PCBERT: Parent and Child BERT for Chinese Few-shot NER

Peichao Lai Feiyang Ye Lin Zhang Zhiwei Chen
Yanggeng Fu Yingjie Wu Yilei Wang
College of Computer and Data Science, Fuzhou University, China
yilei@fzu.edu.cn

Abstract

Achieving good performance on few-shot or zero-shot datasets has been a long-term challenge for NER. The conventional semantic transfer approaches on NER will decrease model performance when the semantic distribution is quite different, especially in Chinese few-shot NER. Recently, prompt-tuning has been thoroughly considered for low-resource tasks. But there is no effective prompt-tuning approach for Chinese few-shot NER. In this work, we propose a prompt-based Parent and Child BERT (PCBERT) for Chinese few-shot NER. To train an annotating model on high-resource datasets and then discover more implicit labels on low-resource datasets. We further design a label extension strategy to achieve label transferring from high-resource datasets. We evaluated our model on Weibo and the other three sampling Chinese NER datasets, and the experimental result demonstrates our approach's effectiveness in few-shot learning.

1 Introduction

NER is a fine-grained sequence labeling task, a slight change in each token will significantly impact the model results. A big challenge of NER is to enhance the performance in low-resource scenarios. There are some prior works (Yang et al., 2017; Lee et al., 2017; Abhishek et al., 2017) that demonstrate that transfer learning can improve the model performance. However, they all rely on similar semantic distribution between source and target datasets, and both datasets should contain rich annotated data. A significant difficulty of few-shot or zero-shot NER is the lack of annotated labels in practical application. Another challenge of Chinese NER is the implicit word boundary, which makes it difficult for the model to distinguish the entity boundary. The lexicon-based approach is a standard solution to solve the above issue. But the performance of traditional lexicon-based models in Chinese few-shot NER is still unsatisfactory.

Recently, prompt-tuning (Lester et al., 2021) on the pre-trained language models (PLMs) has been thoroughly considered for low-resource scenarios because the prompt-tuning process is highly consistent with the target task. Previous work (Cui et al., 2021; Ma et al., 2022; Chen et al., 2021) has demonstrated that prompt-tuning can more effectively enhance the model performance on few-shot NER compared with fine-tuning. However, when the semantic distribution is quite different, using prompt-tuning for semantic transfer learning will decrease model performance, which implies the semantic transfer is unsuitable for NER in the above situation. Besides, the implicit boundaries of Chinese words make the size of the prompt template uncertain and require a higher ability to judge its boundary. Moreover, using inappropriate prompt construction engineering on Chinese few-shot NER datasets can not improve model performance effectively but increases training time.

In this work, we introduce an enhanced lexicon feature and a prompt-based label transfer approach to address the above issues. We leverage the lexicon feature to enhance Chinese word boundary distinction ability in few-shot NER datasets. We further design a label extension strategy to achieve label transferring from high-resource datasets. We propose a Parent and Child BERT(PCBERT) model powered by a label lexicon adapter and a prompt-tuning component to integrate the lexicon feature and the implicit label feature. And it is worth noting that our implementation with a transformer encoder is more efficient than some decoding template-based approaches. We evaluated our model on Weibo(Peng and Dredze, 2015) and the other three samplings of Chinese NER datasets, and the experimental result demonstrates our approach's effectiveness in few-shot learning. Our model outperforms other related work in all experiments and achieves state-of-the-art F1 scores on Weibo.

Proceedings of the 29th International Conference on Computational Linguistics, pages 2199–2209
October 12–17, 2022.

The contributions of this work can be summarized as follows:

1. We introduce a label extension strategy to implement the label transfer learning in few-shot NER, which can effectively enhance the model performance.

2. We propose a new PCBERT model consisting of a P-BERT component and a C-BERT component to integrate the lexicon feature and the implicit label feature.

3. Experimental results verify that our approaches are suitable for Chinese few-shot NER transfer learning and achieve excellent performance on few-shot learning.

2 Preliminaries

2.1 Problems of Few-shot NER

In the few-shot NER tasks, given the high-resource source domain dataset $S = \{P_S, L_S\}$, where the $P_S = \left\{\left(X_S^1, Y_S^1\right), \ldots, \left(X_S^R, Y_S^R\right)\right\}$ is the set of input text and corresponding labels, and $L_S = \{l_1, \ldots, l_m\}$ is the set of entity label categories with size m. Then given the low-resource target domain dataset $T = \{P_T, L_T\}$, the task aims to enhance the model performance in the target domain dataset by utilizing the resources of the source domain dataset. However, the traditional NER transfer learning approaches face two main challenges: the semantic distribution difference between the source and target domains; the same category labels have different definitions in different datasets.

2.2 Label Extension Strategy

Formally, we denote $\mathcal{D}_P(X)$ to represent the semantic space of input text X, and $\mathcal{D}_\mathcal{E}(L)$ represent the semantic distribution that contains label $l \in L$. The correlation between model performance p and the semantic distribution can be explained as:

$$p \propto \frac{\mathcal{D}_P(X_S) \cap \mathcal{D}_P(X_T)}{\mathcal{D}_P(X_S) - \mathcal{D}_P(X_T)} \quad (1)$$

$$p \propto \frac{\mathcal{D}_\mathcal{E}(L_S) \cap \mathcal{D}_\mathcal{E}(L_T)}{\mathcal{D}_\mathcal{E}(L_S) - \mathcal{D}_\mathcal{E}(L_T)} \quad (2)$$

when the semantic space gap between the source domain and the target domain is large, the semantic intersection of S and T is quite limited compared with the semantic difference between S and T (i.e., $\mathcal{D}_P(X_S) \cap \mathcal{D}_P(X_T) \ll \mathcal{D}_P(X_S) - \mathcal{D}_P(X_T)$). The semantic deviation makes the pre-trained model more difficult to fine-tune than the uniform distribution model and even decreases performance in the target domain. Therefore, it is tough to carry out cross-domain semantic migration on few-shot NER datasets.

