DATA PAPER

Few-shot Learning for Named Entity Recognition Based on BERT and Two-level Model Fusion

Yuan Gong, Lu Mao & Changliang Li†

AI Lab, KingSoft Corp, Beijing 100086, China

Keywords: Few-shot learning; Named entity recognition; BERT; Two-level model fusion

Citation: Gong, Y., Mao, L., Li, C.L.: Few-shot learning for named entity recognition based on BERT and two-level model fusion. Data Intelligence 3(4), 568-577 (2021). doi: 10.1162/dint_a_00102
Received: February 14, 2021; Revised: April 2, 2021; Accepted: May 18, 2021

ABSTRACT

Currently, as a basic task of military document information extraction, Named Entity Recognition (NER) for military documents has received great attention. In 2020, China Conference on Knowledge Graph and Semantic Computing (CCKS) and System Engineering Research Institute of Academy of Military Sciences (AMS) issued the NER task for test evaluation, which requires the recognition of four types of entities including Test Elements (TE), Performance Indicators (PI), System Components (SC) and Task Scenarios (TS). Due to the particularity and confidentiality of the military field, only 400 items of annotated data are provided by the organizer. In this paper, the task is regarded as a few-shot learning problem for NER, and a method based on BERT and two-level model fusion is proposed. Firstly, the proposed method is based on several basic models fine tuned by BERT on the training data. Then, a two-level fusion strategy applied to the prediction results of multiple basic models is proposed to alleviate the over-fitting problem. Finally, the labeling errors are eliminated by post-processing. This method achieves F1 score of 0.7203 on the test set of the evaluation task.

1. INTRODUCTION

Named Entity Recognition (NER) [1] is one of the basic tasks in the field of natural language processing. NER is aimed to extract entities from texts, which is widely used in knowledge graph, information extraction, information retrieval, machine translation, and question answering. Because the end-to-end entity recognition methods based on deep learning can avoid manual feature engineering, their performance is far better than the traditional rule-based methods and statistical learning methods, and thus the deep learning methods

† Corresponding author: Changliang Li (Email: lichangliang@kingsoft.com; ORCID: 0000-0002-8705-5025).

have become the mainstream solutions for NER. Among them, BERT [2] has achieved excellent results in NER task because of its strong feature extraction ability.

In recent years, with the development of information technology, the military data such as documents about military equipment and test evaluation present an explosive growth. How to automatically obtain effective information from these military documents has become an urgent problem. As a basic task of military document information extraction, NER for military documents has received great attention. However, due to the difficulty and the cost of data collection and annotation, NER for military documents still needs further research and improvement.

In order to promote the technology of NER in the field of military test and evaluation, China Conference on Knowledge Graph and Semantic Computing (CCKS) and System Engineering Research Institute of Academy of Military Sciences (AMS) released the NER task of test evaluation in 2020, which required the recognition of four types of entities including Test Elements (TE), Performance Indicators (PI), System Components (SC) and Task Scenarios (TS). In this task, due to the particularity and confidentiality of the field, only 400 labeled data were published.

We regard NER as a typical sequence labeling task. In this paper, the BIO (Begin, Inside, Outside) character-level annotation format is used to label the text data. Specifically, let TE denote test elements, PI denote performance indicators, SC denote system components, and TS denote task scenarios. Thus the total number of labels is label_num = 9, including B label, I label of four types of entities, and one O label. For example, Figure 1 shows the BIO labels of the sentence "美军正在测试一款新型电磁导轨炮，可以约7240千米/小时的速度发射弹药." (which means "the U.S. military is testing a new electromagnetic rail gun, which can fire ammunition at a speed of about 7240 km/h.").

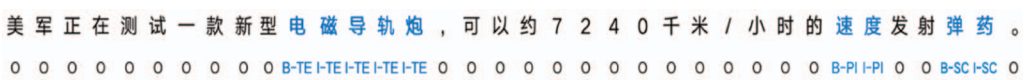

Figure 1. The BIO labels of an example of training data.

In this paper, we proposed a few-shot learning for NER based on BERT and two-level model fusion. In the training phase, we used the basic models, BERT + CRF [3] and BERT + Bi-LSTM + CRF [4], to fine tune on the training data set. In the prediction phase, we first used the fine-tuning results of multiple basic models, then in order to alleviate the over-fitting problem, and we proposed a two-level fusion strategy composed of logit fusion and differentiation fusion to improve the prediction performance of the model. Finally, the labeling errors were eliminated by post-processing. This method achieved F1 score of 0.7203 on the test set of the evaluation task.

