids the drawback that traditional NER cannot use back-translation for applying data augmentation. The success achieved by the above methods illustrates the effectiveness of semi-supervised learning methods on low-resource tasks. This paper will apply the SS-GANs method to solve the low-resource NER problem and propose the Seq-GAN-BERT semi-supervised learning model. Experimental results show that our proposed Seq-GAN-BERT model can effectively utilize unlabeled data and improve model representation ability through adversarial learning, and achieve good performance on low-resource NER task.

3 Seq-GAN-BERT: semi-supervised BERT with GAN for sequence tagging tasks

3.1 problem definition

The goal of the NER is to extract entities from the input texts and accurately determine their types. Given an input sentence $X=(x_1,x_2,...,x_n)$, where n is the length of the sentence, for a given entity type set E, output all entity sets $\varepsilon=\{\varepsilon_i=(a_i,b_i,e_i)\,|\,a_i\le b_i, e_i\in E\}$ contained in X, where $a_i,b_i\in\{0,1,\cdots,n\text{-}1\}$ represents the start and end positions of the extracted entities, and e_i represents the entities in E type (e.g., PER, LOC, etc.).

3.2 Seq-GAN-BERT model

The NER semi-supervised learning model Seq-GAN-BERT proposed in this paper is a kind of SS-GANs based on BERT, which consists of three parts: BERT encoder, generator and discriminator. This semi-supervised learning model incorporates the discriminator design ideas of SS-GANs: in the forward propagation process, the real samples encoded by BERT together with the fake samples generated by the generator are input to the discriminator, and the discriminator will classify the real sample features into k categories while distinguishing the authenticity of the input. The alternating training of the discriminator and the generator is performed to train a k-class discriminator based on semi-supervised learning. The generator generates samples resembling the real data distribution in the adversarial training with the discriminator, and the classification ability of the discriminator is also enhanced. At this time, the back-propagated gradient will optimize the parameters of the BERT encoder, thereby improving the performance of entity classification. The overall model structure proposed in this paper is shown in Figure 1.

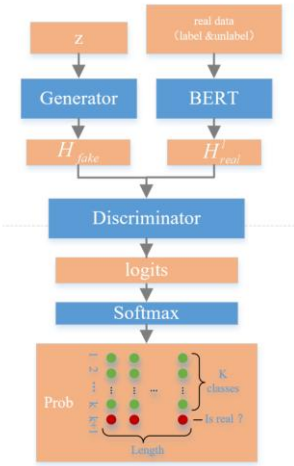


Figure 1: Seq-GAN-BERT model

For the input sentence $X=(x_1,x_2,...,x_n)$, it obtains the semantic vector representation $H^l_{real}=(h^l_0,h^l_1,\cdots,h^l_{n-1})$ through the l-layer self-attention unit of BERT, where l belongs to $[1,L]$. At the same time, the generator obtains the generated sample representation H_{fake} through the input Gaussian noise $z\in N(\mu,\sigma^2)$. After that, the discriminator receives the output representation H^l_{real} of BERT and the generated sample representation H_{fake}, performs $k+1$ classification to obtain the output logits $H_{\log its}\in R^{L\times(k+1)}$, where the first k classes are real samples classes, and the $k+1$ class is fake samples. After $H_{\log its}$ is normalized by the softmax function, the probability distribution sequence $P=(P_1,\ P_2,\cdots,P_n)$ is obtained, where $P_n\in R^{k+1}$, and finally the classification is completed by finding the category corresponding to the maximum probability indexs $Y=\arg\max(P)$. During the whole process, the generator and the discriminator are alternately trained, and the generator generates a more realistic sample representation in the adversarial training with the discriminator. The addition of domain unlabeled corpus enables the discriminator to be more fully trained and the generator to be further strengthened. The BERT encoder learns more general semantic information through this adversarial training, thereby improving the entity recognition effect. And the generator and discriminator in the model will be introduced in detail in the next section.

