applied sciences

MDPI

Article

Label-Guided Data Augmentation for Chinese Named Entity Recognition

Miao Jiang * and Honghui Chen

National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha 410073, China; chenhonghui@nudt.edu.cn
* Correspondence: jiangmiao20@nudt.edu.cn

Abstract: Chinese named entity recognition (NER) is a fundamental natural language processing (NLP) task that involves identifying and categorizing entities in text. It plays a crucial role in applications such as information extraction, machine translation, and question-answering systems, enhancing the efficiency and accuracy of text processing and language understanding. However, existing methods for Chinese NER face challenges due to the disruption of character-level semantics in traditional data augmentation, leading to misaligned entity labels and reduced prediction accuracy. Moreover, the reliance on English-centric fine-grained annotated datasets and the simplistic concatenation of label semantic embeddings with original samples limits their effectiveness, particularly in addressing class imbalances in low-resource scenarios. To address these issues, we propose a novel Chinese NER model, LGDA, which leverages Label-Guided Data Augmentation to mitigate entity label misalignment and sample distribution imbalances. The LGDA model consists of three key components: a data augmentation module, a label semantic fusion module, and an optimized loss function. It operates in two stages: (1) the enhancement of data with a masked entity generation model and (2) the integration of label annotations to refine entity recognition. By employing twin encoders and a cross-attention mechanism, the model fuses sample and label semantics, while the optimized loss function adapts to class imbalances. Extensive experiments on two public datasets, OntoNotes 4.0 (Chinese) and MSRA, demonstrate the effectiveness of LGDA, achieving significant performance improvements over baseline models. Notably, the data augmentation module proves particularly effective in few-shot settings.

Keywords: Chinese named entity recognition; low resource; data augmentation; label semantic fusion

check for updates

Academic Editor: Douglas O'Shaughnessy

Received: 28 January 2025
Revised: 21 February 2025
Accepted: 24 February 2025
Published: 26 February 2025

Citation: Jiang, M.; Chen, H. Label-Guided Data Augmentation for Chinese Named Entity Recognition. *Appl. Sci.* **2025**, *15*, 2521. https://doi.org/10.3390/app15052521

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Chinese named entity recognition (NER) is an important task in natural language processing (NLP) that identifies and categorizes entities in text into predefined categories. This task is critical to applications such as information extraction, machine translation, and question-answering systems, significantly improving the efficiency and accuracy of text processing. Despite its importance, achieving high performance in NER often requires a large number of high-quality annotated data [1]. However, acquiring such data is particularly challenging in low-resource domains or languages, as it demands extensive domain knowledge, expert annotation, and significant cost. This data scarcity often results in poor model performance due to the insufficient learning of hidden feature representations. Data augmentation offers a promising solution by expanding training samples through operations such as deletion, replacement, insertion, and back translation [2].

While data augmentation has proven effective in various NLP tasks, like sentence classification [3], machine translation [4], and dialogue generation [5], its application to NER presents unique challenges. Unlike other tasks where semantic labels remain unchanged, NER requires precise label predictions at the character level. Traditional augmentation methods, such as back translation, can disrupt character-level semantics and lead to misaligned entity labels, thereby reducing prediction accuracy [6]. To address these issues, recent studies have explored incorporating label semantics as additional input signals [7,8] or leveraging fine-grained entity annotation types to guide data augmentation [9]. While these methods improve the model's understanding of text semantics, they often rely on English-centric datasets, which are scarce in other languages and are usually of smaller scale. Additionally, current approaches that concatenate label semantic embeddings with original sample embeddings fail to effectively guide the learning process and may unnecessarily increase input length. Furthermore, low-resource entity recognition tasks are hindered by class distribution imbalances, making it difficult for models to learn from infrequent classes.

To address these challenges, we propose LGDA, a novel Chinese NER model that leverages Label-Guided Data Augmentation. The LGDA model consists of three core components: a data augmentation module, a label semantic fusion module, and an optimized loss function. LGDA operates in two stages. In the first stage, a masked entity generation language model enhances the data by embedding entity labels into the original samples to sharpen Chinese entity boundaries. The model also replaces original entities with newly generated ones, thereby expanding the training set. In the second stage, label annotations are introduced as domain-specific prior knowledge and integrated into the model alongside the augmented and original samples. Twin encoders with shared parameters process the samples and label annotations, and a cross-attention mechanism fuses these inputs, enabling the model to learn representations enriched with label semantics. Additionally, an optimized loss function adaptively assigns higher weights to samples from less distinguishable classes, addressing the long-tail distribution issue in low-resource scenarios. To evaluate the proposed method, extensive experiments were conducted on two public datasets: OntoNotes 4.0 (Chinese) and MSRA. The results demonstrate that LGDA outperforms competitive baselines, achieving state-of-the-art performance. Notably, LGDA excels in few-shot scenarios, where the data augmentation module proves especially effective.

In summary, the main contributions of this paper are as follows:

- We propose a novel Chinese NER model, LGDA, which introduces Label-Guided Data Augmentation to enhance performance in low-resource scenarios.
- We design a two-stage framework that includes a data augmentation module for expanding training data and a label semantic fusion module for integrating label information effectively.
- We validate the proposed method through extensive experiments on two public datasets, OntoNotes 4.0 (Chinese) and MSRA, achieving state-of-the-art performance.

2. Related Work

2.1. Chinese NER

The field of Chinese NER has witnessed substantial progress due to the integration of pre-trained language models and novel architectures. Early advancements include BERT [10], which laid the groundwork by encoding words and adding a conditional random field (CRF) layer for sequence labeling. Lattice LSTM [11] expanded on this by employing a dictionary-based lattice structure to better utilize character and sequence data in Chinese NER. To further refine Chinese character understanding, Glyce-BERT [12]

incorporated pictographic features by using a grid convolutional neural network and enhanced generalization through multi-task learning, while LEBERT [13] adapted the BERT framework with a lexicon adaptation layer, converting sample characters and lexicon information into "character-word pair" sequences. FLAT [14] improved upon lattice structures by adopting a flat Transformer architecture, allowing for more comprehensive character and entity learning. Innovative approaches like BERT-MRC [15] redefined NER as a reading comprehension task, using entity labels as queries to improve entity identification. However, this transformation, which enlarges the sample set and extends the length of text sequences, consequently increases the computational time required for both training and inference.

XLNet-Transformer-R [16] combines XLNet with a Transformer encoder and R-Drop to prevent overfitting, effectively handling long texts and context. MW-NER [17] enriches semantics and reduces segmentation errors by fusing multi-granularity word information with character data through a strong–weak feedback attention mechanism.

While these methods have shown promise, they also have limitations. Many rely heavily on large annotated datasets, which can be scarce for certain languages or domains. Additionally, some methods may struggle with class distribution imbalances, leading to less accurate recognition of infrequent entities. Furthermore, the complexity of certain models can make them computationally expensive and difficult to fine-tune for specific tasks.

2.2. Low-Resource NER

Addressing the low-resource challenge in NER involves enhancing model generalization and learning from limited samples. In low-resource scenarios, the scarcity of annotated data exacerbates challenges such as the long-tail distribution phenomenon, where certain entity types are underrepresented. To tackle this, researchers have explored strategies that build on fine-grained annotations, semi-supervised learning, and the incorporation of external knowledge.

E-strPron [18] stands out by integrating a variety of linguistic features, such as pronunciation and radicals, using cross-Transformers and a soft fusion module to enhance the model's grasp of Chinese characters. These methods collectively demonstrate the versatility and continuous evolution of Chinese NER techniques. Jia et al. [19] introduced a semi-supervised pre-training approach that integrates a word library into the language model, enhancing self-attention mechanisms to improve performance on Chinese NER. This method enriches the model's ability to discern entity boundaries and adapt to data-scarce scenarios. Similarly, other approaches have leveraged multilingual training data to improve generalization [6], particularly by explicitly injecting entity labels into the sentence context. Such methods enhance the model's understanding of entity regularities and establish clearer relationships between entities and labels, thus mitigating the scarcity of annotated data.

