Data, Information and Computing Science
A. Nayyar et al. (Eds.)

doi:10.3233/ATDE250011

A Study on a Domain BERT-Based Named Entity Recognition Method for Faulty Text

Hongjing SHEN[a], Chongjiang TIAN[a], Xin CHEN[b], Jingxin OU[b], Xiaobo HU[c] and Min HAN[a, 1]

[a] *School of Computer and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China*
[b] *Supply Chain Department, China Electronics Technology, Group Corporation 10th, Chengdu, China*
[c] *Development Department, Chengdu Guolong Information Engineering Co., Chengdu, China*

Abstract. Aiming at addressing the issue of highly specialized faulty text and the challenge of sparse character vectors in high-dimensional space resulting from repetitive characters and limited character types, a named entity recognition method based on Domain BERT (DBERT) is proposed. The DBERT model achieves effective dimensionality reduction and refinement of faulty text features by introducing a feature compression strategy. It also undergoes domain-specific pre-training to fully learn and adapt to the unique characteristics and specializations of faulty text. Subsequently, the DBERT model extracts context-related features of characters in the text and combines these features with specific character representations after a weighting operation. Named entity recognition is then performed using a combination of BiLSTM and CRF models. Finally, DBERT-BiLSTM-CRF is compared with LSTM-CRF and BiLSTM-CRF on an automobile maintenance domain dataset, demonstrating superior performance in terms of recall, precision, and F1 score.

Keywords. knowledge graph, knowledge extraction, named entity recognition, faulty text

1. Introduction

With the rapid development of information technology, natural language processing (NLP) technology has been increasingly utilized in various fields. Named Entity Recognition (NER), a fundamental task in NLP, holds significant importance in extracting structured information from unstructured text and building knowledge graphs.

Since Chinese is different from English and other languages that use space delimitation, it has no obvious word boundary separator, and the word boundaries are fuzzy. A Chinese sentence is a continuous sequence of characters, which makes it difficult for the model to learn effective linguistic patterns and word boundaries directly from it. Additionally, faulty text data has its uniqueness. Due to the lack of domain-labeled data and the expensive and specialized nature of data labeling, the named entity recognition task is more challenging compared to general domain data. Faulty text data is mostly manually entered, leading to phenomena such as irregular entry and colloquialization of descriptions. Entities in the text

[1] Corresponding author: Min Han, e-mail addresses: hanmin@swjtu.edu.cn

are highly associated with domain-specific vocabulary, and the density of entity distribution is higher than that of general-purpose domain text. This leads to a low rate of correct recognition of faulty text entities in the methods used for the general-purpose domain.

To address the aforementioned issues, this paper proposes a Domain BERT (DBERT)-based named entity recognition method for erroneous text. The approach allows the DBERT model to comprehensively learn and adjust to the distinctive characteristics and specificities of erroneous text through domain-specific pre-training. It also achieves efficient dimensionality reduction and enhancement of erroneous text features through a feature compression strategy. Building on this foundation, the accuracy of erroneous text named entity recognition (NER) is further enhanced by integrating BiLSTM (Bidirectional Long Short-Term Memory) and CRF (Conditional Random Field) models.

The main contributions of this paper are as follows: (1) This paper proposes an efficient named entity recognition method for Chinese faulty text, which is based on Domain BERT (DBERT) combined with BiLSTM and CRF models. (2) The sparsity and repetitiveness of faulty text features are effectively addressed by a feature compression strategy, significantly improving the accuracy of entity recognition. (3) Through domain-specific pre-training methods, the DBERT model fully learns and adapts to the unique features of the faulty text. (4) By combining BiLSTM and CRF models, the method not only enhances the understanding of text context features but also improves the accuracy of entity boundary recognition through the global normalization function. (5) Evaluated on a dataset in the field of automobile repair, it is compared with models such as LSTM-CRF and BiLSTM-CRF in terms of recall, precision, and F1 score.

