symmetry

MDPI

Article

Low-Resource Named Entity Recognition via the Pre-Training Model

Siqi Chen [1,†,‡], Yijie Pei [1,†], Zunwang Ke [2,*] and Wushour Silamu [2]

[1] Xinjiang Multilingual Information Technology Laboratory, Xinjiang Multilingual Information Technology Research Center, School of Software, Xinjiang University, Urumqi 832001, China; bubble777@stu.xju.edu.cn (S.C.); poppy795@stu.xju.edu.cn (Y.P.)
[2] College of Information Science and Engineering, Xinjiang University, Urumqi 832001, China; wushour@xju.edu.cn
* Correspondence: kzwang@xju.edu.cn
† Yijie Pei and Siqi Chen are contributed equally to this research.
‡ Current address: Xinjiang University, Urumqi 830046, China.

Abstract: Named entity recognition (NER) is an important task in the processing of natural language, which needs to determine entity boundaries and classify them into pre-defined categories. For low-resource languages, most state-of-the-art systems require tens of thousands of annotated sentences to obtain high performance. However, there is minimal annotated data available about Uyghur and Hungarian (UH languages) NER tasks. There are also specificities in each task—differences in words and word order across languages make it a challenging problem. In this paper, we present an effective solution to providing a meaningful and easy-to-use feature extractor for named entity recognition tasks: fine-tuning the pre-trained language model. Therefore, we propose a fine-tuning method for a low-resource language model, which constructs a fine-tuning dataset through data augmentation; then the dataset of a high-resource language is added; and finally the cross-language pre-trained model is fine-tuned on this dataset. In addition, we propose an attention-based fine-tuning strategy that uses symmetry to better select relevant semantic and syntactic information from pre-trained language models and apply these symmetry features to name entity recognition tasks. We evaluated our approach on Uyghur and Hungarian datasets, which showed wonderful performance compared to some strong baselines. We close with an overview of the available resources for named entity recognition and some of the open research questions.

Keywords: named entity recognition; cross-lingual pre-trained language model; low-resource; attention-based fine-tuning; data augmentation

Citation: Chen, S.; Pei, Y.; Ke, Z.; Silamu, W. Low-Resource Named Entity Recognition via the Pre-Training Model. *Symmetry* **2021**, *13*, 786. https://doi.org/10.3390/sym13050786

Academic Editor: Jan Awrejcewicz

Received: 18 March 2021
Accepted: 27 April 2021
Published: 2 May 2021

1. Introduction

With the popularization and rapid development of information technology, natural language processing technology plays a key role in the processing, understanding and applications of text in the face of numerous unstructured text datasets generated on the Internet. Named entity recognition is one of the important basic research tasks, which plays an important role in computer automatic processing and the understanding of natural languages.

Developments in artificial intelligence, people's demand for automatic processing and artificial intelligence's understanding of a text are increasingly. Therefore, automatic extraction of semantic information [1] from natural language texts is becoming more and more important. As a piece of key semantic information in natural language, the named entity recognition task has gradually become an important basic research problem in natural language processing since it was first proposed in the Sixth Message Understanding Conference (MUC-6) in the 1990s. Much research has since been carried out on NER, using both knowledge engineering and machine learning approaches. Named entity recognition

can not only be used as an independent tool in the process of information extraction, but also plays an important role in various research fields of natural language text processing, such as automatic text summarization, automatic answering machines, machine translation, knowledge base construction and machine reading comprehension.

For the named entity recognition model [2] based on a deep neural network, good results have been achieved in many named entity recognition tasks, but its success depends heavily on large-scale training data with tags. For some widely used languages, the acquisition of manually tagged data may be relatively easy, so deep learning-based models tend to succeed in named entity recognition tasks in these languages. However, there are still numerous small languages in the world. Those languages with relatively scarce resources often have little or no tag data. How to realize named entity recognition for these languages has become an important issue in the research field. In existing studies, researchers have begun to improve low-resource word representations via knowledge transfer from high-resource languages using bilingual lexicons [3]. Bari [4] proposed an unsupervised cross-lingual NER model that can transfer NER knowledge from one language to another in a completely unsupervised way without relying on any bilingual dictionary or parallel data. Xie [5] proposed a method that finds translations based on bilingual word embeddings which use the unsupervised transfer of natural language processing models in named entity recognition. Ni and Jian [6] presented two weakly supervised approaches for cross-lingual NER with no human annotation of a target language. Rijhwani and Shruti [7] proposed a method of "soft gazetteers" that incorporates ubiquitously available information from English knowledge bases, such as Wikipedia, into neural named entity recognition models through cross-lingual entity linking.

Since the NER task was proposed by the research community, high-quality manual annotation training data have been created in some languages with numerous speakers (such as English). With these high-quality annotated data, a NER model can be established for low-resource languages with little or no annotated data based on the cross-language migration method. The cross-language NER task is a hot topic in NER research. The statistical methods, represented by conditional random field (CRF), are based on feature rules and semi-supervised learning. As there are a large number of concurrent words in Uyghur texts, there is ambiguity. There is a lack of a unified writing style for names in Uyghur, and some names are written in several ways. The adhesiveness of Uyghur leads to deviations. Words are created by joining affixes with stems. This creates a large number of unregistered people, institutions, and place names, leading to sparse data. Therefore, new ideas and methods are needed to further improve the accuracy of Uyghur named entity recognition. In our study, the pre-training model was used to share the vocabulary layer, and the data enhancement strategy was implemented on the dataset. Thus, the generalization ability of the model was improved effectively.

In this paper, as depicted in Figure 1, we propose a fine-tuning model for low-resource language models that is capable of addressing these issues: *LRLFiT*. First, we propose data augmentation by adding rich high-resource datasets and using the shared vocabulary characteristics of the pre-training model *XLM-R* [8] to fine-tune the cross-lingual pre-training model. Additionally, we confirm that the proposed method is symmetric.

We used *XLM-R* to pre-train a language model on a large cross-lingual corpus. Moreover, we introduce an attention-based fine-tuning strategy that selects relevant semantic and syntactic information from the pre-trained low-resource language model. To evaluate our model, we collected and annotated four corpora for the NER of a low-resource language. The experimental results show *LRLFiT* can significantly improve the performance with a few labeled examples.

