A Transformer-Based Model for Named Entity Recognition in Winning Bid Text

Yalan Ling[1], Zhuangye Luo[1], Feng Zeng[1,*], Xiaowei Xie[1]
{lingyalan2023@163.com, luozhuangye@csu.edu.cn, fengzeng@csu.edu.cn, showay@csu.edu.cn}
[1] School of Computer Science and Engineering, Central South University, Changsha, China
* Corresponding author

Abstract. There are quite a few business information in the winning bid document, which is relatively important for the business transactions in the bidding market. The mining of winning bid text belongs to the relevant matters of the Named Entity Recognition (NER). This paper constructs TransBERT, which uses channel attention and bidirectional encoding representation of character-level features to explore the information in text. The model has deep semantic representation, can capture sequence dependence, and enforce label consistency. In order to solve the problem of entity boundary segmentation errors, the Channel Attention Mechanism (DSENet) is constructed, and the Character-level ConvNet (CharCNN) is introduced to capture character-level semantic information. These methods focus on entity content and localization, and also suppress irrelevant features to enhance recognition. The experimental evaluation shows that the performance of TransBERT is better than the current most advanced baseline model. Its precision has increased by 17.37%, the recall has increased by 7.54%, and the F1-score has increased by 12.84%.

Keywords: Named Entity Recognition (NER) , Bidirectional Encoder Representations from Transformers (BERT) , winning bid text mining

1 Introduction

Named Entity Recognition (NER), a basic part in Natural Language Processing (NLP), aims to identify and classify entities in unstructured text that are indicated by rigid designators such as organizations and prices. In this era of information explosion, NER has become a crucial technology for information extraction and analysis, with its value being particularly prominent in the commercial sector—as exemplified by the parsing of winning bid texts. Due to the diversity of bidding systems across different countries, these documents exhibit significant variations in format and content, commonly characterized by diverse forms, grammatical ambiguities, and implicit semantic information. For instance, the name of winning bidder and the winning price often appear in multiple variants and are embedded within complex sentence structures. As a result, effectively processing such highly non-standardized texts has become an important research topic in the field of NER.

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365295

While Conditional Random Fields (CRF) [1] have been applied to NER tasks, they struggle with long-range dependencies and complex contextual cues, which limits their ability to model deep semantic relationships. Despite the capacity of the Bidirectional Long Short-term Memory (BiLSTM) and CRF combination to identify entity boundaries, it struggles with the accurate recognition of nested, continuous, and irregular entities. Although BERT model can understand context bidirectionally and provide rich semantic representations, it mainly focuses on word-level information while neglecting character-level features, making it less effective in recognizing entities with character features. Moreover, the self-attention mechanism of BERT assigns attention weights to each token, and this allocation is based on the interaction between tokens rather than the importance of features. As a result, the model may focus too much on unimportant features, ignoring key features in entity recognition.

In response to the constraints of existing NER models, we introduce Transaction Bidirectional Encoder Representations from Transformers (TransBERT). We designed the DSENet module to effectively solve the problem of wasting computation resource on irrelevant channel features in traditional channel attention mechanisms. Multi-layer BiLSTM (MBiLSTM) is adopted to learn deeper feature representations for understanding complex sentence structures. To capture the specific traits of the name of the winning bidder and the winning price in winning bid text, the proposed model takes advantage of CharCNN in capturing character-level semantic information. Our experimental evaluations confirm that TransBERT outperforms other models on the winning bid text dataset, significantly boosting the accuracy of NER.

Our work makes the following contributions:

- We present the DSENet module, which enhances the ability of BERT to capture contextual information by emphasizing features related to entity content and entity localization.
- We use MBiLSTM structure instead of single-layer BiLSTM to better capture long-range dependencies in sequences.
- We design the TransBERT model, which combines the advantages of multiple models and solves the problems of character-level feature extraction and important feature weighting.

Following this introduction, the body of the paper is organized into four sections. A review of related work is provided in 2, which is followed by a detailed presentation of our proposed TransBERT model in 3. 4 is devoted to the experimental results, while 5 provides conclusion.

2 Related Work

This section reviews three key technological directions that have significantly influenced the development of NER: traditional NER methods, BERT models, and attention mechanisms.

