



Article

An ERNIE-Based Joint Model for Chinese Named Entity Recognition

Yu Wang ^{1,2} , Yining Sun ^{1,2,*}, Zuchang Ma ¹, Lisheng Gao ¹  and Yang Xu ¹

¹ Anhui Province Key Laboratory of Medical Physics and Technology, Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China; briskyu@mail.ustc.edu.cn (Y.W.); zcma@iim.ac.cn (Z.M.); lsgao@iim.ac.cn (L.G.); yxu@hfcas.ac.cn (Y.X.)

² Science Island Branch of Graduate School, University of Science and Technology of China, Hefei 230026, China

* Correspondence: ynsun@iim.ac.cn

Received: 29 July 2020; Accepted: 14 August 2020; Published: 18 August 2020



Abstract: Named Entity Recognition (NER) is the fundamental task for Natural Language Processing (NLP) and the initial step in building a Knowledge Graph (KG). Recently, BERT (Bidirectional Encoder Representations from Transformers), which is a pre-training model, has achieved state-of-the-art (SOTA) results in various NLP tasks, including the NER. However, Chinese NER is still a more challenging task for BERT because there are no physical separations between Chinese words, and BERT can only obtain the representations of Chinese characters. Nevertheless, the Chinese NER cannot be well handled with character-level representations, because the meaning of a Chinese word is quite different from that of the characters, which make up the word. ERNIE (Enhanced Representation through kNowledge IntEgration), which is an improved pre-training model of BERT, is more suitable for Chinese NER because it is designed to learn language representations enhanced by the knowledge masking strategy. However, the potential of ERNIE has not been fully explored. ERNIE only utilizes the token-level features and ignores the sentence-level feature when performing the NER task. In this paper, we propose the ERNIE-Joint, which is a joint model based on ERNIE. The ERNIE-Joint can utilize both the sentence-level and token-level features by joint training the NER and text classification tasks. In order to use the raw NER datasets for joint training and avoid additional annotations, we perform the text classification task according to the number of entities in the sentences. The experiments are conducted on two datasets: MSRA-NER and Weibo. These datasets contain Chinese news data and Chinese social media data, respectively. The results demonstrate that the ERNIE-Joint not only outperforms BERT and ERNIE but also achieves the SOTA results on both datasets.

Keywords: joint training; named entity recognition; pre-training models; ERNIE; BERT

1. Introduction

Named Entity Recognition (NER), as the fundamental task of Natural Language Processing (NLP), aims to recognize entities with specific meanings from unstructured text, such as the names of people, locations, and organizations [1]. It is the initial step in extracting valuable knowledge from unstructured text and building a Knowledge Graph (KG). The performance of NER may affect downstream knowledge extraction tasks, such as the Relation Extraction (RE) [2]. In the early years, researchers used rule-based or dictionary-based methods for NER tasks [3,4]. However, these methods lack generalization because they are proposed for particular types of entities. Machine learning and deep learning methods emerging in recent years are also used in NER tasks [5,6]. Nevertheless, the performance of these methods often suffers from small-scale human-labelled training

data, resulting in poor generalization capability, especially for rare words. Therefore, it is of interest to know whether the prior semantic knowledge can be learned from large amounts of unlabelled corpora to improve the performance of NER.

Recently, BERT (Bidirectional Encoder Representations from Transformers) [7] achieved state-of-the-art (SOTA) results in various NLP tasks. It can obtain prior semantic knowledge from large-scale unlabelled corpora through pre-training tasks and improve the performance of downstream tasks by transferring this knowledge to them [7]. However, Chinese NER is still a more challenging task for BERT because there are no physical separations between Chinese words. Therefore, BERT can only obtain character-level representations during pre-training. For example, the meaning of the two sentences inputted into BERT in Figure 1 is the same. The token of the English sentence is a word, while the token of the Chinese sentence is a character.



Figure 1. The difference between Chinese and English sentences when inputted into Bidirectional Encoder Representations from Transformers (BERT).

During the procedure of “Masked Language Model (MLM)”, which is a pre-training task, BERT will mask some tokens at random and predict it in order to learn the prior semantic knowledge about the tokens. Therefore, BERT can only learn the character-level representations as the tokens are all Chinese characters. However, the Chinese NER cannot be handled well when only using character-level representations, because in general, the meaning of a Chinese word is quite different from that of the characters, which make up the word. ERNIE (Enhanced Representation through kNnowledge IntEgration) [8], which is an improved pre-training model of BERT, is more suitable for Chinese NER because it is designed to learn language representations enhanced by knowledge masking strategy. Unlike the character-level masking strategy of BERT, which can only learn the character-level representations of Chinese, the knowledge masking strategy of ERNIE consists of entity-level and phrase-level masking strategies and can learn the prior semantic knowledge of Chinese entities and phrases implicitly during pre-training. The model has better generalization and adaptability due to the knowledge masking strategy [8].