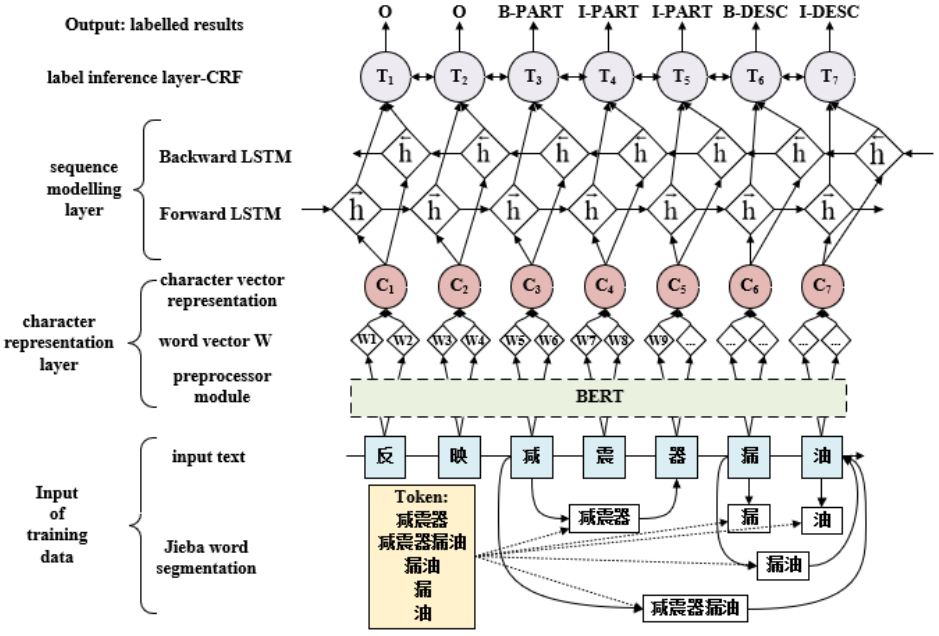


Figure 1. The architecture of the DBERT-BiLSTM-CRF model

2. Related Work

Named Entity Recognition (NER) task, as a fundamental task in natural language processing, aims to identify entities with specific meanings and types from unstructured text. This task is essential for knowledge extraction and can be approached through various methods, including rule-based, statistical-based, and deep learning-based methods.

Rule-based approaches rely on pre-defined rules to recognize specific patterns or keywords in text. Eftimov[1] uses manually developed rules to recognize relevant specific entities. Chiticariu et al.[2] developed a rule-based Named Entity Recognition (NER) system using the GATE framework.

Statistical-based methods, such as the Hidden Markov Model (HMM) and conditional random field (CRF), can learn statistical relationships between entities and contexts based on training data. Toutanova et al.[3] achieved efficient entity recognition using CRF models combined with rich lexical and contextual features. Li et al.[4] utilized the Conditional Hidden Markov Model (CHMM) combined with pre-trained language models to infer potential true labels in noisy observations.

Currently, deep learning-based approaches for Named Entity Recognition (NER) tasks are a mainstream method to effectively capture complex linguistic features and contextual relationships by training on large volumes of textual data. Jia[5] integrated entity information into BERT using Char-Entity-Transformer. Gong[6] propose a hierarchical long short-term memory (HiLSTM) framework. Qi[7] proposes a Chinese medical entity recognition model based on multi-neural network fusion and the improved Tri-Training algorithm. Liu[8] pre-trained the NER-BERT model based on the created dataset.

Luoma[9] suggested a simple method called Contextual Majority Voting (CMV) to merge different sentence predictions. Xiang[10] proposed a model enhanced for Clinical Named Entity Recognition (CMNER) with local and global character representations.

3. Named Entity Recognition Model Based on DBERT-BiLSTM-CRF

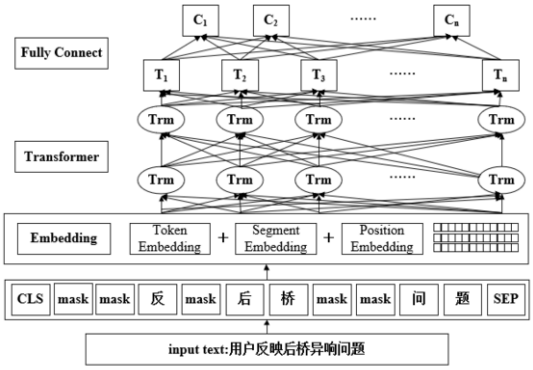
The architecture of the proposed DBERT-BiLSTM-CRF-based named entity recognition model is shown in Fig.1. The model is divided into three parts: the DBERT character representation layer, the BiLSTM sequence modeling layer, and the CRF label inference layer.