In this work, we use label extension to enrich the label features in T. As shown in Equation 2, $\mathcal{D}_\mathcal{E}(L_S) \cap \mathcal{D}_\mathcal{E}(L_T)$ represents the semantic distribution range that implicitly contains the intersection of L_S and L_T. It may include entity labels from S and does not exist in T, making label extension reasonable as T is a low-resource dataset with fewer labels. The label extension can be implemented with an annotation model with fully supervised training on S and annotating on T. However, some issues may impact the label extension accuracy. One is the annotation model performance; another is that the same category labels may explain the different meanings between S and T. These issues can be treated as label noise that affects the target task performance. To address the above issues, we adopt a prompt-based approach with a label fusion layer in our proposed model to reduce the influence of label noise.

3 Method

In this paper, we propose a two-stage model named PCBERT for Chinese few-shot NER, which consists of the Parent and the Child component. Both components are implemented with BERT (Devlin et al., 2019), and we defined them as P-BERT and C-BERT, respectively. The overall model structure of PCBERT is illustrated in Figure 1. The P-BERT is a prompt-based model to extract the implicit label extension features in the target dataset; the C-BERT is a lexicon-based model inspired by the LEBERT (Liu et al., 2021a) and further incorporates multi-label features of each lexicon. In the first stage, the P-BERT pre-trains on the label extension dataset. Then the P-BERT is set to be frozen in the second stage, providing label extension features to fine-tune the C-BERT. The structure and functionality are described in the following.

3.1 P-BERT

The primary function of P-BERT is prompt-tuning on the label extension dataset and providing prompt features for C-BERT. The label extension dataset is constructed by the method mentioned in Section 2.2. The inspiration for prompt-tuning comes from models like GPT-3 (Brown et al., 2020), and T5 (Raffel et al., 2020), which transform the tar-

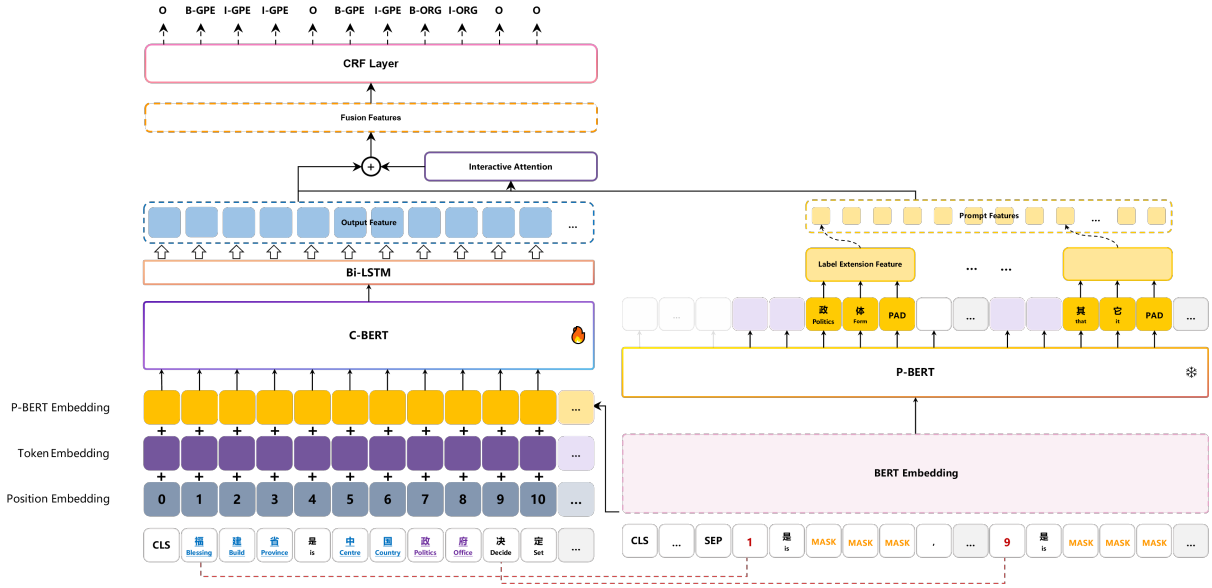

Figure 1: The overall structure of PCBERT. The P-BERT is trained on the label extension dataset in the prompt-tuning stage and provides label extension features for the C-BERT in the fine-tuning stage. While fine-tuning, the P-BERT is set to be frozen.

get task into text-to-text form and directly model text using PLMs. In this work, our prompt-tuning approach is designed toward the target task, consisting of a template function $TP(X, Y)$ that converts the raw input to prompt input. The label in the template input is a textual string instead of an entity category index, which helps leverage implicit knowledge from PLMs and reduces the influence of label noise in the label extension dataset.