MECT [20] introduces a cross-Transformer model that integrates multi-modal data, significantly aiding low-resource Chinese NER tasks. The Multi-Grained [21] method effectively transfers knowledge from larger to smaller models by using an enhanced Viterbi algorithm, while LEAR [8] approaches NER as a question-answering task, proposing label knowledge integration to improve text representation. PCBERT [9] tackles low-sample Chinese NER with a two-stage model. Specifically, P-BERT is a prompt-based model that extracts implicit label extension features from a label extension dataset during the prompt tuning stage and subsequently provides these features to C-BERT during the fine-tuning stage. Meanwhile, C-BERT is a lexicon-based model that incorporates multiple label features for each dictionary entry. Each token is associated with a lexicon set derived from an

external dictionary, and each lexicon is mapped to a corresponding label set. This approach can enhance the model's capacity to accurately determine entity boundaries by selecting the lexicon that best aligns with the correct labels. Although validated experimentally, constructing vocabulary trees and label sets for each input sentence is laborious.

2.3. Data Augmentation

Data augmentation is a widely used technique to address data scarcity, originating in Computer Vision and later adapted to NLP tasks [2]. Traditional data augmentation methods, such as deletion, replacement, insertion, and back translation, typically preserve semantic labels at the sentence level. However, for NER tasks, where labels are assigned to each character or token, these methods often disrupt character-level semantics, resulting in misaligned or inaccurate entity labels.

To counter this issue, researchers have proposed advanced augmentation techniques tailored to NER. For instance, Zhou et al. [6] explored data augmentation through the use of multilingual training data and heuristic rules that explicitly inject entity labels into the sentence context. This approach reduces the risk of label misalignment and enriches the augmented data with meaningful entity information.

In recent years, prompt-based data augmentation methods have been successfully applied to sentence-level low-resource learning tasks, largely due to the careful design of templates and label words. By constructing templates in the form of cloze tests, the likelihood of introducing meaningless or erroneous noise samples can be reduced, thereby improving the quality of the augmented samples [22–24]. Lee et al. [22] systematically analyzed how to construct templates and select examples for prompt learning when applied to NER, proposing the combination of contextually learned task templates for input augmentation. However, when applied to character-level sequence-labeling tasks such as NER, predicting the entity type of each text span in a sentence requires manually designing templates, which is the inefficient and time-consuming process of enumerating all potential entities for template querying. To address this issue, Ma et al. [25] proposed a more convenient method that reconstructs the NER task into a language model problem without the need for templates. Specifically, this work explored rules that help language models automatically search for suitable label words, retaining the prior knowledge of entity labels provided by prompt-based data augmentation methods while avoiding the complex process of template construction.

In low-resource sample data, the long-tail distribution phenomenon is more likely to occur. Current NER methods based on traditional data augmentation merely perform simple augmentation operations on the original samples, which do not yield satisfactory results. Simple sentence-level addition or deletion may disrupt the samples of token-level classification tasks or cause label misalignment, potentially exacerbating the imbalance of the sample distribution.

3. Primary Definition and Model Architecture

To formalize the problem under study, we provide a mathematical description following the conventions of prior research [26–28]. NER is treated as a sequence-labeling task. Given an input sentence $X = \{x_1, x_2, \ldots, x_n\}$, the model predicts a label sequence $Y = \{y_1, y_2, \ldots, y_n\}$, where each label y_i corresponds to the characters x_i and $y_i \in \mathbb{R}$. The predefined set of label types is denoted by $S = \{s_1, s_2, \ldots, s_t\}$, where t is the total number of label types. The label types vary depending on the domain; for instance, general domains may include entities such as person names, locations, and organizations, whereas medical domains might use categories like diseases, symptoms, and medications. In low-resource settings, each label category contains only a limited number of annotated samples, posing

challenges for effective supervised training. In addition, we employ the BIO (Beginning, Inside, Outside) tagging scheme to predict the label of each token. Specifically, each token is classified as "B-entity_type" to denote the beginning of an entity, "I-entity_type" to indicate the inside of an entity, or "O" to signify that the token is outside any entity. The NER process is mathematically expressed in Equation (1):

$$\hat{y} = \arg\max_{j}(\text{soft}\max(f(x_i))), \tag{1}$$

where $f(\cdot)$ denotes the model, x_i represents the input character, and $\hat{y}$ is the predicted output. It is worth noting that if the probabilities of multiple categories output by softmax are equal, the model will randomly select one.

Existing Chinese NER models that incorporate data augmentation methods encounter several challenges, including entity label misalignment in the augmented samples, insufficient utilization of label semantic information, and neglect of long-tail distribution issues in low-resource scenarios. To address these limitations, we propose LGDA, a novel Chinese NER model based on Label-Guided Data Augmentation. The overall framework of LGDA is illustrated in Figure 1. The model comprises three core components: a data augmentation module, a label semantic fusion module, and an optimized loss function.

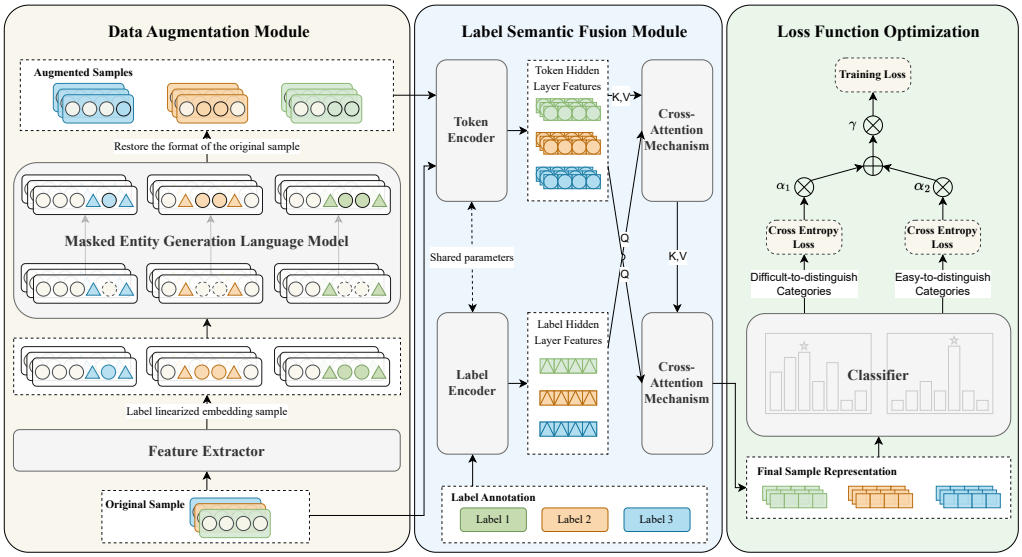


Figure 1. The framework of proposed LGDA model.

Specifically, the LGDA model operates in two stages. In the first stage, a masked entity generation language model is used to enhance sample data, and label annotations guide the model to learn improved sample representations for NER. Data augmentation is achieved by embedding entity labels linearly in the original samples to sharpen the boundaries of Chinese entities (see Section 4.1.1). The masked identifiers are then applied to cover the original entities, replacing them with newly generated ones, thereby expanding the training set (see Section 4.1.2). In the second stage, label annotations are introduced as additional domain-specific prior knowledge and incorporated into the model alongside the augmented and original samples. To integrate label semantics, two parameter-sharing twin encoders encode the samples and label annotations (see Section 4.2.1), with a cross-attention mechanism for fusion (see Section 4.2.2). This approach guides the model to learn sample representations enriched with the semantic information of the labels. To address the long-tail distribution problem in low-resource scenarios, the training loss function is optimized to adaptively assign higher weights to samples from less distinguishable classes (see Section 4.3).

