applied sciences

MDPI

Article

Prompt-Based Word-Level Information Injection BERT for Chinese Named Entity Recognition

Qiang He [1,2], Guowei Chen [1], Wenchao Song [1] and Pengzhou Zhang [1,*]

[1] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100803, China
[2] State Key Laboratory of Media Convergence Production Technology and Systems, Beijing 100803, China
* Correspondence: zhangpengzhou@cuc.edu.cn

Abstract: Named entity recognition (NER) is a subfield of natural language processing (NLP) that identifies and classifies entities from plain text, such as people, organizations, locations, and other types. NER is a fundamental task in information extraction, information retrieval, and text summarization, as it helps to organize the relevant information in a structured way. The current approaches to Chinese named entity recognition do not consider the category information of matched Chinese words, which limits their ability to capture the correlation between words. This makes Chinese NER more challenging than English NER, which already has well-defined word boundaries. To improve Chinese NER, it is necessary to develop new approaches that take into account category features of matched Chinese words, and the category information would help to effectively capture the relationship between words. This paper proposes a Prompt-based Word-level Information Injection BERT (PWII-BERT) to integrate prompt-guided lexicon information into a pre-trained language model. Specifically, we engineer a Word-level Information Injection Adapter (WIIA) through the original Transformer encoder and prompt-guided Transformer layers. Thus, the ability of PWII-BERT to explicitly obtain fine-grained character-to-word relevant information according to the category prompt is one of its key advantages. In experiments on four benchmark datasets, PWII-BERT outperforms the baselines, demonstrating the significance of fully utilizing the advantages of fusing the category information and lexicon feature to implement Chinese NER.

Keywords: named entity recognition; prompt learning; pre-trained language model; BERT adapter

check for updates

Citation: He, Q.; Chen, G.; Song, W.; Zhang, P. Prompt-Based Word-Level Information Injection BERT for Chinese Named Entity Recognition. *Appl. Sci.* **2023**, *13*, 3331. https://doi.org/10.3390/app13053331

Academic Editors: Valentino Santucci and Tobias Meisen

Received: 19 December 2022
Revised: 6 February 2023
Accepted: 27 February 2023
Published: 6 March 2023

Copyright: © 2023 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

Named entity recognition (NER) tries to distinguish the entities' boundary and category labels from unstructured text, which is an important natural language processing task, owing to the application in several domains of information extraction, question answering, machine translation, and so on. NER was formulated as a sequence labeling problem in earlier research, and several of the current works for NER are established based on neural network approaches [1–9]. However, Chinese NER is correlated with word segmentation. In detail, entities' boundaries and segmentation boundaries tend to overlap each other. Therefore, some of the pipeline methods use an intuitive way to segment the sentence before approaching the Chinese NER task. Similar to other pipeline methods, the method mentioned above faces the problem of error propagation. From another aspect, name entities are a considerable solution for out-of-vocabulary segmentation, which may further exacerbate entity recognition errors.

In order to enhance the character-based Chinese NER methods, two groups of models introduce the segmentation information. One of them explicitly models the word information by accelerating the additional word or other relevant features' embedding with the character features [10,11]. As their validity has been verified, pre-trained language embeddings, such as BERT [12], aim to inject the implicit word-level syntactic and semantic

knowledge. The other group modifies the original encoding structure to accommodate the situation where character and word information are both used as input [13–17]. Although some of the recent work [18,19] considers the combination of these two lines of research, the methods do not utilize the syntactic and semantic knowledge extracted by original language model representation in the process of utilizing BERT. On the other side, the category information is not efficiently capitalized. Illustrated in Figure 1, the original text "南京市长江大桥" is used to construct several word-character lattices according to the pre-trained word dictionary and their word embedding. As a result, the words inside the original sequence are "南京 (Nanjing)", "市长 (Mayor)", "长江大桥 (Yangtze River Bridge)", "南京市 (Nanjing City)", "江大桥 (Daqiao Jiang)", "大桥 (Bridge)", and "长江 (Yangtze)". For the token "江大桥 (Daqiao Jiang)", we need to consider the category information to distinguish whether the word is a named entity. While only considering the shallow word segmentation information, the entity may be recognized as "大桥 (Bridge)".

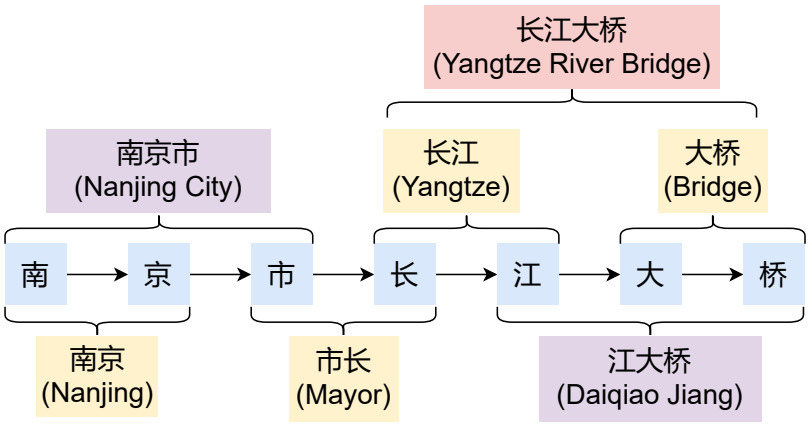

Figure 1. An example of words in a Chinese sentence.

Inspired by the work about BERT Adapter [20–22] and the prompt learning, we propose Prompt based Word-level Information Injection BERT (PWII-BERT) to integrate prompt guided word features inside the pre-trained language model immediately. By preparing the domain-related dictionary in advance, the raw sentence is transformed into a sequence of characters with their paired corresponding words. Specifically, we engineer a Word-level Information Injection Adapter (WIIA) for dynamically fusing the most relevant word-level feature for all the characters through a biaffine mechanism. In addition, we automatically extract the soft prompt for guiding the word-level information injection. The WIIA is inserted between the Transformer encoder and prompt-guided Transformer layers. By using original pre-trained language models, prompt-based word-level information is integrated through fine-tuning both the language model and extra parameters, which is dissimilar to the classical Adapter fixing the BERT parameters. The effectuality of PWII-BERT is verified on four widely used Chinese NER datasets. One of the main advantages of PWII-BERT is its ability to incorporate both category information and lexicon features, which authorize it to catch both syntactic and semantic information for Chinese NER and achieve further improvements compared with several state-of-the-art methods. Furthermore, several comprehensive comparisons with detailed analyses show that PWII-BERT gains a significant improvement in discriminating the boundary and type for each candidate text span.