The contribution of our work is that we proposed a general NER method which can be easily transferred to other scenarios, especially for those with small data set. The traditional way for few-shot learning usually expands training data by pseudo-labeling unlabeled data. Instead, we considered combining logit fusion with differentiation fusion strategy to correct the over fitting problem caused by small samples. The evaluation results showed the effectiveness of the proposed method.

2. RELATED WORK

The main methods of NER include rule-based methods, statistical learning and deep learning:

NER based on rules and dictionaries relies on a lot of prior knowledge, and thus the labor cost is extremely high. In addition, it also has the disadvantages of low efficiency and weak portability [5].

NER based on statistical learning can avoid the need for manual rule construction. The common methods include Maximum Entropy Model [6], Hidden Markov Model [7], Support Vector Machine [8] and Conditional Random Field [9]. However, these methods rely on predefined features. Feature engineering is not only expensive but also related to specific domains, so the generalization and migration ability of the methods is weak [10].

The end-to-end models based on deep learning can avoid manual feature engineering and mine deep features, which is the current research focus. The Recurrent Neural Network and its variant models [11] as well Convolutional Neural Network and its variant models [12] are widely used in NER tasks. In recent years, the pre-trained word embedding technology has received more and more attention [13]. Among them, the BERT pre-trained language model was released by Google AI team in 2018 [2]. In essence, BERT is a feature representation with strong generalization ability trained by self supervised learning on massive unlabeled corpus, which can extract semantic information of text in a deeper level. As a result, the pre-trained BERT model can be fine-tuned with additional output layers to create state-of-the-art models for a wide range of NLP tasks.

3. THE PROPOSED APPROACH

As shown in Figure 2, in the training phase, firstly, the input text was cleaned and pre-processed to correct error and inconsistent data labeling. Then, the basic models, BERT + CRF and BERT + Bi-LSTM + CRF, were used to fine-tune on the pre-processed training data set. In the prediction phase, the input text was also first pre-processed, and then the training results of the basic models were used for prediction. In order to alleviate the over-fitting problem, we proposed logit fusion to improve the quality of prediction results, and differentiation fusion to improve the prediction ability. Finally, erroneous entities, nested entities, and adjacent entities were eliminated by post-processing, and thus the final prediction results were generated.

3.1 Data Pre-processing

Through statistical analysis, we found that there was a lot of noise in the original training data set, such as the spaces, question marks and other characters shown in Figure 3(a), and there were also problems of error and inconsistent labeling in the corpus. In this paper, the text of training data and test data was cleaned by pre-processing, including unifying character encoding, double-byte to single-byte, removing noise characters, and correcting entity position. For example, the pre-processing result of a labeled sentence

