Attention-CNN with Multi-Task Learning for Chinese Named Entity Recognition

Yanhong Fu[1*], Fuwang Chen[2]
[1]School of Preschool Education, Nanyang Vocational College of Agriculture, Nanyang 473061, China
[2]School of Information Engineering, Nanyang Vocational College of Agriculture, Nanyang 473061, China
E-mail: fuyanhong0309@163.com
[*]Corresponding author

Keywords: attention mechanism, convolutional neural networks, chinese named entity recognition, bidirectional encoder representation from transformer, multi-task learning

Received: February 20, 2025

Named entity recognition serves as a cornerstone in natural language processing and has garnered extensive research attention due to its significance in various downstream applications. Owing to the intricate nature of Chinese texts, characterized by complex syntactic structures and the lack of explicit word boundaries, conventional NER methodologies often encounter difficulties in simultaneously optimizing recognition accuracy and computational efficiency. To address this issue, the study proposes a named entity recognition algorithm that integrates attention mechanisms with Convolutional Neural Networks, incorporates into a Transformer-based bidirectional encoder framework for training. A multi-head self-attention mechanism is employed to capture the global semantic information of the text, and multi-task learning is introduced to construct the final model. When evaluated on datasets with sample sizes of 200, 1000, and 3000, the proposed model consistently outperforms the baseline models in terms of precision, recall, and F1 score. Specifically, under the low-resource setting with 200 samples, the model achieves a precision of 98.62%, a recall of 98.10%, and an F1 score of 98.36%. In terms of inference efficiency, the model processes at a speed of 2618 tokens per second. The experimental results indicate that this method can be widely applied in various fields such as information extraction and text understanding, providing strong technical support for related research.

Povzetek: Model AC-MTL združuje pozornostne mehanizme, konvolucijske nevronske mreže in večopravilno učenje za kitajsko prepoznavo imenovanih entitet. Povezuje globalni pomen in lokalne značilnosti, odlikujeta ga robustnost in natančnost.

1 Introduction

Chinese Named Entity Recognition (CNER) is a fundamental task in Natural Language Processing (NLP), aiming to automatically identify entities with specific meanings within Chinese text. As Chinese information technology continues to advance, Named Entity Recognition (NER) has become a crucial technology in various applications, including information extraction, sentiment analysis, knowledge graph construction, intelligent question answering, and machine translation [1]. CNER is increasingly being utilized across various sectors, including finance, healthcare, e-commerce, and law. It provides crucial support for cross-domain data integration, information extraction, and intelligent applications [2]. With the rise of deep learning technologies, NER methods based on Convolutional Neural Networks (CNN), Recurrent Neural Networks, and Transformers have gradually become the mainstream in research [3]. However, CNER still faces significant challenges due to the unique structure of the Chinese language. Traditional NER methods based on dictionaries and machine learning still suffer from low universality and poor cross-regional recognition performance [4]. Bidirectional Encoder Representations from Transformers (BERT), built on the Transformer architecture, effectively captures deep contextual information from text, thereby improving the accuracy and generalization of entity recognition in complex contexts [5]. The Multi-Head Self-Attention Mechanism (MHSA) is particularly well-suited for capturing long-range dependencies and contextual relationships in Chinese text, enhancing NER accuracy by considering the global semantic information across the entire sentence [6]. To address the issues of low generalizability and suboptimal recognition performance in CNER, a novel recognition approach—Attention-Enhanced Convolutional Neural Network (Attention-CNN)—was proposed to improve recognition accuracy and optimize computational efficiency. The study also introduces Multi-Task Learning (MTL) to develop the final CNER hybrid model, named Attention-CNN with Multi-Task Learning for Chinese Named Entity Recognition (AC-MTL). By combining the advantages of MHSA and CNN, this study aims to simultaneously process global semantics and local features. The AC-MTL model provides an effective and feasible new

method to improve the performance and accuracy of CNER.

2 Related works

With the advent of the information age, CNER has emerged as a crucial task in natural language processing, aiming to extract specific types of entities—such as person names, locations, and organizations—from unstructured text [7]. The lack of clear word boundaries in Chinese, combined with semantic ambiguity, nested entities, and long-distance dependencies, has long made CNER a challenging problem. Early research in NER primarily relied on statistical learning approaches such as Conditional Random Fields (CRF) and Hidden Markov Models (HMM), which were heavily dependent on hand-crafted features and lacked generalizability in complex scenarios [8]. Later on, deep learning approaches took center stage. Notably, the BiLSTM-CRF framework proposed by Huang et al. significantly improved sequence labeling performance and became a widely adopted baseline in NER research [9]. In recent years, the development of pre-trained language models has driven substantial advances in NER performance. Devlin et al.'s BERT model, which employs a deep bidirectional Transformer to capture contextual semantics, demonstrated strong performance across various NLP tasks and has been extensively applied to NER [10]. Several BERT-based adaptations have been introduced to better model Chinese-specific linguistic features. For example, Chay-intr et al. introduced a Lattice Attention Encoding (LATTE) method for character-based word segmentation that achieved promising results on standard datasets in Chinese, Japanese, and Thai [11]. These studies underscore BERT's potential for modeling word boundaries, contextual dependencies, and semantic richness in Chinese NER tasks.

Beyond foundational architectures, the integration of attention mechanisms and multi-task learning has become a prominent direction for boosting NER performance. For instance, Patel and Ezeife proposed a novel aspect-based opinion mining system, BERT-MTL, which introduces auxiliary tasks to enable shared representation across multiple subtasks, simultaneously handling aspect term and category extraction. This approach not only improves accuracy but also significantly reduces training time [12]. The GlobalPointer method further overcomes the limitations of CRF in recognizing overlapping entities. Zhai et al. developed a CNER framework that utilizes an Efficient GlobalPointer model to effectively address entity nesting, along with a context shielding window mechanism [13]. These works validate the effectiveness of structural integration strategies in enhancing NER capabilities. In terms of applied CNER, several studies have extended the task to domain-specific text, including medical, agricultural, and railway documents. Models combining CNNs and attention mechanisms have shown promising performance by leveraging convolutional layers for local feature extraction and attention mechanisms for capturing global dependencies [14]. Yang et al. proposed a BERT-based CNER model tailored for complex filtering in COVID-19 epidemiological investigation texts, resulting in notable improvements in both accuracy and F1 score [15]. Zhao et al. introduced a high-performance NER model for agricultural texts by incorporating multi-level glyph feature modeling and self-attention mechanisms. This model achieved an F1 score of 95.56% and enriched target word representations through hierarchical glyph feature learning [16]. A summary and comparison of these studies are provided in Table 1.

Table 1: Structured summary of related work

Author(s)	Dataset / Domain	Method	Key results	Major contribution
Huang et al. [9]	CoNLL/multi-task labeling	BiLSTM-CRF	Multi-task average F1>91%	Introduced a classic deep structured model for sequence labeling; became a standard NER baseline.
Devlin et al. [10]	Multilingual pre-training corpora	BERT: Bidirectional Transformer	Significant improvement in F1	Proposed the BERT pre-trained language model, establishing a new paradigm for NER tasks.
Chayintr et al. [11]	BCCWJ/CTB6/BEST2010	LATTE (Lattice+ GNN+Attention)	Improved word segmentation accuracy	Addressed multi-granularity semantic ambiguity using lattice-based encoding and attention mechanisms.
Patel and Ezeife [12]	SemEval-14 ABSA	BERT-MTL (Multi-task Learning)	Improved multi-task accuracy	Enhanced generalization and training efficiency between subtasks through shared BERT representations.
Zhai et al. [13]	Medical texts/CMeEE and others	Knowledge Distillation+Efficient GlobalPointer	F1 score exceeds existing best results	Proposed an efficient GlobalPointer architecture to handle nested entities and redundant information while balancing accuracy and speed.
Yang et al. [15]	COVID-19 epidemiological texts	BERT+BiLSTM+IDCNN+CRF	F1 score exceeds existing best results	Constructed a multi-level architecture for complex medical text modeling, improving CNER accuracy.
Zhao et al. [16]	Agricultural chinese texts	ALBERT+CNN+BiLSTM+Self-Attention+CRF	F1=95.56%	Enhances the generalization ability of named entity recognition in agricultural texts by leveraging multi-level glyph features.