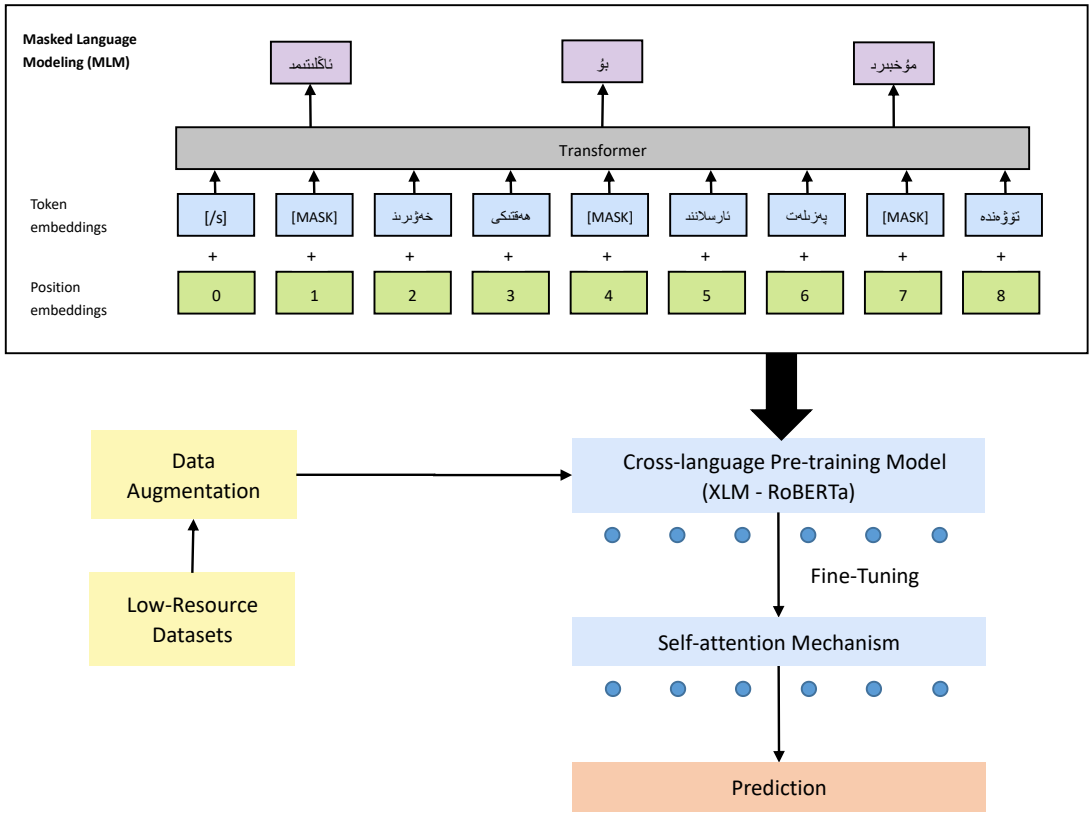


Figure 1. A fine-tuning model for low-resource language models—*LRLFiT*.

As a combined method, our contributions are as follows:

1. We constructed three low-resource agglutinative language datasets for named entity recognition, including Uyghur and Hungarian datasets. We also completed the annotation of Uyghur datasets with the help of Uyghur language experts. Each of the language datasets was divided for 70% training, 15% validation and 15% testing.
2. We propose a fine-tuning strategy for low-resource agglutinative languages—a data augmentation method with symmetry, based on the feature that a cross-language pre-training model shares a vocabulary layer. It is based on the pre-training model for low-resource languages, and there are no fine-tuning resources for low-resource Uyghur and Hungarian at present.
3. We propose a fine-tuning method that adopts a self-attention structure, and our pre-trained model learns the contextualized features from inputs of the dataset. In this task, this method does better to select relevant semantic and feature information from the pre-trained language model. This unique fine-tuning method effectively improves the performance of downstream tasks by using the features of adhesive language, such as context association.

2. Related Works

Named entity recognition is an important task in natural language processing and has a wide range of applications. We summarize the related works on three topics: (1) data augmentation; (2) cross-lingual pre-trained language models; (3) self-attention.

Data augmentation. Data augmentation is an effective method with which to expand the sizes of data samples. The larger the scale and the higher the quality, the better the generalization ability of the model. Data augmentation was created to solve the problem of insufficient data by using original documents to generate similar examples. Wei and Zou [9].

Present ed*EDA*, an easy data augmentation technique used to improve the performance of text classification tasks. For a given sentence in the training set, *EDA* randomly chooses and performs one of the following operations: synonym replacement, random insertion, random swap or random deletion. Random insertion and random exchange will change the grammatical structure of the original sentence, lose some labels and delete words randomly. If some core words are deleted, the data may deviate from the original label in the feature space. There are very few synonyms for words in Uyghur and Hungarian, so replacing synonyms does not add much data. The method of back translation [10] can often increase the diversity of text data. Compared with substitution words, it can sometimes change the syntactic structure and retain semantic information. However, the data generated by the back-translation method depend on the quality of the translation, and most of the translation results are not very accurate. The method of back translation can often increase the diversity of text data. Compared with substitution words, it can sometimes change the syntactic structure and retain semantic information. However, the data generated by the retranslation method depends on the quality of translation, and the existing Uighur and Hungarian translation results are inaccurate. The semi-supervised learning method [11] was proposed to make better use of unlabeled data and reduce the dependence on large-scale annotated datasets.

From the *EDA* results, we can see that the traditional data augmentation method has a certain effect, but it is mainly aimed at small amounts of data. For a deep learning model that craves numerous training data, the traditional method has always been limited. Unsupervised data augmentation (UDA) [12] was proposed to convert data from a high-resource language to a low-resource language, using a bilingual dictionary and an unsupervised machine translation model to expand the machine translation training set for the low-resource language. Dai [13] used data augmentation techniques for sentence-level and sentence-pair NLP tasks and adapted some of them for the NER task. Above all, the method of data enhancement can be used as a powerful tool to solve the problems of data imbalance and missing data quickly when we train the NLP model.