3.3 Generator and discriminator

The generator is mainly used to generate fake samples. It receives a Gaussian vector z during forward propagation and generates a sample matrix after passing through the neural network. The generator parameters are updated as training progress, generating samples closer to the true sample distribution. The generator can be a self-attention mechanism (Ashish Vaswani et al., 2017), CNN(Alex Krizhevsky et al., 2012) and LSTM (Sepp Hochreiter and Jürgen Schmidhuber, 1997), considering that CNN can only extract local features, LSTM cannot run in parallel, this paper chooses multi-head self-attention as the generator:

$$MultiHead(Q,K,V) = Concat(head_1, \ldots, head_h)W^o \quad (1)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

$$Q = K = V = zW_z \quad (3)$$

where W_z is the mapping matrix, the noise z is projected to Q, K and V. W^o, W_i^Q, W_i^k and W_i^v is the learning parameter matrix of the self-attention mechanism.

The discriminator of the model in this paper is the MLP (Multi-Layer Perceptron) classifier, which receives the real sample semantic vector ssequence output from the encoder and the fake sample vectors sequence generated by the generator. And perform $k+1$ classification to each token vector in the sequence. The parameter updates of the upper network parts all depend on their corresponding loss functions: discriminator loss L_D and generator loss L_G. We define the probability distribution of the output of model m as p_m. The discriminator loss L_D is the sum of the supervised and unsupervised losses.

$$L_D = L_{D_{\sup}} + L_{D_{un\sup}} \quad (4)$$

Supervised loss $L_{D_{\sup}}$ is the sum of the cross-entropy losses for the classification of each token

$$L_{D_{\sup}} = -\sum_{n=1}^{l}\{E_{x,y\sim p_d}\log[p_m(y=y\mid x, y\in(1,\ldots,k))]\} \quad (5)$$

The unsupervised loss $L_{D_{un\sup}}$ consists of two parts. The real samples are minimized for its $k+1$ class probability, and the generated samples are maximized.

$$L_{D_{un\sup}} = L_{DR_{un\sup}} + L_{DF_{un\sup}} \quad (6)$$

$$L_{DR_{un\sup}} = -\sum_{n=1}^{l}\{E_{x,y\sim p_d}\log[1-p_m(y=y\mid x, y=k+1)]\} \quad (7)$$

$$L_{DF_{un\sup}} = -\sum_{n=1}^{l}\{E_{x,y\sim G}\log[p_m(y=y\mid x, y=k+1)]\} \quad (8)$$

The loss L_G of the generator also consists of two parts. The mean square error loss $L_{G_{feature\ matching}}$ ensures that the generated samples are as consistent as possible with the real sample distribution, and the other part is the unsupervised loss $L_{G_{un\sup}}$ caused by the discriminator in judging the authenticity of the sample. The loss L_G is the sum of both $L_{G_{feature\ matching}}$ and $L_{G_{un\sup}}$ The formula is

$$L_{G_{feature\ matching}} = \sum_{n=1}^{l}\{\| E_{x\sim p_d} f(x) - E_{x\sim G} f(x)\|_2^2\} \quad (9)$$

$$L_{G_{un\sup}} = -\sum_{n=1}^{l}\{E_{x\sim G}\log[1-p_m(y=k+1\mid x)]\} \quad (10)$$

$$L_G = L_{G_{feature\ matching}} + L_{G_{su\sup}} \quad (11)$$

In the training process, to ensure that the model maintains a good balance between supervised learning and unsupervised learning, and to maximize the classification performance of the model, it is necessary to assign reasonable weights to the supervised loss and the unsupervised loss. The unsupervised loss is adjusted by setting the unsupervised coefficient, and the discriminator loss actually used in the experiment is shown in formula (12).

$$L_D = L_{D_{\sup}} + \eta \times (L_{DR_{un\sup}} + L_{DF_{un\sup}}) \quad (12)$$

We first try to determine the coefficient by using the proportion of the labeled samples in each batch, but through experiments, it is found that such a coefficient setting does not bring good model performance. Since the unsupervised learning in this model is based on supervised learning, supervised learning will directly affect the classification performance, so the relationship between the unsupervised coefficient η and the proportion of labeled samples may be nonlinear. Considering that the unlabeled input data may far exceed the labeled data, the exponent β is introduced to reduce the negative impact of supervised loss. The setting of η is as formula (13).

$$\eta = \left(\frac{n_l}{B}\right)^{\beta} \tag{13}$$

where B is the size of each batch of data, n_l is the number of labeled samples in each batch.