We use vanilla BERT as P-BERT, each input $X = \{x_1, \ldots, x_n\}$ in the label extension dataset is converted into prompt input X_{prompt} with the $TP(X, Y)$. The prompt input consists of the following parts:

$$X_{prompt} = [\text{CLS}]\, X\, [\text{SEP}]\, TP\,(X, Y) \qquad (3)$$

where the first part of X_{prompt} is the origin input X, and the second part is label templates computed by the $TP(X, Y)$. [CLS] and [SEP] are the special token of BERT. Each label template follows the form as "$[Index]is[Z]$", where the index slot $[Index]$ indicates each token position in X, and the label slot $[Z]$ is the Chinese word that represents the label Y. Each label template is concatenated with a comma. Then, the label slot is padded to the same size by the tokenizer to adapt parallel training better and locate the output features. During prompt-tuning, the label slot of each input will be masked with the [MASK] token, and its task goal is to restore the masked label tokens. Then the loss function can be defined with the cross-entropy loss:

$$\mathcal{L}_{prompt} = -\sum_{i} z_i \log\left(p\left(\hat{\mathbf{z}}_\mathbf{i} \mid X\right)\right) \qquad (4)$$

where $z_i \in Z$ and $\hat{\mathbf{z}}_\mathbf{i}$ is the corresponding predicted token.

3.2 C-BERT

Chinese NER tasks are more challenging because the word boundary of sentences is not explicit. Many works (Sui et al., 2019; Li et al., 2020; Zhang and Yang, 2018) have demonstrated that leveraging lexicon information can effectively enhance the model performance. In few-shot NER, the lexicon information is vital in promoting the model to understand token-level semantic information. For each input sequence X, we construct a lexicon tree following the method of (Liu et al., 2021a). As shown in Figure 2, the lexicon set of token x_i can be embedded as $\omega_i = \{\omega_{i1}, \ldots \omega_{im}\}$, where $x_i \in \mathbb{R}^{1\times H}$,$\omega_i \in \mathbb{R}^{m\times H'}$, H is the hidden dimension of each token and H' is the hidden dimension of each word. Moreover, we further introduce a label set for each word. In this work, we adopt a BERT classifier model pre-trained on the high-resource dataset to predict top-k labels embeddings $L_{ij} = \left\{L_{ij}^1, \ldots, L_{ij}^k\right\}$ for ω_{ij}, where $L_{ij} \in \mathbb{R}^{k\times H^*}$, H^* is the hidden dimension of a

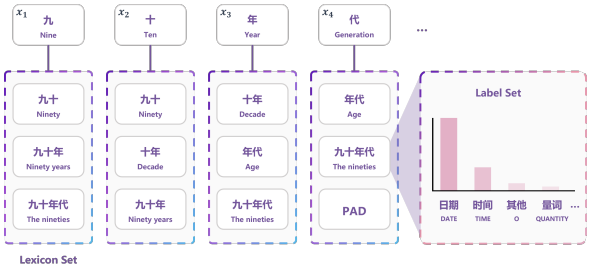

Figure 2: Each token x_i corresponds to a lexicon set, and each lexicon corresponds to a label set.

label string. It is worth noting that each lexicon comes from the external dictionary and is a subset of the input.

A variant of LEBERT is designed to serve as C-BERT in our implementation. As shown in Figure 1, C-BERT's word embedding is the sum of the P-BERT and its word embeddings. We propose a Label Lexicon Adapter (LLA) after the first encoder layer in C-BERT to leverage the lexicon and corresponding labels information. Figure 3. displays the detailed structure of C-BERT, where $H_1^o = \{h_1^o, \ldots, h_n^o\}$ is the set of original output hidden states in the first encoder layer, where the n is the length of the input sequence. In the LLA, the input contains the hidden states H^o from the first encoder layer; the lexicon set ω_i in each token position and corresponding top-k label embedding $L_i = \{L_{i1}, \ldots L_{im}\}$.

We use label attention to compute the relevance between multi-label and lexicon context features, and $\xi_{ij} = [h_i^o; \omega_{ij}]$ is the concatenation between word ω_{ij} and the hidden state h_i^o in position i. Then we transform the multi-label features to align the lexicon context features:

$$\tilde{L}_{ij} = \boldsymbol{W}_2^{\mathrm{L}} \left(\tanh\left(\boldsymbol{W}_1^{\mathrm{L}} L_{ij}^T + \boldsymbol{b}_1^{\mathrm{L}}\right)\right) + \boldsymbol{b}_2^{\mathrm{L}} \quad (5)$$

where $\boldsymbol{W}_1^{\mathbf{L}} \in \mathbb{R}^{(H'+H)\times H^*}$ and $\mathbf{W}_2^{\mathbf{L}} \in \mathbb{R}^{(H'+H)\times(H'+H)}$ are weight matrices; $\boldsymbol{b}_1^{\mathrm{L}}$, $\boldsymbol{b}_2^{\mathrm{L}}$ are biases. The label attention score can be calculated as:

$$\boldsymbol{\alpha}_{ij} = \mathrm{softmax}\left(\xi_{ij} \boldsymbol{W}_{attn}^{\mathrm{L}} \tilde{L}_{ij}\right) \quad (6)$$

where $\boldsymbol{W}_{\mathrm{attn}}^{\mathrm{L}} \in \mathbb{R}^{(H'+H)\times(H'+H)}$ is the label attention weight matrix. The multi-label features can be further computed by the weighted sum:

$$F_{ij}^{\mathrm{L}} = \frac{1}{k} \sum_{t=1}^{k} \alpha_{ij} \tilde{L}_{ij}^t \quad (7)$$