In summary, the development of CNER has evolved from statistical learning methods to deep learning, and further toward integrated pre-trained architectures and multi-task modeling. Deep learning approaches that incorporate self-attention mechanisms and convolutional neural networks have demonstrated superior performance in capturing complex data patterns and modeling global contextual information. The primary challenge at present lies in how to jointly model character-level semantics, word boundaries, and contextual dependencies while achieving accurate entity classification and boundary recognition. To address this, the study proposes the AC-MTL model, which integrates attention mechanisms with CNN structures. This design aims to achieve a better balance between global semantic understanding and local feature extraction, particularly when dealing with complex entities and long-form texts, thereby enhancing the model's adaptability in Chinese named entity recognition tasks.

3 CNER model based on attention mechanism and CNN

3.1 CNER design based on CNN and attention mechanism

With the development of the internet, artificial intelligence has become ubiquitous in people's lives, bringing convenient and intelligent technologies for societal advancement [17]. NER serves the purpose of automatically identifying entities such as person names, organizations, and locations in text [18, 19]. In service, NER needs to accurately and quickly recognize specific entities, whereas traditional NER methods often suffer from insufficient accuracy. CNN, a feedforward neural network that utilizes convolutional operations and a deep architecture, is widely applied in tasks like object detection and image recognition [20]. In NER tasks, CNNs can be employed to extract local contextual features from embedded character sequences. The standard processing pipeline involves four main steps. First, the input Chinese sentence is encoded by a pre-trained language model such as BERT into a two-dimensional embedding matrix $X \in R^{n \times d}$, where n denotes the sentence length and d represents the dimensionality of each character's embedding vector. This embedded sequence is then passed through a one-dimensional convolutional layer. The convolutional layer applies multiple sets of filters with varying kernel sizes—specifically, window sizes of 3, 5, and 7—sliding along the sequence dimension to capture local features at different granularities. Each filter generates a feature map, and all resulting feature maps are concatenated to form a richer representation. Following the convolution operation, a max-pooling layer is applied to reduce the length of the feature maps and retain the most salient features. The pooled output is subsequently fed into a fully connected layer or a CRF layer for final entity label prediction. Unlike the two-dimensional convolution used in image processing tasks, the convolution operation in this model is performed only along the temporal (sequence) dimension, and thus constitutes a one-dimensional convolution. This approach effectively captures local structural features in Chinese, such as radicals, part-of-speech combinations, and character patterns, thereby enhancing the model's ability to understand short-range entity structures. When the convolutional layer processes the sequence, the dimensions are adjusted, and the padding size is shown in Equation (1).

$$paddingSize = \frac{f-1}{2} \tag{1}$$

In Equation (1), f represents that the convolutional kernel size is odd. The formula used to calculate the convolution output size is provided in Equation (2).

$$\begin{cases} w_{out} = \dfrac{w + 2 \times paddingSize - f}{s} + 1 \\ h_{out} = \dfrac{h + 2 \times paddingSize - f}{s} + 1 \end{cases} \tag{2}$$

In Equation (2), w and h represent the width and height of the input image, while s is the stride. To reduce the output dimensions, a pooling operation is performed as shown in Equation (3).

$$\begin{cases} \mathrm{C}_i = \mathrm{Conv1D}(\mathrm{A}_i, \mathrm{W}, b) \\ \mathrm{C}_{\max} = \max(\mathrm{C}_1, \mathrm{C}_2, ..., \mathrm{C}_k) \end{cases} \tag{3}$$

In Equation (3), A_i represents the input processed by MHSA, and C_i is the output after convolution. In the field of natural language processing, CNN is widely used to extract features such as the structural components of Chinese characters. CNN can also handle long Chinese sentences or capture potential word properties. However, since CNN performs better in learning local features and cannot fully consider global semantics, it may encounter issues with inaccurate recognition of Chinese, as its operational scope is limited. The core idea of the attention mechanism is to focus on specific locations while ignoring less important information, similar to how humans focus their attention on specific parts of an object to enhance feature learning from semantic information [19]. Among the different types of attention mechanisms, MHSA has multiple attention heads, and when processing semantic information, it not only extracts local features clearly but also processes them in parallel, allowing global features to be expressed more distinctly [20]. The principle of the MHSA mechanism is shown in Figure 1.

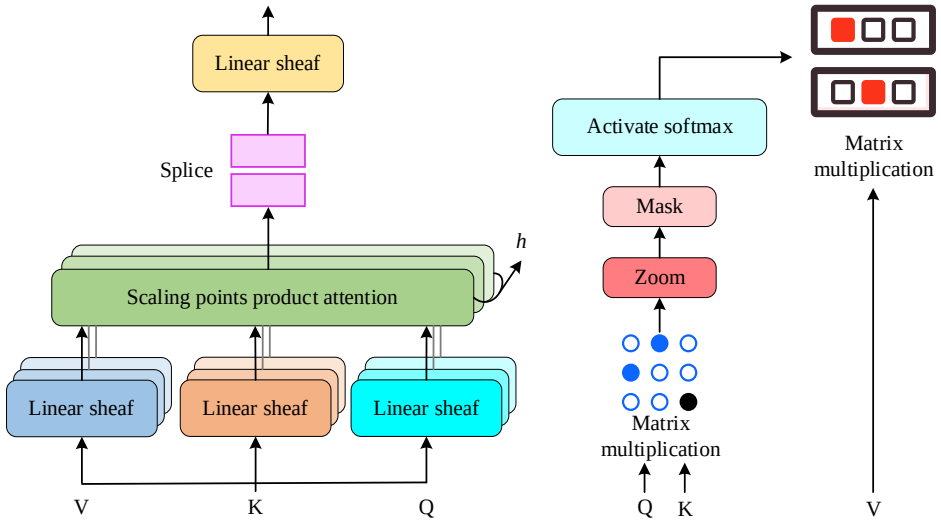


Figure 1: Diagram of the multi-head attention mechanism.

As shown in Figure 1, MHSA is capable of focusing on the most important tasks at the moment by gathering various pieces of information. First, for each sequence, a query vector (Q), a key vector (K), and a value vector (V) are assigned. These vectors are then processed through linear layers for individual linear transformations. After that, they are aggregated into scaled dot-product attention, where the attention distribution is calculated. Subsequently, operations like concatenation are performed, followed by another round of linear transformations. When the Q and K vectors undergo attention via matrix multiplication and masking, the resulting scores are processed by the softmax function. All the outcomes are then added to the V vector, and after the final matrix multiplication, the output is obtained. The attention computation during linear transformations is expressed in Equation (4).

$$\begin{cases} Q = XW^Q \\ K = XW^K \\ V = XW^V \end{cases} \tag{4}$$

In Equation (4), X represents the word vector $X = \{x_1, x_2, x_3, \ldots, x_n\}$. Due to the parallel computing capability of MHSA during CNER, it effectively captures global semantic information. Therefore, the study proposes combining MHSA with CNN to form Attention-CNN, which improves the accuracy of text CNER. The structure of CNER based on Attention-CNN is shown in Figure 2.