The Cross-lingual Pre-trained Language Model. Since the advent of ELMO [14] and *BERT* [15] models, people have paid more and more attention to large-scale pre-training language models in natural language processing. These models have achieved great success in NLP task types, such as text classification, machine translation, text abstract, question, and answer systems. These models are good for a single language. *Multilingual BERT* is pre-trained in the same way as monolingual *BERT*, but instead of being trained only on monolingual English data with an English vocabulary, it is trained on the Wikipedia pages of 104 languages with a shared word vocabulary. These 104 languages lack Uygur. For Uyghur in NER, the effect is poor. To speed up the application of natural language processing (NLP) to more languages, Facebook has open source enhanced *LASER* [16] libraries. *LASER* is the first library to use a single model to deal with a variety of languages (including low-resource languages such as Kabel and Uighur) to NLP zero-sample migration from one language (e.g., English) to several other languages, including languages with extremely limited training data. MultiFiT [17] extends *ULMFiT* [18] to make it more efficient and more suitable for language modeling beyond English: it utilizes tokenization based on subwords rather than words and employs a QRNN [19] rather than an *LSTM*. *XLM* [20] uses a known pre-processing technique (BPE) [21] and a dual-language training mechanism with *BERT* in order to learn relations between words in different languages. The *XLM-R* [8] uses filtered common-crawled data (over 2 TB) to demonstrate that using a large-scale multilingual pre-training model can significantly improve the performance of cross-language migration tasks.

Self-Attention. The attention mechanism imitates the human brain to pay attention to a certain key parts of things and distributes more attention. By calculating the probability distribution of attention, the most critical and important part is highlighted, thereby optimizing the traditional deep learning model. In deep learning, the attention mechanism is widely used in the field of computer vision [22]. Subsequently, Bahdanau [23] introduced

the attention mechanism into the machine translation task by referring to the application of the attention mechanism in the image classification task, making the attention mechanism become a hot topic in the field of natural language processing. With the advances of research, various improved attention mechanism models have achieved good results in text summary generation, text classification, syntactic analysis, sentiment classification, short text dialogue, and other tasks. A deep neural network can effectively learn important feature information from the text and effectively solve the problem of a lack of feature representation, thereby improving the accuracy of NLP tasks such as named entity recognition. Vaswani [24] has introduced the self-attention mechanism to machine translation for capturing global dependencies between input and output, and achieved state-in-the-art performance. Ming Gao [25] proposed an improved method for named entity recognition with an attention mechanism of ID-CNNs-based models. In a language comprehension task, [26] used self-attention to learn long-term dependencies. Tan and Zhixing [27] applied self-attention to the semantic role labeling task and obtained the latest research results.

3. Methodology

In this section, we will explain our methodology. Our training consists of four stages. We first augment the datasets of low-resource languages. Then the language model is pre-trained on a large-scale cross-language text corpus. In addition, the pre-trained model is fine-tuned on unsupervised language modeling tasks. Moreover, we propose a cross-language fine-tuning strategy, namely, the fine-tuning strategy of adding attention mechanism to build our NER model and using discriminative fine-tuning to capture different types of information on different layers.

3.1. Cross-Language Model Pre-Training

We utilized *XLM-R* to model the conditional probability. *XLM-R* uses the same shared vocabulary to process all languages through byte pair encoding (BPE) [21]. Shared contents include the same alphabet or the same anchor tokens, such as digits or proper nouns. This method of sharing dictionaries can greatly improves the alignment of across languages in the embedded space. We learn the BPE splits on the concatenation of sentences sampled randomly from the monolingual corpora. Sentences are sampled according to a multinomial distribution with probabilities. Additionally, sentences are sampled according to a probable multinomial distribution $\{q_i\}_{i=1,2,3...n}$, where:

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \tag{1}$$

when $p_i = \frac{n_i}{\sum_{k=1}^{N} n_k}$ and $\alpha = 0.5$. This distributed sampling method increases the number of tokens associated with low-resource languages and alleviates the bias toward high-resource languages. In particular, this method prevents words in low-resource languages from being split at the character level.

3.2. Data Augmentation

Data augmentation is an effective method with which to expand the sizes of data samples. The larger the scale and the higher the quality, the better the generalization ability of the model. There are several effects of using data enhancement in our model: increasing the amount of trained data, improving the generalization ability of the model, increasing the noise data, and improving the robustness of the model. Based on the cross-language characteristics of the *XLM-R* [8] model, we expanded the data to other languages in the training data, to improve the ability of the model. According to the characteristics of different languages, data enhancement was performed for low-resource agglutinative languages, including Uyghur and Hungarian, in which words are manufactured via stems and concatenated with several suffixes, because stems are used as the representations of content. These two languages have high uncertainty in their writing forms and many

redundant features; these features allow infinite derivatives in their vocabularies. Uyghur has the characteristic of stickiness, and Uyghur has many loan words from Persian. For the Hungarian language, its characteristics and the composition of the tone have some similarities with English. To sum up, we took advantage of the shared vocabulary layer of the pre-training model, and we added some Persian datasets to the Uyghur datasets, and some English named entity recognition datasets to the Hungarian named entity recognition datasets. The specific augmentation is shown in Section 4.1.

3.3. Fine-Tuning

In this section, we will describe the method of fine-tuning, to address a issue that the target domain differs substantially from the pretraining corpus, which adds an additional fine-tuning objective: masked language modeling in the target domain [28]. Briefly, we adopt a two-step approach:

Domain tuning. In the first step, we apply the *XLM-R* objective, which is to maximize the log-probability of randomly masked tokens, to fine-tune the contextualized word embeddings by back propagating, Then, we adopt this training procedure to an equal amount of unlabeled data in the source domain and a dataset that includes all available target domain data, we create ten random maskings of each instance; in each masking, following the original *XLM-R* training procedure o randomly mask 15% of tokens. We then perform three training iterations over this masked data.

Task-specific fine-tuning. Contextualized embeddings is very important for our work. Because they are powerful features for a wide range of downstream tasks. We fine-tune the contextualized word embeddings and learn the prediction model by backpropagating from the labeling objective on the source domain labeled data. The log probability can then be computed by the log softmax,

$$\log p(y_t \mid \mathbf{w}_{1:T}) = \beta_{y_t} \cdot \mathbf{x}_t - \log \sum_{y \in \mathcal{Y}} \exp(\beta_y \cdot \mathbf{x}_t) \tag{2}$$

where the contextualized embedding $\mathbf{x}_t$ captures information from the entire sequence $\mathbf{w}_{1:T} = (w_1, w_2, \ldots, w_T)$, and β_y is a vector of weights for each tag y. To fine-tune the contextualized word embeddings, the model is trained by minimizing the negative conditional log-likelihood of the labeled data.