3.1 Character representation layer

BERT is a deep bidirectional pre-trained language representation model based on the Transformer model. It is designed to better capture semantic and contextual information in sentences. In faulty text data, the character vectors generated exhibit excessive sparsity in high-dimensional space due to frequent repetitive character combinations and a relatively small number of character types. This sparsity can have a negative impact on training effectiveness and model performance. To tackle this issue, this study enhances the Chinese BERT model by adding a fully connected layer after the BERT output layer and incorporating convolutional operations. These modifications help reduce feature dimensions, leading to compressed and accurate feature representation. Consequently, each character's representation is transformed from a 768-dimensional vector in the original BERT model to a more concise 512-dimensional vector. This enhancement not only

decreases vector sparsity but also aids the model in processing and identifying key information in faulty text more efficiently.

The structure of the DBERT model in this paper is illustrated in Fig.2.



Fugure 2. DBERT-based embedding vector generation

In this study, the fault text data, such as "用户反映车辆门窗玻璃无法升降，查由升降开关内触点接触烧触造成，更换开关后排除。" (Users reported that the vehicle's door window glass cannot be raised. This issue may be caused by a burnt contact in the lift switch, which can be resolved by replacing the switch.), contains rich contextual information. Typically, the text following "用户反映" (Users reported) describes the fault phenomenon, the section after "造成" (caused by) explains the cause of the fault, and the content following "更换" (replacing) pertains to the solution. The faulty text data exhibits clear contextual features. Therefore, in this paper, we utilize the DBERT model to conduct domain-specific pre-training on the character features in the text data to comprehensively capture and convey the context of the text, aiming to enhance the accuracy of entity recognition in the faulty text data.

Since Chinese is different from English and other languages that use space delimiters, it has no obvious word boundary separator. The whole sentence is a continuous sequence of characters, which makes it difficult for the model to learn effective linguistic patterns and lexical boundaries directly. To obtain the lexical features in the text more accurately, the text is segmented before word vector embedding. The Jieba lexicalization tool library is used to process the text data.

The input of the DBERT model is a text token after word segmentation[11]. Each token generates corresponding embedding vectors through the DBERT model, which capture the semantic information of the token in its context

$$H = DBERT(tokens) \tag{1}$$

Where H is the hidden state matrix output by DBERT, and each row corresponds to the embedding vector of a token in the input sequence. To obtain the features of each character using the DBERT model, individual characters are used as separating tokens and special tokens ([CLS] and [SEP]) at the beginning and end of each sentence are introduced to conform to the model input specification. For example, for the text "[CLS] 用户反映后桥异响问题 [SEP]". In the model output, the special characters at the beginning and end

are omitted, only the vector of middle characters is retained, and a certain percentage of the vocabulary is randomly replaced with special [mask] markers.

3.2 Sequence modeling layer

Traditional Long Short-Term Memory (LSTM) networks can only process information from the past to the current moment due to their unidirectional nature. To overcome this limitation, the BiLSTM model[12] allows the model to integrate historical and future information by deploying both forward and reverse LSTM networks on the time series. In the BiLSTM model, the time series is inputted into both forward and backward LSTMs, resulting in hidden state outputs for both forward and reverse sequences. Subsequently, the hidden state output sequences in these two directions are combined to form a comprehensive hidden state sequence. At each time step, BiLSTM computes the output by:

Forward LSTM processing sequence:

$$\overrightarrow{h_t} = LSTM\left(\overrightarrow{h_{t-1}}, x_t\right) \tag{2}$$

Backward LSTM processing sequence:

$$\overleftarrow{h_t} = LSTM\left(\overleftarrow{h_{t+1}}, x_t\right) \tag{3}$$

Merge the hidden states in both directions:

$$h_t = \left(\overrightarrow{h_t}; \overleftarrow{h_t}\right) \tag{4}$$

In this way, BiLSTM is able to capture pre- and post-textual information, which enhances the model's overall understanding of the data.