The main findings of this paper can be summarized as follows:

- Through the lexicon feature in the Chinese dictionary, the proposed Word-level Information Injection Adapter effectively utilizes additional word-level features that are well-suited to encourage the representation ability for Chinese NER;
- PWII-BERT applies the soft prompt for guiding the word-level information injection which can take advantage of using the category feature against the prompt automatically extracted;

- On four datasets, our experiments on widely-used benchmark datasets demonstrate the superior performance of PWII-BERT compared to state-of-the-art Chinese NER methods.

The rest of this paper is organized as follows: Section 2 shows the PWII-BERT model and its three key modules: (1) prompt representations; (2) Word-level Information Injection Adapter; and (3) Conditional Random Field (CRF). In Section 3, we describe the datasets and hyperparameters used in our experiments and demonstrate the superior performance of our proposed method. Section 4 offers a thorough review of the existing research on NER methods. In Section 5, we conclude the paper by summarizing our contributions and outlining potential directions for future work.

2. Related Work

2.1. Adapter-Tuning

BERT Adapter [20] is proposed to solve the downstream tasks by only tuning the adapter. In detail, the adapter is added between two layers, and the original pre-trained language model parameters are fixed while training for a specific task. The adapter is first used in neural machine translation [21] with task-specific structures. In addition, MAD-X [23] designs a structure with an adapter for more efficient transfer learning tasks. Furthermore, K-ADAPTER is proposed to inject knowledge into language models with additional training options. As for the proposed method, a Word-level Information Injection Adapter is used for better fusing the lexicon features. Rather than using the original BERT parameters without modification, we employ fine-tuning to adjust them.

2.2. NER

Chinese NER models are divided into three parts: Lexicon-based models, Pre-Trained Model-based models, and Hybrid models. With the aim of enhancing the Chinese NER task with lexicon features, Lexicon-based models encode both words and characters for sequence labeling [13], and some researchers further improve the methods in different aspects, such as training efficiency [18,24] and model degradation [15]. Transformer-based methods reveal remarkable effects for sequence tagging in Chinese tasks. Refs. [25,26] introduce the character features from BERT and obtain better results compared with the classical static word embedding. Some of the recent work [18,19] considers the combination of these two methods above, and the methods do not utilize the syntactic and semantic knowledge extracted by original language model representation in the process of utilizing BERT. ERNIE [27,28] changes random masking to entity masking and word masking for enhancing the entity information. BERT [29] is proposed to utilize the domain knowledge. ZEN [30] modifies the encoder to a multi-layered N-gram encoder. The previous NER methods directly use the word information as extra knowledge. In addition, we use the category prompt to guide the lexicon feature injection, which is a more efficient way. We integrate word-level features at the middle level, enabling BERT to better incorporate and interact with this knowledge and avoid catastrophic forgetting.

2.3. Prompt Learning

With the wide application of large-scale pre-trained language models such as GPT-3 [31], prompt learning received considerable attention. Several research studies [32–37] imply that prompt learning effectively stimulates knowledge from PLMs compared with the transitional fine-tuning tasks, and the reason may be that the prompt learning formulates the task in a similar way to the pre-training task. Thus, prompt learning gains great effectivity on few-shot and out-of-domain tasks. For leveraging the ability of PLMs with less trainable parameters, several studies try to mask the redundant parameters [38,39]. As a typical approach, prefix-tuning [40] obtains a sequence of continuous task-specific vectors for tuning the model with fewer parameters. Apart from the above methods aiming to conduct efficient fine-tuning, our approach pays more attention to achieving efficient category and task-specific knowledge transfer through the prompt-guided Transformer.

Specifically, the proposed method tunes the whole model instead of only training the Adapter. In addition, we use a prompt-guided lexicon injection method which is closer to the NER task to inject lexicon information, rather than relying on the training data.

3. Methods

The structure of our prefix-based word-level information injection BERT is shown in Figure 2. The model is made up of two parts, the prefix part which introduces category information according to the soft prompt, and the word-level information adapter makes a lexicon injection that could obtain the utmost out of both category and lexicon features. PWII-BERT differs from BERT in two key ways. First, PWII-BERT receives the input of character and word features, where original plain text is formulated as a series of character–word pairs. The second difference is the use of an adapter that is inserted between the Transformer layers and the prompt-guided Transformer layers, allowing additional knowledge to be integrated into the pre-trained language model effectively. In the following, we describe (1) a prefix-based category information introduction module; (2) the word-level information injection through the BERT adapter; and (3) the PWII-BERT by applying the above two parts into BERT.

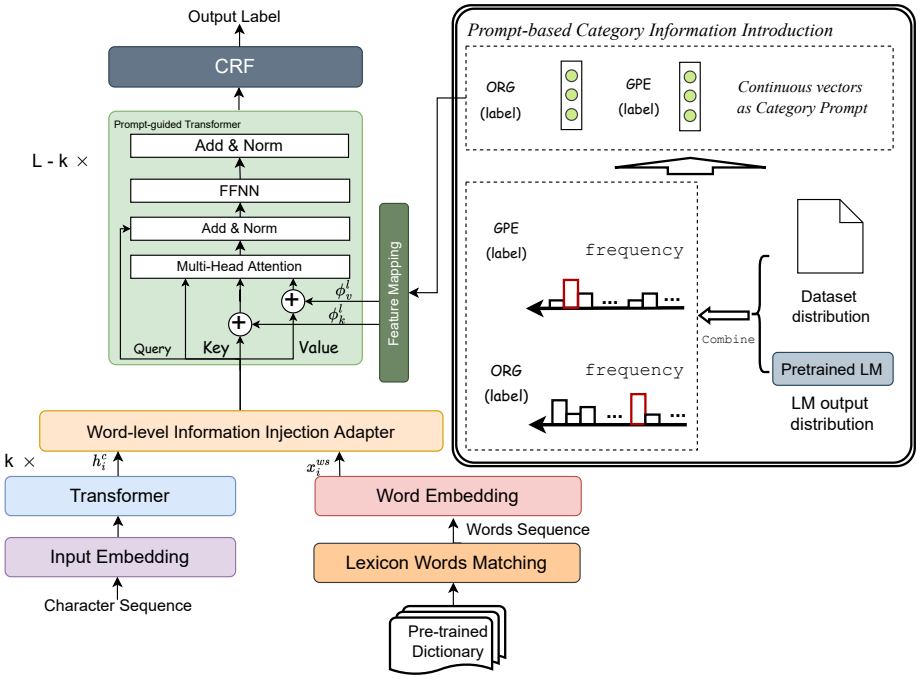

Figure 2. Structure of PWII-BERT.