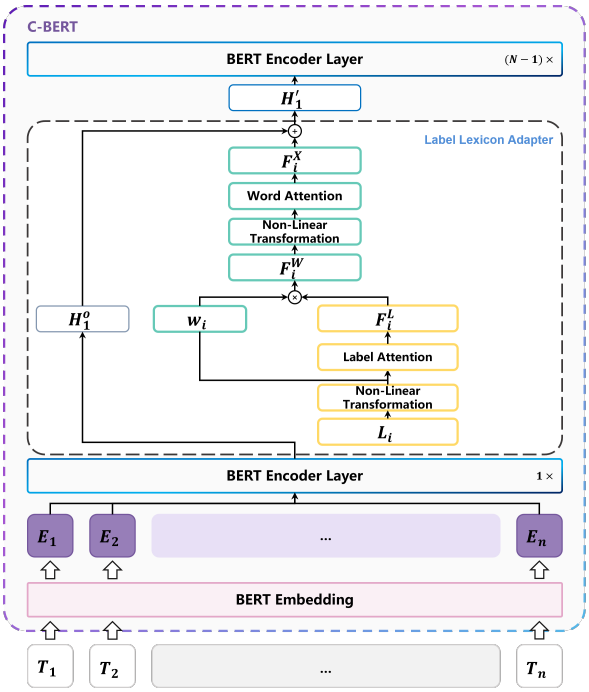

Figure 3: The implementation details of C-BERT.

We fuse features of lexicons with the corresponding label sets to enhance the lexicon representation, and the multi-label features can effectively alleviate the label noise from P-BERT:

$$F_{ij}^{\omega} = \left[\omega_{ij}; F_{ij}^{\mathrm{L}}\right] \quad (8)$$

The computed lexicon features F_i^{ω} are directly injected into the BERT following (Liu et al., 2021a) with the word attention, the lexicon information is calculated by:

$$\widetilde{\omega}_{ij} = \boldsymbol{W}_2^{\omega} \left(\tanh\left(\boldsymbol{W}_1^{\omega} F_{ij}^{\omega^T} + \boldsymbol{b}_1^{\omega}\right)\right) + \boldsymbol{b}_2^{\omega} \quad (9)$$

$$\beta_{ij} = \mathrm{softmax}\left(h_i^o \boldsymbol{W}_{attn}^{\omega} \widetilde{\omega}_{ij}\right) \quad (10)$$

$$F_{ij}^{\mathrm{X}} = \frac{1}{m} \sum_{j=1}^{m} \beta_{ij} \widetilde{\omega}_{ij} \quad (11)$$

where $\boldsymbol{W}_1^{\omega} \in \mathbb{R}^{H\times(H'+H^*)}$, $\mathbf{W}_2^{\omega} \in \mathbb{R}^{H\times H}$ are weight matrices; $\boldsymbol{W}_{attn}^{\omega} \in \mathbb{R}^{H\times H}$ is the word attention weight matrix; and $\boldsymbol{b}_1^{\omega}$, $\boldsymbol{b}_2^{\omega}$ are biases.

Finally, the fusion features of each token are computed by:

$$H_1' = H_1^o + F^{\mathrm{X}} \quad (12)$$

3.3 Interactive Training

During fine-tuning, the primary function of P-BERT is to provide label extension features for C-BERT. We intercept the label templates part of the P-BERT output, and the label extension features $F_i^P = \{f_1, \ldots, f_d\}$ are the label slot part corresponding to each label template, where d is the max size of the label string. Then the prompt feature for each token is computed as:

$$P_i = \frac{1}{d} \sum_{j=1}^{d} f_j \tag{13}$$

We use a bidirectional LSTM (BiLSTM) model to enhance the timing information of C-BERT output:

$$H^{\mathrm{B}} = \mathrm{BiLSTM}\left(H_N^{\mathrm{o}}\right) \tag{14}$$

where $H_N^o = \left\{\tilde{h}_1, \ldots \tilde{h}_n\right\}$ is the C-BERT output hidden states.

To further mitigate the impact of the potential label noise, an interactive attention mechanism is applied to calculate the relevance between the output hidden states of BiLSTM $H^B = \left\{\hat{h}_1, \ldots \hat{h}_n\right\}$ and the prompt features P:

$$\gamma_i = \mathrm{softmax}\left(\hat{h}_i \boldsymbol{W}_{attn}^{\mathrm{P}} P_i^T\right) \tag{15}$$

$$\tilde{P}_i = \sum_{i=1}^{n} \gamma_i P_i \tag{16}$$

where $\boldsymbol{W}_{attn}^{P} \in \mathbb{R}^{H \times H}$ is the interactive attention weight matrix, and the fusion features φ can be calculated as:

$$\varphi_i = \left[\widehat{h}_i; \widetilde{P}_i\right] \tag{17}$$

Finally, fusion features are taken into a Conditional Random Field (CRF) layer and predict the label for each token. And the loss function of fine-tuning can be defined by minimizing the negative likelihood loss as:

$$\mathcal{L} = -\sum_{i} \log(p(\boldsymbol{Y} \mid \boldsymbol{X})) \tag{18}$$

4 Experiments

4.1 Datasets

We investigate the effectiveness of our model on four Chinese NER datasets. Including Weibo (Peng and Dredze, 2015), Ontonotes 5.0 (Weischedel et al., 2011), Resume (Zhang and Yang, 2018) and MSRA (Levow, 2006). The statistics of the target datasets are shown in Table 1, and we randomly sample a small train set from each original dataset during training to simulate the few-shot scene.

Table 1: The statistics of the target datasets.

Dataset	Train	Dev	Test	Entity Types
Weibo	1.4k	0.27k	0.27k	8
Ontonotes	15.7k	4.3k	4.3k	4
Resume	3.8k	0.46k	0.48k	8
MSRA	46.4k	-	4.4k	3

Table 2: The statistics of the high-resource dataset.