3.4. Self-Attention

Here the self-attention mechanism proposed [29] in the literature is compared to the attention mechanism. First of all, as for the scaled dot-product attention mechanism, self-attention only completes the attention calculation within the sequence to find the internal connections within the sequence, making it essentially an attention mechanism that uses the dot product to perform similarity calculations. The point product attention can be calculated as follows:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K and V are all vector forms, which are three matrices obtained from the same input and calculations of different parameters. $Q \in \mathbb{R}^{n \times d_K}$, $\boldsymbol{K} \in R^{m \times d_K}$, $\boldsymbol{V} \in \mathbb{R}^{m \times d_V}$, d_K represent the second dimensions of Q and K. The attention layer is essential to $nn \times d_K$ sequence coding and has become a new sequence $n \times d_V$; $\sqrt{d_k}$ has the adjustment function, preventing the multiplication result from being too large***the inner product control Q; K is not too big, the use of the softmax operation will result to a probability distribution and get a result—V multiplied by the matrix—to ensure the softmax distribution. In addition, considering that an attention mechanism cannot capture important features from multiple perspectives and levels, and in order to capture text context features in multiple dimensions, it is necessary to use a multi-head attention mechanism. In the multi-head attention mechanism, Q, K and V are mapped through the parameter matrix and then

the dot product attention is performed. The process is repeated h times. Finally, the final result is obtained by a linear transformation, so as to obtain more comprehensive feature information.

$$\text{head}_i = \text{attention}\left(QW_i^Q, KW_i^k, VW_i^v\right), \tag{4}$$

$$\text{multi-head}\ (\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{concat}(\ \text{head}_1, \cdots, \ \text{head}_h) \tag{5}$$

Among them, $\boldsymbol{W}_{\mathrm{i}}^{Q} \in \mathbb{R}^{d_K \times \tilde{d}_K}$, $\boldsymbol{W}_{\mathrm{i}}^{k} \in \mathbb{R}^{d_K \times \tilde{d}_K}$, $\boldsymbol{W}_{\mathrm{i}}^{v} \in \mathbb{R}^{d_v \times \tilde{d}_v}$; concat means to concatenate the results each time.

4. Experiments

4.1. Datasets

For our experiments, we made use of two low-resource languages datasets—Uyghur and Hungarian ones, which contain three entities: people, places, and organizations. This Uygur dataset was taken from the Xinhua Uygur channel and Tianshan network, under the guidance of Uyghur experts, to complete the corpus tagging work. The Hungarian dataset was extracted from [30] corpus. Our data were divided into 70% training, 15% validation and 15% test sets. In addition, we used two kinds of high-resource datasets, namely, English [31] and Farsi Wikipedia [32], for augmentation of experimental data according to the characteristics of languages. The details of the datasets are shown in Tables 1–3. The expanded dataset is shown in Table 4.

Table 1. Uyghur entity recognition tagging corpus.

Corpus	Data Sets	LOC	ORG	PER
Train	8489	8400	7922	4298
Vertical	1819	1796	1232	977
Test	1820	1925	1644	818

Table 2. Hungarian entity recognition tagging corpus.

Corpus	Data Sets	LOC	ORG	PER
Train	6701	1344	14303	1050
Vertical	1436	256	3459	215
Test	1436	320	2671	236

Table 3. English and Farsi entity recognition tagging corpora.

Corpus	Data Sets	LOC	ORG	PER
Farsi	14559	18446	14764	17071
English	14041	8297	10027	11135

Table 4. The expanded datasets of the entity recognition tagging corpora.

Corpus	Data Sets	LOC	ORG	PER
Uyghur	23048	16697	17949	15433
Hungarian	21260	19790	29067	18121

4.2. Labeling Method and Model Evaluation Index

In this study, we used the "strict" standard to evaluate the model's results. For the NER task, "strict" means that a predicted entity is judged to be correct only if both the entity span and the entity type are correct [33]. The annotation methods of named entity identification include BIO, BIOE, BIOES, etc. This study adopted the BIO annotation method. B is the beginning of the entity, I is the non-beginning part of the entity and O is

the non-entity part. There are seven types of tags to be predicted, namely, I-PER, I-ORG, I-LOC, B-PER, B-ORG, B-LOC and O [34].

$$P = \frac{T_p}{T_p + F_p} \times 100\% \tag{6}$$

$$R = \frac{T_p}{T_p + F_n} \times 100\% \tag{7}$$

$$F1 = \frac{2PR}{P + R} \times 100\% \tag{8}$$

where T_p is the number of entities correctly identified by the model, F_p is the number of unrelated entities identified by the model and F_n is the number of related entities that the model does not recognize. The annotation of the dataset is shown in Figure 2.

Uyghur	Sentence	ناگلىتىمىز	خەۋىرىنى	ھەقتىكى	بۇ	نارسلانتىڭ	پەزىلەت	مۇخبىرىمىز	تۇۋندە		
	Label	0	0	B-PER	I-PER	0	0	0	0		
Hungary	Sentence	A	f	els	l	r	alakult	a	Friesland	Hung	Kft	.
	Label	0	0	0	0	0	0		B-ORG	I-ORG	I-ORG	0
English	Sentence	Arafat	to	meet	Peres	in	Gaza	after	flight	ban	.	
	Label	B-PER	0	0	B-PER	0	B-LOC	0	0	0	0	
Farsi	Sentence	خوانده است	و	بی سابقه	را	تعطل	این	رضایی	محسن	.		
	Label	B-PER	I-PER	0	0	0	0	0	0	0		

Figure 2. Examples of two dataset annotation methods in Uyghur and Hungarian.

4.3. Baseline Models

In recent years, most of the research on named entity recognition was based on deep learning. We tried to use some traditional machine learning methods for comparison, but they were not satisfactory, so for our next contrast model we used common neural networks or a cross-language pre-training model. As we are the first to use a cross-language pre-training model to perform NER in Uygur and Hungarian. We compared our *LRLFiT* model with several baseline models in different categories, including *CNN-LSTM*, *BILSTM*, *BiGRU* and XLM-R, each of which is described below.

CNN-LSTM: A *CNN* [35] and *LSTM* [36] combination model. The *CNN* can be trained from vocabulary characteristics' shape information, and optimizes the model's parameters by sharing parameters. *LSTM* can come to the conclusions according to the context.

BiLSTM: A long short-term memory network (LSTM) is proposed to solve the problem of long-term dependence of the cyclic neural network due to the extensive sentence processing and too much information. The forward *LSTM* is combined with the backward *LSTM* to form *BiLSTM* [37], and the output of *BiLSTM* is the joint action of two cyclic neural networks in opposite directions, which can predict the probability that each word belongs to different labels.