3.1. Task Formulation

Given an input sentence with n characters $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, the NER task is formulated to recognize the several types of named entity $e_i = \{x_{START(i)}, x_{START(i)+1}, \ldots, x_{END(i)}\}$ that is designated as belonging to the entity class $C \in \mathbb{C}_k$(e.g., "LOC"), where $START(i)$ denotes the start position and $END(i)$ denotes the end position.

3.2. Prompt-Based Category Information Introduction

Due to the distribution differentiation among multiple datasets and prompt-based studies always suffering from the significant impact of prompt engineering, we automatically select proper label words that can primarily represent each class for category information introduction. Similar to the hard prompt methods, the most frequent word of the specific class C is selected to partly represent the class. In detail, the frequency $f(\boldsymbol{e}_i = \boldsymbol{w}, y = C)$ of each entity $\boldsymbol{e}_i$ with label C is calculated, the entity with the most

frequent is assigned as the label word to obtain the category information, and the label word is selected through Equation (1),

$$\mathcal{S}(C) = \arg\max_{w} \mathcal{F}(\boldsymbol{e}_i = \boldsymbol{w}, y = C). \tag{1}$$

In order to additionally bear in mind the statistical knowledge in the entire training set, all the samples with original text X and label Y in the training set are fed into the pre-trained language model and obtain the probability distribution $p(\hat{x}_j = w|X)$ of each word. We define an indicator function $\mathcal{I}_C(\hat{x}_j = w|X, y_j) \rightarrow \{0, 1\}$ to select all the words belonging to the top$k$ predictions of x_j in the predicting sample $\{X, Y\}$. Thus, the frequency in class C is calculated as $\mathcal{F}_C(\hat{x}_j = w|X, y_j) = \sum_j \mathcal{I}_C(\hat{x}_j = w|X, y_j)$, where the top k results of the positions are extracted. The label word of the specific category C is recognized according to Equation (2),

$$\mathcal{S}(C) = \arg\max_{w} \sum_{(X,Y)\in\mathcal{D}} \sum_{j}^{|X|} \mathcal{F}_C(\hat{x}_j = w, y_j = C). \tag{2}$$

To capitalize on both discrete and statistical level aspects, label words are chosen by considering both the data and LM. The representative word could be acquired using the following Equation (3),

$$\begin{aligned} \mathcal{S}(C) = \arg\max_{w} \{ & \sum_{(X,Y^*)\in\mathcal{D}} \sum_{i}^{|X|} \mathcal{F}(x_j = w, y_j = C) \\ & \cdot \sum_{(X,Y)\in\mathcal{D}} \sum_{i}^{|X|} \mathcal{F}_C(\hat{x}_j = w, y_j = C) \} \end{aligned} . \tag{3}$$

The high-frequency label words chosen are likely to be frequently used words overall. However, using these label words can lead to conflicts between different classes while training. To prevent this, after selection, we delete the conflicting words through the process of Equation (4),

$$w = \mathcal{S}(C), \qquad if \qquad \frac{\{(x = w, y = C)}{\sum_k \{(x = w, y = k)} > T_h, \tag{4}$$

where T_h is a threshold, which is a hyperparameter for filtering the words.

All of the entities containing the characters in $\mathcal{S}(C_k)$ are denoted as the representative label entities in category C_k for information injection. Following the prior span-based model, we use a pre-trained language model BERT [41] to obtain the contextualized representation $\boldsymbol{x}_j$ of each character $\mathrm{x}_j \in \mathbf{x}$. Therefore, the features of entity span e_i are disposed as Equation (5),

$$\boldsymbol{p}_h(\mathrm{e}_i) = [\boldsymbol{x}_{START(i)}; \boldsymbol{x}_{END(i)}; \phi(\mathrm{e}_i)], \tag{5}$$

where $\phi(\mathrm{e}_i) \in \mathbb{R}^d$ denotes the one-hot embedding of span length features, and d denotes the longest entity length counted from statistics in the dataset.

Each piece of category information is calculated as the weighted average of all the representation of entities through the trainable parameters λ_i. The representation of these is calculated through Equation (6),

$$\boldsymbol{p}_c(\mathrm{e}_i) = \sum_{i}^{|E_k|} \lambda_i * \boldsymbol{p}_h(\mathrm{e}_i). \tag{6}$$

Specifically, the proposed method obtains a set of trainable embedding matrices $\boldsymbol{P} = \{\boldsymbol{p}_c(\mathrm{e}_1), \boldsymbol{p}_c(\mathrm{e}_2), \ldots, \boldsymbol{p}_c(\mathrm{e}_{|k|})$, where k is the number of category.

3.3. Word-Level Information Injection Adapter

Instead of enhancing the character-level representation by injecting the knowledge and training the framework through multi-task learning, a Word-level Information Injection Adapter is introduced to capture the word-level information. Each word in the sentence is constitutive of two parts of information, the original character semantic representation as well as the additional word-level segmentation feature. The Word-level Information Injection Adapter fuses the feature for predicting the BIO label for each position, which is shown in Figure 3. The adapter is the novel module added between two layers of BERT, and the module is tuned in a different way compared with traditional task-based fine-tuning.

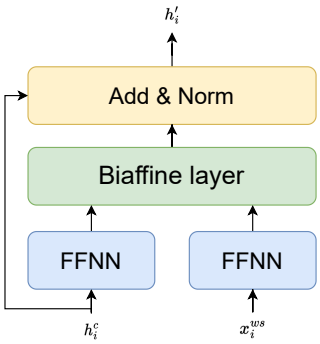

Figure 3. The framework of the Word-level Information Injection Adapter (WIIA). Pairing with the prepossessed corresponding word-level information, the character-level features are the input of WIIA. A biaffine layer is used for fusing the two-level information, and the mixed results are presented with a new eigenvector that is used for subsequent classification tasks after being added with character-level hidden features and layer normalization.