However, the potential of ERNIE has not been fully explored. In order to transfer the prior semantic knowledge to a downstream task (e.g., NER), ERNIE must be fine-tuned over a task-specific dataset. As shown in Figure 2, for a NER task, the representation h_d of a token t_d (for $d = 1, \dots, D$) can be used to classify this token with respect to the target categories, and $h = \{h_1, \dots, h_D\}$ can be regarded as the token-level features. For a text classification task, the representation h_C of the [CLS] is a fixed dimensional pooled representation of the sequence, and h_C can be regarded as the sentence-level feature. Therefore, it would be of special interest to know whether the performance of NER can be improved by utilizing both the token-level and sentence-level features.

In this paper, we aim to improve the performance of Chinese NER by utilizing both the token-level and sentence-level features. The main contributions of this paper can be summarized as follows:

- ERNIE was selected as the pre-training model because of the knowledge masking strategy it has. This masking strategy is more suitable for Chinese NER than BERT [8].
- We propose the ERNIE-Joint, which is an ERNIE-based joint training model for Chinese NER. The learning objective of ERNIE-Joint is to maximize the conditional probability $p(y_i, y_s | x)$ over a unique cost function, where x denotes the input sentence, y_s and y_i denote the results of NER and text classification. In this way, the token-level and sentence-level features can be both utilized.
- In order to use the raw NER datasets and avoid additional annotations, the classification task is performed according to the number of entities in sentences.

- The experiments are conducted on two datasets: MSRA-NER and Weibo. These datasets contain Chinese news data and Chinese social media data, respectively. Experimental results demonstrate that the ERNIE-Joint not only outperforms BERT and ERNIE but also achieves the SOTA results on both datasets. However, given that ERNIE-Joint introduces the cross-entropy errors in classification task when calculating the loss function, its running time will be higher than that of ERNIE, which can be a drawback of ERNIE-Joint.

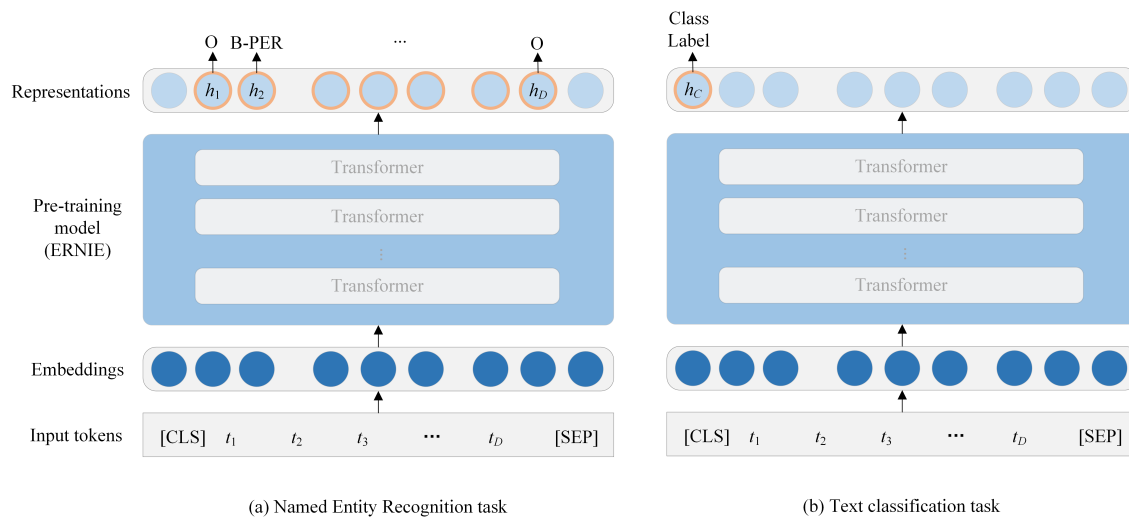


Figure 2. The illustration of fine-tuning Enhanced Representation through kNowledge IntEgration (ERNIE) on Named Entity Recognition (NER) and text classification tasks.

2. Related Work

In this section, we will introduce the related work of NER and pre-training models.

2.1. Named Entity Recognition

The Named Entity Recognition aims to recognize the entities with specific meanings in the text. Rule-based and dictionary-based approaches had played an important role. For example, Friedman et al. [3] developed a general natural language processor that identifies clinical information in narrative reports. Gerner et al. [4] used a dictionary-based approach to identify species names. However, rule-based and dictionary-based methods lack generalization because they are proposed for particular types of entities. Researchers also tried to use machine learning or statistical model like Conditional Random Field (CRF) to recognize entities from unstructured data. Zhang et al. [9] presented a stochastic model to tackle the problem of Chinese NER. Chen et al. [5] used two conditional probabilistic models for the Chinese NER task. Nevertheless, these methods need hand-crafted features, which is time-consuming and laborious. In recent years, deep learning methods attracted increasing attention. These methods can improve the performance of the NER without feature engineering. Researchers mainly adopted Bidirectional Long Short-Term Memory (BiLSTM) with a CRF layer to conduct the NER task [6,10,11]. Some researchers also utilized the attention mechanism. For example, Wei et al. [12] and Wu et al. [13] improved BiLSTM-CRF model with the self-attention mechanism. Yin et al. [14] proposed an advanced BiLSTM-CRF model based on the radical level features and self-attention mechanism. However, the performance of these methods often suffers from small-scale human-labelled training data.