3.3 Labeling Inference layer

In the named entity recognition task, the BiLSTM sequence processing model has the ability to handle long-distance dependent information. However, BiLSTM mainly focuses on extracting features from text and fails to fully consider the constraint relationships between labels. At this point, the inclusion of the CRF layer is particularly important for the NER model. The CRF layer is able to define the transfer probabilities between labels using the transfer matrix, thus enforcing the legitimacy rules in the label sequence. The CRF layer ensures that the model takes into account the correlations between labels when predicting labels, which significantly enhances the model's accuracy in identifying the boundaries of the entities. The CRF layer not only uses the distribution of the model's predicted values for each time step but also considers the transfer probabilities of labels between neighboring time steps. It can be seen as adding a global normalization to the annotation of the predicted sequence, which ensures that the final predicted label sequence is optimal overall.

In the Conditional Random Field (CRF) model, the objective is to predict the most probable output label sequence given an input sequence. CRF accomplishes this by establishing a joint probability distribution over the entire sequence of labels, instead of modeling each label independently.

At its core, it computes the conditional probability, which represents the output sequence given the input sequence, when the output sequence probability of the output sequence. This probability is calculated in equation (5).

$$P(y|x) = \frac{1}{Z(x)} exp\left(\sum_{t=1}^{T}\sum_{k} \lambda_k f_k(y_{t-1}, y_t, x, t)\right) \tag{5}$$

Where $Z(x)$ is a normalization factor that ensures that the probabilities sum to 1. The calculation is shown in equation (6):

$$Z(x) = \frac{1}{Z(x)} exp\left(\sum_{t=1}^{T}\sum_{k} \lambda_k f_k(y_{t-1}, y_t, x, t)\right) \tag{6}$$

$f_k(y_{t-1}, y_t, x, t)$is the feature function, which typically depends on the input features at the current position, the current label, and the previous label. is the weight of the feature function, which is learned through training data.

The probability of all possible labeled sequences is calculated using Conditional Random Fields (CRF), and the most probable sequence is selected as the final prediction.

$$y^* = arg\ max_y p(y|O) \tag{7}$$

Where y^* is the optimal labeled sequence predicted by the model, and O is the given observation sequence, i.e., the output of the BiLSTM layer. The model searches for the sequence with the highest probability of the entire label sequence occurring given the observation sequence.

4. Experiments

4.1 Datasets

This experiment utilizes the fault description text extracted from the claim form in the Multi-core Value Network Collaborative Cloud Service Platform for the Automotive and Components Industry[13] as the corpus to construct the dataset. The corpus mainly includes common fault components, fault phenomena, fault causes, and fault solutions in each system of the automobile.

This paper relies on the Automotive Engineering Handbook and combines the experience of maintenance personnel and expert advice to classify the entity categories into 15 classes.

The common sequence annotation methods are BIO annotation, BMES annotation, and BIOES annotation. To enhance the efficiency of annotating fault text corpus, this paper utilizes the BIO annotation method to create a small-scale dataset for automotive fault maintenance. The dataset comprises 10,779 annotated sentences. For model training and evaluation, the dataset is split into a training set and a test set in an 8:2 ratio. Sample examples are presented in Tab.1.

Table1. Sample Example

Text	Annotation Formatting
后	B- Fault Components
桥	I- Fault Components
壳	I- Fault Components
处	O
严	O
重	O
漏	B - Fault Phenomena
油	I- Fault Phenomena
拆	O
检	O
发	O
现	O
密	B- **Causes of Faults**
封	I- **Causes of Faults**
不	I- **Causes of Faults**
良	I- **Causes of Faults**
导	O
致	O
夹	B-**Solution**
胶	I-**Solution**
密	I-**Solution**
封	I-**Solution**

4.2 Experimental environment and evaluation indicators

The model training work in this experiment was carried out in Python 3.9 and TensorFlow 2.6 environments. During the training phase, the parameter settings of the model are listed in detail in Tab.2.