According to the motivation of injecting the word-level information, the adapter accepts inputs on both sides, the j-th character feature and the paired corresponding word-level segmentation feature. In the processed sequence, the input for the j-th character is represented as (h_j^c, x_j^{ws}), h_j^c denotes the character embedding, and $x_j^{ws} = x_{j1}^w, x_{j2}w, ..., xjm^w$ denotes the paired word embedding. The k-th word in x_j^{ws} is represented by Equation (7),

$$x_{jk}^w = \mathbf{e}^w(w_{jk}), \tag{7}$$

where $\mathbf{e}^w$ is used to map the one-hot word index into word embedding, and w_{jk} is the k-th word. In addition, a nonlinear transformation layer is conducted, and the word-level features are projected according to Equation (8),

$$v_{jk}^w = \mathbf{W}_2(\tanh(\mathbf{W}_1 x_{jk}^w + \mathbf{b}_1)) + \mathbf{b}_2 \tag{8}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_c \times d_w}$, $\mathbf{W}_2 \in \mathbb{R}^{d_c \times d_c}$, and $\mathbf{b}_1$ and $\mathbf{b}_2$ are scaler bias. d_w and d_c are the embedding size and the dimension of hidden size in the language model, respectively.

To select the relevant words from matched ones, we use an additional attention mechanism to combine the character and words. In detail, the character representation at the j-th position is denoted as $v_{jk}^w \in \mathbb{R}^{m \times d_c}$, and m denotes how many words are selected according to the current character. The attention score of each word is calculated through Equation (9),

$$\begin{aligned} h_j^s &= \mathrm{FFNN}_s\left(h_j^c\right) \\ h_{jk}^e &= \mathrm{FFNN}_e\left(v_{jk}^w\right) \\ a_{jk} &= h_j^{s\top} \mathbf{U}_m h_{jk}^e + \mathbf{W}_m\left(h_j^s \oplus h_{jk}^e\right) + \mathbf{b}_m \end{aligned}, \tag{9}$$

where $\mathbf{U}_m$ and $\mathbf{W}_m$ are the trainable parameters of biaffine attention. Accordingly, the final word-level feature is summed by Equation (10),

$$z_j^w = \sum_{k=1}^{m} a_{jk} v_{jk}^w. \tag{10}$$

The fusing feature which combines these two aspects is calculated as Equation (11),

$$\tilde{h}_j = h_j^c + z_j^w. \tag{11}$$

Finally, the dropout layer is used to stay away from overfitting.

3.4. PWII-BERT

PWII-BERT is based on the Transformer architecture [41], which consists of piles of identical building modules. The structure allows PWII-BERT to effectively produce and break down large amounts of data. Moreover, we first add K sets of trainable embedding matrices $\{\theta^{L-K+1}, \theta^{L-K+2}, \ldots, \theta^{L}\}$ for a last K prompt-guided encoder layer, $\theta \in \mathbb{R}^{2\times|K|\times d_c}$, and d_c denotes the hidden size, so the prompt for the last K layers in the l-th layer is calculated as Equation (12),

$$[\phi_k^l; \phi_v^l] = \theta^l \boldsymbol{P}. \tag{12}$$

The input of the original Transformer encoder with n characters is denoted as $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ and denotes the matched lexical words, which is a sub-sequence of the original sentence, as $\mathbf{s}_{cw} = \{ws_1, ws_2, ..., ws_n\}$. The representation of the l-th layer $\boldsymbol{X}^l \in \mathbb{R}^{nd}$ is projected into the three vectors through Equation (13),

$$\boldsymbol{Q}^l = \boldsymbol{X}^l \boldsymbol{W}^Q, \qquad \boldsymbol{K}^l = \boldsymbol{X}^l \boldsymbol{W}^K, \qquad \boldsymbol{V}^l = \boldsymbol{X}^l \boldsymbol{W}^V, \tag{13}$$

where $\boldsymbol{W}^Q$, $\boldsymbol{W}^K$, $\boldsymbol{W}^V \in \mathbb{R}^{d_c \times d_c}$. Following [41], the option of Attention first L layers of Transformer layer acts as Equation (14),

$$\text{Attention}(Q^l, K^l, V^l) = \text{softmax}(\frac{Q^l (K^l)^T}{\sqrt{d_c}}) V^l \tag{14}$$

The options of each Transformer layer act as Equation (15),

$$\begin{aligned} \text{MultiHead}(Q^l, K^l, V^l) &= \text{Concat}(\text{head}_1, ..., \text{head}_\text{h}) W^O \\ \text{where head}_\text{i} &= \text{Attention}(Q^l, K^l, V^l) \\ G &= \text{LN}(H^{l-1} + \text{MultiHead}(H^{l-1}))' \\ H^l &= \text{LN}(G + \text{FFN}(G)) \end{aligned} \tag{15}$$

where $H^l = \{h_1^l, h_2^l, ..., h_n^l\}$ denotes the hidden representation of the l-th layer; MultiHead is the attention option with several attention heads; LN denotes the normalization layer; FFNN is the feed-forward network with nonlinear transformation, where the $W_i^Q \in \mathbb{R}^{d_c \times d_k}$, $W_i^K \in \mathbb{R}^{d_c \times d_k}$, $W_i^V \in \mathbb{R}^{d_c \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_c}$ are trainable parameters for query, key, and value, respectively.

While injecting the word-level information with diminutive changes in the pre-trained language model, we first obtain the output $H^k = \{h_1^l, h_2^l, ..., h_n^l\}$ after l successive Transformer layers. Then, each pair (h_j^l, x_j^{ws}) is passed through the Word-level Information Injection Adapter (WIIA) , which transforms the j_{th} pair into $\tilde{h}_j^l$, and the option is shown as Equation (16),

$$\tilde{h}_j^k = \text{WIIA}(h_j^l, x_j^{ws}). \tag{16}$$

After injecting the word-level information between the l-th layer and $l+1$-th layer, we input $\widetilde{H}^l = \{\tilde{h}_1^l, \tilde{h}_2^l, ..., \tilde{h}_n^l\}$ into the remaining prompt-based Transformer. The attention operation in the following prompt-based Transformer is redefined as Equation (17),

$$Attention^l = softmax(\frac{\boldsymbol{Q}^l[\boldsymbol{\phi}_k^l; \boldsymbol{K}^l]^T}{\sqrt{d_c}})[\boldsymbol{\phi}_v^l; \boldsymbol{V}^l]. \tag{17}$$

After generating the representation of the input sentence and the category prompt-guided module, we combine them and compute attention scores by using the self-attention flow, and the output of the last Transformer layer is denoted as H^L. The refined attention score captures the changes caused by the addition of the guidance prompt.