2.2. Pre-Training Models

The pre-training models aim to learn word embeddings or representations with prior semantic knowledge through pre-training tasks from a large number of unlabelled corpora. Mikolov et al. [15,16] first proposed the Word2Vec model to generate the word embeddings. However,

the non-contextual word embeddings fail to model the polysemous words. Peters et al. [17] proposed ELMO, which learned contextual embeddings according to the internal states of a deep Bidirectional Language Model (BiLM) based on BiLSTM. However, BiLSTM is weaker than the Transformer in feature extraction [18]. Devlin et al. [7] released BERT in 2018, which consists of multi-layer bidirectional Transformer blocks [7,18]. BERT enhances the performance of downstream tasks through fine-tuning and achieves the SOTA results in various NLP tasks [7,19]. Devlin et al. [7] first illustrated how to fine-tune the pre-training model on different NLP tasks, including the NER. Since then, some researchers have conducted the NER task based on BERT. For example, Labusch et al. [20] applied BERT to the NER task in contemporary and historical German text. Taher et al. [21] used BERT to recognize the named entity in Persian. Hakala et al. [22] applied the multilingual BERT to Spanish biomedical NER.

Zhang et al. [8] improved the pre-training tasks of BERT and released ERNIE in 2019. ERNIE masks the entities and phrases during the pre-training procedure to obtain the prior semantic knowledge about them, which is more suitable for Chinese NER. However, the potential of ERNIE has not been fully explored. In Section 3, we will introduce the ERNIE-Joint, which is a joint training model based on ERNIE.

3. Methods

In this section, we first briefly introduce ERNIE, then propose the joint training model ERNIE-Joint.

3.1. ERNIE

ERNIE is an improved pre-training model of BERT and consists of multi-layer Transformer blocks, too. The Transformer can capture the contextual information for each token through self-attention, and generate the contextual embeddings [8]. ERNIE uses 12 Transformer layers, 768 hidden units, and 12 attention heads as well as BERT. The main difference between ERNIE and BERT is masking strategies. As shown in Figure 3, BERT randomly masks the Chinese characters in a sentence during pre-training. The character-level masking strategy can obtain character-level representations, but high-level representations are hard to fully modelled. On the contrary, ERNIE takes an entity or a phrase as one unit, which is usually composed of several characters. All of the characters in the same unit are masked during pre-training. In this way, the prior semantic knowledge of entities and long semantic dependency are implicitly learned, such as the relationship between the two entities “Hefei City” and “Provincial capital” shown in Figure 3.

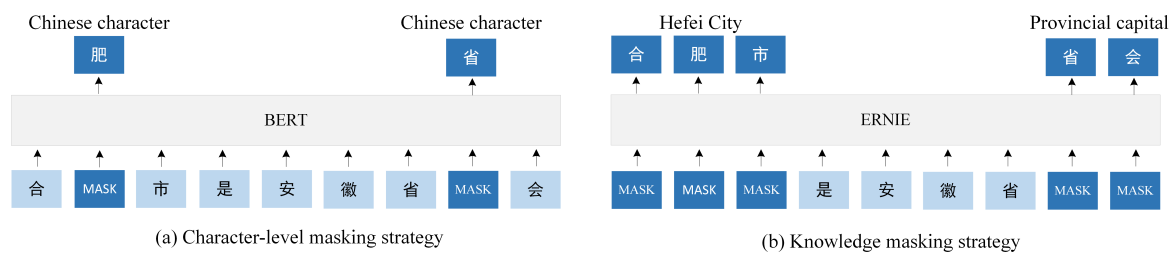


Figure 3. The different masking strategies between BERT and ERNIE.

3.2. ERNIE-Joint

As shown in Figure 4, ERNIE-Joint is an improved model of ERNIE, and the core part of it is still ERNIE. Therefore, the input format of ERNIE-Joint must be exactly the same as that of ERNIE. The input of ERNIE-Joint constructed by summing the following four parts:

- Token IDs: This number means the ID of each token based on the dictionary of ERNIE-Joint.

- Sentence IDs: ERNIE uses this number to determine which sentence the token belongs to. However, all the sentence ID numbers are “0” in this work because we input only one sentence at a time into ERNIE-Joint, not a sentence pair.
- Position IDs: The Transformer cannot obtain position information. Therefore ERNIE-Joint uses position IDs to obtain the order of the tokens.
- Segmentation IDs: This number indicates whether a character belongs to an entity or a phrase. Specifically, “0” means it belongs to the beginning of a Chinese entity or phrase, and “1” means it does not belong to the beginning.