BiGRU: *GRU* (recurrent recurrent unit) is a variant of *LSTM* (long short-term memory) and an improved model of the recurrent neural network (RNN). As a variant of *LSTM*, *GRU* is also very suitable for processing sequence data and can remember the information of previous nodes through the "gate mechanism." In the *BiGRU* neural network, the context information is obtained from front to back and upward at the same time to improve the accuracy of feature extraction [38]. *BiGRU* has the advantages of small dependence on word vectors, low complexity and a fast response time.

mBERT: *Multilingual BERT* [15] is a transformer model pretrained on a large corpus of multilingual data in a self-supervised fashion, *mBERT* follows the same model architecture and training procedure as *BERT*, except that it is pre-trained on concatenated Wikipedia data of 104 languages. For tokenization, *mBERT* utilizes WordPiece embeddings [39] with a 110,000-word shared vocabulary to facilitate embedding space alignment across different languages. This means it was pretrained on the raw texts only, with no humans labeling

them in any way (which is why it can use lots of publicly available data), instead using an automatic process to generate inputs and labels from those texts.

XLM-R: *XLM-R* [8] shows how pretraining multilingual language models at scale leads to significant performance gains for a wide range of cross-lingual transfer tasks. It trains a transformer-based masked language model on one hundred languages, using more than two terabytes of filtered CommonCrawl data. *XLM-R* significantly outperforms *multilingual BERT (mBERT)* [15] in a variety of cross-lingual benchmarks.

4.4. Experiment Setting

To learn cross-language knowledge, we used the *XLM-R* model, which uses a *BERT* architecture [24] and pre-trains on 100 languages, with a hidden layer size of 768, 12 transformer blocks and 12 self-attention heads. We ran these typical models on 4 Tesla K80 GPUs and we used TensorFlow version 2.1 to build the experimental model. In our experiment, Adam [40] optimizer was used; the CUDA version was 11.2. Additionally, the training methods of baseline models are in Table 5. The other super parameters of *LRLFiT* model are shown in the Table 6.

Table 5. The training methods of the other models.

Hyperparameters	*CNN_LSTM*	*BiLSTM*	*BiGRU*	*mBERT*
max sequence length	128	128	128	128
max epochs	10	10	10	10
batch size	16	16	16	32
learning rate	5e-5	5e-5	5e-5	5e-5
gradient accumulation steps	1	1	1	1
warmup proportion	0.0	0.0	0.0	0.1
activation	softmax	softmax	softmax	-

Table 6. The hyperparameters of *LRLFiT*.

Hyperparameters	Uyghur	Hungarian
max sequence length	128	128
max epochs	10	10
batch size	8	16
learning rate	5e-5	6e-5
gradient accumulation steps	1	1
warmup proportion	0.0	0.0
dropout	0.2	0.2

4.5. Results and Analysis

We combined two low-resource datasets to analyze the experimental results in terms of accuracy, recall rate and F1. Combined with the characteristics of the cross-language pre-training model and the experimental performance, we got the evidence that it is better than the previous model. Thus, we present the results of *XLM-R* on the named entity recognition task. Finally, we compare multilingual and monolingual models and present results on low-resource languages.

From Tables 7 and 8, we can see that the accuracy, recall and F1 of our model on the Uyghur and Hungarian datasets were the best, and the F1 scores of the method in this paper reached 95.00% and 96% for the two languages. First, the F1 of *BiLSTM-CRF* was higher than that of *CNN-CRF*. It can be seen that the bi-directional structure of *BiLSTM* has a stronger ability to acquire context sequence features than the one-way structure. The recognition effect of the *BiGRU* model is better than those of the *CNN-LSTM* model and the *BiLSTM* model. The reason is that the average lengths of labeled words in these datasets were generally a little longer than normal, which was conducive to the *BiGRU* network capturing long-distance information in the training process, extracting the entity

features more accurately, obtaining effective information in the entity, and improving the recognition accuracy.

Table 7. Uyghur: named entity recognition results of different algorithm models.

Model	P(%)	R(%)	F1(%)
CNN_LSTM	68.12	64.32	66.51
BiLSTM	66.22	62.13	64.23
BiGRU	68.25	60.66	67.34
mBERT	69.56	62.52	65.58
Xlmr	71.25	72.36	70.42
LRLFiT	**73.69**	**77.54**	**75.35**

Table 8. Hungarian: named entity recognition results of different algorithm models.

Model	P(%)	R(%)	F1(%)
CNN_LSTM	91.89	86.57	89.15
BiLSTM	91.89	84.56	87.25
BiGRU	90.65	86.15	88.52
mBERT	90.87	87.25	88.93
Xlmr	91.46	90.69	90.56
LRLFiT	**92.88**	**91.61**	**92.45**

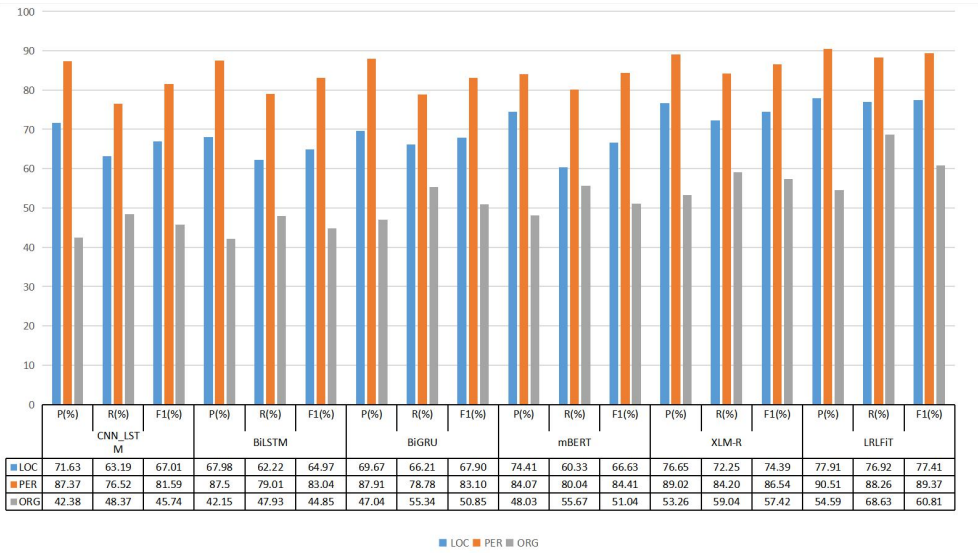
Among the cross-language pre-training models used in the NER field that performed best for Uyghur and Hungarian data, the latest is the Facebook team's *XLM-R*. We provide *Multilingual BERT* training as a reference, and the results are not satisfactory. On the basis of *Multilingual BERT*, a new task—the translation language model—has been added. Therefore, an experimental comparison with *Multilingual BERT* was added to Table 6. Finally, our LRLFiT model outperformed both *XLM-R* and *Multilingual BERT* in every setting. All of those issues point to the importance of carefully collecting corpora to generate pre-trained language models for each language, especially for language models with fewer resources. These models are often under-represented in large multilingual models.