3.5. Training and Decoding

A conditional Random Field (CRF) layer is employed to model the relation between the output labels. Given $H^L = \{h_1^L, h_2^L, ..., h_n^L\}$, a relevant score is calculated as Equation (18),

$$O = \mathbf{W}_o H^L + \mathbf{b}_o. \tag{18}$$

While assigning the output sequence as $\mathbf{y} = \{y_1, y_2, ..., y_n\}$, its probability sequence is calculated as Equation (19),

$$p(\mathbf{y}|\mathbf{s}) = \frac{\exp(\sum_i (O_{i,y_i} + T_{y_{i-1},y_i}))}{\sum_{\tilde{\mathbf{y}}} \exp(\sum_i (O_{i,\tilde{y}_i} + T_{\tilde{y}_{i-1},\tilde{y}_i}))}, \tag{19}$$

where T is used for transit between labels, and $\tilde{\mathbf{y}}$ contains all the candidate label sequences. The CRF is trained by negative log-likelihood loss as Equation (20),

$$\mathcal{L} = -\sum_m \log(p(\mathbf{y}|\mathbf{s})), \tag{20}$$

where m is the index in training set with N instances $\{\mathbf{s}_m, \mathbf{y}_m\}|_{m=1}^N$. Similar to the previous methods, the Viterbi algorithm is used for inference.

4. Results

4.1. Datasets

Four benchmark datasets that are commonly used were the subject of the experiments, including Weibo [42,43], OntoNotes [44], Resume [13], and MSRA [45]. The OntoNotes and MSRA datasets are collected from the news domain, the Weibo NER dataset is drawn from the social media website Sina Weibo (https://www.weibo.com/, accessed on 26 February 2023), and the Resume dataset is collected from Sina Finance (http://finance.sina.com.cn/stock/index.shtml, accessed on 26 February 2023). By using the same data split with [15], we show these statistics of datasets in Table 1.

Table 1. The statistics of the datasets.

Dataset	Form	Train	Dev	Test
MSRA	Sent	46.4 k	-	4.4 k
	Char	2169.9 k	-	172.6 k
Weibo	Sent	1.4 k	0.27 k	0.27 k
	Char	73.8 k	14.5 k	14.8 k
Ontonotes	Sent	15.7 k	4.3 k	4.3 k
	Char	491.9 k	200.5 k	208.1 k
Resume	Sent	3.8 k	0.46 k	0.48 k
	Char	124.1 k	13.9 k	15.1 k

4.1.1. Hyperparameters and Evaluation Metrics

Our model is constructed based on BERT$_{BASE}$ [12], with 12 layers of Transformer, and the initialing model is downloaded from huggingface (https://github.com/huggi

ngface/transformers, accessed on 26 February 2023). The pre-trained word embedding is downloaded from Tencent (https://ai.tencent.com/ailab/nlp/en/download.html, accessed on 26 February 2023). The WIIA is injected between the seventh and eighth layers of BERT_{BASE}. The optimizer used in our method is Adam, and the learning rate for BERT_{BASE} is set to 1×10^{-5}, while the learning rate is set to 1×10^{-4} for other parameters. All of the parameters containing the pre-trained word embedding are fine-tuned during the training of PWII-BERT. The dropout rate used in this paper is set to 0.5. We clip the input sentence with a maximum number of 200, and the mini-batch size is set to 32 for training. The max length of the sequence is set to 200, and the training batch size is 32.

The stander CoNLL Precision (P), Recall (R), and $F1$ score are used as the evaluation metrics, which are shown as

$$P = \frac{T_P}{T_P + F_P} \tag{21}$$

$$R = \frac{T_P}{T_P + F_N} \tag{22}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{23}$$

where T_P is the number of true positive instances, F_P is the number of false positives, and F_N is the number of false negatives.

4.1.2. Baselines

Several of the following NER baselines are illustrated to evaluate the proposed PWII-BERT:

- **Zhang and Yang [13]**. The Chinese NER method using the Lattice LSTM structure;
- **Ma et al. [18]**. A word-enhanced method through soft lexicon features;
- **Liu et al. [15]**. An encoding strategy for encoding the word embedding;
- **Zhu and Wang [46]**. The injection structure through a convolutional attention network;
- **Li et al. [19]**. A flattened Transformer structure for a unified character-word sequence;
- **BERT**. Pre-trained Chinese BERT is used for sequence labeling;
- **Word-BERT**. We deploy a word-level baseline method, the word features, and character features are concatenated, and the LSTM layer and CRF are used for fusion;
- **ERNIE [27]**. ERNIE mask entities for pre-training;
- **ZEN [30]**. Add the N-gram feature into the pre-trained language model.

The proposed PWII-BERT is compared with several models on the Resume dataset, and the results are shown in Table 2. The proposed structure obtains a significant reforming on Recall metrics, and these gains are due to the following reasons. Instead of inserting all the candidate lexicons as an additional feature, the PWII helps the model earn lexicon information by explicitly injecting. The injection fuses the character-level features and word-level features under the guidance of category information. These improved experimental results verified the importance of prompt-guided word information for Chinese NER.

Table 2. Results of P, R, and F1 score on Resume datasets.

Model	P	R	F1
Zhang and Yang [13]	94.81	94.11	94.46
Ma et al. [18]	95.30	95.77	95.53
Liu et al. [15]	95.27	95.15	95.21
Zhu and Wang [46]	95.05	94.82	94.94
PWII-BERT	95.52	96.87	96.19

4.1.3. Overall Performance

As shown in Table 3, we additionally compare our method PWII-BERT with several state-of-the-art models on Chinese NER tasks, and we have the following observations.

Table 3. Results of F1 score on Weibo, Ontonotes, MSRA, and Resume datasets. † denotes the results we reproduce for the method according to the code (https://github.com/yangjianxin1/LEBERT-NER-Chinese, accessed on 26 February 2023. The other results come from Liu et al. [47]).

Model	Weibo	Ontonotes	MSRA	Resume
Zhang and Yang [13]	63.34	75.49	92.84	94.46
Ma et al. [18]	69.11	81.34	95.35	95.53
Liu et al. [15]	65.30	75.79	93.50	95.21
Zhu and Wang [46]	59.31	73.64	92.97	94.94
Ding et al. [16]	59.50	75.20	94.40	-
Li et al. [19]	68.07	80.56	95.46	95.78
Liu et al. [47] †	70.92	81.53	95.38	96.03
BERT	67.27	79.93	94.71	95.33
BERT+Word	68.32	81.03	95.32	95.46
ZEN	66.71	79.03	95.20	95.40
ERINE	67.96	77.65	95.08	94.82
PWII-BERT	72.06	81.64	95.70	96.19

(1) It can be observed that PWII-BERT outperforms all the baselines and other lexicon-based models on four widely used datasets. Our model has an average $F1$ improvement of 0.43% for the F1 score over the previous state-of-the-art method. The experimental results demonstrate the effectiveness of the ability to incorporate both category information and lexicon features.

