Causality-aware Deconfounded NER Network for Chinese Named Entity Recognition

Anonymous ACL submission

Abstract

Chinese Named Entity Recognition (Chinese NER) faces challenges such as ambiguous entity boundaries and limited classification accuracy, mainly due to the lack of clear word boundaries and the strong coupling between semantic features and interfering features. To address the interference caused by confounding factors like domain bias, this paper proposes a Causality-aware Deconfounded NER network (CDNER). By integrating multi-granularity feature extraction, causal deconfounding, and prototype learning into the CRF (Conditional Random Field) model, the network enhances the model's recognition accuracy and outputs more robust entity representations. Experimental results demonstrate that CDNER achieves performance close to the current state-of-the-art, especially excelling in complex text environments.

1 Introduction

Named Entity Recognition (NER) is crucial for many practical applications in Natural Language Processing (NLP). Chinese Named Entity Recognition faces several unique challenges, such as domain bias, annotation noise, significant granularity differences, and ambiguous boundaries of nested entities. To address these issues, this paper proposes the CDNER model by combining causal inference and multi-granularity feature enhancement techniques. It captures contextual information of features at different granularities using multiple parallel convolutional sub-networks, then separates interfering features through a confounder encoding module. Meanwhile, it maintains the prototype vector of each entity category through a prototype learning module, which enhances the class discriminability of semantic features. When outputting the entity recognition results, the model combined the boundary prediction with the CRF layer: the boundary head outputs boundary probabilities and offsets to optimize entity span localization, while the CRF layer achieves global consistency in sequence labeling through label transition constraints. In summary, the main contributions of this paper are as follows: it initially explores the performance improvement of removing confounders on the Chinese Named Entity Recognition task, and based on this, an effective multi-technology coordinated Chinese NER method is designed to achieve accurate entity capture in complex contextual environments.

2 Related Work

The Named Entity Recognition(NER) task serves as a crucial downstream task for numerous Natural Language Processing tasks, such as entity relation extraction(Cheng et al., 2021), entity linking(Gu et al., 2021), entity coreference resolution(Clark and Manning, 2016), and knowledge graph construction(Ji et al., 2022). Traditional Chinese NER methods mainly include sequence labeling(Huang et al., 2015; Lample et al., 2016; Ma and Hovy, 2016) and span classification(Li et al., 2021a; Yu et al., 2020). Huang et al(Huang et al., 2015)were the first to employ a Bidirectional Long Short-Term Memory(BiLSTM) model combined with a Conditional Random Field(CRF) for label prediction and annotation. Due to its excellent performance across multiple datasets, the BiLSTM-CRF architecture has been adopted in many subsequent related studies(Lample et al., 2016; Ma and Hovy, 2016) and has become a classic model.

In recent years, the integration of deep learning-based pre-trained language models such as ELMo(Peters et al., 2018) and BERT(Devlin et al., 2019) has improved the performance of Chinese NER tasks, yet several limitations remain: The FLAT model(Li et al., 2020) converts the lattice structure into a flat structure consisting of spans. Each span corresponds to a character or latent word and its position in the original lattice. breaking

through the efficiency bottleneck of the lattice structure. However, it ignores the nested associations of entities, leading to the loss of nested correlations between fine-grained vocabulary and entities. The RICON model(Gu et al., 2022) propose a simple but effective method for investigating the regularity of entity spans in Chinese NER, dubbed as Regularity-Inspired reCOgnition Network. Addresses the contradiction between utilizing internal entity rules and balancing boundary localization by designing a dual-module collaborative architecture. Nevertheless, the correlation between its two branches may cause errors in the rule branch to affect boundary judgment, and it lacks an explanation for the causality of entity boundary decisions. The LEBERT model(Liu et al., 2021) integrates external lexicon knowledge into BERT layers directly by a Lexicon Adapter layer, breaking the limitation of traditional methods that simply perform shallow concatenation of lexicon features and BERT features without fully leveraging the deep semantics of pre-trained models. However, its performance relies on the quality of the constructed lexicon. The W2NER model(Li et al., 2022) present a novel alternative by modeling the unified NER as word-word relation classification, namely W2NER. The architecture resolves the kernel bottleneck of unified NER by effectively modeling the neighboring relations between entity words with Next-Neighboring-Word(NNW) and Tail-Head-Word-*(THW-*) relations. However, this word-word relation increases the cost of data annotation. The BOPN model(Tang et al., 2023) predicts the boundary offsets between candidate spans and their nearest entity spans. such as sample imbalance and insufficient association modeling, through boundary offset prediction. Yet, this model fails to optimize the differentiation of entity types, making it prone to confusing similar entity types. The GNER model(Ding et al., 2024), a Generative NER system that shows improved zero-shot performance across unseen entity domains. However, the inherent flaw in its negative instance generation mode may cause the model to over-learn features of negative samples while ignoring those of positive samples.