4. Approach

4.1. Data Augmentation Module

4.1.1. Label Linearization Embedding

Data augmentation is a direct and effective approach to addressing the low-resource problem [29]. However, in NER tasks, which rely heavily on character-level semantics, traditional data augmentation methods often produce augmented samples that misalign with the original labels [30]. To address this issue, we propose linearizing and embedding entity labels in the original samples. This approach not only constrains the generation of new samples during augmentation but also sharpens the boundaries of Chinese entities. The specific process is defined as follows:

$$X = \{x_1, x_2, \ldots, x_n\} \overset{linearization}{\longrightarrow} X' = \{x_1, s_1, x_2, s_2, \ldots, x_n, s_n\}, \tag{2}$$

where x_n is an entity and its corresponding label type is s_n. The label is inserted before and after the corresponding character and used as input for the subsequent masked entity generation language model. For sequences shorter than the model's maximum input length limitation, we apply padding to extend them to the desired length. Conversely, for sequences exceeding this limitation, truncation is employed.

For instance, as illustrated in Figure 2, the character sequence and its corresponding label sequence from the original annotated sample are first linearized. Subsequently, entity-type labels with semantic meaning in the label sequence (i.e., labels other than the *O* type) are embedded into the character sequence.

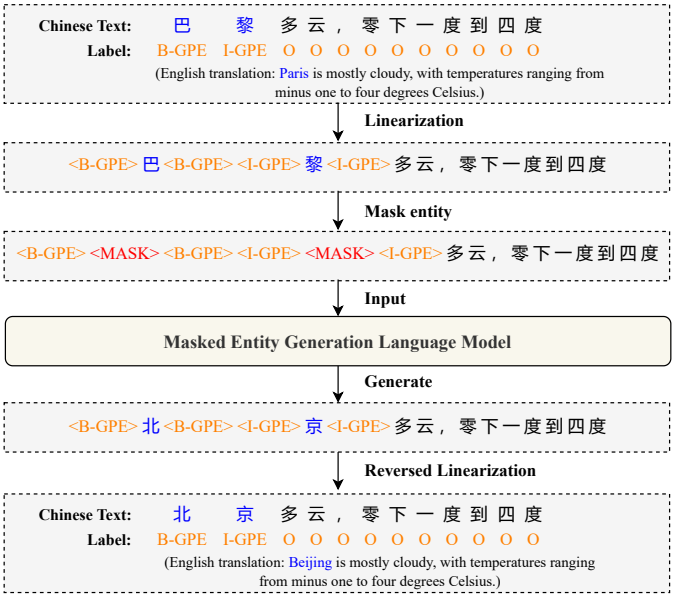


Figure 2. An example of data augmentation. "B-GPE" means "the beginning of GPE type entity", where GPE means "Geo-Political Entity". "I-GPE" means "the inside of GPE type entity".

4.1.2. Masked Entity Generation Language Model

The linearized sample X' is input into the masked entity generation language model, which is a sequence-to-sequence generative model. First, characters in entities other than the *O* type are randomly replaced with the masking identifier [Mask]. Secondly, the model maximizes the probability between the masked entity characters and the linearized sequence X'. After training and fine-tuning, the model is able to use the unmasked original

sample context and the linearized embedded label semantic information to predict the masked characters, thereby generating new entities. The process is shown in Formula (3):

$$\max_{\theta} \log p_{\theta}(\ X|X^{'}) \approx \sum_{i=1}^{n} k_i \log p_{\theta}(x_i|X^{'}), \tag{3}$$

where θ represents the parameters of the masked entity generation language model. X denotes the original sample, and x_i is a character in X. k_i is a 0–1 integer; if $k_i = 1$, it indicates that x_i is replaced by the masking identifier [Mask]; if $k_i = 0$, it indicates that the character remains unchanged. n represents the number of characters in X'.

Since the model is pre-trained on the original training corpus, the predicted masked entity characters are likely to be the entity characters in the original sample. To enable the model to generate enriched and coherent augmented samples that are different from the original samples, we raise the threshold of candidate entity probabilities, increase the set of candidate entity characters, and use a random strategy to select candidate entity characters from this set. In the experiment, we randomly select one of the first five candidate entities and repeat the process for three rounds. To reduce noise from the augmented samples, we train a NER model on gold-standard samples and use it to label augmented sentences. We retain only those sentences where the model's predictions align with the original labels. Additionally, we apply a Gaussian distribution to sample the entity character masking rate, thereby obtaining a dynamically changing entity masking rate and enhancing the diversity of augmented samples. As we will demonstrate in Section 6.6, the model is capable of generating correct and coherent augmented samples that are related to the entity labels in the original samples. Furthermore, we examine the effect of augmented samples on the recognition performance across various classes in Appendix A.1.

As shown in Figure 2, after the prediction of the masked entities is completed, the newly generated samples are linearized. To maintain consistency with the original training samples, they need to be converted back into the format of the original samples. The final augmented samples, along with the original samples, serve as inputs for the subsequent label semantic fusion stage.

4.2. Label Semantic Fusion Module

4.2.1. Twin Encoders

Labels in named entity recognition tasks are typically highly generalized representations of entity categories. However, when humans understand entities, they usually pre-judge the domain in which the entity resides and attribute more domain-related concrete meanings to abstract entity labels. For example, in the Chinese text "每个人在各自的岗位上发光发热", the entity "发热" is generally not labeled as an entity in news domains, but in medical domains, it is an important symptom-type entity.

Furthermore, the same entity may have different specific meanings in different domains. For instance, "苹果" is categorized as a fruit in the food domain but as a company or brand in the financial domain. To enable the model to simulate human thought processes as closely as possible, we introduce label annotations as additional prior knowledge. In practical applications, label annotations are also easy to obtain and replace, which can enhance the model's domain generalizability. The specific label annotations are shown in Table 1 below.

Compared with the original training set, the scale of the corpus for labels and their label annotation is quite small, which is insufficient to train a good encoder from scratch. Therefore, inspired by twin networks, we propose the use of twin encoders. This twin structure can effectively cope with small-scale data and class data imbalance. Specifically, the encoder for label annotation, called the label encoder, is structurally similar to that of

the original training samples, namely, the token encoder, and shares model parameters. The shared encoder effectively captures the representation of label annotations without introducing additional parameters. Moreover, during inference, the learned label encoder generates label representations only once. These representations can be cached, thereby eliminating the need for the label encoder to recalculate them. We compared the training and inference times of our model with those of other methods. More details are reported in Appendix A.2.

Table 1. Domain-extended meanings of labels.

Label	Annotation
PER (Person)	Names and fictional characters
LOC (Location)	Natural landscape locations such as mountain ranges, rivers, etc.
ORG (Organization)	Organizations, including companies, government parties, schools, governments, and news agencies
GPE (Geo-Political Entity)	Geopolitical entities divided by geographical location, including countries, cities, and towns

Given the input sample sequence X and the label annotation sequence L, the encoders extract their hidden layer features, as shown in Formula (4):

$$H_X = \{h_{x_1}, h_{x_2}, \ldots, h_{x_n}\} = PLM(x_1, x_2, \ldots, x_n), \tag{4}$$

$$H_L^{s_j} = \left\{h_{l_1}^{s_j}, h_{l_2}^{s_j}, \ldots, h_{l_m}^{s_j}\right\} = PLM\left(l_1^{s_j}, l_2^{s_j}, \ldots, l_m^{s_j}\right),$$

where PLM refers to the neural network model used by the encoder, which can be replaced based on the task and the characteristics of the dataset. In this paper, the RoBERTa model is adopted. To verify that the improvements of the model are independent of the RoBERTa model, we supplemented relevant experiments in Appendix A.3. In the formulas, $H_X \in \mathbb{R}^{n \times d}$, where n represents the length of the sample sequence X and d is the vector dimension of the encoder. $H_L^{s_j} \in \mathbb{R}^{|S| \times m \times d}$, where $H_L^{s_j}$ denotes the hidden layer features of the label annotation for category s_j; $s_j \in S$, where $|S|$ is the number of elements in the entity label-type set, i.e., the number of label types, and m represents the length of the label annotation.