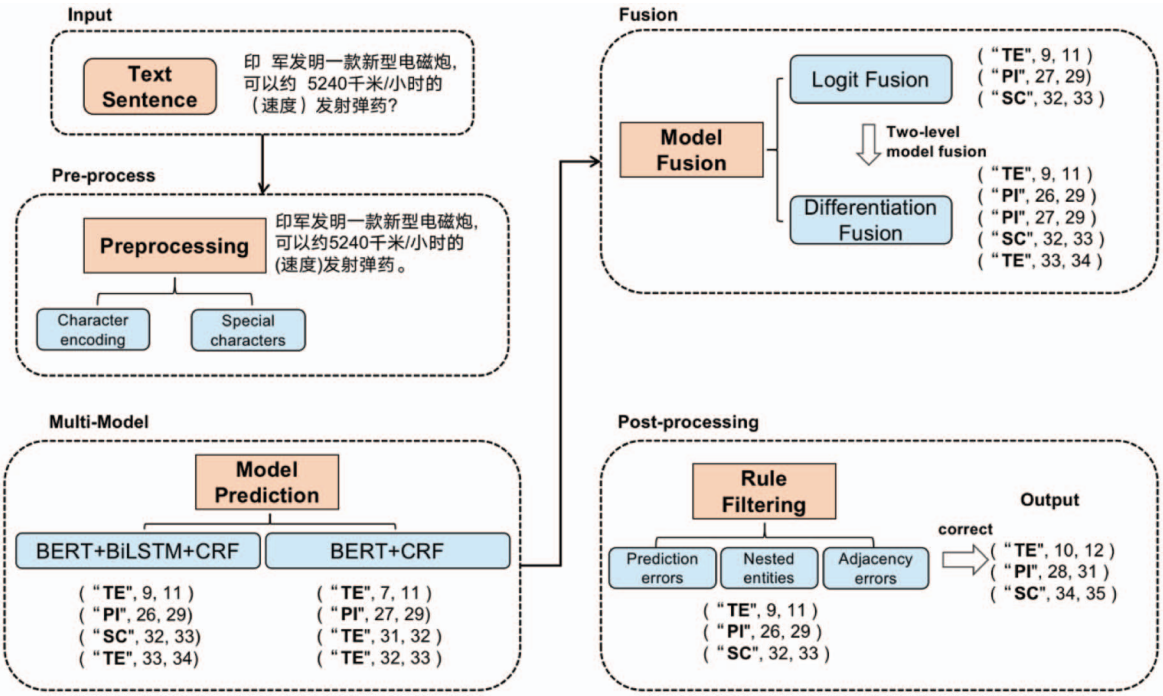

Figure 2. The flow diagram of prediction stage.

"印 军发明一款新型电磁炮，可以约 5240千米/小时的（速度）发射弹药?\r\n\r\n" (which means "The Indian Army invented a new type of electromagnetic gun that can fire ammunition at a speed of about 5240 km/h.") is shown in Figure 3.

Example :

```
{
    "originalText": "印 军发明一款新型电磁炮,可以约 5240千米/小时的（速度）发射弹药?\r\n\r\n",
     "entities": [{"label_type": "TE","overlap": 0,"start_pos": 10,"end_pos": 12},
     {"label_type": "PI","overlap": 0,"start_pos": 28,"end_pos": 31},
     {"label_type": "SC","overlap": 0,"start_pos": 34,"end_pos": 35}]
}
```

（a）Raw data instance

```
{
    "originalText": "印军发明一款新型电磁炮,可以约5240千米/小时的(速度)发射弹药。",
     "entities": [{"label_type": "TE","overlap": 0,"start_pos": 9,"end_pos": 11},
     {"label_type": "PI","overlap": 0,"start_pos": 26,"end_pos": 29},
     {"label_type": "SC","overlap": 0,"start_pos": 32,"end_pos": 33}]
}
```

（b）Pre-processed data instance

Figure 3. Examples of data pre-processing results.

3.2 Basic Models

When selecting basic models, we have tried the models of BERT + CRF, BERT + Bi-LSTM + CRF, BERT + BIGRU + CRF [14], and BERT + IDCNN + CRF [15]. We finally selected BERT + CRF and BERT + Bi-LSTM + CRF as the basic NER models owing to their prediction ability.

3.2.1 BERT + CRF

BERT was used to output vector representation of deep features, and CRF was used as downstream task layer to generate sequence labeling results. Through the fine-tuning of BERT on training data, the vector representation combined the linguistic knowledge contained in the pre-trained model with the task knowledge contained in the NER training data. Besides, CRF can capture the conditional transition probability between different tags, so as to alleviate the logic error in entity tag sequence in the prediction process, such as I tag following O tag.

3.2.2 BERT + Bi-LSTM + CRF

Based on the BERT+ CRF model, a Bi-LSTM layer was added to the encoding layer between BERT and CRF. The Bi-LSTM can further transform and map the feature vectors output by BERT to extract more diverse context features.

3.3 Model Fusion

BERT + CRF model and BERT + Bi-LSTM + CRF model always face the over-fitting problem when training data are small. Therefore, this paper proposed a two-level fusion strategy applied to the prediction stage to improve the performance of the model.

3.3.1 Logit Fusion

In the prediction phase, for a specific input text, the output of the encoding layer of a basic model is a logit matrix M, and its dimension is *max_seq_length* *label_num*, where *max_seq_length* is the maximum length of the text and *label_num* is the number of NER tags.

Let the logit matrixes of the two models be M_1 and M_2, respectively. Based on that, the weighted fusion result of the logits is as follows (Equation (1)):

$$M = \alpha M_1 + \beta M_2 \tag{1}$$

where α and β are real numbers, which represent the weight given to M_1 and M_2, respectively. α and β are assigned empirically. Specifically, the basic model with better performance will be given higher weight to enhance its influence in the fusion results. Furthermore, the logit fusion of the above two basic models can be extended to multiple basic models.