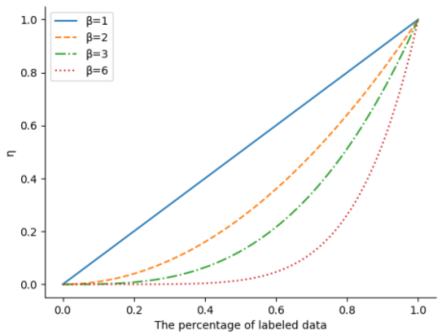


Figure 2: Curves of unsupervised coefficients η corresponding to different parameters β

As shown in figure 2. When the proportion of labeled samples is constant, the larger the index β is, the smaller the η is, and the model will be less negatively affected by the unsupervised loss, and more attention will be paid to the supervised classification. We will analyze the effect of η on model performance in subsequent experimental sections.

4 Experiment

In this section, firstly, we introduce the dataset and evaluation metrics used to test the model's performance. Secondly, relevant experimental details and hyperparameters are reported. Thirdly, the results of different experiments are analyzed from different perspectives, including comparing the baseline model, the choice of other generators, the impact of hyperparameters. Finally, the effectiveness of the proposed model is further verified on the part-of-speech tagging task.

4.1 Datasets and Evaluation Metrics

Our model is evaluated on four datasets, CLUENER, WeiboNER, Laptop14 and CoNLL-2003. To ensure that the experimental results are convincing, we selected both Chinese and English datasets, of which CLUENER and WeiboNER are Chinese datasets, and CoNLL-2003 and Laptop14 are English datasets. CLUENER(Liang Xu et al., 2020) is a fine-grained dataset, WeiboNER (Nanyun Peng and Mark Dredze, 2015) is a social domain dataset, CoNLL-2003 (Erik F. Tjong Kim Sang and Fien De Meulder, 2003) is a news dataset, and Laptop14 (Maria Pontiki et al., 2014) is a dataset of laptop reviews. The statistics are shown in Table 1. Furthermore, we adopt precision, recall and micro F1 as evaluation metrics for entity extraction. All reported results are averaged by running 10 experiments with random initialization.

Dataset	Train sentences	Test sentences	Entity types
CLUENER	10748	1343	10
WeiboNER	1350	270	7
CoNLL-2003	14986	3465	4
Laptop14	2741	304	3