4.2.2. Cross-Attention Mechanism

After obtaining the hidden layer features of the sample and label annotation sequences, existing methods mostly adopt direct multiplication or concatenation [7,22,24]. The characteristic of these methods is that they do not fully learn the semantic information of the labels and also increase the length of the input sequence. We propose a new fusion strategy that uses a cross-attention mechanism, allowing the model to focus on the most critical parts of the sample and ignore other redundant parts while integrating label semantic knowledge into the sample representation. At the same time, this fusion strategy can also maintain the fixed length and structure of the sample.

Specifically, we use the hidden layer features of the label annotation $H_L^{s_j}$ as the query state of the attention mechanism, with the key and value states being the hidden layer features of the sample character sequence H_X. As shown in Formulas (5) and (6), d is the model dimension, and W^Q, W^K, and W^V are all learnable parameters with dimensions $\mathbb{R}^{d \times d}$.

$$C_j' = Attention\left(H_L^{s_j} W^Q, H_X W^K, H_X W^V\right), \tag{5}$$

$$Attention(Q, K, V) = \text{soft}\max\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{6}$$

After obtaining the weighted sample representation C'_j, it is used as the key and value states of the cross-attention mechanism, with H_X as the query state. The attention mechanism is executed again to obtain the final sample representation C_j. The specific process is shown in Formula (7):

$$C_j = Attention\left(H_X W^Q, C'_j W^K, C'_j W^V\right). \tag{7}$$

4.3. Loss Function Optimization

In low-resource scenarios, the issue of imbalanced sample categories is prominent, with the "O type" entities dominating the sample sequence labels. Predicting the labels for these entity characters is relatively simple compared with other entity types with richer semantics. However, traditional cross-entropy loss functions do not distinguish between the difficulty levels of category discrimination, which can cause the model to focus on the easily distinguishable "O type" throughout the training process, failing to fully learn other entity types in the samples.

Therefore, we introduce a modulation factor γ to enable the model to focus more on the difficult-to-distinguish entity types, achieving the effect of adaptive weight tuning. At the same time, we also adopt an uncertainty weighting strategy, adding a regularization term to further balance the loss weights of easy-to-distinguish and difficult-to-distinguish categories. The training loss process of the model is shown in Formulas (8) and (9):

$$\hat{p} = \text{soft}\max(WC + b), \tag{8}$$

$$\mathcal{L} = \sum_{j}^{t} -\alpha_j (1 - \hat{p}^j)^\gamma \log(\hat{p}^j) \mathbb{1}(j == y). \tag{9}$$

Firstly, Formula (8) applies a linear layer to the sample representation C, where W and b are learnable parameters. Formula (9) represents the final loss function, where α_j is a hyperparameter representing the loss weight for different label categories, $\mathbb{1}(\cdot)$ is the indicator function, j denotes the j-th category, and y represents the label category. When $j = y$, the value of $\mathbb{1}(j == y)$ is 1; otherwise, it is 0. γ is a hyperparameter with a value range of $[0, 5]$. When its value is zero, the above formula becomes the common cross-entropy loss function (we found that $\gamma = 2$ achieves the best result in our experiments; Section 6.5). $(1 - \hat{p}^j)^\gamma$ can reduce the loss contribution of easy-to-distinguish categories. Specifically, if $\hat{p}^j$ approaches 1, meaning that j is an easy-to-distinguish category, then $(1 - \hat{p}^j)^\gamma$ approaches 0, and the contribution of this category to the loss is small; if $(1 - \hat{p}^j)^\gamma$ approaches 1, it indicates a difficult-to-distinguish category, and its loss weight increases. By introducing this modulation mechanism, the performance of the model can be significantly enhanced, and we will present detailed experimental results in Section 6.3.

For the entire training process of the LGDA model, we provide an explanation in the pseudocode shown in Algorithm 1. In Algorithm 1, given the training dataset $\mathbb{D}_{train}$, model f, and encoder network PLM, the goal is to find the model parameters θ that optimize performance. Initially, in the data augmentation module, a mini-batch of original samples (X, Y) is extracted from the training data and linearized to obtain the sample X'; then, entities in X' are randomly masked to obtain X_{mask}, and the encoder network is fine-tuned. Subsequently, the masked entity generation language model is used to predict the masked characters, thereby generating augmented samples X_{aug}. In the label semantic fusion module, the labels Y are first converted into label annotations L; then, twin encoders are used to obtain the hidden layer features of L and the sample X, and finally, the cross-attention mechanism is utilized to integrate label semantic knowledge into the sample representation.

Algorithm 1 Pseudocode of LGDA model

Require: Dataset $\mathbb{D}_{train}$, Model f, Encoder Network PLM
Ensure: Model Parameters θ
$\mathbb{D}_{aug} = \emptyset$,
for each mini-batch $(X, Y) \in \mathbb{D}_{train}$ **do**
  Obtain the original label linearization X' of X;
  Randomly mask entities in X' to get X_{mask};
  Fine-tune $PLM(X_{mask})$;
end for
for each mini-batch $(X, Y) \in \mathbb{D}_{train}$ **do**
  Obtain the original label linearization X' of X;
  Randomly mask entities in X' to get X_{mask};
  Generate new text data $X_{aug} = PLM(X_{mask})$;
  $\mathbb{D}_{aug} = \mathbb{D}_{aug} \cup (X_{aug}, Y)$;
  $\mathbb{D}_{aug} = Filter(\mathbb{D}_{aug})$;
end for
for each mini-batch $(X, Y) \in \mathbb{D}_{aug} \cup \mathbb{D}_{train}$ **do**
  Convert labels Y to label annotation L;
  Input L into the label encoder network to get its hidden layer features H_L;
  Input X into the text encoder network to get its hidden layer features H_X;
  Use cross-attention mechanism to obtain the final sample representation C through Equations (5)–(7);
  Input into the final linear layer and calculate the $softmax$ to get the final prediction probability;
  Calculate the total loss $\mathcal{L}$ through Equation (9);
  Update parameters θ to minimize the total loss $\mathcal{L}$.
end for

5. Experiments Setting

5.1. Datasets and Evaluation Metrics

To validate the efficacy and superiority of the method proposed in this paper, the model's performance was evaluated on two widely used Chinese public datasets: OntoNotes 4.0 (Chinese) (https://catalog.ldc.upenn.edu/LDC2011T03, accessed on 27 January 2025) and MSRA (https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/MSRA, accessed on 27 January 2025). The datasets utilized in the experiments of this paper are a mix of data from various domains, including news, broadcasting, and online logs. The same entity may correspond to different labels across different domains, and using such datasets can better verify the effectiveness of the label semantic fusion module introduced in the proposed method.

Specifically, the OntoNotes 4.0 (Chinese) dataset comprises a total of 24,297 sample sentences and 28,001 annotated entities, covering multiple domains, such as news, broadcasting, and online logs. The types of named entities include Person (PER), Organization (ORG), Location (LOC), and GPE. As for the MSRA dataset, it is a news domain dataset publicly released by Microsoft, totaling 50,732 sample sentences and 80,884 annotated entities, which specifically includes three types of named entities: PER, ORG, and LOC. An overall analysis of the OntoNotes 4.0 (Chinese) and MSRA datasets is shown in Table 2. Detailed statistics for each type of entity in the OntoNotes 4.0 (Chinese) and MSRA datasets are presented in Tables 3 and 4, respectively. Additionally, we employ three widely used evaluation metrics in the field of named entity recognition: precision (P), recall (R), and F1 score (F1).

Table 2. The statistical data of sentences in the OntoNotes 4.0 (Chinese) and MSRA datasets.

Dataset	Tr Set	Val Set	Test Set
OntoNotes 4.0 (Chinese)	15,650	4301	4346
MSRA	41,729	4637	4366

Table 3. Statistics on the number of various named entities in OntoNotes 4.0 (Chinese).