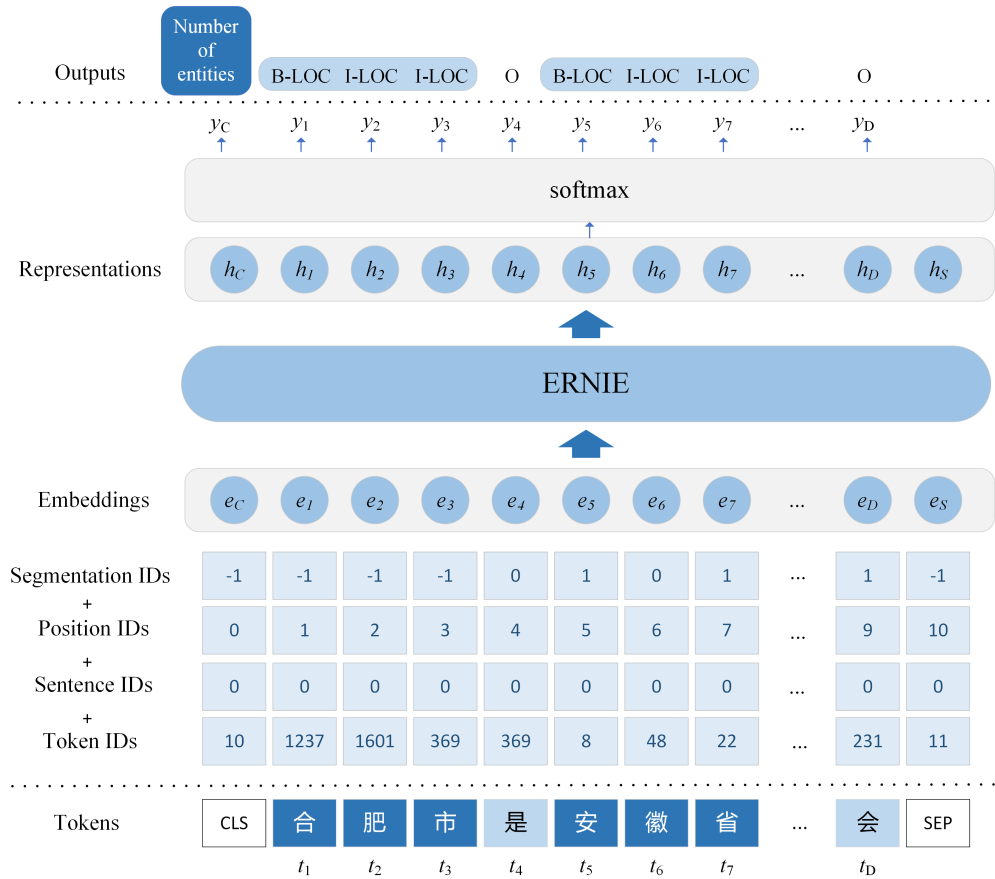


Figure 4. The architecture of ERNIE-Joint.

As mentioned before, the input of ERNIE-Joint is identical with that of ERNIE. Therefore, it is necessary to insert the special token [CLS] as the first token of a sequence and [SEP] as the final token. Given an input token sequence $t = \{t_1, \dots, t_D\}$, the output of ERNIE-Joint is $h = \{h_C, h_1, \dots, h_D\}$ where h_C is the representation of [CLS], and the h_d corresponds to t_d , being $d = 1, \dots, D$.

For the NER task, the token-level categories probabilities for the token t_d can be obtained through:

$$P_d = \text{softmax}(\mathbf{W}_s h_d + \mathbf{b}_s), \quad (1)$$

where $\mathbf{W}_s \in \mathbb{R}^{N \times H}$ and $\mathbf{b}_s \in \mathbb{R}^N$, that is, the token-level classifier matrix and bias. H is the dimension of final hidden state and N is the number of token-level categories. The category for token t_d can be obtained by:

$$y_d = \text{argmax}(P_d). \quad (2)$$

The loss function for one token is:

$$loss = \sum_{n=1}^N p(y_d^n) \log[q(y_d^n)], \quad (3)$$

where $p(y_d^n)$ denotes the probability distribution of correct labels and $q(y_d^n)$ denotes the probability distribution of predicted labels.

For the text classification task, the sentence-level category probabilities for token [CLS] can be obtained through:

$$P_c = \text{softmax}(\mathbf{W}_c h_c + \mathbf{b}_c), \quad (4)$$

where $\mathbf{W}_c \in \mathbb{R}^{M \times H}$ and $\mathbf{b}_c \in \mathbb{R}^M$, that is, the sentence-level classifier matrix and bias. M is the number of sentence-level categories. The category can be obtained by:

$$y_c = \text{argmax}(P_c). \quad (5)$$

The loss function for the text classification task is:

$$loss = \sum_{m=1}^M p(y_c^m) \log[q(y_c^m)], \quad (6)$$

where $p(y_c^m)$ denotes the probability distribution of correct labels and $q(y_c^m)$ denotes the probability distribution of predicted labels.

In order to fine-tune the ERNIE-Joint model by joint training both the two tasks, we define a unique cost function. Giving the input sentence \mathbf{x} , the learning objective for joint training is to maximize the following conditional probability:

$$p(\mathbf{y}_i, \mathbf{y}_s | \mathbf{x}) = p(y_c | \mathbf{x}) \prod_{d=1}^D p(y_d | \mathbf{x}), \quad (7)$$

where \mathbf{y}_i denotes the correct number of the entities in the input sequence, and $\mathbf{y}_s = \{y_1, \dots, y_D\}$ denotes the correct sequence for the NER. Then, the optimization goal for joint training is to minimize this cost function:

$$cost = \frac{1}{D+1} \left\{ \sum_{d=1}^D \sum_{n=1}^N p(y_d^n) \log[q(y_d^n)] + \sum_{m=1}^M p(y_c^m) \log[q(y_c^m)] \right\}. \quad (8)$$

Moreover, sentences are classified based on emotion or intent for common text classification tasks. However, additional annotations are necessary if we use a raw NER dataset for joint training based on the indicators above. Therefore, we introduce the number of entities in a sentence for the text classification task, and additional annotations can be avoided in this way. However, the distribution of categories in the training set, validation set, and test set must be similar when re-labelling. The re-labelling results of the two NER datasets are presented in Section 4.