3.3.2 Differentiation Fusion

Differentiation fusion aims at multi-group prediction results, which can be fused by intersection, union or voting. We chose union as the second level fusion strategy, so that multi-group results can complement each other and improve the recall of prediction. However, this fusion, at the same time, may cause some conflict problems such as nested entities.

3.4 Post-processing

As mentioned above, the differentiation fusion may cause the problem of nested entities, and the prediction results of basic models may contain some errors. In order to improve the accuracy of the prediction results, the correction rules as follows were used for post-processing:

(1) Aiming at the problem of nested entities in prediction results, we kept longer entities and removed the nested ones;
(2) Aiming at the problem of adjacent entities in the prediction results, we considered the categories of the entities. If their categories were the same, they would be merged into one long entity; otherwise all adjacent entities would be retained;
(3) We deleted the entities with obvious errors in the prediction results, such as entities with incomplete brackets, or entities ending with ',' and other punctuations.

4. EXPERIMENTS AND ANALYSIS

4.1 Data Set

CCKS2020 NER task for test evaluation contained four types of entities, including TE, PI, SC and TS. The official organization provided 400 training data. In the process of model training, for the needs of model optimization and hyper-parameter selection, we randomly selected 90% samples from 400 training data as training set and the rest as validation set.

4.2 Experimental Setup

For basic models, we mainly trained two versions of BERT+ CRF, namely BERT + CRF-1 and BERT + CRF-2, and one version of BERT + Bi-LSTM + CRF. Theoretically, the introduction of more new models was conducive to learning more diversified feature representation, so as to improve the expression effect of model integration. The basic parameters of each model were shown in Table 1.

Table 1. The parameter setting of basic models.

Model	BERT+CRF-1	BERT+CRF-2	BERT+Bi-LSTM+CRF
max_seq_length	320	300	300
label_num	9	9	9
batch_size	4	4	4
dropout_rate	0.4	0.5	0.3
bi-lstm units	/	/	128
hidden_size (BERT)	1024		
learning_rate	Adjusted dynamically, adjusted every 10 epochs, 5e-5, 3e-5, 2e-5, 1e-5, 5e-6 and 1e-6		
crf_lr_multiplier	100 times of learning-rate of BERT layer		
optimization	Adam		
epoch	60		

In this paper, we chose the large Chinese version of roberta_wwm[①] as the basic BERT pre-trained language model, which contained 24 block layers, 16 multi-head attention layers and outputed 1,024 dimensional feature vectors. The learning rate of model training was adjusted dynamically, and the learning rate was adjusted every 10 epochs. The model trained 60 epochs. CRF layer and BERT layer were trained with different learning rates. In this method, the learning rate of CRF layer was 100 times of that of BERT layer, and Adam optimization algorithm was used for iterative training. In the logit fusion stage, two basic models fusion and three basic models fusion were used, and the weight parameters were (1.1, 0.9) and (0.4, 0.3, 0.3), respectively.

4.3 Results Analysis

In order to further analyze the effectiveness of the proposed strategy in practical application, we compared the experimental results of online test data sets with logit fusion strategy, differentiation fusion strategy and post-processing correction strategy, as shown in Table 2.

Table 2. On-line test results.

Method	F1 score
BERT+CRF-1	0.683
BERT+CRF-2	0.686
BERT+Bi-LSTM+CRF	0.699
Logits-2 (BERT+CRF-1 and BERT+Bi-LSTM+CRF)	0.705
Logits-3 (3 basic model merge)	0.708
Differentiation fusion (Logits-2+Logits-3)	0.714
Rule post-processing (Based on two-level fusion results)	0.720

① https://github.com/ymcui/Chinese-BERT-wwm

It can be seen from the experimental results in Table 2 that in the scene of small sample NER, the two-level fusion strategy proposed in this paper was significantly improved compared with the basic models based on BERT. As shown in Figure 2, due to the lack of training data, the prediction results of the basic model may arise various problems, such as boundary errors of entities and type prediction problems. Thus, we considered using the logit fusion strategy to correct the problems caused by small samples. Compared with the F1 score of the basic model recognition results, after the first level logit fusion, the F1 score was improved by about 0.83%. Considering the difference of the prediction results of different models, we adopted the method of fusion differentiation fusion to improve the recall rate. As a result, after the second-level union fusion, the F1 score was improved by about 1.52%. However, fusion differentiation may cause the problem of nested entities, so we designed a rule of filtering for post-processing correction to improve the accuracy of the prediction results. The F1 score of the final online result of the method proposed reached 0.7203.