2.1 Traditional NER Methods

The field of NER can trace its origins to approaches utilizing manually constructed rules and dictionaries. These methods often rely on manual feature engineering, particularly in classical machine learning models like the Support Vector Machine (SVM) and the Hidden Markov Model

(HMM). These methods can capture sequence information to some extent. However, they are inadequate for dealing with complex texts. In the study by Kong et al.[2], two Convolutional Neural Network (CNN) architectures were adopted to extract the text features of the visual information of character shapes and the speech information. Since Long Short-Term Memory (LSTM)[3] appeared, due to its own strong ability to model sequential functions, it has been widely applied in the NER task. Although there are various methods in various fields of the NER task, the recognition efficiency and generalization ability still need to be broken through.

2.2 BERT Models

There are changes in the NER field, and the breakthrough of deep learning has become the main reason. Devlin et al. proposed the BERT model [4], which uses bidirectional Transformers to pre-train text corpora and obtain deep bidirectional representations. After word vector training, it is dynamic and its ability is enhanced. The BERT-CNN architecture proposed by Alyoubi et al.[5] proposes a new framework for sentence representation learning and text classification. Chen et al. [6] construct a BERT-CRF model, directly inputting BERT's token embeddings into the CRF layer for efficient NER. Focusing on privacy, The investigation by Muraliharan et al. [7] centered on sensitive information detection in documents, employing a BERT-LSTM hybrid within an NLP framework to address this challenge. The BERT-BiLSTM model proposed by Keremu et al.[8] rigorously evaluated sentiment classification performance using the constructed Uyghur text dataset, and the results showed good performance. On a dataset collected from a real power plant, Li et al. [9] reported excellent results in NER tasks by employing the BERT-BiLSTM-CRF model. Although existing technologies have improved performance to some extent, there is still improvement space in the generalization ability and feature expression of models when dealing with unstructured text in specific domains.

2.3 Attention Mechanisms

Attention mechanisms now have a pervasive presence within NER research, with the core concept of extracting key information from complex input data, focusing on the most important parts of the current task, and constructing correlations between them. By constructing a multi-channel graph attention network (MCGAT), Zhao et al. [10] modeled the relative positions of characters and words. Using word frequency statistics and mutual pointwise information, significant performance gains were realized. Further advancing this direction, Wang et al. [11] proposed Polymorphic Graph Attention Network (PGAT) to capture the dynamic and multi-dimensional correlation between characters and words, making character representations have advantages. In a parallel development, Yu et al. [12] created the Modular Attention Network (MAttNet), equipping models with the flexibility to process expressions encompassing different information types in end-to-end frameworks. Although attention mechanisms have achieved success in the field of NER, they do not adequately consider the interdependence between feature channels, thus neglecting the importance of feature level recalibration.

3 Methods

Aiming to automatically identify the name of the winning bidder and the corresponding winning price entity within unlabeled and unstructured winning bid documents, we propose the TransBERT model shown in Fig. 1. TransBERT uses BERT to obtain rich contextual information features, syntactic structures, and semantic representations. At the same time, TransBERT utilizes the DSENet module to emphasize entity content and entity localization, and uses the CharCNN module to capture character-level features, integrating CharCNN and DSENet to form the CD module. The model fuses the outputs of BERT and CD to obtain global features. This rich representation is fed into MBiLSTM. Finally, CRF performs NER labeling on the sequences of BiLSTM. The detailed design of TransBERT is presented as follows.

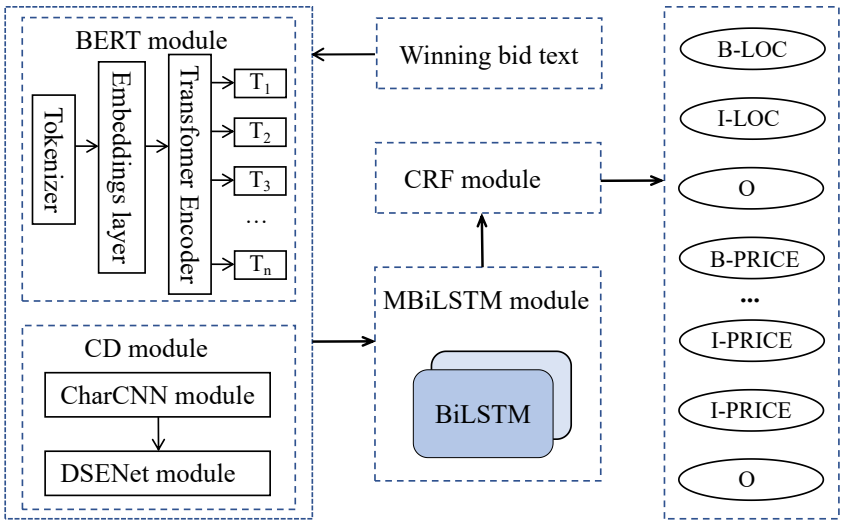

Fig. 1. TransBERT module.