Subset	Train	Dev	Test	Entity Types
CLUENER	10.7k	1.34k	1.34k	10
CNERTA	38.5k	4.44k	4.44k	5
RenMinRiBao	50.7k	4.63k	4.63k	4
Others	27.0k	2.83k	2.83k	10
Sum	**126.9k**	**13.2k**	**13.2k**	**18**

Besides, we construct a high-resource dataset to implement the label extension. The high-resource dataset is integrated with multiple datasets, including CLUENER (Xu et al., 2020), CNERTA (Sui et al., 2021), RenMinRiBao (Xia et al., 2005), and datasets from unknown sources. The high-resource dataset covers plenty of data and labels, and it can accurately support the label expansion on the low-resource datasets. The statistics of the high-resource dataset are shown in Table 2.

4.2 Experimental Settings

We implement the PCBERT based on the Transformers (Wolf et al., 2020) BERT with 12 layers of transformer in this work. The encoder hidden dimension H of P-BERT and C-BERT is 768; the word embedding dimension of the lexicon H' and label string H^* are both set as 200.

We use the Adam optimizer in all experiments. Before training all the target datasets, we first train a pre-labeled model on the high-resource dataset to annotate the extension entity labels for each train set and generate the label extension train set. Then our P-BERT is trained on the label extension train set. The learning rate of prompt-tuning is set as 1e-4. During fine-tuning on the original train set, the P-BERT is set as frozen, and we use an initial learning rate of 1e-5 for the C-BERT and 1e-2 for other parameters. We sample the same size from all datasets for few-shot learning, the max sequence

length is set as 150, and we train a maximum epoch number of 20 in all datasets.

To evaluate our proposed PCBERT, we compare it with the following approaches:

BERT. (Devlin et al., 2019) The BERT model with a token classifier is the baseline of the BERT-based NER approach.

BERT-LC. Based on the vanilla BERT, we further add a BiLSTM-CRF layer behind the BERT output layer to better compare with our proposed PCBERT.

Lattice LSTM. (Zhang and Yang, 2018) A lexicon-based Chinese NER approach is implemented with a lattice-structure LSTM model.

FLAT. (Li et al., 2020) An enhanced lattice-structured NER approach. By constructing a flat structure Transformer to fully leverage the lattice information and utilize the parallelism of GPUs.

LEBERT. (Liu et al., 2021a) A lexicon enhanced the Chinese sequence labeling model. Integrating external lexicon knowledge into BERT with a Lexicon Adapter layer.

LEBERT-LC. Based on the vanilla LEBERT, we further add a BiLSTM layer behind the BERT output layer in LEBERT to better compare with our proposed PCBERT.

4.3 Overall Results

We randomly sample different samples from the dataset in Table 1 to simulate NER in the few-shot scenario. The train set sampling sizes K are 250, 500, 1000, and 1350 (the max size of Weibo is 1350), respectively. We use the standard F1-score evaluation metrics to compare the performance.

Table 3 illustrates the experimental results of the Chinese few-shot NER. Our model outperforms all related approaches when K is 250 and achieves the best result on all the samples of Weibo and Ontonotes. Besides, our model performance in Weibo at K=250 outperforms other approaches at K=1350, demonstrating that our approach achieves excellent performance on the few-shot dataset.

The experimental results also indicate that all models' performance in different datasets is quite different even under the same sample size. We speculate that it is related to the semantic environment quality of the dataset rather than the number of entity types. Furthermore, our PCBERT shows more significant advantages on Weibo and Resume datasets with worse semantic environment quality.

Table 3: The overall results on Chinese Few-shot NER.

Dataset	Methods	K=250	K=500	K=1000	K=1350
Weibo	BERT	56.42	62.21	61.27	61.21
	BERT-LC	65.10	71.14	72.03	72.45
	Lattice LSTM	40.37	49.54	53.80	58.27
	FLAT	51.42	56.95	58.70	64.27
	LEBERT	65.83	67.12	70.34	69.12
	LEBERT-LC	66.92	71.11	71.80	73.42
	PCBERT	**73.52**	**73.49**	**76.58**	**77.88**
Ontonotes	BERT	63.85	69.50	71.33	72.42
	BERT-LC	65.69	73.54	74.97	77.19
	Lattice LSTM	39.71	45.46	54.54	57.48
	FLAT	49.01	46.35	49.34	57.44
	LEBERT	69.48	69.01	73.78	74.84
	LEBERT-LC	70.26	69.89	73.83	76.01
	PCBERT	**74.42**	**75.62**	**78.33**	**81.52**
Resume	BERT	53.80	62.64	69.36	70.65
	BERT-LC	92.26	**94.66**	95.16	**96.41**
	Lattice LSTM	85.63	89.60	92.01	93.13
	FLAT	84.62	90.77	92.97	87.79
	LEBERT	89.15	92.56	94.02	95.19
	LEBERT-LC	91.60	93.03	**95.40**	95.16
	PCBERT	**93.42**	94.01	94.96	95.97
MSRA	BERT	68.44	72.28	81.21	82.28
	BERT-LC	79.01	83.13	87.84	89.32
	Lattice LSTM	54.69	63.61	74.27	76.31
	FLAT	59.62	70.20	80.79	64.95
	LEBERT	79.11	85.18	87.77	89.35
	LEBERT-LC	80.92	**86.09**	**88.11**	88.70
	PCBERT	**81.08**	85.25	87.88	**89.72**

4.4 Analysis and Discussion

Ablation Study

We analyze the impact of each module in our PCBERT by designing several experiments. Table 4 presents the performance comparison between PCBERT and other ablation models. First, we observe a performance decline when removing the P-BERT component, demonstrating that P-BERT plays a vital role in model performance. We then observe that its results outperform LEBERT and LEBERT-LC on Weibo and Ontonotes when K is less than or equal to 500, which verifies that multi-label features can improve the model performance in the few-shot scenario. Moreover, after removing the label extension strategy (LEA) by using the original annotated dataset to train the model, the performance also decreases, indicating that the label extension strategy is effective in our approach.