Existing studies have achieved breakthroughs in areas such as entity rule utilization, span modeling, efficiency optimization, and knowledge integration. However, limitations persist in aspects like confounder interference(Zhang et al., 2024; Zeng et al., 2025), the depth of integration between multi-granularity(Su et al., 2024) and entity type

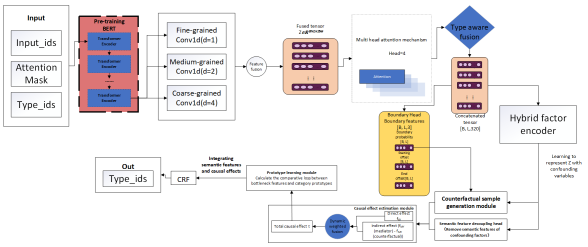

Figure 1: Architecture Diagram of the CDNER Model.The CDNER model consists of three sub-modules, namely the multi-granularity feature extraction module, the causal deconfounding module, and the prototype contrastive learning module.

information, and the collaboration between prototype learning(Li et al., 2021b) and CRF. These challenges provide directions for improvements in the design of the CDNER model proposed in this paper.

3 Methodology

CDNER model receives text sequence input, which is encoded by BERT. Then, it extracts features via Multi-Scale Convolution(Multi-Scale-Conv). These extracted features are concatenated with the original features for causal deconfounding and semantic decoupling. From the deconfounded semantics, entity prototypes are extracted for prototype contrastive learning, and the contrastive loss is calculated. Finally, sequence decoding and output are performed through the CRF layer. The overall architecture of the model is illustrated in Figure. 1.

3.1 Data Processing

In this section, the model first reads the raw BIO-format data stored in text documents. The data follows a specific structure: each line corresponds to one character and one label, with the character and label separated by a Tab; each sample segment is separated by a blank line.

Next, the read_ner_data function stores the text and labels of the samples separately, forming sub-sample pairs in the format of (character, label). Subsequently, a custom NERDataset class is responsible for data preprocessing and format conversion, which includes converting characters into tokens and assigning indices to them, aligning the lengths of labels, and constructing an entity type set to support type-aware modeling. Finally, attention masks are generated to filter out invalid samples and mark the positions of valid text.

3.2 Multi-Scale Feature Extraction Module

Three parallel convolutional sub-networks with 128 output channels are designed in this paper to process features of different scales. For fine-, medium-, and coarse-grained features, the model specifies three different dilation rates (denoted as d for short hereafter), where $d = [1, 2, 4]$, corresponding to the fine, medium, and coarse scales respectively. A convolution operation is performed for each dilation rate d. The formula for the convolution operation is as follows1:

$$Y_d = \text{GELU}\left(\text{LayerNorm}\left(\text{Conv1D}_d(X)\right)\right) \quad (1)$$

where $X \in \mathbb{R}^{B\times 768\times L}$ denotes the input tensor, Conv1D_d represents the 1D convolution operation with a dilation rate of d, LayerNorm stands for the layer normalization operation, GELU (Gaussian Error Linear Unit) denotes the Gaussian error linear unit activation function, and $Y_d \in \mathbb{R}^{B\times 128\times L}$ represents the output tensor.

Subsequently, the features extracted by the three sub-networks are concatenated together through feature fusion. The concatenation formula is as follows2:

$$Y = [Y_1, Y_2, Y_4] \in \mathbb{R}^{B\times L\times 384} \quad (2)$$

where B denotes the batch size (Batch Size), and L denotes the sequence length (Sequence Length).