3.1 BERT Module

The BERT module serves as the contextual backbone of our architecture, leveraging its bidirectional transformer encoder to capture deep contextual information from the input text. The input text is first WordPiece tokenized, and the composite representation of each token is formed by the sum of three different embedding types: Token Embeddings, Segment Embeddings, and Position Embeddings. Token Embeddings encode the semantics of words. Segment Embeddings are used to distinguish sentence boundaries. Position Embeddings preserve the order of tokens in the text. Add three vectors together, and then feed the final embedded vector into the bidirectional transformer encoder, which relies on self-attention mechanisms to capture global dependencies.

3.2 CD Module

The CD module consists of CharCNN and DSENet. CharCNN is used to extract character-level features from text, while DSENet is used to enhance the representation of entity content and entity localization. The integration of CharCNN and DSENet effectively addresses the feature extraction and representation challenges inherent in the proposed approach.

CharCNN Module Although the BERT model can capture contextual information of vocabulary, it may not be effective in processing some morphologically rich vocabulary, especially entities containing spelling variants, abbreviations, or special symbols. By introducing CharCNN[13], the model can learn character-level features to better handle the above situations and improve its ability to recognize complex vocabulary. CharCNN is a machine learning method used to process string data. The process flow of CharCNN is shown in Fig. 2. Firstly, CharCNN segments the text into a sequence of characters, and Char Embedding converts these characters into embedding vectors. This conversion process functions like a search, where each character has a unique index used by the model to find the matching embedding vector. These embedding vectors are then stacked into a three-dimensional tensor. Subsequently, multiple one-dimensional convolutional layers ($Conv1d$) are used to convolve this tensor, with each layer employing differently sized kernels to capture multi-scale features. Max-pooling operates on each convolutional output by selecting the most prominent activations, thereby maintaining critical feature information. After these features are integrated, they form a comprehensive character-level feature. Finally, through Linear layer, these features are transformed into one-dimensional space to prepare for subsequent tasks.

DSENet Module Although BiLSTM can capture long-range dependencies in sequences, in some cases, these dependencies may still be insufficient. DSENet has improved Squeeze-and-excitation networks (SENet)[14] by learning the importance weights of feature channels and reweighting the output of the CharCNN layer, emphasizing important features and suppressing unimportant ones, making the model more flexible in capturing long-range dependencies in sequences, especially for entities with larger spans. As shown in Fig. 3, the process of DSENet includes Squeeze, Excitation, and Scale. Firstly, for the input feature $X \in H' \times W' \times C'$, feature extraction F_{tr} is used to obtain the feature $U \in H \times W \times C$, and U is separated into two independent weight vectors U_1 and U_2, one for entity content and the other for entity localization. Features U_1 and U_2 are transformed into vectors of $Z_1 \in 1 \times 1 \times C_1$ and $Z_2 \in 1 \times 1 \times C_2$ through Squeeze operation F_{sq}, respectively, that is, average pooling is performed on features U_1 and U_2. The purpose of this step is to compress the information of the channel, helping the model understand the global importance of entity content and entity localization throughout the sequence. The excitation operation applied to Z_1 and Z_2 is structurally composed of two fully connected layers with ReLU and Softmax serving as their activation functions, respectively. This process conducts dimensionality reduction followed by dimensionality enhancement, learning the correlations between different channels and yielding weight vectors via the sigmoid function. The final stage involves feature modulation through channel-wise reweighting, where the original input U is element-wise multiplied by the derived attention weights to generate the refined output $\tilde{X}$. This step adjusts the feature values of each channel, emphasizing important