Table 2. Experimental parameter settings

Parameters	(be) worth
Embedding Size	768
Number of Transformer Layers	12
Learning rate	4e-5
Batch size	64
Epoch	30
Dropout	0.5

To verify the accuracy of the model for the entity labeling task, this study evaluates the model using a test set. The performance evaluation metrics of the model include accuracy (P), recall (R), and F1 value (F1), with detailed formulas for these metrics provided in Equation (8).

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN} \tag{8}$$
$$F1 = \frac{2PR}{P + R}$$

Where TP denotes the number of positive class samples correctly identified as positive class, FP denotes the number of negative class samples incorrectly identified as positive class, and FN denotes the number of positive class samples incorrectly identified as negative class. F1 is the precision rate (P) and recall (R) of the reconciled mean. These measures assess the model's ability to correctly identify the positive class and to recognize all positive classes, respectively.

4.3 Experimental results and analysis

To evaluate the performance of the DBERT-BiLSTM-CRF model proposed in this paper for faulty text-named entity recognition, this experiment establishes comparison groups, which include the independent LSTM-CRF model, the BiLSTM-CRF model, and the generalized BERT-BiLSTM-CRF model. To address potential data imbalance issues resulting from small sample sizes of certain entity categories in the annotated corpus, the models in the comparison are tested on all entity types in the dataset as well as only on entity types with sufficient data, such as faulty pieces, faulty phenomena, faulty causes, and solutions, respectively.

Table 3. Training results for all entity recognition

Model	Indicators and Scores		
	Precision	***Recall***	***F1-Measure***
LSTM-CRF	78.34	72.36	75.20
BiLSTM-CRF	82.08	77.61	79.77
BERT-BiLSTM-CRF	85.21	83.30	84.24
DBERT-BiLSTM-CRF(ours)	**88.53**	**86.84**	**87.68**

In this experiment, two sets of experiments were designed to evaluate and compare the effects of different dataset configurations on the performance of the named entity recognition model. The first set of experiments used the complete labeled entity dataset to train the model, and the training results are presented in Tab.3. The second set of experiments focused on a specific entity category with richer data for training, and the training results are listed in Tab.4. The experimental results indicate that entities in the training set that are unevenly distributed or have insufficient data in some parts can constrain the model's performance.

Table 4. Training results for main entity recognition

Model	Indicators and Scores		
	Precision	***Recall***	***F1-Measure***
LSTM-CRF	80.04	75.63	77.77
BiLSTM-CRF	83.15	78.33	80.68
BERT-BiLSTM-CRF	86.38	84.77	85.56
DBERT-BiLSTM-CRF(ours)	**89.26**	**87.59**	**88.41**

With the same corpus data, it can be observed that the DBERT-BiLSTM-CRF model demonstrates significant advantages in all performance metrics. This model structure utilizes the DBERT module to deeply understand the contextual meaning of erroneous text words, while the BiLSTM-CRF combination further enhances the ability to accurately predict entity boundaries in text. By incorporating the DBERT module, the F1 score of the BiLSTM-CRF framework improves from 79.77% to 87.68%. This enhancement is attributed to BiLSTM-CRF's capacity to optimize the correlation between characters and adjacent tags based on the rich contextual features provided by DBERT, thereby accurately outputting the globally optimal tag sequence.

In this study, the changes in F1 scores of various models within the first 20 Epochs are compared. The BiLSTM-CRF model integrated with DBERT demonstrates satisfactory performance in the initial stages and maintains this level of performance stability in the subsequent Epochs. In contrast, other models typically need more training iterations to achieve performance stability, as illustrated in Fig.3. This suggests that the DBERT-BiLSTM-CRF model offers a notable advantage in training efficiency.