To further analyze the impact of the label extension strategy, we replace the label extension dataset with the high-resource dataset to train the P-BERT (LEB). The results in Table 4 show a severe model performance decrease when directly adopting the high-resource dataset for prompt-tuning. Furthermore, the phenomenon becomes more prominent when the sample size K becomes smaller. And we observed there are different decrease degrees in

Table 4: Results of the Ablation Study on Chinese Few-shot NER.

Dataset	Methods	K=250	K=500	K=1000	K=1350
Weibo	PCBERT	**73.52**	**73.49**	**76.58**	**77.88**
	-P-BERT	67.28	71.85	70.02	72.66
	-LEA	67.06	70.31	71.88	72.73
	-LEB	61.95	67.01	68.62	69.33
Ontonotes	PCBERT	**74.42**	**75.62**	**78.33**	**81.52**
	-P-BERT	72.94	72.42	72.55	74.66
	-LEA	69.13	72.10	74.24	72.62
	-LEB	62.23	66.07	68.86	70.09
Resume	PCBERT	**93.42**	94.01	**94.96**	**95.97**
	-P-BERT	91.18	92.99	94.41	95.41
	-LEA	91.28	**94.33**	**94.96**	95.55
	-LEB	87.17	91.64	92.97	93.96
MSRA	PCBERT	**81.08**	85.25	**87.88**	**89.72**
	-P-BERT	80.59	**85.50**	86.95	87.88
	-LEA	82.77	84.32	86.20	84.32
	-LEB	79.09	81.36	83.61	84.75

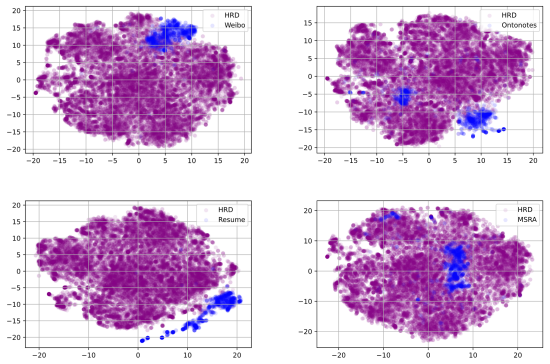

Figure 4: t-SNE visualization of each sampled train set and the high-resource dataset.

different datasets. For example, the performance on Weibo decreased by 11.57% (K=250), while MSRA decreased by only 1.99% (K=250). We present the visualization of semantic distribution between each sampled (K=1350) train set and the high-resource dataset (HRD) in Figure 4. The sentence-level representation is obtained from the BERT embedders onto a 2-dimensional space using t-SNE (Van der Maaten and Hinton, 2008). It can be concluded from Figure 4 and Table 4 that when the semantic space gap between the source dataset and the target dataset increases, transfer training directly from the source dataset will decrease the model performance.

Impact of Feature Injection

The results in Table 3 and Table 4 have demonstrated that the injected lexicon and multi-label features in C-BERT can effectively enhance the model performance. We speculate that multi-type lexicon or multi-label features injection can improve

Table 5: Comparison between LEBERT with random initial lexicon embeddings (LEBERT-RW) and original LEBERT.

Dataset	Methods	K=250	K=500	K=1000	K=1350
Weibo	LEBERT	65.83	67.12	70.34	69.12
	LEBERT-RW	64.08	67.16	68.89	70.42
Ontonotes	LEBERT	69.48	69.01	73.78	74.84
	LEBERT-RW	66.65	71.41	73.93	75.96
Resume	LEBERT	89.15	92.56	94.02	95.19
	LEBERT-RW	90.77	93.44	94.77	95.68
MSRA	LEBERT	79.11	85.18	87.77	89.35
	LEBERT-RW	79.34	83.83	88.74	88.59

the model's perception of fine-grained information and judgment of entity boundaries. Moreover, we further adopt LEBERT with random initial lexicon embeddings (LEBERT-RW) to compare the original LEBERT on four datasets. As shown in Table 5, the performance of LEBERT-RW is similar to LEBERT, which indicates that the boundary information introduced by feature injection is more critical to the model than the semantic distribution of the word embeddings.

Impact of Label Extension

To further analyze the impact of the label extension strategy, we evaluate the PCBERT performance when each extension label is removed from the label extension train set of Weibo (K=1350). Figure 6 illustrates the results, sorted in descending order according to each metric. We can conclude that, in most cases, removing an extension label will cause the model performance to decrease. It also shows that in the Weibo dataset, introducing any extension label will bring the final performance improvement in prompt-tuning, which indirectly indicates that our prompt-based PCBERT can effectively suppress the label extension noise.