In order to solve the problem of low-resource data with fewer annotations, a solution based on a cross-language model was proposed to establish a reliable model and make accurate predictions according to the existing knowledge. The model works as a neural network with self-attention mechanism characteristics.

As can be seen in Figure 3, through the comparative analysis of PER, ORG and LOC, the F1 score of *LRLFiT* reached 0.7535 for Uyghur language. Then it identified the entity types PER, ORG and LOC, and the F1 scores reached 0.8029, 0.9458 and 0.8847, respectively.

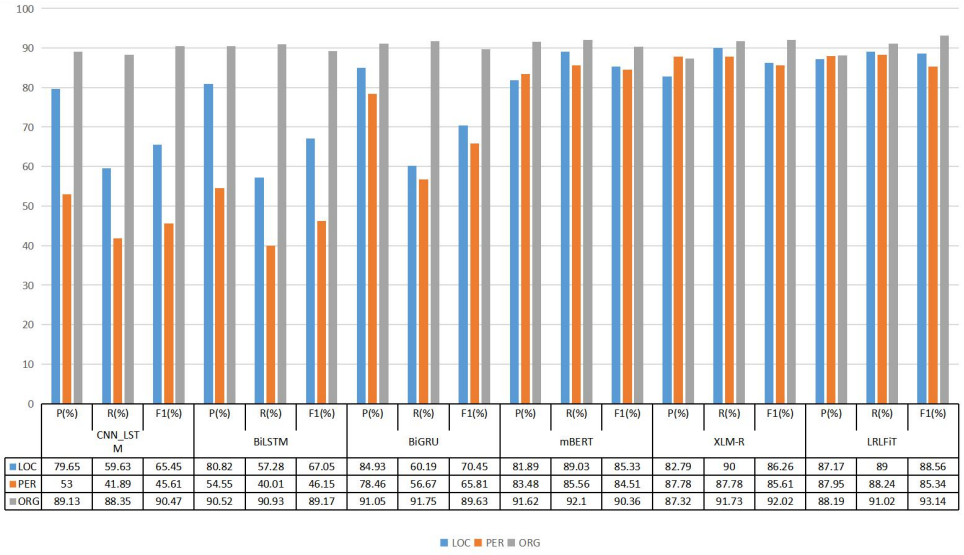
As can be seen in Figure 4, the F1 score of *LRLFiT*, reached 0.9288 in Hungarian. It identified the PER, ORG and LOC entity types, with the F1 scores being 0.8029, 0.9458 and 0.8847, respectively.

Taken together, these results suggest that unlike a deep neural network model, while using the selected *XLM-R* training models of the application for the characteristic of migration, first and with almost unlimited text, it studies the context of every input sentence using implicitly learned universal grammar; second, it can be taught from open fields of knowledge transferred downstream of the named entity recognition task, to improve the low-resource tasks. Its language processing for low-resource languages is also very good. Additionally, the pre-training model with fine-tuning mechanism has good scalability, We can see that the pre-training model has achieved the best results in the downstream tasks, and the F1 value has improved a lot.



	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
	CNN_LSTM			BiLSTM			BiGRU			mBERT			XLM-R			LRLFiT		
LOC	71.63	63.19	67.01	67.98	62.22	64.97	69.67	66.21	67.90	74.41	60.33	66.63	76.65	72.25	74.39	77.91	76.92	77.41
PER	87.37	76.52	81.59	87.5	79.01	83.04	87.91	78.78	83.10	84.07	80.04	84.41	89.02	84.20	86.54	90.51	88.26	89.37
ORG	42.38	48.37	45.74	42.15	47.93	44.85	47.04	55.34	50.85	48.03	55.67	51.04	53.26	59.04	57.42	54.59	68.63	60.81

Figure 3. The performance comparison of identifying PER, ORG and LOC entity types in Uyghur on *LRLFiT*.



	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
	CNN_LSTM			BiLSTM			BiGRU			mBERT			XLM-R			LRLFiT		
LOC	79.65	59.63	65.45	80.82	57.28	67.05	84.93	60.19	70.45	81.89	89.03	85.33	82.79	90	86.26	87.17	89	88.56
PER	53	41.89	45.61	54.55	40.01	46.15	78.46	56.67	65.81	83.48	85.56	84.51	87.78	87.78	85.61	87.95	88.24	85.34
ORG	89.13	88.35	90.47	90.52	90.93	89.17	91.05	91.75	89.63	91.62	92.1	90.36	87.32	91.73	92.02	88.19	91.02	93.14

Figure 4. The performance comparison of identifying PER, ORG and LOC entity types in Hungarian on *LRLFiT*.

To thoroughly verify the effect of different *LRLFIT* versions on entity recognition effect, we used two versions (base, large) for comparison, as shown in Figure 5—which shows the scores of the different algorithms in Uyghur and Hungarian. In the Uyghur language, the results showed that the overall effect became better and better with the increases in the size and the number of parameters of the model pre-training corpus. The results showed that *XLM-Roberta-Large* had the best effect. The P, R and F1 scores were 0.7431, 0.7897 and 0.7709, respectively. The results were 0.62% percent, 1.43% and 1.74% higher than the best results for for *Multilingual BERT*.