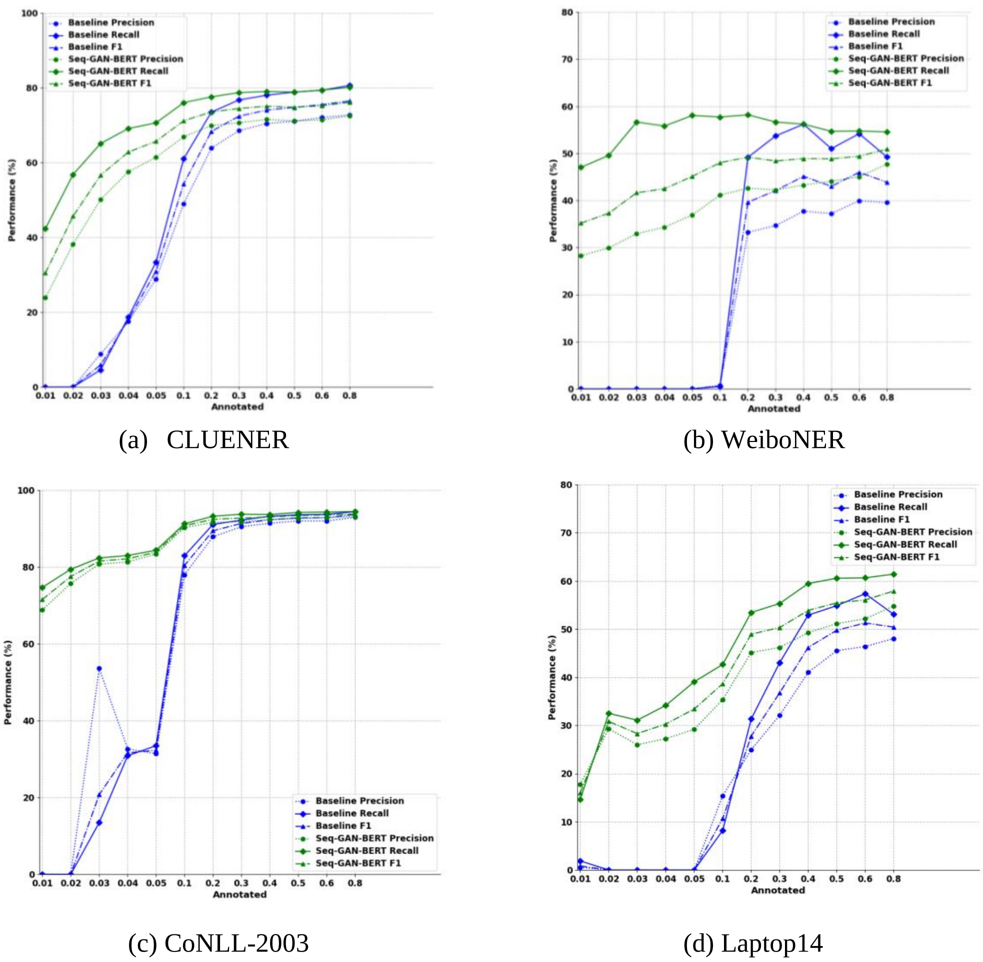
Table 1: The statistics of NER datasets

4.2 Datasets and Evaluation Metrics

In our experiment, a 12-layer, 768-dimension BERT model was used, the batch size is 64, the maximum text length is 128, and the learning rate is set to 2e-5. When constructing a small sample training set, we have made certain rules and restrictions to ensure that the randomly selected labeled sample data contains all pre-given entity types. Training epoch is set to 6 on WeiboNER and Laptop14 datasets and set to 3 on other datasets.

4.3 Main results

We select BERT-Softmax (Jing Li et al., 2022) as the baseline for comparison in the experiment. To pay attention to the performance of the Seq-GAN-BERT model under a small amount of labeled data, we test the model's performance with different amounts of labeled data. The specific operation is randomly selecting a certain ratio of samples from the training set as labeled samples and removing the labels from the remaining data as unlabeled samples. The ratio of labeled data is increased from 1% for our experiments. When the ratio is 1, it means that all labeled data is used, and no unlabeled data is used. Figures 3(a)-3(d) correspond to the experimental results on the CLUENER, WeiboNER, Laptop14 and



(a) CLUENER

(b) WeiboNER

(c) CoNLL-2003

(d) Laptop14

Figure 3: Performance comparison of Seq-GAN-BERT and the baseline model in NER

CoNLL-2003 datasets, respectively. The abscissa in the figure is the ratio of labeled data, and the ordinate is the performance of the model. As is shown in figure 3, our model achieves good performance on low-resource named entity recognition, and when the labeled data is insufficient, our model has a significant advantage over the baseline model.

CLUENER: As shown from Figure 3(a), there are only 107 labeled data, that is, 1% of the total, the F1 of the baseline is close to 0, and our model can still learn useful information from the data and classify some samples correctly. Seq-GAN-BERT stays ahead baseline until the labeled sample ratio reaches 0.5. After the labeled sample ratio reaches 0.5, the performance of the baseline is comparable to our model.

WeiboNER: As shown in Figure 3(b), Seq-GAN-BERT always leads the baseline regardless of the proportion of labeled data. This indicates that our model performs well on small sample tasks.

CoNLL-2003: As can be seen from Figure 3(c), when the ratio of labeled data is small, Seq-GAN-BERT has an overwhelming advantage over the baseline, and this advantage is maintained until the ratio of labeled samples reaches 0.3. Although the performance of the baseline gradually approaches, our model has always been ahead.

Laptop14: In Figure 3(d), we observe similar outcomes with WeiboNER dataset. Our model consistently outperforms the baseline models.