4. Experiments and Results

In this section, we will introduce the datasets for joint training and show the experimental results. The experiments were performed with PaddlePaddle, which is a framework of deep learning. For hardware, we used an eight-core CPU and an NVIDIA Tesla V100 GPU.

4.1. Datasets

There are two kinds of datasets be used for the experiments: MSRA-NER and Weibo. These datasets contain Chinese news data and Chinese social media data, respectively. The MSRA-NER

dataset of SIGHAN Bakeoff 2006, which carries precise annotations from the news field and is provided by Levow et al. [23], contains three kinds of entity types: PER (Person), ORG (Organization), and LOC (Location) as shown in Table 1.

Table 1. The summary of the MSRA-NER.

| Entities | Training Set | Validation Set | Test Set |
|----------|--------------|----------------|----------|
| PER | 8144 | 884 | 1864 |
| ORG | 9277 | 984 | 2185 |
| LOC | 16,571 | 1951 | 3658 |
| Total | 33,992 | 3819 | 7707 |

The Weibo dataset includes 1890 messages sampled from Sina Weibo between November 2013 and December 2014. This dataset is annotated with four types: PER (Person), ORG (Organization), LOC (Location), and GPE (Geo-Political), including named and nominal mentions. This dataset is divided into the training set, validation set and test set as He et al. [24]. The summary is listed in Table 2.

Table 2. The summary of the Weibo.

| Named entities | Training set | Validation set | Test set |
|------------------|--------------|----------------|----------|
| PER | 574 | 90 | 111 |
| ORG | 183 | 47 | 39 |
| LOC | 56 | 6 | 19 |
| GPE | 205 | 26 | 47 |
| Nominal entities | Training set | Validation set | Test set |
| PER | 766 | 208 | 170 |
| ORG | 42 | 5 | 17 |
| LOC | 51 | 6 | 9 |
| GPE | 8 | 1 | 2 |
| Total | 1885 | 389 | 414 |

As mentioned before, in order to make the two NER datasets available for the classification task, we re-label the sentences in the datasets based on the number of entities they contain. Table 3 shows the results after re-labelling. For example, a sentence in the MSRA-NER is labelled as category A if it contains no entities and as category B if it contains one or two entities. We try to make the distribution of each label similar in the training set, validation set, and test set.

Table 3. The summary of re-labelling.

| Datasets | Labels | Training Set | Validation Set | Test Set | Number of Entities |
|----------|--------|--------------|----------------|----------|--------------------|
| MSRA-NER | A | 8153 | 918 | 1832 | 0 |
| | B | 8008 | 860 | 1699 | 1,2 |
| | C | 4073 | 540 | 1105 | ≥ 3 |
| Weibo | A | 612 | 123 | 107 | 0,1,2 |
| | B | 738 | 147 | 163 | ≥ 3 |

4.2. Hyper-Parameters

For hyper-parameters, we adjust them according to the performance on the validation set. The hyper-parameters used in this paper are listed in Table 4. Moreover, we use Adam as the optimizer.

Table 4. Hyper-parameters.

| Hyper-Parameters | Values |
|------------------|--------------------|
| Batch size | 32 |
| Epoch | 6 |
| Learning rate | 5×10^{-5} |
| Weight decay | 0.01 |

4.3. Results

In this section, we will show the experimental results on the two datasets. We introduce the precision, recall and F1-score to evaluate the performance. The precision value refers to the ratio of correct entities to predicted entities. The recall value is the proportion of the entities in the test set that are correctly predicted. The F1-score is calculated according to the following formulation:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (9)$$

Tables 5 and 6 show the results of different datasets. Moreover, we use “baseline” to indicate the Bidirectional Gate Recurrent Unit (BiGRU) with a CRF layer as Zhu et al. [25].

4.3.1. MSRA-NER Dataset

Table 5 shows the experimental results of diverse models on the MSRA-NER dataset, which is provided by Levow et al. [23]. The training set, validation set, and test set used to test the models in Table 5 are all identical. In the first block, we give the performance of previous methods. Chen et al. [5], Zhang et al. [26], and Zhou et al. [27] exploited multi-prototype embeddings and leveraged rich hand-craft features for the Chinese NER task, and Zhang et al. [26] obtained the F1-score of 91.18%. Dong et al. [28] applied a BiLSTM-CRF model which utilizes radical features and achieved the F1-score of 90.95%. Yang et al. [29] obtained the F1-score of 91.67% by proposing a CNN-BiRNN-CRF model, which incorporates stroke features. Cao et al. [30] utilized adversarial transfer learning to conduct the NER task. Zhu et al. [25] investigated a Convolutional Attention Network (CAN) for Chinese NER task. The model Zhang et al. proposed achieved the highest F1-score of 93.18% in the first block, but the result heavily depends on the quality of external lexicon data. The results of the baseline, BERT, ERNIE, and our model are listed in the second block. The pre-training models (BERT and ERNIE) outperform all the previous methods and baseline model without additional features. However, the ERNIE-Joint we proposed achieves the SOTA result with the F1-score of 94.20%.