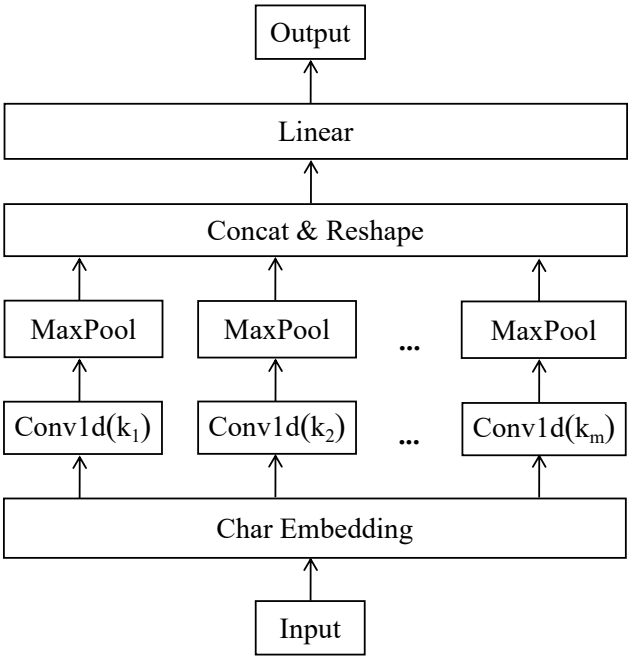

Fig. 2. CharCNN module.

channel information while suppressing unimportant channels.

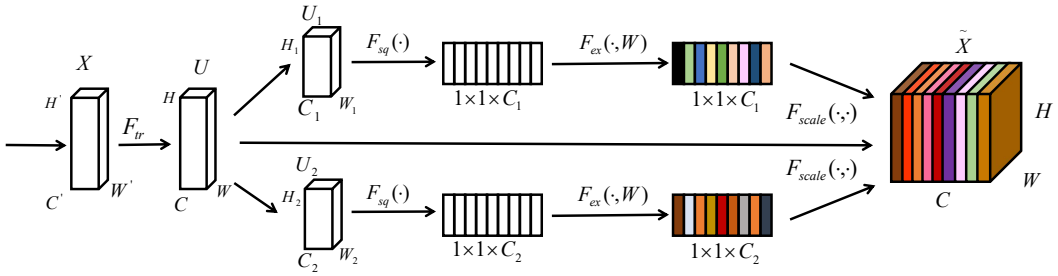

Fig. 3. DSENet module.

3.3 MBiLSTM Module

MBiLSTM consists of multiple layers of BiLSTM, each layer containing forward and backward LSTM, which propagate information through the sequence in opposing directions. MBiLSTM can input the hidden states of the frst forward and backward propagations into deeper bidirectional cycles, enabling each position to obtain richer and more stable contextual information. This module reduces the attenuation of key information during the transmission process. Unlike BiLSTM passively transmitting information from both ends of a sequence to the middle, MBiLSTM can fuse and abstract the extracted local features, thereby enhancing its ability to model long-range dependencies. The MBiLSTM module is shown in Fig. 4.

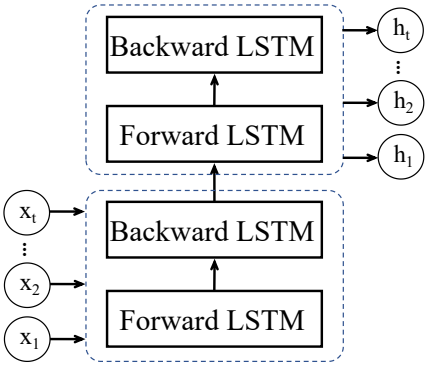

Fig. 4. MBiLSTM module.

The LSTM cell structure is the basic module of the forward and backward layer, which is described in detail in Fig. 5. The calculation relies on three gating mechanisms: forget gate (f_t), input gate (i_t), and output gate (o_t). These components act as information filters, controlling the retention, integration, and emission of data between time steps to maintain long-distance contextual associations. The key variables include hidden states (h_{t-1}, h_t), input (x_t) and cell states (C_{t-1}, C_t,

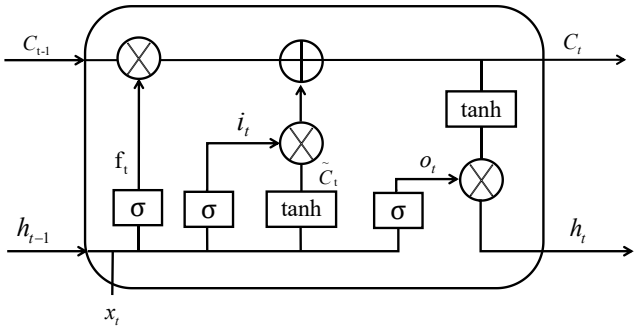

Fig. 5. LSTM cell structure.