Category	Tr Set	Val Set	Test Set
PER	4391	1652	1864
LOC	928	483	491
GPE	4960	3166	3452
ORG	3088	1649	1877
Totally	**13,367**	**6950**	**7684**

Table 4. Statistics on the number of various named entities in MSRA.

Category	Tr Set	Val Set	Test Set
PER	16,105	1510	1973
LOC	3486	1657	2877
ORG	19,481	1090	1331
Totally	**70,446**	**4257**	**6181**

5.2. Baseline Models

This section compares the proposed LGDA model with recent baseline models related to Chinese named entity recognition tasks to validate the superiority of the proposed model. A summary of the baseline models is as follows:

- **BERT (2018)** [10]: BERT, a pre-trained language model with a labeled classifier, fully encodes words and further adds a CRF label decoding layer.
- **Lattice LSTM (2018)** [11]: A dictionary-based Chinese named entity recognition method using a lattice structure that fully utilizes the information of characters and character sequences.
- **BERT-MRC (2019)** [15]: It reconstructs the named entity recognition task as a reading comprehension task, with entity labels as query questions and answers as entities, promoting entity extraction with this query prior knowledge.
- **MECT (2021)** [20]: A cross-Transformer model based on multi-modal data embeddings that integrates the structural information of Chinese characters to improve the effect of low-resource Chinese named entity recognition.
- **Multi-Grained (2021)** [21]: A multi-grained distillation method for named entity recognition, using an improved Viterbi algorithm to construct pseudo-labels and transfer knowledge from teacher models to lightweight models.
- **LEAR (2021)** [8]: It views the named entity recognition task as a question-answering task, using preset questions for entity extraction, and proposes using label knowledge to improve text representation.
- **PCBERT (2022)** [9]: A two-stage model for low-sample Chinese named entity recognition: the first stage uses P-BERT with prompt templates to extract implicit label features, and the second stage fine-tunes a lexicon-based C-BERT with label expansion functions.
- **XLNet-Transformer-R (2023)** [16]: A model that combines XLNet and a Transformer encoder with relative positional encodings to process long text and contextual information and uses R-Drop to prevent overfitting.
- **MW-NER (2023)** [17]: A model for Chinese NER that fuses multi-granularity word information by using a strong–weak feedback attention mechanism and combines it with character information via two fusion strategies to enrich semantics and reduce segmentation errors.

- **E-strPron (2024)** [18]: A method that integrates pronunciation, radical, strokes, writing order, and lexicon features, using two cross-Transformers to capture attention and a soft fusion module to combine mutual attention.

5.3. Research Questions

To thoroughly examine the performance of the proposed LGDA model, we conduct experiments around the following research questions:

RQ1 Compared with the baseline models, can the LGDA model achieve better results in Chinese named entity recognition tasks?
RQ2 How does the LGDA model perform under the few-shot setting?
RQ3 Which module in the LGDA model contributes the most to improving the model's performance?
RQ4 What is the impact of label annotation on the LGDA model?
RQ5 How do different hyperparameter settings affect the LGDA model?

5.4. Parameter Settings

All experiments in this paper are conducted in a Python 3.8.16 environment using the deep learning toolkit PyTorch 1.8.0 (https://pytorch.org, accessed on 27 January 2025). The RoBERTa-large [31] model, which has 24 hidden layers, outputs tensor dimensions of 1024, and has 16 self-attention heads and a total of 355 M parameters, is used as the encoder in this paper. The Adam optimizer [32] is used throughout the experiments. In the data augmentation module, a batch size of 30 and a learning rate of 1×10^{-5} are used, with training for 30 epochs using the Adam optimizer. During training, 70% of the entity characters are randomly masked. In the label semantic fusion module, learning rates of 2×10^{-5} and 1×10^{-5} are used for the OntoNotes 4.0 (Chinese) and MSRA datasets, respectively, with a weight decay of 5×10^{-4} for 400 epochs of training. A learning rate scheduling strategy is adopted, where the learning rate is adjusted every 100 epochs, decaying to 0.25 of its original value. For the hyperparameters in Equation (9), α_i and γ are set to 1 and 2, respectively. All experiments in this paper are conducted on a single device with a 2.60 GHz Intel(R) Xeon(R) Platinum CPU and an Nvidia RTX 3090 GPU, using Nvidia apex 0.1 (https://github.com/NVIDIA/apex, accessed on 27 January 2025) for mixed precision training acceleration.

6. Results and Discussion

6.1. Overall Model Performance

To address **RQ1**, experiments on the overall performance of the model were conducted on the datasets OntoNotes 4.0 (Chinese) and MSRA. Table 5 presents the complete experimental results of the LGDA model and baseline models on three evaluation metrics. To facilitate direct comparison, we underline and bold the best-performing baseline model and the best results for each column of metrics.

From Table 5, it can be seen that LGDA outperforms the baseline models in most metrics on the OntoNotes 4.0 (Chinese) dataset, with improvements in the F1 metric ranging from 0.06% to 25.23% and the highest improvements in the P and R metrics being 12.20% and 38.43%, respectively. However, E-strPron has a higher P metric than the LGDA model due to the fusion of rich features. Similarly, XLNet-Transformer-R has a higher R metric than the LGDA model. The reason for this is the use of R-Drop to prevent overfitting, which can help improve recall by enhancing the model's generalization ability. On the MSRA dataset, the LGDA model performs better than the baseline models in all metrics, with improvements in the F1 metric ranging from 0.31% to 11.21% and improvements in the P and R metrics of 0.35% to 6.57% and 0.28% to 15.83%, respectively. In short, the F1

metric of the models proposed in this paper surpasses both of these methods, indicating that the LGDA model utilizes label semantic knowledge more effectively.

Table 5. Model performance comparison on the OntoNotes 4.0 (Chinese) and MSRA datasets, with the best-performing baseline models for each metric indicated by underlines and the best results emphasized in bold.

Model	OntoNotes 4.0 (Chinese)			MSRA		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
BERT + CRF	72.84	59.72	65.63	90.57	83.06	86.65
LatticeLSTM	76.35	71.56	73.88	93.57	92.79	93.18
BERT-MRC	82.98	81.25	82.11	96.18	95.12	95.75
MECT	77.57	76.27	76.92	94.55	94.09	94.32
Multi-grained KD	77.94	74.23	76.05	93.90	92.11	92.99
LEAR	81.23	83.07	82.14	96.18	95.42	95.82
PCBERT	-	-	81.57	-	-	89.72
XLNet-Transformer-R	79.34	**83.16**	81.21	92.07	94.01	93.03
MW-NER	-	-	81.51	-	-	95.47
E-strPron	**84.90**	77.96	81.28	96.18	95.94	96.06
LGDA	81.72	82.67	**82.19**	**96.52**	**96.21**	**96.36**

- **Conclusion 1:** The proposed model, LGDA, performs better than the baseline models in the majority of metrics and achieves superior results in Chinese named entity recognition tasks.

6.2. Few-Shot Experiment

To address **RQ2**, we conduct few-shot experiments, and the performance of the model on the two datasets OntoNotes 4.0 (Chinese) and MSRA is shown in Table 6. To more intuitively demonstrate the superiority of the proposed model under the few-shot setting, the best baseline model LEAR is chosen for comparison. The experimental results of LEAR compared in this paper come from the related few-shot exploration experiments in that research work [8]. The experimental results show that whether under the one-shot or five-shot setting, the improvement in the model LGDA proposed in this paper is very significant. Specifically, in the one-shot scenario, LGDA can absolutely improve the F1 metric by 15.48–16.14 compared with the baseline model; in the five-shot scenario, the absolute improvement is 11.33–30.56. When the data augmentation module is removed, taking the 1-shot setting on the OntoNotes 4.0 (Chinese) dataset as an example, the model shows absolute decreases of 9.95, 8.04, and 11.28 in the F1, P, and R metrics, respectively. This indicates that in the few-shot scenario, data augmentation significantly enhances the model's recognition effectiveness.