Table 5. The results on the MSRA-NER.

| Models | Precision/% | Recall/% | F1-Score/% |
|-------------------|--------------|--------------|--------------|
| Chen et al. [5] | 91.22 | 81.71 | 86.20 |
| Zhang et al. [26] | 92.20 | 90.18 | 91.18 |
| Zhou et al. [27] | 91.86 | 88.75 | 90.28 |
| Dong et al. [28] | 91.28 | 90.62 | 90.95 |
| Yang et al. [29] | 92.04 | 91.31 | 91.67 |
| Cao et al. [30] | 91.30 | 89.58 | 90.64 |
| Zhang et al. [11] | 93.57 | 92.79 | 93.18 |
| Zhu et al. [25] | 93.53 | 92.42 | 92.97 |
| Baseline [25] | 92.54 | 88.20 | 90.32 |
| BERT [7] | - | - | 92.60 |
| ERNIE [8] | 93.35 | 94.82 | 94.08 |
| ERNIE-Joint | 93.58 | 94.82 | 94.20 |

4.3.2. Weibo Dataset

We also compared the model we proposed with previous methods on Weibo dataset, which consists of Chinese social media text. Table 6 shows the F1-score for named entities, nominal entities, and both (Overall) on the Weibo dataset, which is provided by He et al. [24]. The training set, validation set, and test set used to test the models in Table 6 are all identical. The results of previous methods are listed in the first block. Peng et al. [31] proposed a jointly model which achieves F1-score of 56.05%. Peng et al. [32] also trained the NER task with Chinese Word Segmentation task and improved the F1-score to 58.99%. He et al. [33] proposed a unified model which can utilize cross-domain learning and semi-supervised learning. This model improved the F1-score from 54.82% to 58.23% compared to another model they proposed [24]. As mentioned before, Zhang et al. [11] introduced a lattice structure and obtained the F1-score of 58.79%. This result is slightly better than that of Cao et al. [30], who utilized the adversarial transfer learning. Zhu et al. [25] investigated a Convolutional Attention Network (CAN) to conduct this NER task and obtained the highest F1-score of the methods in the first block. In the second block of Table 6, we also give the results of the baseline, BERT, ERNIE, and the ERNIE-Joint. The baseline model achieves the F1-score of 53.80%. BERT and ERNIE improved the F1-score significantly, but the model we proposed achieves the F1-score of 69.08%, which is the highest result among existing models.

Table 6. The results on the Weibo.

| Models | Named Entities | Nominal Entities | Overall |
|-------------------|----------------|------------------|--------------|
| Peng et al. [31] | 51.96 | 61.05 | 56.05 |
| Peng et al. [32] | 55.28 | 62.97 | 58.99 |
| He et al. [24] | 50.60 | 59.32 | 54.82 |
| He et al. [33] | 54.50 | 62.17 | 58.23 |
| Cao et al. [30] | 54.34 | 57.35 | 58.70 |
| Zhang et al. [11] | 53.04 | 62.25 | 58.79 |
| Zhu et al. [25] | 55.38 | 62.98 | 59.31 |
| Baseline [25] | 49.02 | 58.80 | 53.80 |
| BERT | 60.70 | 69.99 | 68.12 |
| ERNIE | 60.88 | 70.98 | 68.20 |
| ERNIE-Joint | 62.33 | 71.17 | 69.08 |

4.3.3. Run Time Test

In order to test the performance of ERNIE-Joint at runtime, we also compared the running time of ERNIE-Joint and ERNIE on the MSRA-NER dataset. As shown in Figure 5, in general, ERNIE-Joint does require more running time in every epoch than ERNIE.

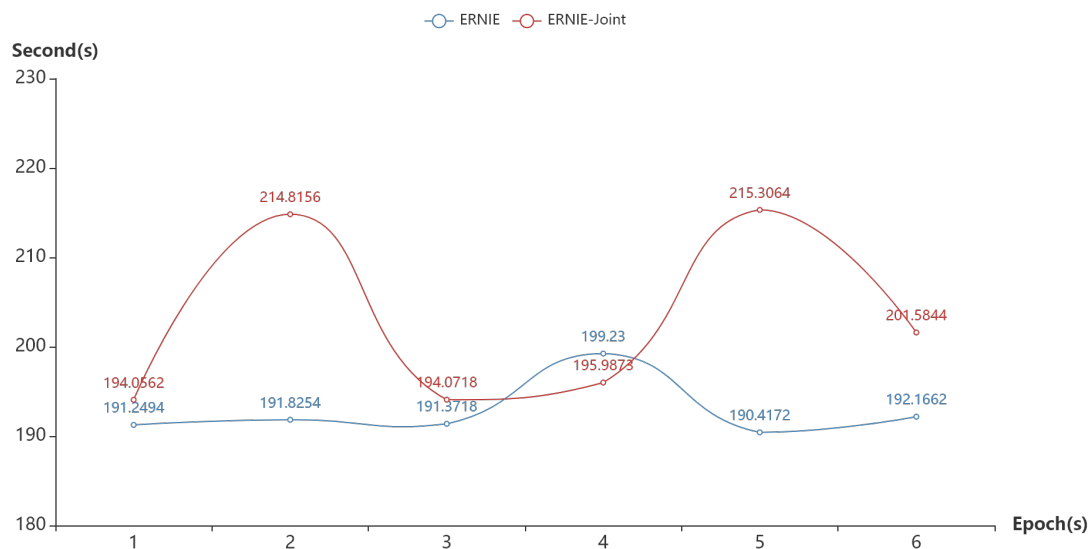


Figure 5. The running time of ERNIE-Joint and ERNIE.