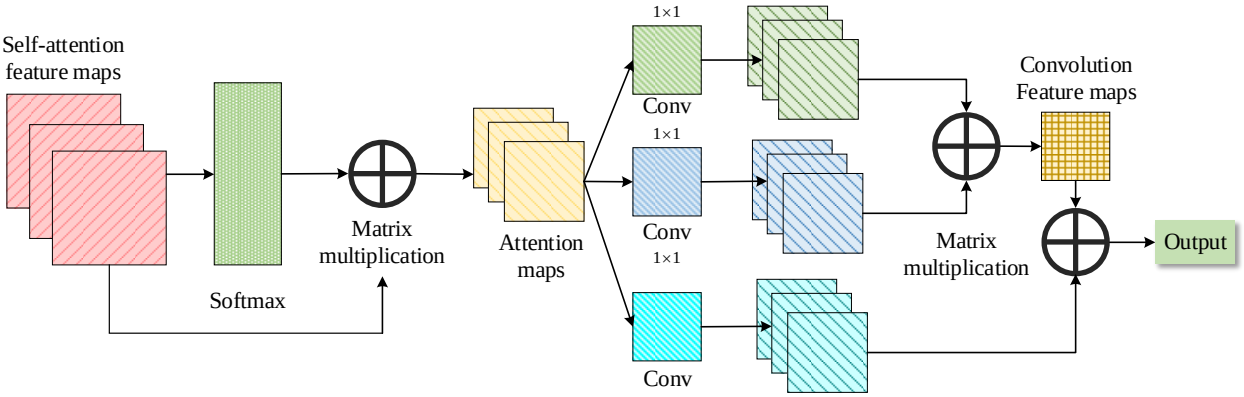


Figure 2: Structure of CNER based on Attention-CNN.

As shown in Figure 2, Attention-CNN structure in CNER has MHSA at the front end, which identifies Chinese entities in the text. Its multiple heads perform parallel computation of attention scores, and after matrix multiplication and softmax function mapping, the output of the MHSA feature map is obtained. This output is then used as the input for CNN, where convolution operations are applied with different kernels, followed by max pooling to reduce dimensions. The process of convolution and pooling is repeated, and after further matrix and function calculations, the final recognition result is output through the maximum probability at the fully connected layer. The more optimal results are selected, and CNN is used to further extract the optimal solution, resulting in the best overall output. Additionally, position encoding solves the problem of lacking sequential order information when the model processes

words at different positions. The expression is shown in Equation (5).

$$\begin{cases} \mathrm{PE}(pos, 2i) = \sin\left(\dfrac{pos}{10000^{\frac{2i}{d}}}\right) \\ \mathrm{PE}(pos, 2i+1) = \cos\left(\dfrac{pos}{10000^{\frac{2i}{d}}}\right) \end{cases} \tag{5}$$

In Equation (5), pos represents the position, and i represents the dimension. Further clarification of the scaled dot-product attention is shown in Equation (6).

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\mathrm{V} \tag{6}$$

In Equation (6), d_k represents the key dimension, which is used for scaling. The combined formulation of MHSA is given in Equation (7).

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, ..., \mathrm{head}_h)W^O \tag{7}$$

In Equation (7), head represents the output of each head.

3.2 Attention-CNN model design for CNER

After completing the design of the Attention-CNN algorithm, the study proceeds to apply it for modeling CNER in text. Existing NER models predominantly focus on surface-level lexical recognition and often fail to capture the deeper semantic features inherent in Chinese characters. This work leverages the parallel computation capability of the attention mechanism to extract global semantic features of Chinese characters and further utilizes CNNs to perform high-precision local extraction of salient features, thereby improving both the accuracy and efficiency of Chinese NER. The character-level embeddings are trained within the BERT framework to enhance the expressive capacity of Chinese representations. BERT is a bidirectional language model capable of performing classification, question answering, and other natural language processing tasks [21, 22]. In this study, the BERT-Base Chinese model is adopted along with its built-in WordPiece tokenizer, which segments the original Chinese character stream into subword units and maps them to vocabulary indices. No additional stopword filtering is applied, and all function words and grammatical particles are retained during training to preserve the full semantic context. The expression for each character vector after BERT training is shown in Equation (8).

$$\begin{cases} e_b = \left(e_{b1}, e_{b2}, \ldots, e_{bn}\right) = BERT\left(s_1, s_2, \ldots, s_n\right) \\ S = \left\{s_1, s_2, \ldots, s_n\right\} \end{cases} \tag{8}$$

In Equation (8), S represents a sentence, n represents the length of a sentence, and s_n and e_b represent the low-dimensional character vector and the character vector obtained after training, respectively. Then, the Attention-CNN CNER model will be established, and the model architecture is shown in Figure 3.

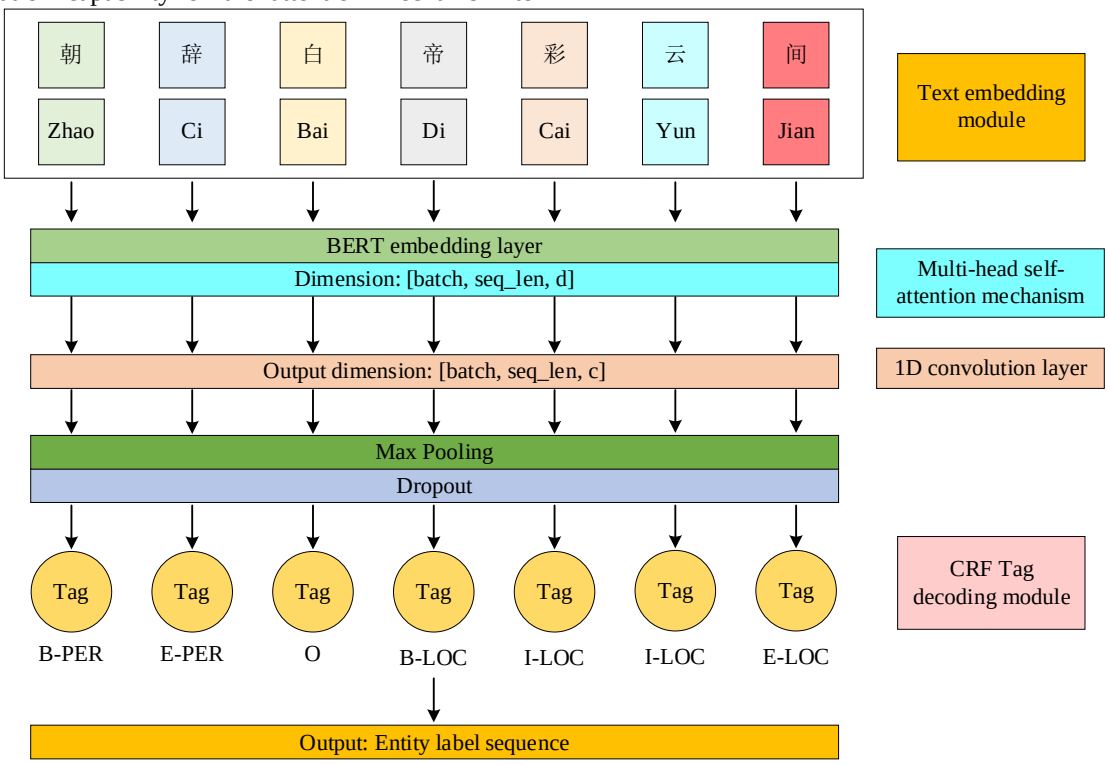


Figure 3: Attention-CNN model architecture for CNER.