Sentence Length

Figure 5 shows the F1-score trend of all baselines and PCBERT on the four datasets in Table 1 with the sampling size of 250. As shown in the results, we discover that PCBERT significantly improves performance in all sentence length intervals of the Weibo and Ontonotes datasets. Comparing the results of LEBERT and LEBERT-LC, it can be observed that adding the BiLSTM layer improves performance in the sampled Weibo and MSRA datasets. One potential reason is that the BiLSTM has a better awareness of directionality and short-distance information. To achieve more stable performances, we add the BiLSTM layer behind the

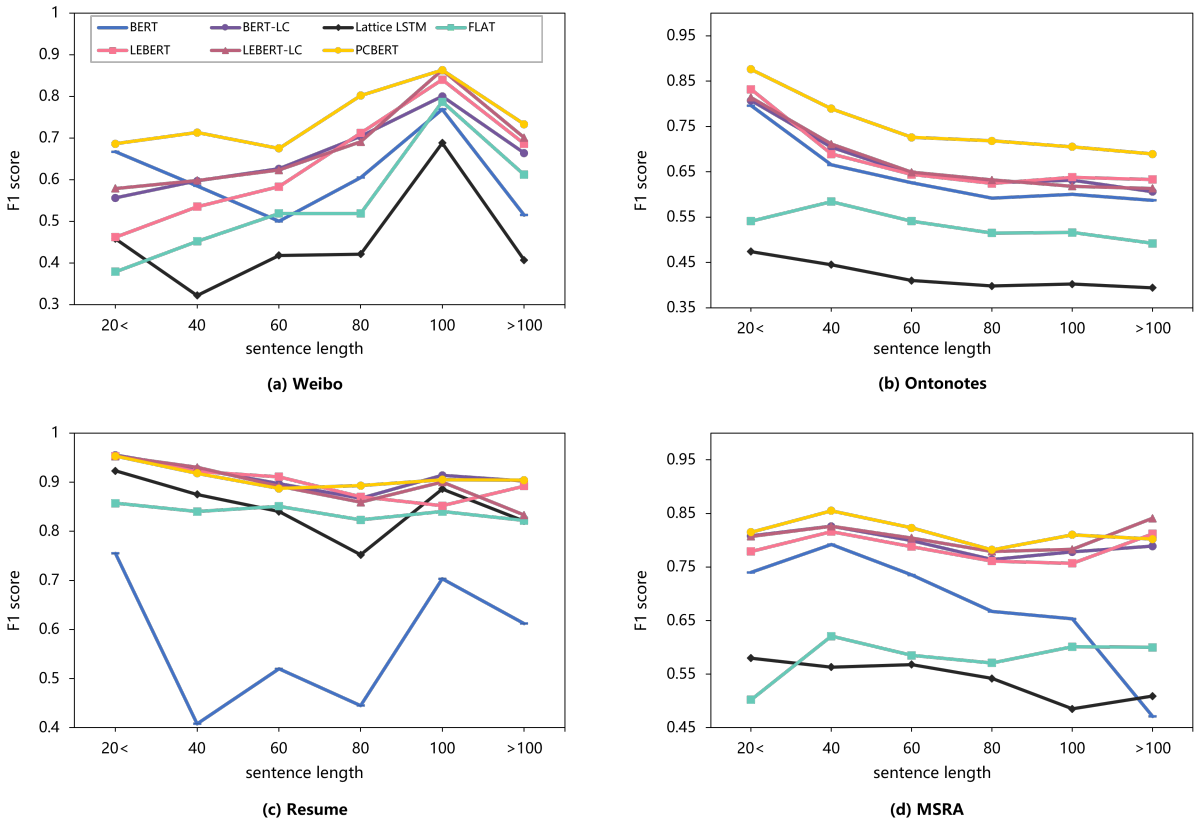

Figure 5: F1-scores against the sentence length.

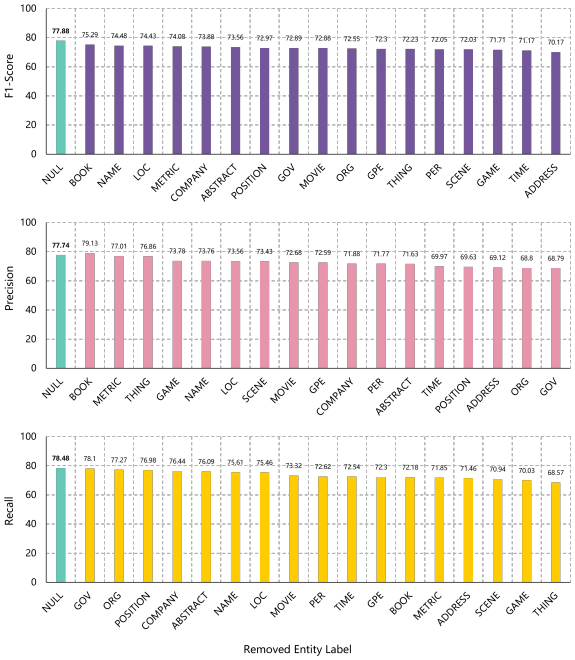

Figure 6: F1-score, Precision, and Recall comparison of PCBERT on the Weibo dataset when removing each extension entity label, where NULL indicates the original label extension train set.

C-BERT.

5 Related Works

Chinese NER

NER is a fine-grained sequence labeling task. With the advent of PLMs, the benchmark of Chinese NER has been dramatically improved. Pre-trained models based on large-scale corpus (Devlin et al., 2019; Lewis et al., 2020; Radford et al., 2019) provide excellent semantic representation for Chinese NER and are used by many works. Some work adds a softmax on PLMs (Yang, 2019) and achieves significant performance; others (Peters et al., 2018; Zheng et al., 2021; Nan et al., 2021) take PLMs as the backbone model to further enhance the original model performance.

Despite the remarkable achievements of PLMs, most existing models still need to be improved in judging Chinese word boundaries. Lexicon-based approaches (Zhang and Yang, 2018; Ma et al., 2020; Gui et al., 2019; Zhao et al., 2020) can effectively alleviate this issue. In particular, many lexicon-based works like Lex-BERT (Zhu and Cheung, 2021) need a high-quality vocabulary with entity-type information. (Zhang and Yang,

2018) proposed the Lattice LSTM approach to leverage all potential words in each segment and only need word vectors, which provided great inspiration for the later work. Recently many works (Xiao et al., 2019; Sperber et al., 2019; Zhang et al., 2019a,b) presented lattice-based transformers to promote parallel computing performance and fuse the PLMs representation into the model. However, most lattice-based transformers only fuse dictionary features in external input sequences without integrating them into the PLM structure. (Liu et al., 2021a) proposed LEBERT integrates lexicon knowledge into BERT layers and achieved state-of-the-art performance in multiple Chinese NER datasets.