5. CONCLUSION AND FUTURE WORK

This paper proposed a few-shot learning for NER based on BERT and two-level model fusion, which can effectively alleviate the over-fitting problem in the process of deep model when training data are small, and improve the prediction performance of basic models. Finally, the F1 score of the evaluation task is 0.7203. In the future, we will focus on how to better solve the problem of entity recognition with small training data, and focus on improving the accuracy and generalization of the NER models.

ACKNOWLEDGEMENTS

We thank China Conference on Knowledge Graph and Semantic Computing (CCKS) and System Engineering Research Institute of Academy of Military Sciences (AMS) for data support.

AUTHOR CONTRIBUTIONS

C.L. Li (lichangliang@kingsoft.com) is the leader of the team, who conceived the original idea of the proposed method. Y. Gong (gongyuan@kingsoft.com) and L. Mao (maolu@kingsoft.com) designed and performed the experiments, optimized the models and discussed the results. L. Mao took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research.

DATA AVAILABILITY STATEMENT

All the data are available in the Science Data Bank repository, https://doi.org/10.11922/sciencedb.01073, under an Attribution 4.0 International (CC BY 4.0).

REFERENCES

[1] Li, J., et al.: A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 1–20 (2020). Available at: https://doi.org/10.1109/TKDE.2020.2981314. Accessed 10 May 2021

[2] Devlin, J., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[3] Pang, N., et al.: Transfer learning for scientific data chain extraction in small chemical corpus with joint BERT-CRF model. In: BIRNDL SIGIR, pp. 28–41 (2019)

[4] Guan, G., Zhu, M.: New research on transfer learning model of named entity recognition. Journal of Physics: Conference Series 1267, No. 012017 (2019)

[5] Lei, Z., Yi, Z.: Big data analysis by infinite deep neural networks. Journal of Computer Research and Development 53(1), 68 (2016)

[6] Borthwick, A., Grishman, R.: A maximum entropy approach to named entity recognition. PhD dissertation, New York University (1999)

[7] Zhou, G., Su, J.: Named entity recognition using an HMM-based chunk tagger. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 473–480 (2002)

[8] Ju, Z., Wang, J., Zhu, F.: Named entity recognition from biomedical text using SVM. In: The 5th International Conference on Bioinformatics and Biomedical Engineering, pp. 1–4 (2011)

[9] Zhang, S., Zhang, S., Wang, X.: Automatic recognition of Chinese organization name based on conditional random fields. In: 2007 International Conference on Natural Language Processing and Knowledge Engineering, pp. 229–233 (2007)

[10] Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models. arXiv preprint arXiv:1910.11470 (2019)

[11] Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)

[12] Collobert, R., et al.: Natural language processing (almost) from scratch. Journal of Machine Learning Research 12(76), 2493–2537 (2011)

[13] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

[14] Cai, Q.: Research on Chinese naming recognition model based on BERT embedding. In: 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), pp. 1–4 (2019)

[15] Wang, Z., et al.: Named entity recognition method of Brazilian legal text based on pre-training model. Journal of Physics: Conference Series 1550, No. 032149 (2020)

AUTHOR BIOGRAPHY

Yuan Gong received his M.E. degree in Robot Science and Engineering from Northeastern University, China, in 2020. He is now an algorithm engineer in AI Lab of KingSoft Corp, Beijing. His research interests include knowledge graph and information extraction.

Lu Mao obtained her PhD degree in Environmental Sciences from Peking University, China, in 2019. She is currently an algorithm engineer in AI Lab, KingSoft Corp, Beijing. Her research interests focus on few-shot information extraction.
ORCID: 0000-0002-2703-7327

Chang-Liang Li received his PhD degree in Pattern Recognition and Intelligence Systems from Institute of Automation, Chinese Academy of Sciences, China, in 2015. He is currently the principal of AI Lab, KingSoft Corp, Beijing. His research interests include knowledge graph and machine translation.
ORCID: 0000-0002-8705-5025