5. Discussion

Firstly, using pre-training models can improve the performance of Chinese NER significantly without utilizing any external resources. The pre-training models have learned abundant prior semantic knowledge from the pre-training corpora (e.g., Baidu News) [8], which can also be known as the “source domain”. The task-specific semantic knowledge will also be obtained during fine-tuning from the training set of the downstream task, which can also be known as the “target domain”. The whole process can be regarded as transfer learning. However, the deep learning models only obtain the semantic knowledge from the “target domain”. The training process is done from scratch, whether it is the baseline model (BiGRU-CRF) or other deep learning models.

Secondly, ERNIE outperforms BERT in Chinese NER task. As mentioned before, BERT can only obtain the character-level representations of Chinese through the character-level masking strategy during pre-training. However, the knowledge masking strategy of ERNIE can learn the relationship between Chinese entities and long semantic dependency implicitly. Therefore, the representations of tokens generated by ERNIE contain the prior semantic knowledge of entities and phrases, which can make the model has better generalization and adaptability.

Thirdly, the performance of Chinese NER can be improved when utilizing both the sentence-level and token-level features. ERNIE only uses the token-level features when conducting the NER task. The representation of [CLS] can be regarded as the sentence-level feature, but it does not participate in the computation of the cost function. The ERNIE-Joint model we proposed utilizes both the sentence-level and token-level features by a unique cost function. The experimental results show that introducing sentence-level features through joint training can improve the performance of the NER task. Moreover, the classification method we proposed makes the raw NER datasets suitable for text classification tasks, and these datasets can be applied to a joint training model without additional annotations.

Finally, given that ERNIE-Joint introduces the cross-entropy errors in classification task when calculating the loss function, its running time will be higher than that of ERNIE. As shown in Figure 5, the running time of ERNIE-Joint is higher than that of ERNIE in each epoch except for the fourth epoch. In the fourth epoch, the running time of ERNIE may be affected by the running environment of a computer. The relatively high running time can be regarded as the drawback of our model.

6. Conclusions

In this paper, we enhance the performance of Chinese NER through an ERNIE-based joint model called ERNIE-Joint. We choose ERNIE as the pre-training model we used because of the knowledge masking strategy it has during pre-training procedure. Knowledge masking strategy can obtain the prior semantic knowledge of entities or phrases, which is more suitable for Chinese NER because there is no segmentation between Chinese characters, and BERT can only obtain character-level representations. Moreover, The ERNIE-Joint, which is a joint training model, can utilize both the sentence-level and token-level features when performing the NER task through a unique cost function. In order to use the raw NER dataset for joint training and avoid additional annotations, text classification task is performed according to the number of entities in sentences. The experiments are conducted on two datasets—MSRA-NER and Weibo. These datasets contain Chinese news data and Chinese social media data, respectively. The results demonstrate that ERNIE-Joint not only outperforms BERT and ERNIE but also obtained the SOTA results on both datasets.

For future work, firstly, we will test the performance of ERNIE-Joint in specific domains, such as sports medicine. Experiments will be carried out on a sports medicine related dataset labelled by ourselves. Secondly, we will test ERNIE-Joint on a multilingual dataset and observe its performance in other languages. Thirdly, In order to build a Chinese KG from unstructured documents, we need to extract the relationship of entities recognized by ERNIE-Joint. Considering that the RE can also be regarded as a classification task, we will try to use ERNIE-Joint for this task.

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

Funding: This research was funded by the major special project of Anhui Science and Technology Department grant number 18030801133, and Science and Technology Service Network Initiative grant number KFJ-STS-ZDTP-079.

Acknowledgments: The authors would like to thank the AISTudio platform for providing computing resources.

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

Abbreviations

The following abbreviations are used in this manuscript:

| | |
|--------|---------------------------------------------------------|
| BERT | Bidirectional Encoder Representations from Transformers |
| BiGRU | Bidirectional Gate Recurrent Unit |
| BiLM | Bidirectional Language Model |
| BiLSTM | Bidirectional Long Short-Term Memory |
| CAN | Convolutional Attention Network |
| CNN | Convolutional Neural Networks |
| CRF | Conditional Random Field |
| ERNIE | Enhanced Representation through kNowledge IntEgration |
| ELMO | Embeddings from Language Models |
| KG | Knowledge Graph |
| MSRA | Microsoft Research Asia |
| NER | Named Entity Recognition |
| NLP | Natural Language Processing |
| RE | Relation Extraction |
| SOTA | State-of-the-art |