(2) Compared with the LEBERT(Liu et al. [47]), our model has an improvement of 1.14% for the F1 score on the Weibo dataset. In addition to the method in which LEBERT uses Adapter to introduce the lexicon features, we additionally add category prompt in the Transformer layers after Adapter to guide the category information injection. Meanwhile, in the process of introducing the vocabulary, we modified the fusion method for fusing the character-level and word-level information. This observation supports the fact that category information used in the guidance of the lexicon injection model makes up for the application gap between the lexicon feature and the NER task.

(3) Except for the Weibo dataset, the proposed method slightly improves on other datasets and has an average $F1$ improvement of 1.01% over it; the possible reason for this phenomenon is the difference in the amount of training sets. The Weibo dataset has only 1400 training samples, and the amount of training data in Resume is nearly three times the amount in Weibo. In addition, the number of training samples in MSRA and Ontonotes is more than 10 times that of the number in Weibo. Since the prompt is obtained automatically in the full training set, PWII-BERT has achieved greater improvement in the dataset with a small amount of data, and its guiding effect is smaller in the dataset with a larger amount of training samples.

4.1.4. Ablation Study

To evaluate the contribution of individual modules of PWII-BERT, each component is removed for evaluation, and the experiments are shown in Table 4. The results show that the performance dropped by 0.27%, 0.41%, 0.25%, and 0.06%, indicating that the CRF decoder captures the relationship inside the output tags and improves. Second, to verify the effectiveness of the category prompt, we remove prompt-guided Transformer structure and use the same Transformer mechanics before and after WIIA. The average decline of the F1 score is 0.49%, demonstrating that the approach of introducing the category information additionally is effective and actually assists the integration of the lexicon feature. Third, we remove the WIIA component to test the influence of lexicon injection. It obtains worse

results by 0.69% on the Weibo dataset, which proves the validity of word-level information from pre-trained word embedding. To sum up briefly, the task relevance of CRF decoder is lower with the addition of category and lexical information, which leads to a very limited influence on the NER results compared with the others.

Table 4. Ablation Study on Weibo, Ontonotes, MSRA, and Resume datasets. Deletion of CRF denotes using softmax as a sequence labeling decoder instead of CRF, deletion of prompt denotes using the original Transformer structure before and after WIIA, and deletion of WIIA denotes removing the WIIA and only using the prompt-based Transformer.

Model	Weibo	Ontonotes	MSRA	Resume
PWII-BERT	72.06	81.64	95.70	96.19
w/o CRF	71.79	81.23	95.45	96.13
w/o prompt	70.71	80.99	95.33	95.12
w/o WIIA	71.10	80.66	95.35	95.70

4.1.5. Effectiveness Study

We also compare the results on each category to analyze the influence of category information. WII-BERT denotes using the original Transformer structure before and after WIIA. "CONT", "RACE", and "LOC" on the Resume dataset, and "GPE.NOM" are dropped while the F1 score becomes 100.00% or 0.00%. From Tables 5–8, the improvement occurs in most of the categories, and only four of them obtain an unexpected decrease. On Resume and MSRA datasets, PWII-BERT breaks down the effect for different categories and obtains an enhancement after adding the prompt-based category information. On the Ontonotes dataset, the category of "GPE", the effect is reduced after adding the category prompt. However, PWII-BERT achieves similar results with the other, with only a 0.05% gap. Due to the imbalance of training set and the limited training samples in the Weibo dataset, only part of the categories obtain a boost in the F1 score. In addition, more than half of the categories achieved better results. It can be observed that PWII-BERT does not specifically enhance the categories with a stable pattern, which proves that prompt-based category information improves the ability of the model to perceive the category from the perspective of the NER task.

Table 5. Results of P, R, and F1 score on the Resume dataset.

	PWII-BERT			WII-BERT		
	P	R	F1	P	R	F1
ORG	93.36	96.56	94.93	93.25	94.94	94.09
TITLE	96.23	96.23	96.23	93.99	95.45	94.72
EDU	97.35	98.21	97.78	98.20	97.32	97.76
Overall	95.52	96.87	96.19	94.54	95.70	95.12

Table 6. Results of P, R, and F1 score on the Weibo dataset.

	PWII-BERT			WII-BERT		
	P	R	F1	P	R	F1
GPE.NAM	78.18	91.49	84.31	68.25	91.49	78.18
LOC.NAM	64.29	47.37	54.55	52.33	42.11	47.06
LOC.NOM	33.33	22.22	26.67	25.00	11.11	15.38
PER.NAM	79.28	79.28	79.28	74.58	79.28	76.86
PER.NOM	72.00	74.12	73.04	71.11	75.29	73.14
ORG.NAM	58.33	53.85	56.00	64.52	51.28	57.14
ORG.NOM	56.25	52.94	54.55	60.00	52.94	56.25
Overall	72.15	71.98	72.07	69.72	71.74	70.71

Table 7. Results of P, R, and F1 score on the Ontonotes dataset.

	PWII-BERT			WII-BERT		
	P	**R**	**F1**	**P**	**R**	**F1**
GPE	83.62	83.23	83.43	81.50	85.56	83.48
LOC	53.04	44.67	48.50	45.05	40.98	42.92
ORG	72.90	74.46	73.68	70.88	74.52	72.65
PER	92.37	96.27	94.28	90.57	97.67	93.99
Overall	81.51	81.78	81.65	79.13	82.94	80.99

Table 8. Results of P, R, and F1 score on the MSRA dataset.

	PWII-BERT			WII-BERT		
	P	**R**	**F1**	**P**	**R**	**F1**
NR	97.11	97.40	97.88	97.64	97.38	97.56
NS	96.59	95.76	96.35	96.63	95.27	95.95
NT	91.47	93,00	92.23	91.62	91.55	91.59
Overall	95.75	95.65	95.70	95.72	94.93	95.33

4.1.6. Demarcation at Different Layers

To explore the influence of different layers to divide the classical Transformer and prompt-guided Transformer, the position of the boundary layer is modified from $k \in \{1, 2, 3, \ldots, 11\}$, which means we use k layers for classical Transformer and add another $12 - k$ layers for prompt-guided Transformer after the Word-level Information Injection Adapter. The results on Weibo dataset demonstrating that the medium layer achieves better performance are shown in Figure 4. This phenomenon can be due to two aspects: (1) The shallow layers extract basic linguistic features, and modifying the pre-trained model early may affect the effect of extraction; (2) The higher layers promote less layered interaction from prompt and lexicon features to BERT, resulting in insufficient use of information.