The fused feature Y is processed through a fusion layer for further fusion, and the fused tensor Z is output. The fusion formula is as follows3:

$$Z = \text{GELU}\left(\text{LayerNorm}(W \cdot Y + b)\right) \quad (3)$$

where $Z \in \mathbb{R}^{B\times L\times 256}$, $W \in \mathbb{R}^{384\times 256}$ are weight matrices, and $b \in \mathbb{R}^{256}$ is a bias vector.

3.3 Causal Deconfounding Module

As the core innovative component of the model, this module reduces the impact of confounding factors through front-door adjustment and counterfactual generation. First, the representation of confounding factors is extracted, and the extraction formula is as follows4,5:

$$Z_{\text{lowdim}} = \text{GELU}\left(\text{LayerNorm}\left(W_1 \cdot H_{\text{bert}} + b_1\right)\right) \quad (4)$$

$$Z = W_2 \cdot Z_{\text{lowdim}} + b_2 \quad (5)$$

where $H_{\text{bert}} \in \mathbb{R}^{B\times L\times 768}$ denotes the fused feature representation.$W_1 \in \mathbb{R}^{768\times 256}$ and $b_1 \in \mathbb{R}^{256}$ are the parameters of the linear layer.$Z_{\text{lowdim}} \in \mathbb{R}^{B\times L\times 128}$ denotes the low-dimensional confounding factor representation.$W_2 \in \mathbb{R}^{128\times 768}$ and $b_2 \in \mathbb{R}^{768}$ are the parameters of the projection layer.$Z \in \mathbb{R}^{B\times L\times 768}$ denotes the final confounding factor representation.

After that, an intervention mask is generated via the Bernoulli distribution based on the boundary probability and the cosine similarity of the confounding factor Z. The formula is as follows6,7,8,9:

$$\text{sim}_{ij} = \frac{Z_i \cdot Z_j}{\|Z_i\|\|Z_j\|} \quad \text{(cosine similarity)} \quad (6)$$

$$\overline{\text{sim}}_i = \frac{1}{L}\sum_{j=1}^{L} \text{sim}_{ij} \quad (7)$$

$$P(\text{intervene}_i) = \text{boundary}_{prob_i} \times \overline{\text{sim}}_i \quad (8)$$

$$\text{intervention}_{mask_i} \sim \text{Bernoulli}\left(P(\text{intervene}_i)\right) \quad (9)$$

After that, based on the mask, the tokens of the original input sequence are replaced, and counterfactual samples are generated to provide comparative data for causal effect estimation. The generation formula is as follows10,11:

$$\begin{aligned} X_{\text{cf}} = X \odot (1 - \text{intervention}_{mask}) \\ + \text{pad}_{token} \odot \text{intervention}_{mask} \end{aligned} \quad (10)$$

$$M_{\text{cf}} = \text{attention}_{mask} \odot (1 - \text{intervention}_{mask}) \quad (11)$$

where X is the original input sequence.X_{cf} is the counterfactual input sequence.M_{cf} is the counterfactual attention mask.

Then, the front-door adjustment module is used to separate the direct causal effect (where the input directly affects the output) and the indirect causal effect (where the input indirectly affects the output through intermediate variables). The mediating variables are calculated by encoding multi-scale fused features. The formula is as follows12,13:

$$M = \text{GELU}\left(\text{LayerNorm}\left(W_3 \cdot F_{\text{context}} + b_3\right)\right) \quad (12)$$

$$M_{\text{cf}} = \text{GELU}\left(\text{LayerNorm}\left(W_3 \cdot F_{\text{cf}} + b_3\right)\right) \tag{13}$$

where $F_{\text{context}} \in \mathbb{R}^{B\times L\times 256}$ is the multi-scale fused feature of the original sample.$F_{\text{cf}} \in \mathbb{R}^{B\times L\times 256}$ is the multi-scale fused feature of the counterfactual sample.$W_3 \in \mathbb{R}^{256\times 128}$ and $b_3 \in \mathbb{R}^{128}$ are the parameters of the mediating variable encoder.M and $M_{\text{cf}} \in \mathbb{R}^{B\times L\times 128}$ are the mediating variables.

The calculation formula for the direct effect(DE) is as follows14:

$$DE = f_{XY}^{\text{direct}}(X) \tag{14}$$

where X is the input feature.$f_{XY}^{\text{direct}}(\cdot)$ is the direct effect mapping function implemented by a neural network.