References

- Wen, Y.; Fan, C.; Chen, G.; Chen, X.; Chen, M. A Survey on Named Entity Recognition. In Proceedings of the International Conference in Communications, Signal Processing, and Systems, Urumqi, China, 20–22 July 2019; pp. 1803–1810.
- He, C.; Tan, Z.; Wang, H.; Zhang, C.; Hu, Y.; Ge, B. Open Domain Chinese Triples Hierarchical Extraction Method. *Appl. Sci.* **2020**, *10*, 4819. [[CrossRef](#)]
- Friedman, C.; Alderson, P.O.; Austin, J.H.; Cimino, J.J.; Johnson, S.B. A general natural-language text processor for clinical radiology. *J. Am. Med. Inform. Assoc.* **1994**, *1*, 161–174. [[CrossRef](#)]
- Gerner, M.; Nenadic, G.; Bergman, C.M. LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinform.* **2010**, *11*, 85. [[CrossRef](#)]
- Chen, A.; Peng, F.; Shan, R.; Sun, G. Chinese named entity recognition with conditional probabilistic models. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, 22–23 July 2006; pp. 173–176.
- Lyu, C.; Chen, B.; Ren, Y.; Ji, D. Long short-term memory RNN for biomedical named entity recognition. *BMC Bioinform.* **2017**, *18*, 462. [[CrossRef](#)] [[PubMed](#)]
- Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
- Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; Liu, Q. ERNIE: Enhanced language representation with informative entities. *arXiv* **2019**, arXiv:1905.07129.
- Zhang, H.P.; Liu, Q.; Yu, H.K.; Cheng, X.; Bai, S. Chinese named entity recognition using role model. *Int. J. Comput. Linguist. Chin. Lang. Process.* **2003**, *8*, 29–60.
- Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
- Zhang, Y.; Yang, J. Chinese ner using lattice lstm. *arXiv* **2018**, arXiv:1805.02023.
- Wei, H.; Gao, M.; Zhou, A.; Chen, F.; Qu, W.; Wang, C.; Lu, M. Named entity recognition from biomedical texts using a fusion attention-based BiLSTM-CRF. *IEEE Access* **2019**, *7*, 73627–73636. [[CrossRef](#)]
- Wu, G.; Tang, G.; Wang, Z.; Zhang, Z.; Wang, Z. An Attention-Based BiLSTM-CRF Model for Chinese Clinic Named Entity Recognition. *IEEE Access* **2019**, *7*, 113942–113949. [[CrossRef](#)]
- Yin, M.; Mou, C.; Xiong, K.; Ren, J. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism. *J. Biomed. Inform.* **2019**, *98*, 103289. [[CrossRef](#)] [[PubMed](#)]
- Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
- Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1188–1196.
- Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; 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; pp. 5998–6008.
- Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-trained models for natural language processing: A survey. *arXiv* **2020**, arXiv:2003.08271.
- Labusch, K.; Kulturbesitz, P.; Neudecker, C.; Zellhöfer, D. BERT for Named Entity Recognition in Contemporary and Historical German. In Proceedings of the 15th Conference on Natural Language Processing, Erlangen, Germany, 8–11 October 2019.
- Taher, E.; Hoseini, S.A.; Shamsfard, M. Beheshti-NER: Persian named entity recognition Using BERT. *arXiv* **2020**, arXiv:2003.08875.
- Hakala, K.; Pyysalo, S. Biomedical Named Entity Recognition with Multilingual BERT. In Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, Hong Kong, China, 4 November 2019; pp. 56–61.
- Levow, G.A. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, 22–23 July 2006; pp. 108–117.
- He, H.; Sun, X. F-Score Driven Max Margin Neural Network for Named Entity Recognition in Chinese Social Media. *arXiv* **2016**, arXiv:1611.04234.

25. Zhu, Y.; Wang, G.; Karlsson, B.F. CAN-NER: Convolutional attention network for Chinese named entity recognition. *arXiv* **2019**, arXiv:1904.02141.
26. Zhang, S.; Qin, Y.; Hou, W.J.; Wang, X. Word segmentation and named entity recognition for sighthan bakeoff3. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, 22–23 July 2006; pp. 158–161.
27. Zhou, J.; Qu, W.; Zhang, F. Chinese named entity recognition via joint identification and categorization. *Chin. J. Electron.* **2013**, *22*, 225–230.
28. Dong, C.; Zhang, J.; Zong, C.; Hattori, M.; Di, H. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In *Natural Language Understanding and Intelligent Applications*; Springer: Cham, Switzerland, 2016; pp. 239–250.
29. Yang, F.; Zhang, J.; Liu, G.; Zhou, J.; Zhou, C.; Sun, H. Five-stroke based CNN-BiRNN-CRF network for Chinese named entity recognition. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Hohhot, China, 26–30 August 2018; pp. 184–195.
30. Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Liu, S. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 182–192.
31. 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.
32. Peng, N.; Dredze, M. Improving named entity recognition for chinese social media with word segmentation representation learning. *arXiv* **2016**, arXiv:1603.00786.
33. He, H.; Sun, X. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.



© 2020 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 (<http://creativecommons.org/licenses/by/4.0/>).