$\tilde{C}_t$). The learnable parameters include weight matrices (W_f, W_i, W_C, W_o) and bias terms (b_f, b_i, b_c, b_o). Its structure is expressed as follows in Eq:

$$f_t = \sigma(W_f \cdot [h_{t-1,x_t}] + b_f) \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1,x_t}] + b_i) \tag{2}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1,x_t}] + b_c) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1,x_t}] + b_o) \tag{5}$$

$$h_t = o_t * tanh(C_t) \tag{6}$$

3.4 CRF Module

CRF is a discriminative probabilistic model for sequence labeling. It captures the mutual influence of adjacent labels to improve the accuracy of prediction. When predicting, CRF calculates the probabilities of all possible position label sequences for the input sequence, and selects the one with the highest probability as the prediction result. After obtaining the label vectors for each position, use MBiLSTM, and use CRF and the Viterbi algorithm, considering the emission scores and transition scores, to decode the optimal label sequence, and use it as the final output. Let the input sequence be $X = \{x_1, x_2, \ldots, x_n\}$, with $Y = \{y_1, y_2, \ldots, y_n\}$ denoting its corresponding label sequence. The transition score from label y_i to y_{i+1} is represented by $A_{y_i, y_{i+1}}$. These scores are stored in a transition matrix and are learned as model parameters during the training process. Calculate the emission score using Eq (7); $score(X,Y)$ is the sum of the transition score and the emission score in Eq (8). Use the Viterbi algorithm to determine the final label assignment and the path y^* and the maximum cumulative score is given by Eq (9).

$$P(y|X) = \frac{exp\{score(X,y)\}}{\sum_{y'} exp\{score(X,y')\}} \tag{7}$$

$$score(X,y) = \sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \tag{8}$$

$$y^* = arg\max score(X,Y') \tag{9}$$

4 Experiments

In the empirical evaluation, TransBERT and current state-of-the-art methods are compared and analyzed on real-world data.

4.1 Experimental Settings

Dataset and Annotation Methods Our data comes from the public resource trading website and construct an experimental dataset of winning bid texts. Randomly divide 1753 annotated texts into training-validation-test set with a ratio of 8:1:1. We use the BIO method to process the data. The winning bid text contains 5 entity types of 2 categories: the name of the winning bidder and the winning price. The entity types are in Table 1.

Table 1: Entity Type

Type of annotations	Instruction
B-LOC	The begin of the name of winning bidder entity
I-LOC	The inside of the name of the winning bidder entity
B-PRICE	The begin of the winning bid price entity
I-PRICE	The inside of the winning bid price entity
O	The outside of any entity

Parameter Configuration Implementation is based on BERT [15]. It has 12 layers, 768 hidden units and 12 attention heads. The model file is approximately 400 MB in size and requires about 3.5 GB of GPU VRAM during operation. Training is carried out using Python 3.9 and PyTorch 2.0.0 in an environment with Windows 11 system and equipped with RTX 4060 GPU (with 8GB memory). For all experiments, a fixed set of hyperparameters was used, detailed as follows: the LSTM hidden layer size is 128, the maximum sequence length is 512, the batch size is 12, the learning rate of BERT in the AdamW optimizer is 3×10^{-5}, and the learning rate of CRF is 3×10^{-3}.

Evaluation Indicators Our evaluation framework contains 3 key indicators: precision (P), recall (R), and F1-score ($F1$). P is used to measure whether the positive example classification is reliable; R is used to measure the proportion of actual positive examples that are correctly found. $F1$ balances these two indicators. Here, TP corresponds to correctly identified entities, FP to spurious positive predictions, and FN to missed entities. These metrics are formally defined as follows:

$$P = \frac{TP}{TP+FP} \times 100\% \tag{10}$$

$$R = \frac{TP}{TP+FN} \times 100\% \tag{11}$$

$$F1 = \frac{2 \times P \times R}{P+R} \times 100\% \tag{12}$$

4.2 Performance Comparison

We compare TransBERT with the other six baseline models, including BERT[15], BERT-CNN[5], BERT-CRF[6], BERT-LSTM[7], BERT-BiLSTM[8], and BERT-BiLSTM-CRF[9]. As shown in Fig. 6, compared with BERT[15], TransBERT has precision, recall and F1-score improved by 17.37%, 7.54% and 12.84%, respectively. This performance gain can be attributed to the fact that TransBERT builds upon the BERT-BiLSTM-CRF architecture, thereby enhancing its inherent strength in capturing contextual information. Moreover, the CD module combines the advantages of CharCNN and DSENet, thus improving the accuracy of entity prediction. Finally, MBiLSTM can capture deeper contextual information in sequences compared to BiLSTM, as entities often rely on contextual information of the entire sentence.