In Figure 3, the Attention-CNN model consists of four modules. First, the Chinese character and pinyin vectors are input into the Chinese semantic feature embedding module, where the four tones in Chinese are represented by [1, 4] for tone values. Then, the attention mechanism is used to globally and parallelly compute the

attention scores for the features such as the Chinese characters and pinyin, allowing more accurate extraction of global semantic features. Next, in the CNN phase, the Chinese semantic features undergo convolution operations, and the final output of the Chinese character labels is sent to the CRF for decoding. In the CRF, PER denotes person names, B indicates the beginning of a label, I represent the intermediate stage of the label, and E signifies the end of the label, O denotes non-entity tokens such as prepositions, while LOC represents location names. The expression for the vector after the character feature fusion operation is shown in Equation (9).

$$E_i = concat\left(\left[e_i^c, e_i^p\right]\right) \tag{9}$$

In Equation (9), e_i^c and e_i^p represent the corresponding character and pinyin vectors of Chinese. The output expression obtained after h attention heads of Multi-Head Self-Attention undergo linear transformation is shown in Equation (10).

$$\begin{cases} h_i = f\left(W_i^q q, W_i^k k, W_i^v v\right) \\ Multihead\left(h\right) = W_o \begin{bmatrix} h1 \\ \vdots \\ hn \end{bmatrix} \end{cases} \tag{10}$$

In Equation (10), W and f represent the learnable parameter matrix and scaled dot-product attention, respectively. Next, convolution operations are performed to connect the convolutional layers, followed by information fusion, as shown in Equation (11).

$$\begin{cases} S' = Conv1D\left(hs_T^P\right) \\ S'' = GeLU\left(Maxpool1D\left(S + S'\right)\right) \end{cases} \tag{11}$$

In Equation (11), S and S' represent the 1D convolution and max pooling, respectively, while S'' represents the entity scoring matrix after convolution processing. After updating the weights, the results obtained by parallel computations in the MHSA are concatenated, as shown in Equation (12).

$$Multi-Head\left(Q,K,V\right) = \left(Head_j \oplus \ldots \oplus Head_h\right) \tag{12}$$

After the computation in Equation (12), higher-precision text content features are obtained. The optimal sequence decoded by the CRF is shown in Equation (13).

$$y^* = \arg\max_y \sum_{i=1}^{n} \left(A_{y_{i-1}, y_i} + P_{i, y_i}\right) \tag{13}$$

In Equation (13), A denotes the transition matrix and P represents the emission score matrix. The Attention-CNN CNER model progressively extracts entity-related features from the text while effectively capturing global key semantic information, thereby enhancing the accuracy of label classification. In specific textual domains, the NER task is often inherently related to other tasks; however, traditional models typically focus on single-task learning. NER naturally correlates with tasks such as entity type classification and sentiment polarity detection. To improve the model's comprehension of semantic nuances, this study adopts a multi-task learning framework, which shares parameters in both the attention and convolutional layers while jointly optimizing multiple related tasks. Therefore, in the final model, MTL is introduced to improve the recognition accuracy, even with limited data. The MTL framework is shown in Figure 4 [23, 24].

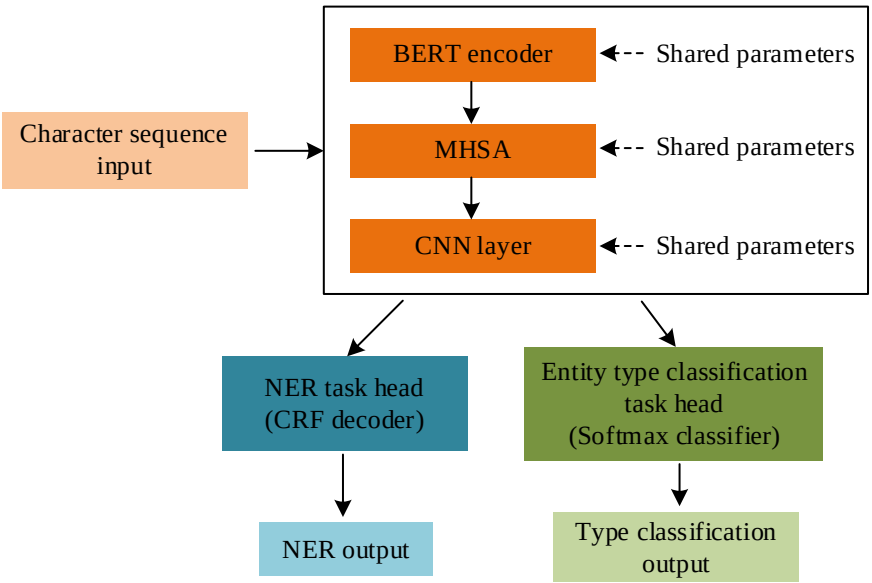


Figure 4: MTL framework with a shared backbone and task-specific branches.

As shown in Figure 4, the MTL module uses soft parameter sharing technology. Although tasks share the underlying feature extraction parts of the convolutional layer and attention mechanism, tasks such as NER, sentiment analysis, and text classification typically have specific output layers and task goals. Therefore, they require independent, task-specific parameters. This design allows for the use of shared lower-level feature

extraction capabilities while ensuring that the individual requirements of each task are met, rather than using hard parameter sharing where all tasks would use the same network layers and weights. When constructing the final model, a Dropout layer is typically added within the CNN framework to prevent overfitting. The Dropout operation is expressed in Equation (14).

$$C_{dropout} = C_{max} \cdot \text{Dropout}(p) \quad (14)$$

In Equation (14), p represents the dropout probability. The operational flow of the AC-MTL document CNER model, which combines Attention-CNN and MTL techniques, is shown in Figure 5.

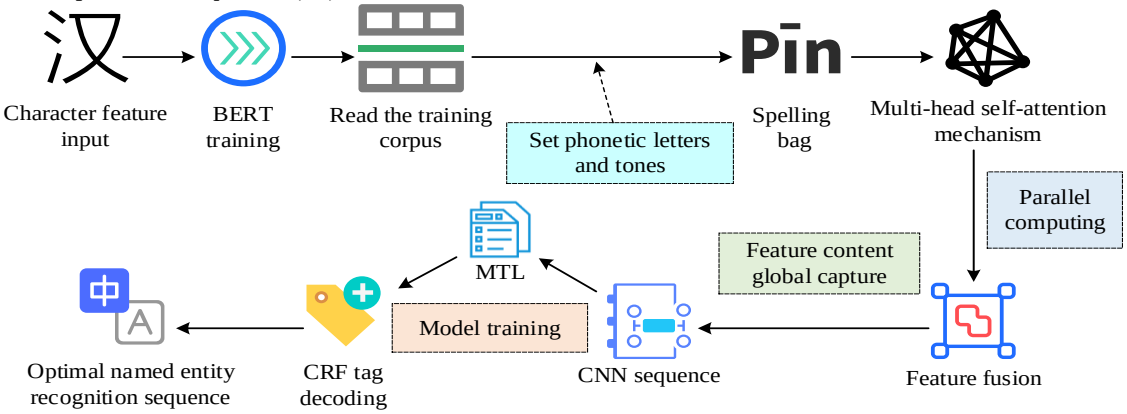


Figure 5: Flowchart of AC-MTL operation combining Attention-CNN and MTL techniques.

As shown in Figure 5, when the AC-MTL model is running, the first step is to input Chinese character features, which are pre-trained in the BERT module. After that, the trained character features proceed to the next step of corpus reading, where pinyin letters and tone values need to be set. Then, during the Attention phase, parallel computation of attention scores, dot-product scaling, and other operations are performed, followed by feature fusion. After Attention extracts the global semantic features, the data proceeds to the CNN sequence stage. Next, after convolution and max-pooling operations, feature combinations and transformations are processed through the fully connected layer. The model is then trained with multi-task learning, followed by CRF label decoding to obtain the final optimal Chinese named entity sequence.