- **Conclusion 2:** In the few-shot scenario, the LGDA model still significantly outperforms the baseline models, with the data augmentation module playing the most significant role.

Table 6. Few-shot experiments on the OntoNotes 4.0 (Chinese) and MSRA datasets.

Settings	Model	OntoNotes 4.0 (Chinese)			MSRA		
		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
1-shot	LEAR	-	-	7.28	-	-	0.39
	LGDA w/o Data Augmentation	15.09	11.13	12.81	3.74	13.35	5.85
	LGDA	**23.13**	**22.41**	**22.76**	**12.70**	**23.63**	**16.53**
5-shot	LEAR	-	-	41.32	-	-	26.22
	LGDA w/o Data Augmentation	42.41	46.16	44.20	39.44	48.13	43.35
	LGDA	**44.88**	**63.67**	**52.65**	**50.98**	**64.08**	**56.78**

6.3. Ablation Study

To address **RQ3**, we test the performance effects of removing each of the three components of the LGDA model—data augmentation module, label semantic fusion module, and loss function optimization. Table 7 presents the results of the ablation study. It is evident that removing each module leads to a decline in performance to varying degrees, indicating that all three modules play an indispensable role in enhancing the model's entity recognition capabilities. Specifically, on the MSRA dataset, removing the data augmentation module results in the most severe model degradation, with decreases of 0.49%, 0.93%, and 0.05% in the F1, P, and R metrics, respectively. This suggests that applying data augmentation to named entity recognition in low-resource scenarios is effective and plays the most significant role in improving recognition accuracy. On the OntoNotes 4.0 (Chinese) dataset, the removal of the label semantic fusion module has the greatest impact, with the F1 metric dropping by 1.36%, indicating that label semantic knowledge can enhance the model's ability to distinguish positive samples. The role of loss function optimization is primarily reflected in the enhancement in the P metric; by adaptively tuning the weights, the model can learn from both positive and negative samples sufficiently, achieving improved recognition precision. Without this module, the P metric would decrease by 1.98% and 0.53% on the OntoNotes 4.0 (Chinese) and MSRA datasets, respectively.

Table 7. Ablation study results on the OntoNotes 4.0 (Chinese) and MSRA datasets.

Model	OntoNotes 4.0 (Chinese)			MSRA		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
LGDA	81.72	82.67	82.19	96.52	96.21	96.36
- Data augmentation module	80.98	81.92	81.45	95.63	96.16	95.89
- Label semantic fusion module	80.08	82.12	81.09	95.75	96.25	96.00
- Loss function optimization	80.13	82.35	81.22	96.01	95.80	95.91

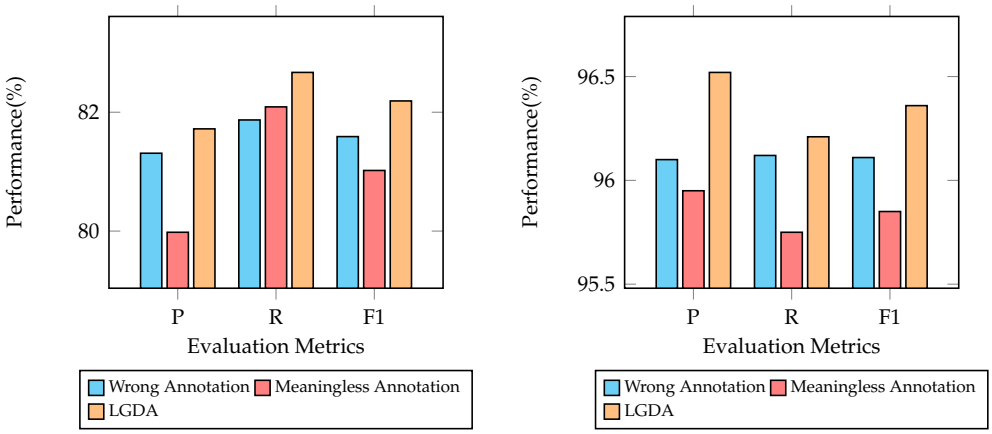
- **Conclusion 3:** In low-resource scenarios, the data augmentation module contributes the most to improving model performance; for scenarios with fewer entity types, the label semantic fusion module of the proposed model plays the most significant role in enhancing recognition effectiveness.

6.4. Label Annotation Experiment

To address **RQ4**, we conduct exploratory experiments in label annotation on the datasets OntoNotes 4.0 (Chinese) and MSRA. As shown in Figure 3, incorrect or meaningless label annotation can lead to varying degrees of performance decline in the model. Figure 3a presents the results on the OntoNotes 4.0 (Chinese) dataset, where incorrect label annotations cause the metrics F1, P, and R to decrease by 0.74%, 0.50%, and 1.41%,

respectively, while meaningless label annotations result in decreases of 1.44%, 2.18%, and 0.71% in the metrics F1, P, and R, respectively. Figure 3b shows similar results on the MSRA dataset. This indicates that label annotations indeed play a positive role in improving the model's recognition accuracy, and the label semantic fusion module proposed in this paper can use labels to guide model training. At the same time, these experimental results also reflect the importance of selecting label annotations, as their quality can directly impact the model.

- **Conclusion 4:** Correct label annotation contributes to enhancing effectiveness in named entity recognition.



(**a**) OntoNotes 4.0 (Chinese) (**b**) MSRA

Figure 3. Analysis of the impact of label annotation on the OntoNotes 4.0 (Chinese) and MSRA datasets.

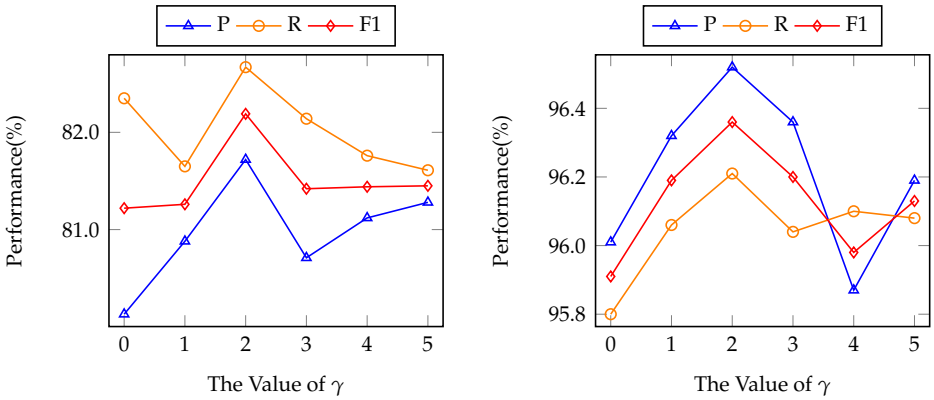
6.5. Hyperparameter Experiment

To address **RQ5**, this study investigates the impact of the γ parameter introduced in the loss function optimization module of this paper within its range of $[0, 5]$ by examining the cases where γ is 0, 1, 2, 3, 4, and 5. The performance of the LGDA model on the OntoNotes 4.0 (Chinese) and MSRA datasets is shown in Figure 4. From Figure 4a, it can be observed that the values of the model's metrics are positively correlated with the value of γ at the beginning. The model performs best when $\gamma = 2$, and beyond this threshold, an increase in γ leads to progressively worse results across all metrics. The reason is that a higher γ increases the model's focus on difficult-to-distinguish samples; however, in practice, easy-to-distinguish samples constitute the majority of the training data. If the model overemphasizes difficult-to-distinguish samples, it may fail to learn from easy-to-distinguish samples adequately, leading to a decline in performance. On the OntoNotes 4.0 (Chinese) dataset, the differences in model performance for the F1, P, and R metrics between the best and worst cases are 1.19%, 1.98%, and 1.30%, respectively. On the MSRA dataset, similar results are observed, as shown in Figure 4b, where the model also performs best when $\gamma = 2$, with the extreme differences in the F1, P, and R metrics being 0.47%, 0.68%, and 0.43%, respectively. Therefore, in the other exploratory experiments of this paper, 2 is chosen as the value for γ.