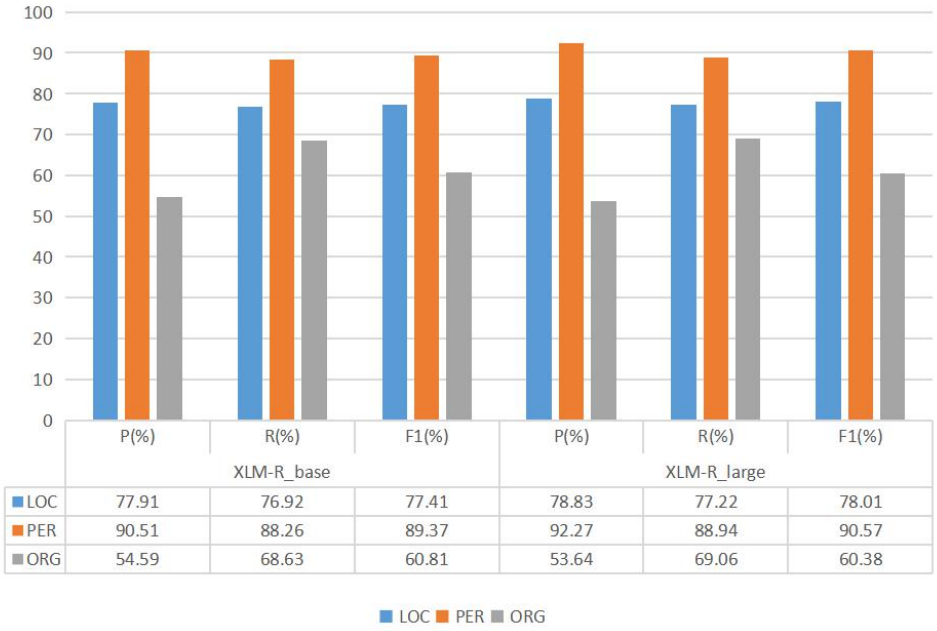


Figure 5. The performance comparison of identifying PER, ORG and LOC entity types in Uyghur on *LRLFiT*.

In the Hungarian language, as shown in Figure 6, the results showed that the overall effect became better and better with the increases in the size and number of parameters of the model pre-training corpus. The results showed that *XLM-Roberta-Large* had the best effect, and the P, R and F1 scores were 0.9389, 0.9269 and 0.9382, respectively. The results were 1.01%, 1.08% and 1.37% higher than the best results for *Multilingual BERT*.

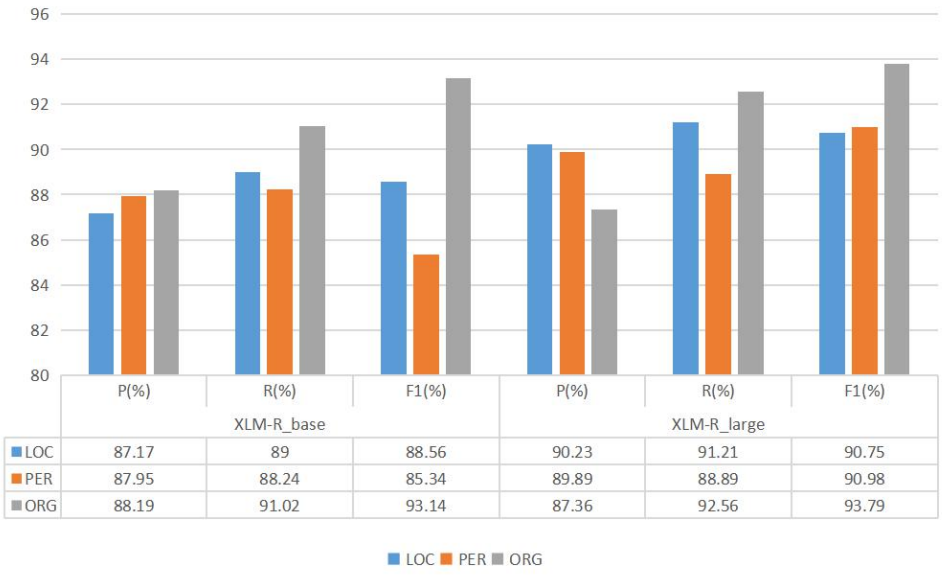


Figure 6. The performance comparison of identifying PER, ORG and LOC entity types in Hungarian on *LRLFiT*.

4.6. Ablation Study

To evaluate the contributions of key factors in our method, we used the *LRLFiT* model for a comparison of ablation experiments for training and testing on datasets of the Uyghur and Hungarian languages. The methods with and without data augmentation, and with and without attention mechanisms were trained. As shown in Figure 7.

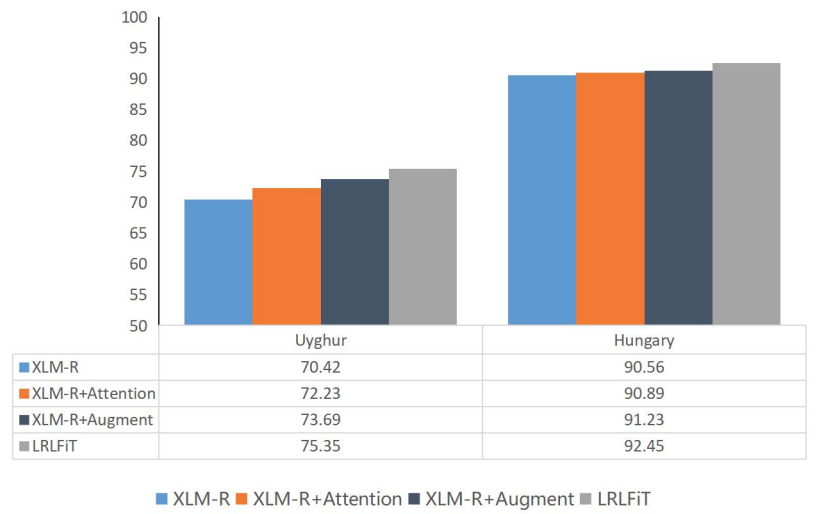


Figure 7. Two ablation studies of the Uyghur and Hungarian languages using *LRLFiT*.

The effect of augmentation. We compare the F1 values between the constructed dataset and the dataset without data augmentation. The experimental results show that the data augmentation method greatly improves the performance of our task, and can effectively take advantage of the feature of vocabulary sharing in the cross-language pre-training model, enabling us to obtain more effective data.

The effect of self-attention. We can observe that our model improves the F1 value by adding a fine-tuning of attention. After adding self-attention to the named entity recognition task, the F1 value was significantly improved. The self-attention mechanism can capture contextual information from multiple different subspaces to better understand the sentence structure, so that entities can be correctly identified without the introduction of the self-attention mechanism.