4.3 Ablation Experiment

The Impact of BiLSTM Layers on the Performance Due to the effectiveness of BiLSTM in capturing semantic associations between long sequences, and also alleviating the vanishing or exploding of the gradient, we conducted experiments on the number of BiLSTM layers in BERT-BiLSTM-CRF. The detailed experimental results are summarized in Table 2.

Table 2 indicates that the BERT-BiLSTM-CRF model achieves its optimal performance in overall precision, recall, and F1-score when configured with 2 BiLSTM layers. Given the observed decline in evaluation metrics with additional BiLSTM layers, we accordingly set this parameter to 2 in the following experiments.

Table 2: Performance of the BERT-MBiLSTM-CRF model with varying numbers of BiLSTM layers

Number of layers	P(%)	R(%)	F1(%)
1	87.31	92.86	90.00
2	**87.43**	**96.63**	**91.80**
3	88.72	92.06	90.36

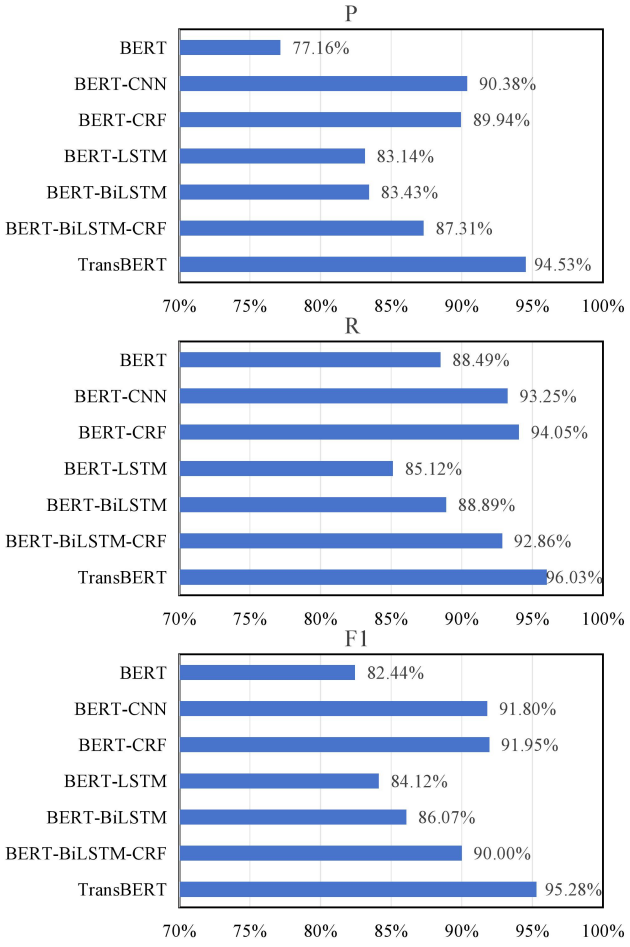

Fig. 6. Comparison of P, R, and F1 between different models on the winning bid text dataset.

The Effectiveness of CharCNN and DSENet Compare the impact of CharCNN and DSENet on the overall performance using the model ablation method.

- Without CharCNN: Analyzing the impact of DSENet without using the CharCNN of TransBERT and evaluating the performance.
- Without DSENet: Analyzing the impact of CharCNN without using the DSENet of TransBERT and evaluating the performance.

The experimental results in Table 3 show that the integrated CharCNN and DSENet framework is superior to any single-module setup. The key findings are as follows:

- It is crucial that DSENet assigns adaptive weights to features, which enables the model to place attention on areas with more information. This explains why the performance of the model will significantly decrease without the DSENet module.
- The extraction of character-level features can achieve better results by enhancing the understanding of entity information in the winning bid text. So deleting CharCNN will lead to a decrease in model performance.