The calculation formula for the indirect effect(IE) is as follows15:

$$IE = f_{MY}(M) - f_{MY}(M_{\text{cf}}) \tag{15}$$

where $f_{MY}(\cdot)$ is the mapping function from mediating variables to outputs.

The direct and indirect effects are dynamically weighted according to the confounding factors, and the final total causal effect (TAU) is obtained to eliminate the impact of confounding factors. The calculation formula is as follows16,17,18,19:

$$\tau = \sigma(\bar{Z}), \quad \bar{Z} = \text{mean}(Z_{\text{lowdim}}) \tag{16}$$

$$W_{DE} = \tau \cdot DE \tag{17}$$

$$W_{IE} = (1-\tau) \cdot IE \tag{18}$$

$$TAU = W_{DE} + W_{IE} \tag{19}$$

where τ is the weight coefficient, which is obtained by mapping the confounding factor Z through the σ function.Z_{lowdim} is the low-dimensional confounding factor representation.W denotes the effect weight.

3.4 Prototype Learning Module

This module enhances the distinguishability between entities of different categories through contrastive learning. For each entity category, the model maintains a prototype vector. During model initialization, the model pre-creates a learnable prototype vector for each entity category. The prototype vectors are defined as learnable parameters and initialized using the K-aiming normal distribution.

Subsequently, a small neural network called bottleneck_extractor is employed to extract bottleneck features from the deconfounded contextual features. The structure of this neural network consists of two linear layers with LayerNorm and GELU activation in between.

Subsequently, the extracted bottleneck features and prototypes undergo L_2 normalization. The normalization formula is as follows20,21:

$$f_{\text{norm}} = \frac{f}{\|f\|_2} \tag{20}$$

$$p_{\text{norm}} = \frac{p}{\|p\|_2} \tag{21}$$

A similarity matrix S for subsequent computations is constructed. S is a matrix of size $i \times j$, and for each element S_{ij} in the matrix, the following holds 22:

$$S_{ij} = \frac{f_i \cdot p_j}{\|f_i\|\|p_j\| \cdot \tau} \tag{22}$$

Meanwhile, positive and negative sample pairs are constructed. For positive sample pairs, the sample features of the same category should be close to their corresponding prototypes; for negative sample pairs, it means that the sample features of different categories should be far from each other's prototypes. Masks for positive samples and negative samples are created to extract their positive sample similarity and negative sample similarity, and the prototype contrastive loss is calculated. The formula is as follows23:

$$L_{\text{contrast}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(\frac{S_{i,y_i}}{\tau}\right)}{\sum_{j=1}^{C}\exp\left(\frac{S_{i,j}}{\tau}\right)}\right) \tag{23}$$

where N is the number of valid samples,C is the number of categories,y_i is the true label of the i-th sample,S_{ij} is the similarity between the i-th sample and the j-th prototype.

3.5 Loss Function Design

During the model training process, the Adam optimizer is employed along with gradient scaling technology to enhance training stability. In the

training process, the total loss consists of CRF loss L_{crf}, confounding factor regularization loss $L_{\text{confounder\_reg}}$, prototype contrastive loss L_{contrast}, and span offset loss L_{offset}. Multi-loss joint optimization is performed with different weights to further improve the overall performance of the model.

The loss function formula is as follows24:

$$\text{Loss} = L_{\text{crf}} + 0.1 \cdot L_{\text{confounder\_reg}} + 0.05 \cdot L_{\text{contrast}} + 0.03 \cdot L_{\text{offset}} \tag{24}$$

4 Experiments

The environment configuration used in this experiment is as follows: Python 3.9.23, torch 2.7.0+cu128, and NVIDIA A10 GPU; the pre-trained model adopted is bert-base-chinese; the maximum length of text sequences is set to 50, the batch size is set to 32, and the number of training epochs is set to 30; the Adam optimizer is used, with the learning rate set to 2e-5.