4 Performance analysis of CNER model integrating attention and CNN with MTL

4.1 Comprehensive evaluation of AC-MTL on Chinese NER Tasks

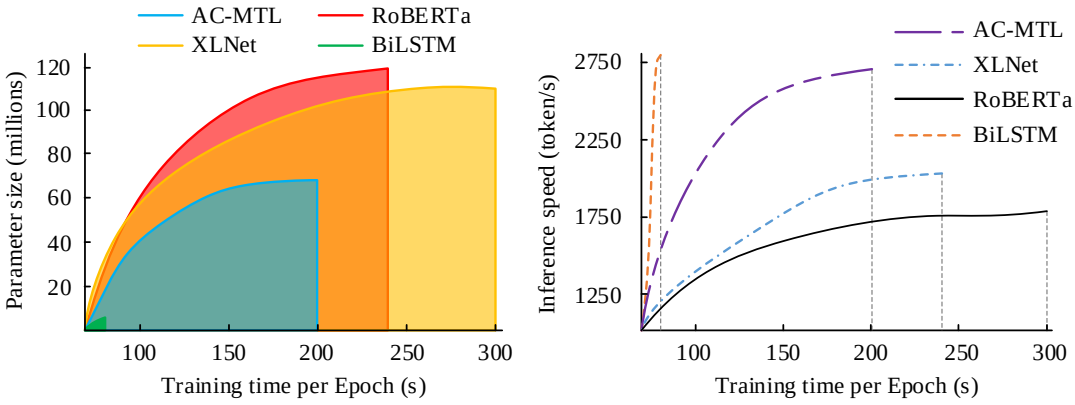
To evaluate the effectiveness of the proposed AC-MTL model, experiments were conducted on a workstation equipped with an Intel Xeon Gold 6248R processor, 128 GB of memory, and an NVIDIA Tesla V100 GPU. The operating system used was Ubuntu 18.04, with PyTorch 1.8.1 as the deep learning framework, CUDA version 11.1, and driver version 450.80.02. The hyperparameters of the AC-MTL model were set based on prior empirical studies and experimental validation. The learning rate was set to $5e^{-5}$, a common starting value for BERT fine-tuning which ensured stable convergence without needing extensive tuning. The batch size was set to 32 to balance training efficiency and GPU memory constraints. The number of training epochs was set to 50, with early stopping applied to prevent overfitting. The dropout rate was set to 0.1, which was the standard value used in Transformer architectures to prevent overfitting. Additionally, the initial weight of the primary task in the multi-task loss function was set to 0.7 to emphasize its central role. This value was empirically validated to deliver favorable performance across multiple experimental settings. The Adam optimizer was employed due to its fast convergence and stability, making it a mainstream choice for deep learning tasks and particularly suitable for Transformer-based text modeling.

To assess the model's performance in Chinese named entity recognition, AC-MTL was compared against baseline CNER models based on BiLSTM, RoBERTa, and XLNet architectures. For a fair comparison, RoBERTa and XLNet were fine-tuned by adding a CRF decoding layer and training with BIO-labeled sequences on the same dataset, to meet the requirements of the NER task. The experiments utilized the Weibo dataset and a subset of the Microsoft Research Asia (MSRA) dataset. The MSRA subset contained 200 samples specifically selected to evaluate performance under low-resource conditions. A stratified sampling strategy was adopted to divide each dataset, ensuring that the distribution of named entity labels was consistent across the training set (70%), validation set (15%), and test set (15%). Four models were used for Chinese NER, and their performance was measured by precision, recall, and F1 score. The results are presented in Table 2.

Table 2: Overall performance comparison of CNER models

Sample size	Model	Precision (%)	Recall (%)	F1 Score (%)
200	AC-MTL	98.62	98.10	98.36
	XLNet	96.21	95.89	95.54
	RoBERTa	88.53	85.37	86.92
	BiLSTM	82.48	82.64	82.73
1000	AC-MTL	98.40	97.90	98.15
	XLNet	97.16	96.29	96.64
	RoBERTa	88.24	88.71	89.44
	BiLSTM	81.76	81.46	82.17
3000	AC-MTL	98.54	98.19	98.79
	XLNet	97.46	95.27	96.26
	RoBERTa	86.89	87.04	87.28
	BiLSTM	82.40	81.67	81.64

As shown in Table 2, under sample sizes of 200, 1000, and 3000, the AC-MTL model consistently maintained a leading position across all three-performance metrics: precision, recall, and F1 score. In the low-resource setting with only 200 samples, AC-MTL achieved a precision of 98.62%, recall of 98.10%, and an F1 score of 98.36%, significantly outperforming XLNet, RoBERTa, and BiLSTM, showing its strong adaptability to limited data. When the sample size increased to 1000, AC-MTL still maintained the highest precision at 98.40% and recall at 97.90%, resulting in an F1 score of 98.15%, which was notably higher than RoBERTa's 89.44% and BiLSTM's 82.17%. It was also worth noting that although XLNet showed some improvement in recall under medium- to high-resource settings, its overall precision stability and combined performance stayed below AC-MTL's. This suggested that AC-MTL achieved a better balance between high-accuracy recognition and error tolerance. Subsequently, the study evaluated the model's runtime efficiency, as illustrated in Figure 6.



(a) Model size of different models (b) Inference time of different models

Figure 6: Inference speed and parameter size comparison among CNER models.

As shown in Figure 6(a), AC-MTL maintained high performance while keeping its parameter size at 64 million, which was significantly smaller than RoBERTa's 118 million and XLNet's 105 million—representing a model compression rate exceeding 40% relative to both. In contrast, although BiLSTM had the smallest parameter size, it lacked deep semantic modeling capability and thus exhibited functional limitations. Through the integration of modular structures and task-guided mechanisms, AC-MTL effectively reduced redundant parameters while preserving both global semantic understanding and local feature representation, achieving a well-balanced trade-off between structural compactness and expressive power. In Figure 6(b), the inference efficiency of each model was further compared using token-level processing speed as the evaluation metric. AC-MTL reached a throughput of 2618 tokens per second, demonstrating significantly faster inference than RoBERTa and XLNet, and approaching the speed of the lightweight BiLSTM model. This improvement was primarily attributed to the introduction of convolutional modules and task decoupling optimizations within the encoding structure of AC-MTL, which collectively enhanced computational efficiency during inference. Overall, AC-MTL exhibited superior performance in both parameter compactness and inference speed. These are two key factors for real-world deployment, making it a practical and deployable solution for resource-constrained environments.

4.2 Ablation study: validating the structural design of AC-MTL

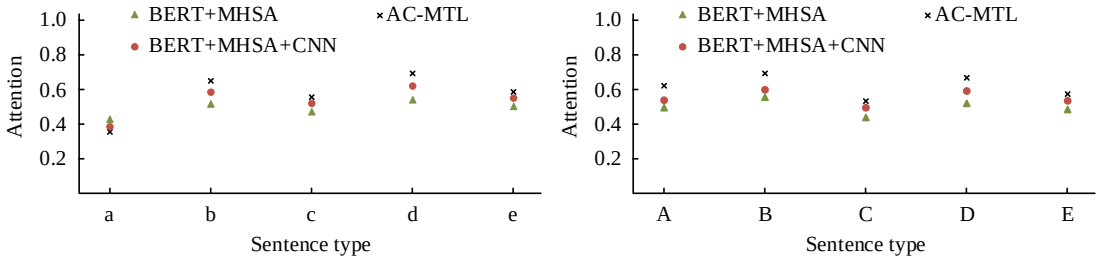
To verify the actual contribution of each core structural module within the AC-MTL model to overall

performance, a systematic ablation study was conducted to compare the effects of different component combinations, as detailed in Table 3.