Table 3: Experimental results of module effectiveness ablation on the winning bid text dataset

Models	P(%)	R(%)	F1(%)
Without CharCNN	88.64	94.44	91.45
Without DSENet	80.94	92.66	86.40
TransBERT	**94.53**	**96.03**	**95.28**

Effect of CharCNN and DSENet placement on model performance In order to evaluate the impact of the positions of CharCNN and DSENet on the model performance, two architecture variants are constructed for comparison.

CharCNN-DSENet (CD Module). The position of CharCNN is before the DSENet.

DSENet-CharCNN (DC Module). The position of CharCNN is after the DSENet.

As shown in Table 4, the CharCNN before DSENet has better performance than the subsequent reverse version CharCNN.

5 Conclusion

In this work, we propose TransBERT, a neural network model designed for named entity recognition in winning bid texts. By integrating CharCNN and DSENet, the model achieves richer and

Table 4: Experimental results of module position ablation on the winning bid text dataset

Position	P(%)	R(%)	F1(%)
CharCNN-DSENet	**94.53**	**96.03**	**95.28**
DSENet-CharCNN	93.55	95.04	94.29

more fine-grained semantic representations, thereby significantly enhancing the accuracy and robustness of entity recognition. Specifically, the model employs BERT to encode the context bidirectionally and capture comprehensive semantic information. The MBiLSTM module is used to model long-distance sequence dependencies, while CharCNN extracts character-level features to improve the model's sensitivity to the internal structure of entities. DSENet adaptively assigns weights to different feature channels, effectively strengthening the representation of entity boundaries and content. Finally, the CRF layer applies grammatical constraints to optimize the label sequence, ensuring the validity and consistency of the output. Experimental results demonstrate that TransBERT can accurately identify target entities in winning bid texts, validating its effectiveness and practical utility for domain-specific information extraction tasks. Overall, this study not only provides a high-performance method for entity recognition but also offers technical support for the automated analysis and processing of winning bid texts. For future work, we plan to extend TransBERT into an end-to-end information extraction system and further explore its application to multilingual scenarios, supporting more general intelligent analysis and information processing tasks.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, et al. Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 1529-37.

[2] Kong B, Liu S, He L, Jia L, Liang Y. CSMA-CNER: Multi-modal Chinese NER task with Cross-and Self-Modality Attention. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). IEEE; 2024. p. 1-6.

[3] Cai R, Zheng Y, Maimaiti M. Leveraging feature fusion for improved NER for tourism field. In: 2023 3rd International Conference on Frontiers of Electronics, Information and Computation Technologies (ICFEICT). IEEE; 2023. p. 124-8.

[4] Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

[5] Alyoubi KH, Alotaibi FS, Kumar A, Gupta V, Sharma A. A novel multi-layer feature fusion-based BERT-CNN for sentence representation learning and classification. Robotic Intelligence and Automation. 2023;43(6):704-15.

[6] Chen SS, Hwang RH, Sun CY, Lin YD, Pai TW. Enhancing Cyber Threat Intelligence with Named Entity Recognition Using BERT-CRF. In: GLOBECOM 2023-2023 IEEE Global Communications Conference. IEEE; 2023. p. 7532-7.

[7] Muralitharan J, Arumugam C. Privacy BERT-LSTM: a novel NLP algorithm for sensitive information detection in textual documents. Neural Computing and Applications. 2024:1-16.

[8] Keremu F, Li L. Sentiment classification model for Uyghur language texts based on BERT BiLSTM. In: 2024 9th International Symposium on Computer and Information Processing Technology (ISCIPT). IEEE; 2024. p. 364-7.

[9] Li C, Li M, Lu H, Pan J, Qin B, Wang H. A Computational Framework for Effective Representation and Extraction of Knowledge Graph for Power Plant Maintenance and Overhaul. In: 2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE; 2023. p. 1269-74.

[10] Zhao Y, Meng K, Liu G. A Multi-Channel Graph Attention Network for Chinese NER. In: International Conference on Neural Information Processing. Springer; 2021. p. 203-14.

[11] Wang Y, Lu L, Wu Y, Chen Y. Polymorphic graph attention network for Chinese NER. Expert Systems with Applications. 2022;203:117467.

[12] Yu L, Lin Z, Shen X, Yang J, Lu X, Bansal M, et al. Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 1307-15.

[13] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. Advances in neural information processing systems. 2015;28.

[14] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 7132-41.

[15] Cui Y, Che W, Liu T, Qin B, Yang Z. Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021;29:3504-14.