In terms of model parameter configuration, CDNER is based on the Chinese BERT Base pre trained model (12 layers of Transformer, 768 dimensional hidden layers, 12 attention heads, with basic parameters of about 102M), and adds three new modules: multi granularity feature extraction, causal unmixing, and prototype contrastive learning, with a total parameter size of about 107.8M. In terms of computational budget, the MSRA dataset takes 5 GPU hours for 30 rounds of training, the WeiBoNER dataset takes 0.5 GPU hours for 30 rounds of training, and 1.5 GPU hours for hyperparameter tuning and ablation experiments, resulting in a total computational cost of 7 GPU hours.

Experimental verification was conducted on the following two widely used Chinese Named Entity Recognition (NER) datasets as Table 1:

The MSRA and WeiBoNER datasets used in this study are publicly available benchmark resources in the field of Chinese named entity recognition. Their original publication has gone through a legal data authorization process, and this study follows academic sharing standards to use the publicly available data, without involving direct consent communication with data objects.

When evaluating model performance, the general metric **F1-score** is used as the benchmark to assess model performance, and the calculation is based on **macro-average** to balance the bias caused by uneven sample sizes across different categories.

The calculation method of F1-score is as follows25:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{25}$$

where Precision(P) = TP / (TP + FP) represents precision rate, which is used to measure the accuracy of the model in "predicting entities".Recall(R) = TP / (TP + FN) represents the recall rate, which is used to measure the completeness of the model in "capturing entities".

4.1 Comparative Experiments

To comprehensively evaluate the effectiveness of the proposed model (CDNER), After conducting all experiments five times, take the average value, the following models are selected for comparison an Table 2:

4.2 Analysis of Comparative Experiment Results

Experiments show that the proposed model achieves excellent performance on different types of datasets: it reaches an F1-score of 0.8885 on the WeiBoNER dataset, and outperforms models such as BERT+CRF, FLAT, LEBERT, and GNER on the MSRA dataset—with only a 0.059 gap in F1-score compared to the optimal model. This performance advantage benefits from the synergistic effect of modules including multi-granularity feature extraction, causal deconfounding, and prototype learning. These modules enable the model to not only deliver outstanding performance in complex and informal text environments like that of the WeiBoNER dataset but also achieve good performance in formal text environments such as the MSRA dataset.

4.3 Ablation Experiment

To verify the contribution of each module to the model's performance, ablation experiments were designed for each component in the model. Each ablation experiment was conducted under the same experimental settings to accurately evaluate the impact of each module on the final performance. All experiments were repeated three times, and the average value was taken to reduce errors caused by randomness.The experimental results of the model body and ablation variants are shown in Figure. 2.

Ablation experiments show that in harsh data environments—characterized by a large number of abbreviations, colloquial expressions, and noisy

Dataset	Training Set Size	Test Set Size
MSRA (Levow, 2006)	∼46,000 sentences	∼4,300 sentences
WeiBoNER (Peng and Dredze, 2015)	∼1,800 sentences	∼300 sentences

Table 1: Relevant Information of Datasets for Experiments

Model Name	WeiBoNER-F1	MSRA-F1
BERT-CRF (Chang et al., 2021)	0.6837	0.9340
FLAT (Li et al., 2020)	0.6342	0.9435
LEBERT (Liu et al., 2021)	0.7095	0.9570
W²NER (Li et al., 2022)	0.7232	0.9610
RICON (Gu et al., 2022)	0.5888	0.9614
BOPN (Tang et al., 2023)	0.7292	0.9639
GNER (Ding et al., 2024)	0.4278	0.9456
Our Model(CDNER)	0.8885	0.9580

Table 2: Comparison of F1 values between the baseline model used for comparison and this model on the WeiBoNER dataset and MSRA dataset

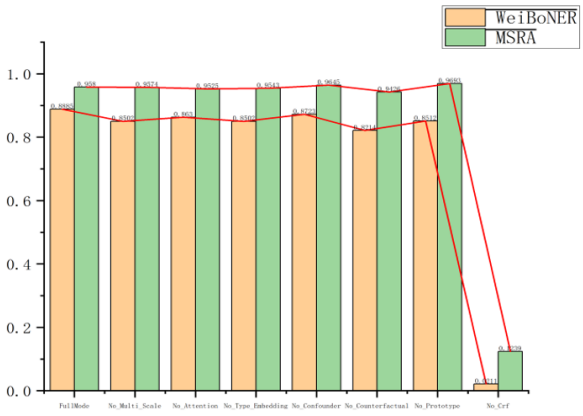

Figure 2: Results and Performance Comparison of Each Ablation Experiment on Two Public Chinese Named Entity Recognition (NER) Datasets: WeiBoNER and MSRA