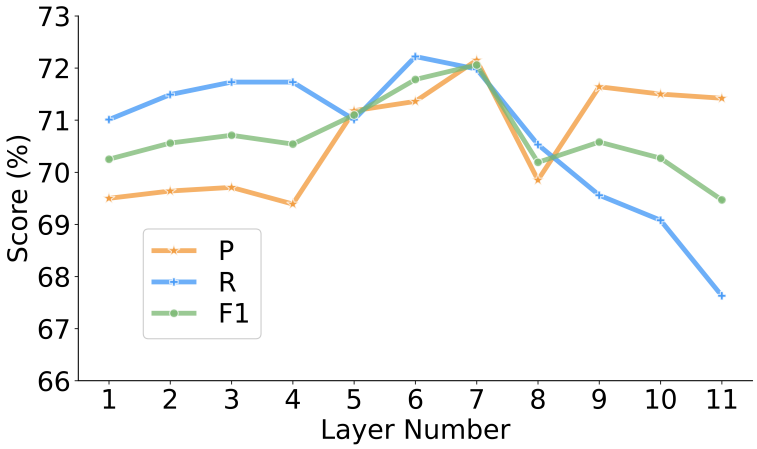

Figure 4. Results of PWII-BERT and WIIA are applied at different layers.

4.1.7. Case Study

To further verify whether our model obtains the lexicon information as prompt guided, we conducted an additional case study. As shown in Table 9, PWII-BERT is compared with the original BERT model. For the first sentence "广西壮族自治区政府 (Guangxi Bourau Autonomous Region Government)", the model guided by category prompt makes the error of recognizing the whole word as "GPE", which may be caused by the word "广西壮族自治区政府" occurring in the pre-trained dictionary and misleading the model. PWII-BERT classifies the sentence as their category "ORG" guided. In addition, the second sentence "佳能大连办公室设备有限公司 (Canon Dalian Office Equipment Co., Ltd.)" is divided into two parts according to the results of WII-BERT, which is another phenomenon of

dictionary-misleading. The whole entity does not appear in the predefined dictionary, thus misclassifying the word "佳能" as "GPE".

Table 9. Case study of the proposed method. WII-BERT denotes using the original Transformer structure before and after WIIA.

Sentence	广西壮族自治区政府 (Guangxi Bourau Autonomous Region Government)												
Matched Words	广西, 壮族, 自治区, 壮族自治区, 自治区政府, 政府, 区政府, 广西壮族自治区政府, 广西壮族自治区 Guangxi Province, Bourau, Autonomous region, Bourau Autonomous Region, Autonomous Region Government, Government, District Government, Guangxi Bourau Autonomous Region Government, Guangxi Bourau Autonomous Region												
Characters	广	西	壮	族	自	治	区	政	府				
Gold Labels	B-GPE	M-GPE	M-GPE	M-GPE	M-GPE	M-GPE	E-GPE	B-ORG	E-ORG				
WII-BERT	B-GPE	M-GPE	M-GPE	M-GPE	M-GPE	M-GPE	M-GPE	M-GPE	E-GPE				
PWII-BERT	B-GPE	M-GPE	M-GPE	M-GPE	M-GPE	M-GPE	E-GPE	B-ORG	E-ORG				
Sentence	佳能大连办公室设备有限公司 (Canon Dalian Office Equipment Co., Ltd.)												
Matched Words	佳能, 办公室, 大连, 设备有限公司, 有限公司, 公司 Canon, Office, Dalian City, Dalian Equipment Co., Ltd., Company Limited., Company												
Characters	佳	能	大	连	办	公	室	设	备	有	限	公	司
Gold Labels	B-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	E-ORG
WII-BERT	B-GPE	E-GPE	O	O	O	O	B-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	E-ORG
PWII-BERT	B-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	M-ORG	E-ORG

5. Conclusions

In this paper, we proposed a Prompt-based Word-level Information Injection BERT (PWII-BERT) to integrate prompt-guided lexicon information into a pre-trained language model to improve the performance of Chinese NER. PWII-BERT explicitly aligns the correlation features in character and corresponding words according to the category prompt. In experiments on four Chinese NER datasets, PWII-BERT outperforms the baseline models, demonstrating the significance of fully utilizing the advantages of fusing the category information and word-level features to implement Chinese NER.

Author Contributions: Conceptualization, P.Z. and Q.H.; methodology, Q.H. and G.C.; software, Q.H.; validation, W.S. and Q.H.; formal analysis, P.Z., W.S. and Q.H.; investigation, Q.H.; resources, G.C., P.Z., W.S. and Q.H.; data curation, Q.H.; writing—original draft preparation, Q.H.; writing—review and editing, G.C., P.Z., W.S. and Q.H.; visualization, W.S. and G.C.; supervision, P.Z.; project administration, G.C., P.Z., W.S. and Q.H.; funding acquisition, G.C., P.Z., W.S. and Q.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (No. 2020AAA0108701).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

NER	Named Entity Recognition
PWII-BERT	Prompt-based Word-level Information Injection BERT
WIIA	Word-level Information Injection Adapter
PLMs	Pre-trained Language Models
CRF	Conditional Random Field