Prompt-tuning

With the emergence of GPT-3 (Brown et al., 2020), the target-task-oriented pre-training form attracted a lot of attention (Schick and Schütze, 2021). Prompt-tuning (Lester et al., 2021) can be regarded as a new template-based pre-training paradigm. Unlike fine-tuning, the downstream task of prompt-tuning is homologous to pre-training. Prompt-tuning is more dependent on the prior distribution of the model, while fine-tuning is more dependent on the posterior distribution (Qiu et al., 2020).

Designing appropriate prompt templates for different tasks is crucial in prompt-tuning performance (Liu et al., 2021b). There is no universal template for all NLP tasks. (Jiang et al., 2020; Yuan et al., 2021; Haviv et al., 2021) proposed discrete prompts to disassemble and replace sentence components for text inference tasks; and (Gao et al., 2021; Ben-David et al., 2021) designed the generation prompt to build generated templates by automatically extracting semantic information from sentences.

In NER tasks, the model requires more specific semantic fine-grained information. Therefore, prompt templates construction approaches for other natural language understanding tasks can not work out well on NER tasks. (Ma et al., 2022) put forward a template-free approach to complete the entity template using the word vector mean of the same entity in the dataset. And (Chen et al., 2021) use an encoder-decoder model to translate the NER task into a prompt-based generation task.

6 Conclusion

In this paper, we propose a Parent and Child BERT for Chinese few-shot NER tasks and achieve state-of-the-art results on the Weibo dataset. Our model consists of P-BERT and C-BERT, where P-BERT is a prompt-based model for providing richer semantic information, and C-BERT is a lexicon-based model. The experimental results demonstrate that our PCBERT effectively improves the performance on the Chinese few-shot NER task. In the future, we will further analyze the performance improvement of label extension strategy in domain-specific datasets.

7 Acknowledgments

This work is supported by the Natural Science Foundation of Fujian Province, PR China (2022J01120).

References

Abhishek Abhishek, Ashish Anand, and Amit Awekar. 2017. Fine-grained entity type classification by jointly learning representations and label embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 797–807.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2021. Pada: A prompt-based autoregressive approach for adaptation to unseen domains. *arXiv preprint arXiv:2102.12206*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xiang Chen, Ningyu Zhang, Lei Li, Xin Xie, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021. Lightner: A lightweight generative framework with prompt-guided attention for low-resource ner. *arXiv preprint arXiv:2109.00720*.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online*

Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1835–1845. 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 *NAACL-HLT (1)*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019. Cnn-based chinese ner with lexicon rethinking. In *IJCAI*, pages 4982–4988.

Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. Bertese: Learning to speak to bert. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3618–3623.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2017. Transfer learning for named-entity recognition with neural networks.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059. Association for Computational Linguistics.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuan-Jing 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.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021a. Lexicon enhanced chinese sequence labelling using bert adapter. *arXiv preprint arXiv:2105.07148*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021b. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuan-Jing Huang. 2020. Simplify the usage of lexicon in chinese ner. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5951–5960.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Linyang Li, Qi Zhang, and Xuanjing Huang. 2022. Template-free prompt tuning for few-shot NER. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5721–5732. Association for Computational Linguistics.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9683–9695.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554.

ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. 2018. Deep contextualized word representations. naacl-hlt.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Matthias Sperber, Graham Neubig, Ngoc-Quan Pham, and Alex Waibel. 2019. Self-attentional models for lattice inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1185–1197.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2019. Leverage lexical knowledge for chinese named entity recognition via collaborative graph network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3830–3840.

Dianbo Sui, Zhengkun Tian, Yubo Chen, Kang Liu, and Jun Zhao. 2021. A large-scale Chinese multimodal NER dataset with speech clues. 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)*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

YingJu Xia, Hao Yu, and Fumihito Nishino. 2005. The Chinese named entity categorization based on the people's daily corpus. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 10, Number 4, December 2005: Special Issue on Selected Papers from CLSW-5*, pages 533–542.

Fengshun Xiao, Jiangtong Li, Hai Zhao, Rui Wang, and Kehai Chen. 2019. Lattice-based transformer encoder for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3090–3097.

Liang Xu, Qianqian Dong, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. 2020. Cluener2020: Fine-grained name entity recognition for chinese. *arXiv preprint arXiv:2001.04351*.

Haiqin Yang. 2019. Bert meets chinese word segmentation. *arXiv preprint arXiv:1909.09292*.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Pei Zhang, Niyu Ge, Boxing Chen, and Kai Fan. 2019a. Lattice transformer for speech translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6475–6484.

Pei Zhang, Niyu Ge, Boxing Chen, and Kai Fan. 2019b. Lattice transformer for speech translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6475–6484.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564.

Xiaoyan Zhao, Min Yang, Qiang Qu, and Yang Sun. 2020. Improving neural chinese word segmentation with lexicon-enhanced adaptive attention. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1953–1956.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. 2021. PRGC: potential relation and global correspondence based joint relational triple extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6225–6235. Association for Computational Linguistics.

Wei Zhu and Daniel Cheung. 2021. Lex-bert: Enhancing bert based ner with lexicons. *arXiv preprint arXiv:2101.00396*.