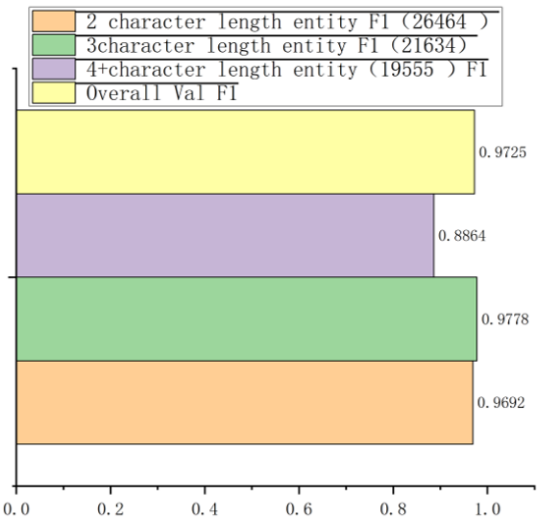

Figure 3: Performance Statistics of the Model on Entities of Different Lengths in the MSRA Dataset

data—where entity recognition is relatively difficult, each module of the model makes a positive contribution to the model's performance and improves its entity recognition capability. However, in environments with more standardized language expressions, relatively clear entity boundaries and types, and high annotation quality, the performance improvement of each module is limited; among them, the confounding factor modeling module and the prototype learning module may even lead to a decline in model performance. This is because the MSRA dataset has high data quality, standardized sequence label annotations, and few confounding factors. In this case, confounding factor modeling instead introduces unnecessary constraints, resulting in overfitting, and the model performs better when this module is removed. Meanwhile, the MSRA dataset has a balanced label distribution and sufficient samples, and the strong constraints of prototype learning may limit the model's ability to capture fine-grained differences—such as the subtle distinction between similar entities.The experimental results of entities with different lengths are shown in Figure. 3.

As can be seen from the figure, the model achieves excellent performance on entities of different lengths. Under comprehensive evaluation, the model demonstrates good overall performance in recognizing various types of entities. The

Data	P	R	F1
MSRA	0.9650	0.9521	0.9580
WeiBoNER	0.9569	0.8571	0.8885

Table 3: Various Performance Metrics of the Proposed Model on Two Public Chinese Named Entity Recognition (NER) Datasets: WeiBoNER and MSRA

multi-granularity feature extraction module in the model can effectively capture the features of 3-character entities at fine-grained, medium-grained, and coarse-grained levels. This enables the model to conduct more sufficient feature learning for such entities and achieve more accurate recognition.

Although 2-character entities are similar in length to 3-character entities, their relatively few characters lead to insufficient expression of some features, resulting in an F1-score slightly lower than that of 3-character entities. Entities with 4 or more characters contain abundant information and have complex structures. When the model processes them, the interference from confounding factors is more severe. Despite the model being equipped with a deconfounding mechanism, it is still difficult to completely eliminate such interference, which limits the recognition effect.

For 1-character entities, the single character carries too little information, making it hard for the model to accurately determine their categories based solely on this information. Even when combined with contextual information, the effective information remains relatively insufficient. This affects the precision and recall of recognition, ultimately leading to a low F1-score.

5 Conclusions

When conducting experiments on two public datasets, the average performance metrics of the proposed model over five runs are as follows Table 3:

The use of MSRA and WeiBoNER public datasets in this study is consistent with their expected use as benchmark data for Chinese NER tasks; Meanwhile, the CDNER model constructed in this study is only used in academic research scenarios and has not been deployed or utilized outside of the research environment.

A novel Chinese Named Entity Recognition (NER) model, CDNER, which integrates causal inference and multi-granularity feature fusion techniques, is proposed by the authors. As can be seen from the experimental results, the model can correctly identify entity types in complex text data. This advantage is attributed to two key factors: first, causal deconfounding and front-door adjustment effectively reduce the interference of confounding noise in the text on the model's entity prediction; second, multi-granularity convolution and self-attention mechanisms can efficiently fuse features of entities with different lengths, thereby improving the overall performance of the model.