5. Conclusions and Future Work

Low-resource NER is a very important yet challenging problem in language processing. In this paper, a cross-lingual learning approach for NER was proposed, focusing on dissimilar languages—Uyghur and Hungarian. The proposition was *LRLFiT*, an effective fine-tuning method for language models that can be applied to low-resource language NER tasks. Moreover, we can choose an attention-based fine-tuning strategy that better selects the context information from the pre-trained language model to provide a meaningful and favorable feature for NER tasks. Specifically, we used a multilingual vocabulary to solve the problem of few resources, performed data enhancements for different languages and then modeled with cross-language pre-training models. Experiments on two languages, namely, Uyghur and Hungarian, demonstrated the effectiveness of our model for low-resource language NER.

In the future, the proposed approach could be applied to other low-resource (ethnic minority) languages, such as Kazakh and Kyrgyz. We will also extend our architecture to other NLP tasks, such as event extraction and sentiment analysis. We hope that our results will catalyze new developments in low-resource agglutinative languages tasks for NLP.

Author Contributions: Datesats, Z.K. and W.S.; Methodology, S.C. and Y.P.; Writing—original draft, S.C. and Y.P.; Writing—review and editing, S.C. and Y.P.; data curation, S.C.; Supervision, Z.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China under grant 2017YFC0820700; in part by the National Language Commission Research Project under grant ZDI135-96; and in part by the China Academy of Electronics and Information Technology, National Engineering Laboratory for Risk Perception and Prevention (NEL-RPP).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are very thankful to the editor and referees for their valuable comments and suggestions for improving the paper.

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

References

1. Bharadwaj, A.; Mortensen, D.; Dyer, C.; Carbonell, J. Phonologically Aware Neural Model for Named Entity Recognition in Low Resource Transfer Settings. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 1462–1472.
2. Tsai, C.T.; Mayhew, S.; Roth, D. Cross-Lingual Named Entity Recognition via Wikification. In Proceedings of the CoNLL, Berlin, Germany, 1 January 2016; pp. 219–228.
3. Feng, X.; Feng, X.; Qin, B.; Feng, Z.; Liu, T. Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge Transfer. *IJCAI* **2018**, *1*, 4071–4077.
4. Bari, M.S.; Joty, S.; Jwalapuram, P. Zero-resource cross-lingual named entity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 37, pp. 7415–7423.
5. Xie, J.; Yang, Z.; Neubig, G.; Smith, N.A.; Carbonell, J. Neural cross-lingual named entity recognition with minimal resources. *arXiv* **2018**, arXiv:1808.09861.
6. Ni, J.; Dinu, G.; Florian, R. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. *arXiv* **2017**, arXiv:1707.02483.
7. Rijhwani, S.; Zhou, S.; Neubig, G.; Carbonell, J. Soft Gazetteers for Low-Resource Named Entity Recognition. *arXiv* **2020**, arXiv:2005.01866.
8. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv* **2019**, arXiv:1911.02116.
9. Wei, J.; Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv* **2019**, arXiv:1901.11196.
10. Shleifer, S. Low resource text classification with ulmfit and backtranslation. *arXiv* **2019**, arXiv:1903.09244.
11. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C. Mixmatch: A holistic approach to semi-supervised learning. *arXiv* **2019**, arXiv:1905.02249.
12. Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.T.; Le, Q.V. Unsupervised data augmentation for consistency training. *arXiv* **2019**, arXiv:1904.12848.
13. Dai, X.; Adel, H. An analysis of simple data augmentation for named entity recognition. *arXiv* **2020**, arXiv:2010.11683.
14. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
15. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
16. Artetxe, M.; Schwenk, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 597–610. [CrossRef]
17. Eisenschlos, J.M.; Ruder, S.; Czapla, P.; Kardas, M.; Gugger, S.; Howard, J. MultiFiT: Efficient multi-lingual language model fine-tuning. *arXiv* **2019**, arXiv:1909.04761.
18. Howard, J.; Ruder, S. Universal language model fine-tuning for text classification. *arXiv* **2018**, arXiv:1801.06146.
19. Bradbury, J.; Merity, S.; Xiong, C.; Socher, R. Quasi-recurrent neural networks. *arXiv* **2018**, arXiv:1611.01576.
20. Lample, G.; Conneau, A. Cross-lingual language model pretraining. *arXiv* **2019**, arXiv:1901.07291.
21. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv* **2015**, arXiv:1508.07909.
22. Zhao, B.; Wu, X.; Feng, J.; Peng, Q.; Yan, S. Diversified visual attention networks for fine-grained object classification. *IEEE Trans. Multimed.* **2017**, *19*, 1245–1256. [CrossRef]
23. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
25. Gao, M.; Xiao, Q.; Wu, S.; Deng, K. An Improved Method for Named Entity Recognition and Its Application to CEMR. *Future Internet* **2019**, *11*, 185. [CrossRef]
26. Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; Zhang, C. Disan: Directional self-attention network for rnn/cnn-free language understanding. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 56.
27. Tan, Z.; Wang, M.; Xie, J.; Chen, Y.; Shi, X. Deep semantic role labeling with self-attention. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 33.

28. Han, X.; Eisenstein, J. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. *arXiv* **2019**, arXiv:1904.02817.
29. Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; Liu, Q. ERNIE: Enhanced language representation with informative entities. *arXiv* **2019**, arXiv:1905.07129.
30. Szarvas, G.; Farkas, R.; Felföldi, L.; Kocsor, A.; Csirik, J. A highly accurate Named Entity corpus for Hungarian. In Proceedings of the LREC, Genoa, Italy, 22–28 May 2006; pp. 1957–1960.
31. Sang, E.F.; De Meulder, F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv* **2003**, arXiv:cs/0306050.
32. Schwenk, H.; Chaudhary, V.; Sun, S.; Gong, H.; Guzmán, F. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv* **2019**, arXiv:1907.05791.
33. Yao, L.; Huang, H.; Wang, K.W.; Chen, S.H.; Xiong, Q. Fine-Grained Mechanical Chinese Named Entity Recognition Based on ALBERT-AttBiLSTM-CRF and Transfer Learning. *Symmetry* **2020**, *12*, 1986. [CrossRef]
34. Sheng, J.; Wumaier, A.; Li, Z. POISE: Efficient Cross-Domain Chinese Named Entity Recognization via Transfer Learning. *Symmetry* **2020**, *12*, 1673. [CrossRef]
35. Ma, X.; Hovy, E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv* **2016**, arXiv:1603.01354.
36. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neur. Netw.* **2005**, *18*, 602–610. [CrossRef] [PubMed]
37. Nguyen, T.H.; Grishman, R. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 365–371.
38. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259.
39. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
40. Da, K. A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.