Table 3: Ablation results of AC-MTL on key structural components

Model architecture	Precision (%)	Recall (%)	F1-score (%)
BERT+CNN	95.14	94.12	94.75
BERT+MHSA	96.23	94.85	95.32
BERT+MHSA+CNN	97.37	96.22	96.74
AC-MTL	99.54	98.19	98.79

As shown in Table 3, the AC-MTL model achieved the highest performance when all structural components were retained, with a precision of 99.54%, a recall of 98.19%, and an F1 score of 98.79%. Compared to the model with only the BERT+MHSA+CNN structure, the F1 score increased by 2.05 percentage points, indicating that the multi-task learning mechanism significantly improved overall performance. In contrast, simplified models that retained only the CNN or MHSA module yielded F1 scores of 94.75% and 95.32%, respectively—substantially lower than the full AC-MTL configuration. This suggested that relying solely on local feature extraction or global semantic modeling was insufficient and that the synergy of module integration was critical for optimal performance. The study further evaluated attention scores across three model structures using two sentence segments. Segments a, b, c, d, and e corresponded to the Chinese sentences: He is a Beijinger. He graduated from Beijing Jiaotong University. He still works in Beijing. The pace of development in Beijing is fast. I also want to study in Beijing. Segments A, B, C, D, and E represented Chinese sentences: Innovation is the core driving force behind enterprise development. Only through continuous exploration of new technologies and new models could one stand out in the fierce market competition. In the field of scientific research, innovation meant breaking free from the constraints of conventional thinking and capturing every spark of inspiration that could lead to transformative change with keen insight. The essence of education lay in cultivating innovative talents. Through diversified curricula and practical activities, students' creativity and spirit of exploration could be effectively stimulated. The sustainable development of cities could not proceed without the integration of innovative concepts. From the application of green energy to the construction of intelligent transportation systems, the power of innovation was evident everywhere. Cultural heritage required innovative expression. By leveraging digital technology and interdisciplinary fusion, traditional culture could be revitalized and given new life in the modern era. The detailed comparative results were illustrated in Figure 7.



(a) The attention weight assigned to the token Beijing(Chinese)

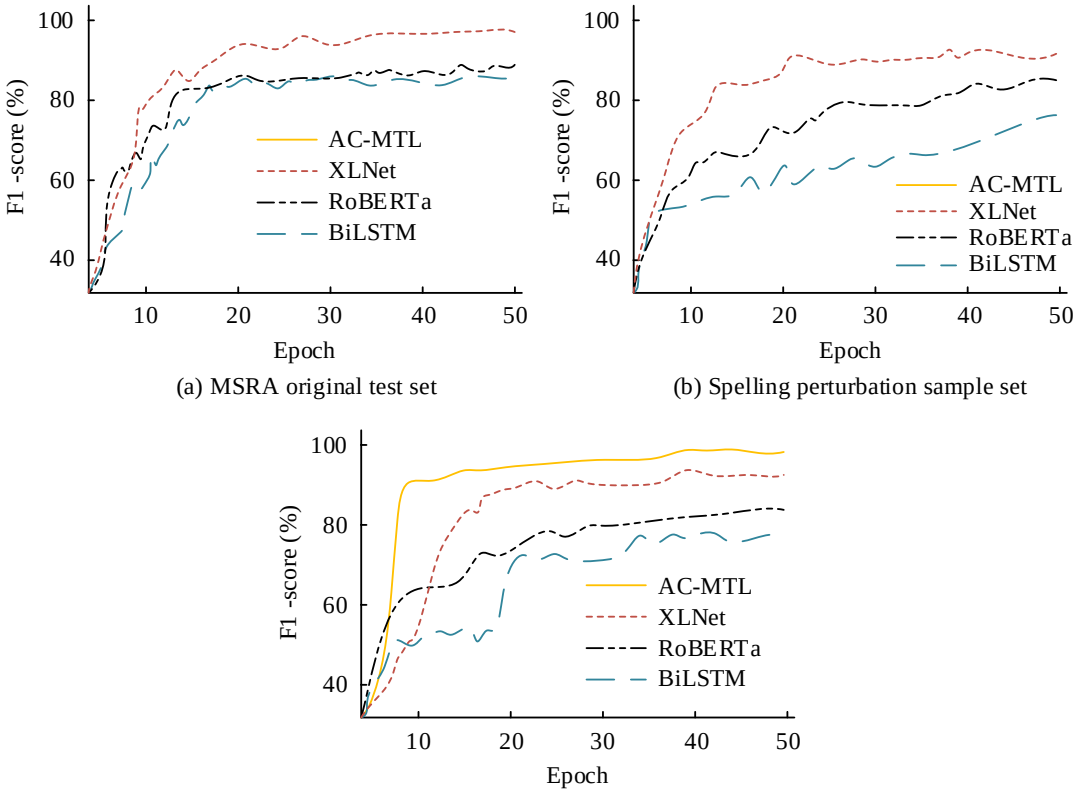
(b) The attention weight assigned to the token Innovation(Chinese)

Figure 7: Attention score comparison for tokens under different model architectures.

In Figure 7(a), the AC-MTL model consistently achieved higher attention scores across all positions compared to the other two baseline models. Notably, it reached 0.63 in sentence b and 0.67 in sentence d, demonstrating more precise semantic recognition of nested entities and core thematic terms. In contrast, the BERT+MHSA model exhibited relatively uniform attention distribution toward the token Beijing, lacking focused differentiation, while the addition of CNN introduced some improvement but still fell short of the structural enhancement achieved by AC-MTL. Overall, AC-MTL demonstrated stronger discriminative capacity and contextual understanding in allocating attention to high-frequency geographical terms under polysemous conditions. In Figure 7(b), AC-MTL exhibited the strongest semantic focus in all contexts. Specifically, it scored 0.68 in sentence B ("scientific thinking") and 0.66 in sentence D ("sustainable development"), surpassing BERT+MHSA in both cases. This indicated that the model had a superior ability to capture the semantic salience of abstract policy-related terms within complex syntactic structures. Notably, even in peripheral semantic scenarios such as "mode of expression," AC-MTL maintained a relative advantage, whereas BERT+MHSA achieved only 0.46. These overall trends suggested that AC-MTL possessed enhanced contextual aggregation and semantic stability when dealing with abstract, highly context-dependent lexical disambiguation, thereby validating the effectiveness of its structural design in recognizing semantically ambiguous words.

4.3 Interpretability and robustness analysis of AC-MTL

To further evaluate the stability and interpretability of the AC-MTL model in practical applications, the study conducted robustness analysis under various types of perturbation scenarios. The model was tested on datasets with noise-injected samples derived from the original corpus, and the resulting F1 scores were shown in Figure 8.