Compared with the OntoNotes 4.0 (Chinese) dataset, on the MSRA dataset, the model is less affected by γ. The reason for this is that the MSRA dataset is larger and has a more balanced sample distribution; hence, changing the value of γ has a smaller impact on the experimental results. In contrast, the OntoNotes 4.0 (Chinese) dataset has a smaller sample size but a greater number of categories, with a more uneven distribution of easy and difficult samples. Therefore, the loss function optimization module plays a more significant

role, and the model performance is more affected by the value of γ. Experimental validation has shown that when $\gamma = 2$, the model achieves the best results on both datasets.

- **Conclusion 5:** When the hyperparameter $\gamma = 2$, the model achieves a balance in focusing on easy- and difficult-to-distinguish samples, resulting in optimal performance.



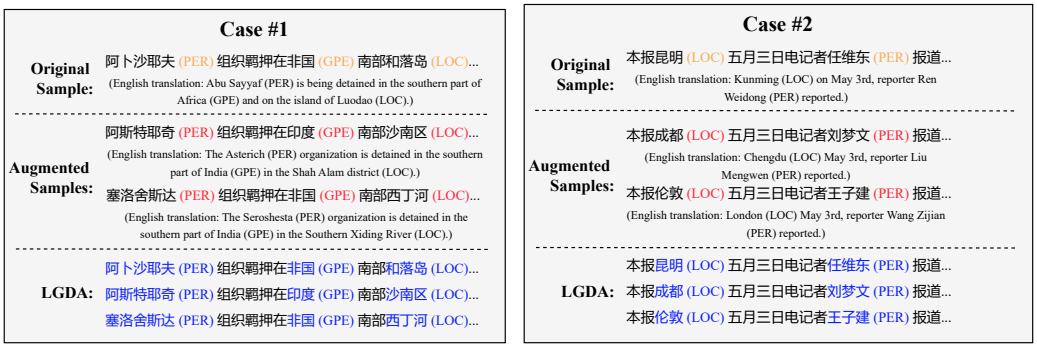
(**a**) OntoNotes 4.0 (Chinese) (**b**) MSRA

Figure 4. The impact of γ values on model performance on the OntoNotes 4.0 (Chinese) and MSRA datasets.

6.6. Case Study

To provide a clearer illustration of the LGDA model's capabilities, we further analyze the augmentation data. Figure 5 presents two specific examples of data augmentation generated by the LGDA model, along with the corresponding recognition results. Case #1 and Case #2 are drawn from the OntoNotes 4.0 (Chinese) and MSRA datasets, respectively. In these examples, entity types in the original samples are marked in orange, entity types in the augmented samples are highlighted in red, and the entities and their types identified by the LGDA model are indicated in blue.

In Figure 5a, Case #1, the original text "阿卜沙耶夫组织羁押在非国南部和落岛" contains entities of the types "PER" ("阿卜沙耶夫"), "GPE" ("非国"), and "LOC" ("和落岛"). After processing by the LGDA model, multiple augmented samples are generated, which not only increases the scale of training data but also enhances the richness of the samples. It can be seen from the figure that the augmented samples produced by the LGDA model ensure the consistency of entity semantics and entity labels and do not change the length of the entities. The figure also shows the recognition results of the LGDA model, which can correctly identify the boundaries and types of entities. Similarly, in Figure 5b, Case #2, the model proposed in this paper generates multiple augmented samples with clear entity boundaries and complete semantics, such as the entity "昆明" of type "LOC" in the original sample being replaced by "成都" and "伦敦" in the augmented samples. In summary, the LGDA model can effectively address the low-resource issue through the sample augmentation module, which is beneficial for improving accuracy in entity recognition.



(**a**) The case from OntoNotes 4.0 (Chinese). (**b**) The case from MSRA.

Figure 5. The case study of LGDA.

7. Conclusions

In this paper, we propose LGDA, a novel approach to addressing the challenges of NER in low-resource scenarios. The framework consists of three key components: (1) a data augmentation module utilizing a masked entity generation strategy to enhance the limited dataset, (2) a label semantic fusion module incorporating a cross-attention mechanism to capture fine-grained label information, and (3) an adaptive loss function designed to mitigate the long-tail distribution problem. Through this two-stage architecture, LGDA effectively enriches sample representations and focuses on hard samples, thereby addressing the fundamental limitations of scarce annotated data. Extensive experimental results on the OntoNotes 4.0 (Chinese) and MSRA datasets demonstrate that LGDA significantly outperforms existing baselines, particularly in few-shot learning scenarios. These findings validate the effectiveness of our proposed approach in enhancing NER performance under low-resource conditions. Future research directions include extending LGDA to multilingual applications and exploring the integration of additional semantic features to further improve model robustness.

Author Contributions: Methodology, M.J.; Writing—original draft, M.J.; Visualization, M.J.; Supervision, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Hunan Provincial Innovation Foundation for Postgraduate: XJQY2024045.

Data Availability Statement: The data presented in this study are openly available at https://catalog.ldc.upenn.edu/LDC2011T03 and https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/MSRA (accessed on 27 January 2025).

Acknowledgments: We express our gratitude to the National Key Laboratory of Information Systems Engineering for providing the experimental facilities.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Supplementary Experiments

Appendix A.1. The Effect of Augmented Samples

We conducted experiments on the MSRA dataset to examine the effect of augmented samples on the recognition performance across various classes, and the results are presented in Figure A1 below. After incorporating augmented samples into the model, the F1 scores for entity recognition of the PER, LOC, and ORG types were improved by 0.43%, 0.87%, and 0.54%, respectively. It is evident that the use of augmented samples has enhanced our model's recognition performance across various categories, particularly for "LOC"-type entities, which are the least numerous.

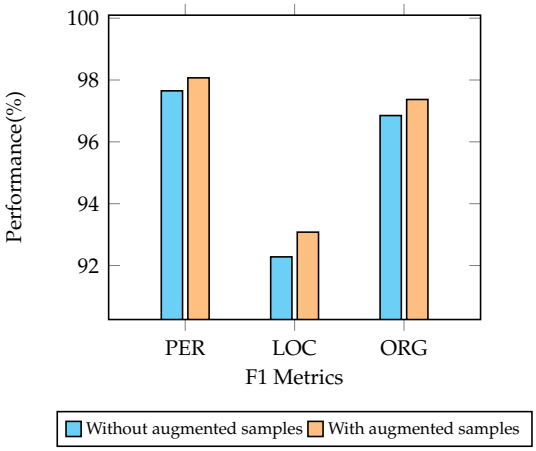


Figure A1. The effect of augmented samples on the MSRA dataset.

Appendix A.2. Efficiency Evaluation of Model

We assessed the efficiency of different models by comparing their training and inference time costs (in seconds). Specifically, training refers to the time cost required for a single training epoch, while inference refers to the time cost for the model to generate all prediction results on the test set. We choose BERT [10] and BERT-MRC [15] as our baselines for efficiency evaluation. The results are shown in Table A1. On both the OntoNotes 4.0 (Chinese) and MSRA datasets, BERT+CRF demonstrates the lowest training and inference times. Nonetheless, its F1 scores are relatively modest, standing at 65.63 and 86.65, respectively (see Section 6.1). BERT-MRC transforms the named entity recognition task into a reading comprehension problem by treating entity labels as questions and answers as entities. This method expands the sample set and sequence lengths, thereby increasing the computational time for training and inference. The LGDA model strikes a balance between efficiency and performance. While it does entail a moderate increase in computational cost relative to BERT, it is likely to offer enhanced capabilities when compared with both the BERT+CRF and BERT-MRC models. It is worth noting that our model has almost no difference in inference time compared with BERT+CRF, which also proves that although we use a twin encoder, the label encoder does not increase the computational load too much during the inference process.

Table A1. The efficiency evaluation of different methods. $(\cdot)$ indicates the relative efficiency compared with BERT+CRF.