References

1. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural Architectures for Named Entity Recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 260–270. [CrossRef]
2. Ma, X.; Hovy, E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 1064–1074. [CrossRef]
3. Chiu, J.P.C.; Nichols, E. Named Entity Recognition with Bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [CrossRef]
4. Gui, T.; Zhang, Q.; Huang, H.; Peng, M.; Huang, X. Part-of-Speech Tagging for Twitter with Adversarial Neural Networks. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2411–2420. [CrossRef]
5. Wang, Y.; Sun, Y.; Ma, Z.; Gao, L.; Xu, Y. An ERNIE-based joint model for Chinese named entity recognition. *Appl. Sci.* **2020**, *10*, 5711. [CrossRef]
6. Yang, L.; Fu, Y.; Dai, Y. BIBC: A Chinese Named Entity Recognition Model for Diabetes Research. *Appl. Sci.* **2021**, *11*, 9653. [CrossRef]
7. Syed, M.H.; Chung, S.T. MenuNER: Domain-adapted BERT based NER approach for a domain with limited dataset and its application to food menu domain. *Appl. Sci.* **2021**, *11*, 6007. [CrossRef]
8. Chen, S.; Pei, Y.; Ke, Z.; Silamu, W. Low-resource named entity recognition via the pre-training model. *Symmetry* **2021**, *13*, 786. [CrossRef]
9. Gao, Y.; Wang, Y.; Wang, P.; Gu, L. Medical named entity extraction from chinese resident admit notes using character and word attention-enhanced neural network. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1614. [CrossRef] [PubMed]
10. Goldberg, Y. Assessing BERT's Syntactic Abilities. *arXiv* **2019**, arXiv:1901.05287.
11. Hewitt, J.; Manning, C.D. A Structural Probe for Finding Syntax in Word Representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 4129–4138. [CrossRef]
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
13. Zhang, Y.; Yang, J. Chinese NER Using Lattice LSTM. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1554–1564. [CrossRef]
14. Yang, J.; Zhang, Y.; Liang, S. Subword Encoding in Lattice LSTM for Chinese Word Segmentation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 2720–2725. [CrossRef]
15. Liu, W.; Xu, T.; Xu, Q.; Song, J.; Zu, Y. An Encoding Strategy Based Word-Character LSTM for Chinese NER. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 2379–2389. [CrossRef]
16. Ding, R.; Xie, P.; Zhang, X.; Lu, W.; Li, L.; Si, L. A Neural Multi-digraph Model for Chinese NER with Gazetteers. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1462–1467. [CrossRef]
17. Higashiyama, S.; Utiyama, M.; Sumita, E.; Ideuchi, M.; Oida, Y.; Sakamoto, Y.; Okada, I. Incorporating Word Attention into Character-Based Word Segmentation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 2699–2709. [CrossRef]
18. Ma, R.; Peng, M.; Zhang, Q.; Wei, Z.; Huang, X. Simplify the Usage of Lexicon in Chinese NER. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5951–5960. [CrossRef]
19. Li, X.; Yan, H.; Qiu, X.; Huang, X. FLAT: Chinese NER Using Flat-Lattice Transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6836–6842. [CrossRef]
20. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-Efficient Transfer Learning for NLP. In Proceedings of the 36th International Conference on Machine Learning, ICML, Long Beach, CA, USA, 10–15 June 2019; Proceedings of Machine Learning Research. Chaudhuri, K., Salakhutdinov, R., Eds.; Volume 97, pp. 2790–2799.

21. Bapna, A.; Firat, O. Simple, Scalable Adaptation for Neural Machine Translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 1538–1548. [CrossRef]
22. Wang, R.; Tang, D.; Duan, N.; Wei, Z.; Huang, X.; Ji, J.; Cao, G.; Jiang, D.; Zhou, M. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In Proceedings of the Findings of the Association for Computational Linguistics, Online Event, 1–6 August 2021; Volume ACL/IJCNLP 2021; pp. 1405–1418. [CrossRef]
23. Pfeiffer, J.; Vulic, I.; Gurevych, I.; Ruder, S. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020; pp. 7654–7673. [CrossRef]
24. Gui, T.; Ma, R.; Zhang, Q.; Zhao, L.; Jiang, Y.; Huang, X. CNN-Based Chinese NER with Lexicon Rethinking. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019; pp. 4982–4988. [CrossRef]
25. 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. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, Vancouver, BC, Canada, 8–14 December 2019; pp. 2742–2753.
26. Hu, Y.; Verberne, S. Named Entity Recognition for Chinese biomedical patents. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 627–637. [CrossRef]
27. Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; Wu, H. ERNIE: Enhanced Representation through Knowledge Integration. *arXiv* **2019**, arXiv:1904.09223.
28. Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Tian, H.; Wu, H.; Wang, H. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, 7–12 February 2020; pp. 8968–8975.
29. 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. [CrossRef]
30. Diao, S.; Bai, J.; Song, Y.; Zhang, T.; Wang, Y. ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020; pp. 4729–4740. [CrossRef]
31. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual, 6–12 December 2020.
32. Schick, T.; Schütze, H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, 19–23 April 2021; Merlo, P., Tiedemann, J., Tsarfaty, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 255–269.
33. Schick, T.; Schmid, H.; Schütze, H. Automatically Identifying Words That Can Serve as Labels for Few-Shot Text Classification. In Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), 8–13 December 2020; pp. 5569–5578. [CrossRef]
34. Shin, T.; Razeghi, Y.; IV, R.L.L.; Wallace, E.; Singh, S. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, 16–20 November 2020; Webber, B., Cohn, T., He, Y., Liu, Y., Eds.; Association for Computational Linguistics: Stroudsburg PA, USA, 2020; pp. 4222–4235. [CrossRef]
35. Han, X.; Zhao, W.; Ding, N.; Liu, Z.; Sun, M. PTR: Prompt Tuning with Rules for Text Classification. *arXiv* **2021**, arXiv:2105.11259.
36. Poth, C.; Pfeiffer, J.; Rücklé, A.; Gurevych, I. What to Pre-Train on? Efficient Intermediate Task Selection. 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; Moens, M., Huang, X., Specia, L., Yih, S.W., Eds.; Association for Computational Linguistics: Stroudsburg PA, USA, 2021; pp. 10585–10605.
37. Ben-David, E.; Oved, N.; Reichart, R. PADA: Example-based Prompt Learning for on-the-fly Adaptation to Unseen Domains. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 414–433. [CrossRef]
38. Frankle, J.; Carbin, M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019; Available online: https://openreview.net/ (accessed on 26 February 2023).
39. Sanh, V.; Wolf, T.; Rush, A.M. Movement Pruning: Adaptive Sparsity by Fine-Tuning. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual, 6–12 December 2020.
40. Li, X.L.; Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 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, 1–6 August 2021; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 4582–4597. [CrossRef]

41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
42. Peng, N.; Dredze, M. Named Entity Recognition for Chinese Social Media with Jointly Trained Embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 548–554. [CrossRef]
43. Peng, N.; Dredze, M. Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 149–155. [CrossRef]
44. Weischedel, R.; Pradhan, S.; Ramshaw, L.; Palmer, M.; Xue, N.; Marcus, M.; Taylor, A.; Greenberg, C.; Hovy, E.; Belvin, R.; et al. *Ontonotes Release 4.0. LDC2011T03*; Linguistic Data Consortium: Philadelphia, PA, USA, 2011.
45. Levow, G. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. In Proceedings of the Fifth Workshop on Chinese Language Processing, SIGHAN@COLING/ACL 2006, Sydney, Australia, 22–23 July 2006; pp. 108–117.
46. Zhu, Y.; Wang, G. CAN-NER: Convolutional Attention Network for Chinese Named Entity Recognition. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 3384–3393. [CrossRef]
47. Liu, W.; Fu, X.; Zhang, Y.; Xiao, W. Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Virtual Event, 1–6 August 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 5847–5858. [CrossRef]

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.