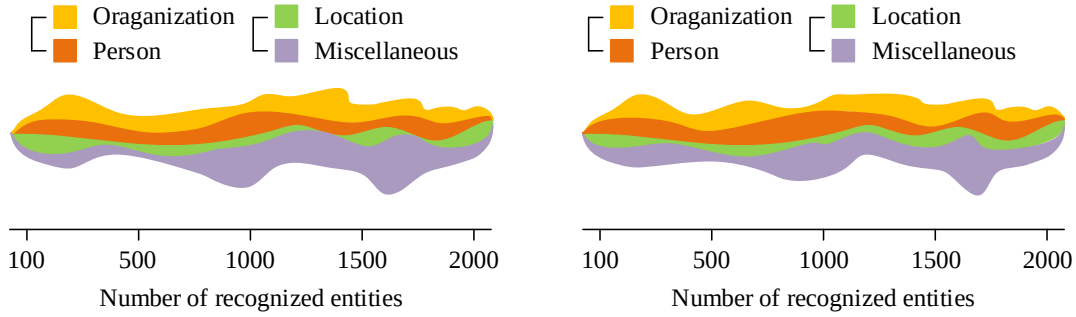
(a) MSRA original test set

(b) Spelling perturbation sample set

(c) Weibo short texts

Figure 8: Robustness of AC-MTL under noisy and informal input conditions.

In Figure 8(a), during training on the standard MSRA test set, the F1 score of the AC-MTL model increased rapidly within the first five epochs and stabilized, eventually converging at 98.79%, significantly outperforming XLNet (96.26%) and RoBERTa (87.28%). This demonstrated both a faster convergence speed and a higher performance ceiling. Notably, AC-MTL reached its major performance plateau by epoch 10, whereas the baseline models required at least 20 epochs to approach a similar level. This indicating that AC-MTL's structural design was more efficient in capturing semantic features and entity boundaries. In Figure 8(b), despite larger fluctuations during training on the spelling-perturbed dataset, AC-MTL maintained strong stability and noise resistance, with a final F1 score of 96.54%, substantially higher than other models. In Figure 8(c), on the Weibo short-text dataset, AC-MTL almost fully converged after just four epochs and stabilized at an F1 score of 98.35%. In contrast, XLNet achieved only 92.78% on this dataset and showed considerable volatility throughout training, reflecting its limited adaptability to unstructured and contextually ambiguous language. Supported by multi-task learning signals, AC-MTL exhibited superior contextual modeling capabilities, allowing it to maintain high recognition accuracy and convergence stability even under fragmented input conditions. Figure 9 presented a validation of AC-MTL's performance in identifying different thematic categories in legal text cases.



(a) AC-MTL (b) Actual Statistics

Figure 9: Distribution of recognized thematic entities in legal texts.

As shown in Figure 9(a), the vertical axis represented the number of entities mentions identified as belonging to each thematic category, while the theme river illustrated the aggregation trend of topic recognition as the text progressed. The thematic stream recognized by the AC-MTL model across 2,000 legal text cases was shown, with each colored band represented a different theme category identified by the model, including organizations, person names, locations, and domain-specific terms. The AC-MTL model demonstrated the ability to accurately identify and distinguish between different types of entities, with a smooth distribution of recognized themes that effectively covered a wide range of entity categories present in the text. Figure 9(b) presented the actual thematic distribution across the 2,000 legal text cases. The distribution generated by the AC-MTL model closely matched the true distribution, with minimal fluctuations between the two. As the number of cases increased, the recognition trends became increasingly aligned. This indicated that the AC-MTL model had strong recognition capabilities and was able to extract and differentiate themes from complex texts with high accuracy, further underscoring its effectiveness in the context of legal document analysis. Finally, to further investigate the limitations of the model, an error analysis was conducted by categorizing 100 misclassified samples produced by the AC-MTL model. The distribution of common misclassification types was presented in Figure 10.

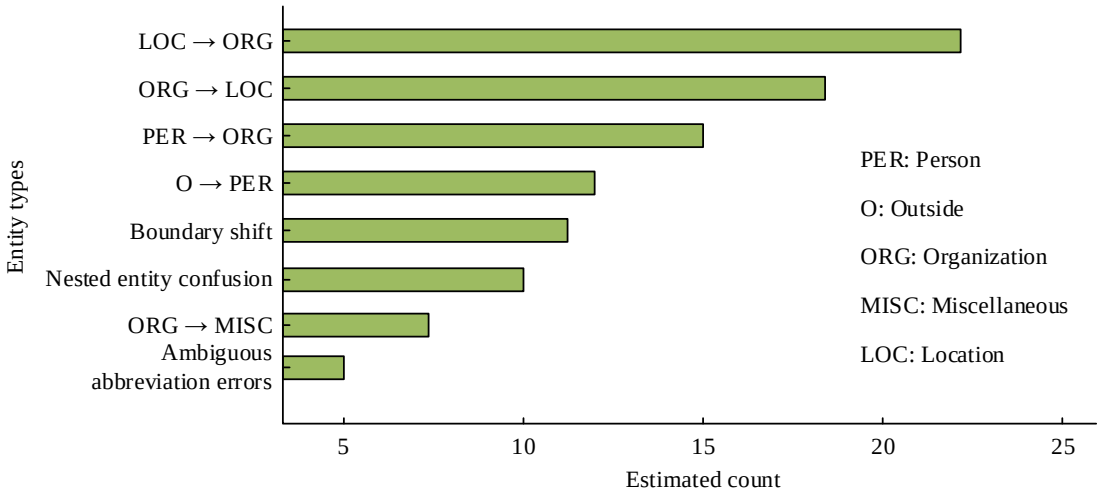


Figure 10: Distribution of common misclassification types in AC-MTL output.

As illustrated in Figure 10, the most frequent misclassification types involved confusion between location and organization entities (LOC ⇔ ORG), which reflected the semantic overlap in Chinese place names and institutional titles. Errors related to person names also occurred, especially when user handles or role-based nicknames were interpreted as named entities. Boundary-related mislabeling and nested entity conflicts occurred in complex expressions, suggesting that further improvements in fine-grained boundary detection might be necessary.

5 Discussion

The proposed AC-MTL model demonstrated outstanding performance in CNER tasks, particularly in maintaining high robustness and accuracy when faced with limited data resources and noisy textual environments. Experimental results confirmed that the integration of attention mechanisms with convolutional neural networks, along with the adoption of a multi-task learning strategy, effectively compensated for the deficiencies of traditional methods in modeling long-range dependencies while enhancing local feature extraction capabilities. This reflected the model's

structural design in terms of both scientific rigor and engineering practicality. Firstly, performance evaluations indicated that AC-MTL consistently outperformed mainstream baseline models such as XLNet, RoBERTa, and BiLSTM across datasets of varying sizes. Even in a low-resource scenario with only 200 samples, the model achieved an F1 score of 98.36%, showing strong generalization capability in data-scarce conditions. Secondly, in terms of inference efficiency, AC-MTL attained a processing speed of 2618 tokens per second, approaching that of the lightweight BiLSTM while significantly surpassing both XLNet and RoBERTa, thereby highlighting its computational advantage for real-world deployment.

The effectiveness of individual modules within the AC-MTL architecture was further validated through ablation experiments. While combinations such as BERT+CNN or BERT+MHSA showed some recognition ability, they fell short in holistic semantic modeling and precise feature localization. Only through the synergistic integration of BERT, MHSA, and CNN—each enhanced by a multi-task learning framework—enabled the model to achieve substantial performance gains. This soft parameter-sharing MTL framework enabled information sharing across multiple subtasks such as entity boundary recognition and type classification, significantly enhancing semantic discrimination capability. Visualization of attention weights revealed that AC-MTL was particularly adept at capturing syntactic and semantic cores when processing polysemous and abstract lexical items (e.g., “Beijing” or “innovation”), showing clearer focus compared to BERT+MHSA and BERT+CNN structures. Moreover, the model’s stable performance on Weibo short texts and spelling-perturbed corpora demonstrated its adaptability to unstructured input, making it suitable for real-world applications such as social media analysis and legal document mining.