Model	OntoNotes 4.0 (Chinese)		MSRA	
	Training	**Inference**	**Training**	**Inference**
BERT + CRF	58.3 (×1.0)	5.2 (×1.0)	167.8 (×1.0)	5.1 (×1.0)
BERT − MRC	258.1 (×4.4)	19.7 (×3.8)	626.7 (×3.7)	14.5 (×2.8)
LGDA	116.4 (×2.0)	5.8 (×1.1)	314.4 (×1.87)	5.4 (×1.1)

Appendix A.3. Impact of Pre-Trained Language Model

To rule out the possibility that the enhancements in model performance are due to the pre-trained prior knowledge embedded in the RoBERTa model, we replaced RoBERTa with BERT and conducted experiments on the OntoNotes 4.0 (Chinese) and MSRA datasets. The results are presented in Table A2 below. It can be observed that the model performance experiences a slight decrease, amounting to an absolute decline of only 0.06–0.07. Moreover, the model's effectiveness after the decline still holds a significant advantage. Especially on the MSRA dataset, the LGDA model, when equipped with BERT, achieves a performance

that is absolutely 0.23 higher than the best baseline. This is sufficient to demonstrate that the efficacy of the method we propose is not contingent upon the use of RoBERTa.

Table A2. The impact of the pre-trained language model. The best results among the baseline models are underlined.

Model	OntoNotes 4.0 (Chinese)			MSRA		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
LEAR	81.23	83.07	<u>82.14</u>	96.18	95.42	95.82
E-strPron	84.90	77.96	81.28	96.18	95.94	<u>96.06</u>
LGDA (with RoBERTa)	81.72	82.67	82.19	96.52	96.21	96.36
LGDA (with BERT)	81.67	82.59	82.13	96.40	96.19	96.29

References

1. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 50–70. [CrossRef]
2. Dai, X.; Adel, H. An analysis of simple data augmentation for named entity recognition. *arXiv* **2020**, arXiv:2010.11683.
3. Min, J.; McCoy, R.T.; Das, D.; Pitler, E.; Linzen, T. Syntactic data augmentation increases robustness to inference heuristics. *arXiv* **2020**, arXiv:2004.11999.
4. Gao, F.; Zhu, J.; Wu, L.; Xia, Y.; Qin, T.; Cheng, X.; Zhou, W.; Liu, T.Y. Soft contextual data augmentation for neural machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5539–5544.
5. Hou, Y.; Liu, Y.; Che, W.; Liu, T. Sequence-to-sequence data augmentation for dialogue language understanding. *arXiv* **2018**, arXiv:1807.01554.
6. Zhou, R.; Li, X.; He, R.; Bing, L.; Cambria, E.; Si, L.; Miao, C. MELM: Data augmentation with masked entity language modeling for low-resource NER. *arXiv* **2021**, arXiv:2108.13655.
7. Ma, J.; Ballesteros, M.; Doss, S.; Anubhai, R.; Mallya, S.; Al-Onaizan, Y.; Roth, D. Label semantics for few shot named entity recognition. *arXiv* **2022**, arXiv:2203.08985.
8. Yang, P.; Cong, X.; Sun, Z.; Liu, X. Enhanced language representation with label knowledge for span extraction. *arXiv* **2021**, arXiv:2111.00884.
9. Lai, P.; Ye, F.; Zhang, L.; Chen, Z.; Fu, Y.; Wu, Y.; Wang, Y. PCBERT: Parent and Child BERT for Chinese Few-shot NER. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 2199–2209.
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
11. Zhang, Y.; Yang, J. Chinese NER using lattice LSTM. *arXiv* **2018**, arXiv:1805.02023.
12. Meng, Y.; Wu, W.; Wang, F.; Li, X.; Nie, P.; Yin, F.; Li, M.; Han, Q.; Sun, X.; Li, J. Glyce: Glyph-vectors for chinese character representations. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 2746–2757.
13. Liu, W.; Fu, X.; Zhang, Y.; Xiao, W. Lexicon enhanced Chinese sequence labeling using BERT adapter. *arXiv* **2021**, arXiv:2105.07148.
14. Li, X.; Yan, H.; Qiu, X.; Huang, X. FLAT: Chinese NER using flat-lattice transformer. *arXiv* **2020**, arXiv:2004.11795.
15. Li, X.; Feng, J.; Meng, Y.; Han, Q.; Wu, F.; Li, J. A unified MRC framework for named entity recognition. *arXiv* **2019**, arXiv:1910.11476.
16. Ji, W.; Zhang, Y.; Zhou, G.; Wang, X. Research on Named Entity Recognition in Improved transformer with R-Drop structure. *arXiv* **2023**, arXiv:2306.08315.
17. Liu, T.; Gao, J.; Ni, W.; Zeng, Q. A Multi-Granularity Word Fusion Method for Chinese NER. *Appl. Sci.* **2023**, *13*, 2789. [CrossRef]
18. Wang, Y.; Chen, Q.; Zhou, Q. E-strPron: Chinese named entity recognition based on enhanced structure and pronunciation features. In Proceedings of the 2024 43rd Chinese Control Conference (CCC), Kunming, China, 28–31 July 2024.
19. Jia, C.; Shi, Y.; Yang, Q.; Zhang, Y. Entity enhanced BERT pre-training for Chinese NER. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 6384–6396.
20. Wu, S.; Song, X.; Feng, Z. MECT: Multi-metadata embedding based cross-transformer for Chinese named entity recognition. *arXiv* **2021**, arXiv:2107.05418.
21. Zhou, X.; Zhang, X.; Tao, C.; Chen, J.; Xu, B.; Wang, W.; Xiao, J. Multi-Grained Knowledge Distillation for Named Entity Recognition. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 5704–5716.

22. Lee, D.H.; Kadakia, A.; Tan, K.; Agarwal, M.; Feng, X.; Shibuya, T.; Mitani, R.; Sekiya, T.; Pujara, J.; Ren, X. Good examples make A faster learner: Simple demonstration-based learning for low-resource NER. *arXiv* **2021**, arXiv:2110.08454.
23. Li, D.; Hu, B.; Chen, Q. Prompt-based Text Entailment for Low-Resource Named Entity Recognition. *arXiv* **2022**, arXiv:2211.03039.
24. Park, E.; Jeon, D.; Kim, S.; Kang, I.; Na, S.H. LM-BFF-MS: Improving Few-Shot Fine-tuning of Language Models based on Multiple Soft Demonstration Memory. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Dublin, Ireland, 22–27 May 2022; pp. 310–317.
25. Ma, R.; Zhou, X.; Gui, T.; Tan, Y.; Zhang, Q.; Huang, X. Template-free prompt tuning for few-shot NER. *arXiv* **2021**, arXiv:2109.13532.
26. Li, X.; Zhang, H.; Zhou, X.H. Chinese clinical named entity recognition with variant neural structures based on BERT methods. *J. Biomed. Inform.* **2020**, *107*, 103422. [CrossRef] [PubMed]
27. Xu, L.; Li, S.; Wang, Y.; Xu, L. Named Entity Recognition of BERT-BiLSTM-CRF Combined with Self-attention. In Proceedings of the International Conference on Web Information Systems and Applications, Chengdu, China, 1–3 August 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 556–564.
28. Ji, B.; Li, S.; Gan, S.; Yu, J.; Ma, J.; Liu, H. Few-shot Named Entity Recognition with Entity-level Prototypical Network Enhanced by Dispersedly Distributed Prototypes. *arXiv* **2022**, arXiv:2208.08023.
29. Kumar, V.; Choudhary, A.; Cho, E. Data augmentation using pre-trained transformer models. *arXiv* **2020**, arXiv:2003.02245.
30. Liu, L.; Ding, B.; Bing, L.; Joty, S.; Si, L.; Miao, C. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. 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), Virtual, 1–6 August 2021; pp. 5834–5846.
31. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
32. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.