To sum up, the experimental results on the above four datasets in different fields strongly verify the effectiveness and superiority of our proposed Seq-GAN-BERT model on the low-resource named entity recognition task.

4.4 Experiment analysis

The effect of different generators on the overall performance of the model: In the above experiments, the generator in our model uses the self-attention mechanism with a better theoretical effect by default. This section also explores the impact of the generator on the performance of the model when using other neural networks. In the experiment, the ratio of annotated samples is set to 0.1. As shown in Figure 4, when the generator is

LSTM, the model's performance is close to the experimental results of the generator with self-attention, which shows that our proposed semi-supervised learning model has stable performance and universality. Considering that self-attention can be calculated in parallel, we recommend using self-attention as a generator in practical project usage.

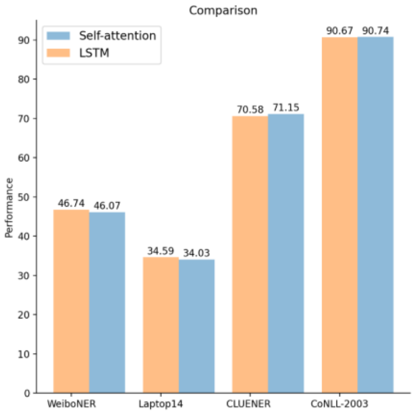


Figure 4: Model performance when generators are different networks

Influence of η in discriminator loss on model performance: During the experiment, the unsupervised loss coefficient η of the discriminator is critical. If the coefficient is too large, the gradient generated by the unsupervised loss will disturb the internal weight of the model. In this part, experiments are conducted to study the impact of unsupervised coefficients η on model performance under low-resource conditions. The set of low-resource proportions with labeled data is {0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.2, 0.3, 0.4 , 0.5}.Taking the CLUENER as the experimental dataset, set the value set of β to {1, 2, 3, 6}, and the unsupervised loss coefficient η can be calculated from Equation 1. The experimental results are shown in Figure 5.

When β=1, the model performs the worst, this is because in the process of backpropagation, a substantial unsupervised coefficient will cause the model to be disturbed by excessive unsupervised gradients, and the semantic representation inside the model is easily diverging, thus affecting the classification accuracy of the model. When $\beta = 2$ or $\beta = 3$, the model can achieve relatively better performance. When β is larger: $\beta = 6$, the model results drop slightly, which indicates that the model has not fully learned the important general representation of unlabeled data.

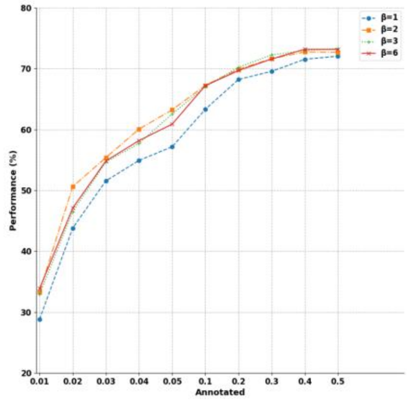


Figure 5: The effect of different coefficients η on model performance

After the above experiments, we can conclude that by setting the unsupervised coefficient η , the supervised learning and unsupervised learning in the model can be better balanced, which is beneficial to guide the model to update and iterate in a more favorable direction improve the performance of the model.