This study only conducted experiments based on publicly shared benchmark datasets and did not involve ethical review procedures such as human participant recruitment and sensitive data collection. Therefore, the approval process of the ethics review committee was not carried out.

6 Limitation and Future Work

Meanwhile, the low recall rate of the model on the WeiBoNER dataset also reflects its limited ability to capture complex or non-standard entities. This is due to two reasons: on one hand, the WeiBoNER dataset is highly colloquial and casual, and the expression of entities in colloquial text is non-standard, making it difficult for the model to capture entity boundaries when dealing with such entities; on the other hand, a large number of emerging entities or non-standard names in the dataset are not fully covered by the training data, leading to difficulties in matching class prototypes of the prototype learning module and subsequent missed detection of some entities.

CDNER's training datasets (MSRA, WeiBoNER) cover formal and social media texts but carry linguistic variation and group representation biases. Underrepresented samples of dialects, ethnic minority languages, and region/occupation-specific entities reduce the model's accuracy for marginalized groups, reinforcing mainstream information advantages and excluding minority language speakers. Though its causal deconfounding module mitigates annotation noise, it may not fully eliminate spurious correlations between sensitive attributes (e.g., gender, geography) and entity categories, leading the model to learn and perpetuate biases, raising fairness concerns.

As a high-performance Chinese NER tool, CDNER faces malicious use and dual-use risks. It could be exploited to mass-extract social media users' personal info (names, locations, occupations), causing privacy leaks, fake profiles, or tar-

geted fraud. Monitoring specific groups' online speech with it may infringe on freedom of expression and privacy. While delivering value in legitimate scenarios like public opinion analysis and knowledge graph construction, erroneous recognition (e.g., mislabeling ordinary content as sensitive) may unjustly harm individuals, and deliberate misuse will amplify information security risks, undermining digital trust and safety.

The architecture of the CDNER model is relatively complex. The coordinated operation of multiple components results in high performance overhead in training and computation. Future work will focus on three aspects: simplifying the model, integrating external knowledge, and optimizing the causal inference mechanism in Chinese Named Entity Recognition.

References

Yuan Chang, Lei Kong, Kejia Jia, and Qinglei Meng. 2021. Chinese named entity recognition method based on bert. In *2021 IEEE International Conference on Data Science and Computer Application (ICDSCA)*, pages 294–299.

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 2021. HacRED: A large-scale relation extraction dataset toward hard cases in practical applications. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2819–2831, Online. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuyang Ding, Juntao Li, Pinzheng Wang, Zecheng Tang, Yan Bowen, and Min Zhang. 2024. Rethinking negative instances for generative named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3461–3475, Bangkok, Thailand. Association for Computational Linguistics.

Yingjie Gu, Xiaoye Qu, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Xiaolin Gui. 2021. Read, retrospect, select: An MRC framework to short text entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12920–12928. AAAI Press.

Yingjie Gu, Xiaoye Qu, Zhefeng Wang, Yi Zheng, Baoxing Huai, and Nicholas Jing Yuan. 2022. Delving deep into regularity: A simple but effective method for Chinese named entity recognition. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1863–1873, Seattle, United States. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *ArXiv*, abs/1508.01991.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *SIGHAN@COLING/ACL*.

Jing Li, Aixin Sun, and Yukun Ma. 2021a. Neural named entity boundary detection. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1790–1795.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973, Vancouver, Canada (Virtual conference). AAAI Press.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. 2021b. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842, Online. Association for Computational Linguistics.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021. Lexicon enhanced Chinese sequence labeling

using BERT adapter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5847–5858, Online. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Conference on Empirical Methods in Natural Language Processing*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Yuling Su, Hong Zhao, Yifeng Zheng, and Yu Wang. 2024. Few-shot learning with multi-granularity knowledge fusion and decision-making. *IEEE Transactions on Big Data*, 10(4):486–497.

Minghao Tang, Yongquan He, Yongxiu Xu, Hongbo Xu, Wenyuan Zhang, and Yang Lin. 2023. A boundary offset prediction network for named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14834–14846, Singapore. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.

Yan Zeng, Ruichu Cai, Fuchun Sun, Libo Huang, and Zhifeng Hao. 2025. A survey on causal reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):5942–5962.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644.