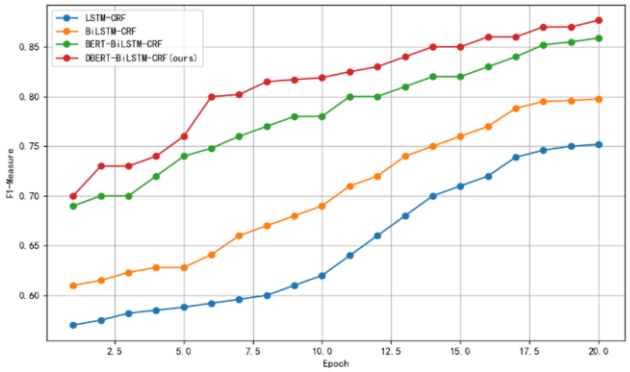


Figure 3. Change in F1 scores for the first 20 Epochs

In summary, the named entity recognition model based on the DBERT-BiLSTM-CRF structure demonstrates a significant performance improvement in terms of recall, precision, and F1 score compared to the baseline model. This validates the model's applicability in the task of named entity recognition of faulty text.

5. Conclusion

In this paper, we propose a named entity recognition method based on Domain BERT (DBERT). Firstly, the DBERT model achieves effective dimensionality reduction and

refinement of faulty text features through a feature compression strategy. It fully learns and adapts to the unique features and specialties of faulty text through domain-specific pre-training. Afterward, the DBERT model extracts context-related features of the characters in the text, combines these features with the specific representation of the characters after a weighting operation, and encodes the features bidirectionally using BiLSTM. Finally, it outputs the label sequence with the highest probability as the final prediction result using CRF. Experiments conducted on the automotive repair domain dataset demonstrate that the method achieves higher recall, precision, and F1 scores through the feature compression strategy and domain-specific pre-training. The model shows better recognition for entity types with sufficient sample sizes, while the recognition ability for entity types with fewer samples needs optimization. Therefore, subsequent enhancements of the model's feature extraction ability for text will be pursued to improve the recognition effect for entity types with insufficient sample sizes.

References

[1] Eftimov, Tome, et al.A rule-based named-entity recognition method for knowledge extraction of evidence based dietary recommendations[J].PLOS ONE,2017,12(6):e0179488.

[2] Chiticariu L, Li Y, Reiss F. Rule-based information extraction is dead! long live rule-based information extraction systems![C]//Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 827-832.

[3] Toutanova K, Klein D, Manning C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network[C]//Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics. 2003: 252-259.

[4] Li Y, Shetty P, Liu L, et al. Bertifying the hidden markov model for multi-source weakly supervised named entity recognition[J]. arXiv preprint arXiv:2105.12848, 2021.

[5] Jia C, Shi Y F, Yang Q R, Zhang Y. 2020. Entity Enhanced BERT Pre-training for Chinese NER. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6384–6396, Online. Association for Computational Linguistics.

[6] GONG C,Li Z H, et al.Hierarchical LSTM with char-subword-word tree-structure representation for Chinese named entity recognition[J].Science China(Information Sciences),2020,63(10):74-88.

[7] Qi R L,Lv P T, et al.Research on Chinese Medical Entity Recognition Based on Multi-Neural Network Fusion and Improved Tri-Training Algorithm[J].APPLIED SCIENCES-BASEL,2022,12(17):8539-8539.

[8] Liu Z H, Jiang F J, Hu Y X, et al. NER-BERT: a pre-trained model for low-resource entity tagging[J]. arXiv preprint arXiv:2112.00405, 2021.

[9] Luoma J, Pyysalo S. Exploring cross-sentence contexts for named entity recognition with BERT[J]. arXiv preprint arXiv:2006.01563, 2020.

[10] Xiang Y,Liu W, et al.Local and global character representation enhanced model for Chinese medical named entity recognition[J].JOURNAL OF INTELLIGENT & FUZZY SYSTEMS,2023,45(3):3779-3790.

[11] Zhao S, Zhu L C, Wang X H, et al. Centerclip: Token clustering for efficient text-video retrieval[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022: 970-981.

[12] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs[J]. Advances in neural information processing systems, 2017, 30.

[13] B. Y. Li, L. F. Sun, S. Y. Wang, R. Tian. Information support system of cloud service platform for automobile industry chain [J]. Computer integrated manufacturing system, 2015, 21(10): 2787-2797. DOI:10.131 96/j.cims.2015.10.028.