4.5 The application of part-of-speech tagging task on the model

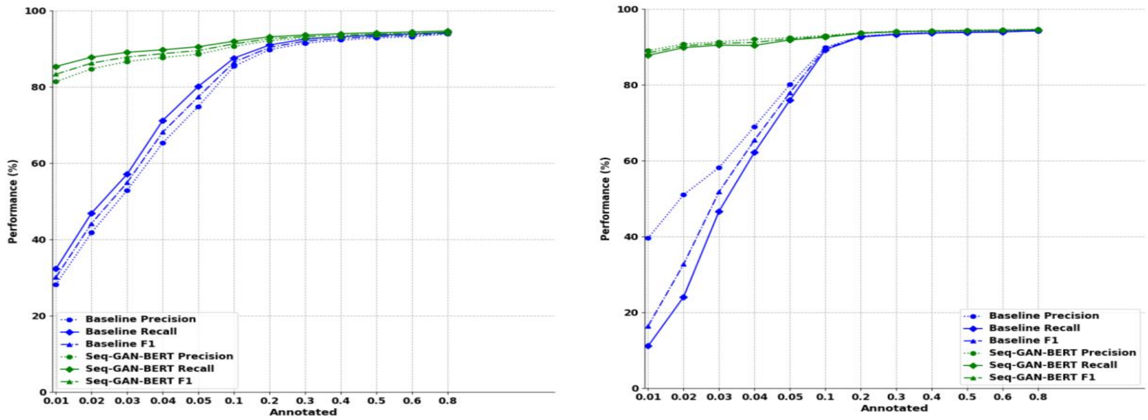
To further verify the superiority of the Seq-GAN-BERT dmoel on low-resource sequence tagging tasks, we apply the mdoel to part-of-speech tagging tasks for experiments. We also choose BERT-Softmax as the baseline.

Dataset: The CoNLL-2003 and RenMinRiBao datasets were selected for part-of-speech tagging experimental evaluation. The CoNLL-2003(Erik F. Tjong Kim Sang and Fien De Meulder. 2003) [31]dataset have been introduced in Section 4.1, and this part of the experiment uses its part-of-speech tagging data. RenMinRiBao[1] is a Chinese news dataset. The statistics of the dataset are shown in Table 2.

Dataset	Train sentences	Test sentences	POS types
CoNLL-2003	14986	3465	46
RenMinRiBao	16279	3000	46

Table 2: The statistics of POS datasets

[1]https://www.heywhale.com/mw/dataset/5ce7983cd10470002b334de3/content



(a) CoNLL-2003 (b) RenMinRiBao

Figure 6: Performance comparison of Seq-GAN-BERT and the baseline in POS task

POS Experiment: The training epoch is set to 3, the data batch size is 64, and the maximum text length is 128, and the learning rate is 2e-5. The experimental results are shown in Figure 6. Figure 6(a) and Figure 6(b) show the experimental results on the datasets CoNLL-2003 and RenMinRiBao, respectively.

As shown in Figure 6(a), Seq-GAN-BERT has better performance relative to the baseline in the few-shot part-of-speech tagging task of the CoNLL-2003 dataset. Our model consistently maintains a significant advantage when the ratio of labeled samples is less than 0.3. When the ratio of labeled samples is greater than 0.3, the Seq-GAN-BERT lead is relatively reduced, but the baseline has not outperformed our model from start to finish. As shown in Figure 6(b), the experimental performance of the RenMinRiBao dataset is similar to that of the CoNLL-2003 dataset. When the ratio of labeled samples is less than 0.5, our model has always maintained a significant advantage. When the ratio of labeled examples is greater than 0.5, Seq-GAN-BERT still leads to the baseline. The experimental results further verify the generality and superiority of our model, and Seq-GAN-BERT can effectively solve the low-resource sequence labeling task.

5 Conclusion

In this paper, we propose the semi-supervised learning model Seq-GAN-BERT for low-resource NER. The proposed model effectively utilizes unlabeled data to improve its small sample learning ability by integrating adversarial generative networks and achieves good performance on low-resource NER. The discriminator and generator in the adversarial generative network are trained alternately. When distinguishing the authenticity of the samples and classifying the samples accurately, the two losses are designed to update the parameters of the BERT model, thereby improving the classification ability of the model. In particular, we also tried two different generators: self-attention and LSTM. Experimental results show that our Seq-GAN-BERT model has significant advantages on low-resource named entity recognition and has better performance on traditional part-of-speech tagging relative to the baseline. We will explore small sample learning for reading comprehension and dialogue question answering tasks in the future.

References

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E.Hinton. 2012. *ImageNet Classification with Deep Convolutional Neural Networks.* In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. Pages 1097-1105.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need.* In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. *DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks.* In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6045–6057. Association for Computational Linguistics.