Nevertheless, certain limitations remained. In contexts with highly sparse information or pronounced semantic ambiguity, the model still suffered from inaccurate boundary detection or entity type confusion. Additionally, although AC-MTL exhibited strong generalization, its reliance on large-scale pre-trained models like BERT posed challenges for deployment in resource-constrained environments, necessitating further compression and optimization. In conclusion, AC-MTL excelled in both theoretical design and empirical performance, offering an efficient, robust, and extensible approach to Chinese named entity recognition. Given its modular architecture and strong performance in capturing both global semantics and local features, the AC-MTL model held significant potential for adaptation across multilingual NER tasks and domain-specific applications such as biomedical text mining, cross-lingual knowledge extraction, and low-resource language processing, where robust entity recognition remained a persistent challenge.

6 Conclusion

To address the limitations of existing methods in handling complex textual environments, this study proposes a Chinese named entity recognition approach that integrates attention mechanisms with convolutional neural networks, and further designs the AC-MTL model by incorporating BERT and multi-task learning techniques for legal document entity recognition. On the standard MSRA test set, the AC-MTL model achieved an F1 score of 98.79%, and on a spelling-perturbed sample set, it reached an F1 score of 96.54%, both outperforming the baseline models XLNet and RoBERTa. When applied specifically to legal document cases, the thematic distribution recognized by the model across 2,000 samples closely matched the actual distribution, demonstrating its strong potential for domain-specific applications and generalization. Although the current method performs well in recognizing named entities in long-form texts, it may still encounter errors in scenarios with high semantic ambiguity or sparse contextual information. Future optimization may proceed in two directions: first, by incorporating larger and more domain-adapted pre-trained language models for targeted fine-tuning; and second, by exploring the integration of external knowledge graphs or entity linking mechanisms to enhance its practical applicability in tasks such as question answering, information extraction, and sentiment analysis.

References

[1] Abdullah M H A, Aziz N, Abdulkadir S J, Alhussian H S A, Talpur N. Systematic literature review of information extraction from textual data: recent methods, applications, trends, and challenges. IEEE Access, 2023, 11(1): DOI: 10535-10562. 10.1109/ACCESS.2023.3240898

[2] Pan X, Xue Y. Advancements of artificial intelligence techniques in the realm about library and information subject—A case survey of latent Dirichlet allocation method. IEEE Access, 2023, 11(2): 132627-132640. DOI: 10.1109/ACCESS.2023.3334619

[3] Shishehgarkhaneh M B, Moehler R C, Fang Y, Hijazi A A, Aboutorab H. Transformer-Based named entity recognition in construction supply chain risk management in Australia. IEEE Access, 2024, 12(3): 41829-41851. DOI: 10.1109/ACCESS.2024.3377232

[4] Almutiri T, Nadeem F. Markov models applications in natural language processing: a survey. I.J. Information Technology and Computer Science, 2022, 2(1): 1-16. DOI: 10.5815/ijitcs.2022.02.01

[5] Eker K, Pehlivanoğlu M K, Eker A G, Syakura M A, Duru N. A comparison of grammatical error correction models in English writing. IEEE Access, 2023, 56(13): 218-223. DOI: 10.1109/UBMK59864.2023.10286642

[6] Yu Z, Shi X, Zhang Z. A multi-head self-attention transformer-based model for traffic situation prediction in terminal areas. IEEE Access, 2023, 11(7): 16156-16165. DOI: 10.1109/ACCESS.2023.3245085

[7] Xiong W. Web News Media retrieval analysis integrating with knowledge recognition of semantic grouping vector space model. Informatica, 2024, 48(5): 41-54. DOI: 10.31449/inf. v48i5.5377

[8] Peng F, McCallum A. Information extraction from research papers using conditional random fields. Information Processing & Management, 2006, 42(4): 963-979. DOI: 10.1016/j.ipm.2005.09.002

[9] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991, 2015. DOI: 10.48550/arXiv.1508.01991

[10] Devlin J, Chang M W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), Minneapolis, Minnesota, 2019: 4171-4186. DOI: 10.18653/v1/N19-1423

[11] Chayintr T, Kamigaito H, Funakoshi K, Okumura M. Latte: Lattice attentive encoding for character-based word segmentation. Journal of Natural Language Processing, 2023, 30(2): 456-488. DOI: 10.5715/jnlp.30.456

[12] Patel M, Ezeife C I. BERT-based multi-task learning for aspect-based opinion mining. International conference on database and expert systems applications, Springer International Publishing, Cham, 2021: 192-204. DOI: 10.1007/978-3-030-86472-9_18

[13] Zhai Z, Fan R, Huang J, Xiong, Zhang L, Wan J, Zhang L. A named entity recognition method based on knowledge distillation and efficient Global Pointer for Chinese medical texts. IEEE Access, 2024, 12(4): 83563-83574. DOI: 10.1109/ACCESS.2024.3405997

[14] Ranjan R, Daniel A K. CoBiAt: A sentiment ClassificatiCobiat: A sentiment classification model using hybrid Convnet-Dual-lstm with attention mechanismon model using hybrid ConvNet-Dual-LSTM with attention mechanism. Informatica, 2023, 47(4): 523-536. DOI: 10.31449/inf. v47i4.3911

[15] Yang C, Sheng L, Wei Z, Wang W. Chinese named entity recognition of epidemiological investigation of information on COVID-19 based on BERT. IEEE Access, 2022, 10(3): 104156-104168. DOI: 10.1109/ACCESS.2022.3210119

[16] Zhao P, Wang W, Liu H, Han M. Recognition of the agricultural named entities with multifeature fusion based on Albert. IEEE Access, 2022, 10(9): 98936-98943. DOI: 10.1109/ACCESS.2022.3206017

[17] Gonçalves T, Rio-Torto I, Teixeira L F, Cardoso J S. A survey on attention mechanisms for medical applications: are we moving toward better Algorithms? IEEE Access, 2022, 10(7): 98909-98935. DOI: 10.1109/ACCESS.2022.3206449

[18] Yang Z, Ma J, Chen H, Zhang J, Chang Y. Context-aware attentive multilevel feature fusion for named entity recognition. IEEE transactions on neural networks and learning systems, 2022, 35(1): 973-984. DOI: 10.1109/TNNLS.2022.3178522

[19] Biswas S, Poornalatha G. Opinion mining using multi-dimensional analysis. IEEE Access, 2023, 11(5): 25906-25916. DOI: 10.1109/ACCESS.2023.3256521

[20] Haque M Z, Zaman S, Saurav J R, Haque S, Islam M S, Amin M R. B-ner: A novel bangla named entity recognition dataset with largest entities and its baseline evaluation. IEEE Access, 2023, 11(1): 45194-45205. DOI: 10.1109/ACCESS.2023.3267746

[21] Liu Y, Wen F, Zong T, Li T. Research on joint extraction method of entity and relation triples based on hierarchical cascade labeling. IEEE Access, 2022, 11(3): 9789-9798. DOI: 10.1109/ACCESS.2022.3232493

[22] Rafi T H, Ko Y W. HeartNet: Self multihead attention mechanism via convolutional network with adversarial data synthesis for ECG-based arrhythmia classification. IEEE Access, 2022, 10(7): 100501-100512. DOI: 10.1109/ACCESS.2022.3206431

[23] Chen X, Cong P, Lv S. A long-text classification method of Chinese news based on BERT and CNN. IEEE Access, 2022, 10(5): 34046-34057. DOI: 10.1109/ACCESS.2022.3162614

[24] Shen Y, Liu Q, Fan Z, Liu J, Wumaier A. Self-supervised pre-trained speech representation based end-to-end mispronunciation detection and diagnosis of Mandarin. IEEE Access, 2022, 10(6): 106451-106462. DOI: 10.1109/ACCESS.2022.3212417