Christian Szegedy, Wojciech Zaremba, Dumitru Erhan Ian Goodfellow Ilya Sutskever, Joan Bruna, and Rob Fergus. 2014. *Intriguing properties of neural networks.* In ICLR.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2019. *Kernel-based generative adversarial networks for weakly supervised learning.* In AI*IA 2019 – Advances in Artificial Intelligence, pages 336–347, Cham. Springer International Publishing.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. *GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples.* In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Pages 2114–2119, Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. *Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition.* In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 142–147.

He Huang, Philip S. Yu and Changhu Wang. 2018. *An Introduction to Image Synthesis with Generative Adversarial Nets.* Computer Science. arXiv: 1803.04469v2.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* In Proceedings of NAACL-HLT pages 4171–4186, Minneapolis, Minnesota, Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. *EDA: Easy data augmentation techniques for boosting performance on text classification tasks.* In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6382-6388, Hong Kong, China. Association for Computational Linguistics.

Jason Weston, Fr´ed´eric Ratle, Hossein Mobahi, and Ronan Collobert. 2008. *Deep learning via semi-supervised embedding.* In Proceedings of the 25th international conference on Machine learning. Pages 1168–1175.

Jing Li, Aixin Sun, Jianglei Han and Chenliang Li. 2022. *A Survey on Deep Learning for Named Entity Recognition.* IEEE Transactions on Knowledge and Data Engineering (Volume: 34, Issue: 1). pages 50-70.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. *Dual Adversarial Neural Transfer for Low-Resource Named Entity Recognition.* In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages: 3461-3471, Florence, Italy. Association for Computational Linguistics.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. *Learning to paraphrase for question answering.* In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Liang Xu, Yu tong, Qianqian Dong, Yixuan Liaom, Cong Yu, Yin Tian, Weitang Liu, Lu Li, Caiquan Liu, and Xuanwei Zhang. *CLUENER2020: Fine-grained Named Entity Recognition Dataset and Benchmark for Chinese.* Computation and Language. arXiv: 2001.04351[cs.CL].

Nanyun Peng and Mark Dredze. 2015. *Named entity recognition for chinese social media with jointly trained embeddings.* In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Pages 548-554, Lisbon, Portugal. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. *SemEval-2014 Task 4: aspect based sentiment analysis.* In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 27-35, Dublin, Ireland. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. *Improving neural machine translation models with monolingual data.* In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germ. Association for Computational Linguistics.

Rui Wang. Ricardo Henao. 2021. *Unsupervised Paraphrasing Consistency Training for Low Resource Named Entity Recognition.* Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pages 5303-5308. Association for Computational Linguistics.

Samuli Laine, Timo Aila. 2017. *Temporal Ensembling for Semi-Supervised Learning.* International Conference on Learning Representations.

Sepp Hochreiter, and Jürgen Schmidhuber. 1997. *Long short-term memory.* Neural ComputationVolume 9Issue 8. pages 1735–1780.

Thomas N.Kipf and Max Welling. 2017. *Semi-supervised classification with graph convolutional networks.* Machine Learning, arXiv:1609.02907.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. 2016. *Improved techniques for training gans.* In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 2234–2242. Curran Associates, Inc.

Ting Chen, Xiaohua Zhai, and Marvin Ritter. 2019. *Self-Supervised GANs via Auxiliary Rotation Loss.* Conference on Computer Vision and Pattern Recognition. Pages 12154-12163.

Xiang Dai and Heike Adel. 2020. *An analysis of simple data augmentation for named entity recognition.* In Proceedings of the 28th International Conference on Computational Linguistics, pages 3861-3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xiang Zhang, Junbo Zhao,and Yann LeCun. 2015. *Characterlevel convolutional networks for text classification.* In Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1, Pages 649–657.

Yaosheng Yang, Meishan Zhang, Wenliang Chen, Wei Zhang, 2 Haofen Wang, and Min Zhang. 2018. *Adversarial Learning for Chinese NER from Crowd Annotations.* The Thirty-Second AAAI Conference on Artificial Intelligence. Pages 1627-1634.

Zhilin Yang, William W.Cohen, and Ruslan Salakhutdinov. 2016. *Revisiting semi-supervised learning with graph embeddings.* In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. Pages 40–48.