symmetry

MDPI

Article

POISE: Efficient Cross-Domain Chinese Named Entity Recognization via Transfer Learning

Jiabao Sheng [1,†], Aishan Wumaier [2,*,†] and Zhe Li [2,*,†]

1 College of Software, Xinjiang University, Urumqi 832001, China; jiabao@stu.xju.edu.cn
2 College of Information Science and Engineering, Xinjiang University, Urumqi 832001, China
* Correspondence: hasan1479@xju.edu.cn (A.W.); lizhe@stu.xju.edu.cn (Z.L.)
† Jiabao Sheng, Aishan Wumaier, and Zhe Li contributed equally to this work.

Received: 9 September 2020; Accepted: 8 October 2020; Published: 13 October 2020

Abstract: To improve the performance of deep learning methods in case of a lack of labeled data for entity annotation in entity recognition tasks, this study proposes transfer learning schemes that combine the character to be the word to convert low-resource data symmetry into high-resource data. We combine character embedding, word embedding, and the embedding of the label features using high- and low-resource data based on the BiLSTM-CRF model, and perform the feature-transfer and parameter-sharing tasks in two domains of the BiLSTM network to annotate with zero resources. Before transfer learning, we must first calculate the label similarity between two different domains and select the label features with large similarity for feature transfer mapping. All training parameters of the source domain in the model are shared during the BiLSTM network processing and CRF layer. In addition, we also use the method of combining characters and words to reduce the problem of word segmentation across domains and reduce the error rate in label mapping. The results of experiments show that in terms of the overall F1 score, the proposed model without supervision was superior by 9.76 percentage points to the general parametric shared transfer learning method, and by 9.08 and 12.38 percentage points, respectively, to two recent high–low resource learning methods. The proposed scheme improves performance in terms of transfer learning between the high- and low-resource data and can identify the predicted data in the target domain.

Keywords: named entity recognition; neural networks; cross-domain; parameter sharing; few resources

1. Introduction

Named entity recognition (NER) is used to recognize named entities such as persons, places, and organizations. NER is an important task in research on the use of natural language processing (NLP). Due to different text types and named entities, a key challenge for an NER task is to transfer knowledge features from one domain to another, which is often referred to as transfer learning [1]. Transfer learning can be used in several settings, notably for such low-resource languages [2] and low-resource domains as biomedical corpora [3] and tourism corpora [4].

Transfer learning can improve task performance by taking advantage of features from similar labels in these cases. Even on datasets with low-resource labels, it can sometimes yield an improvement over state-of-the-art results [5–7].

Some researchers have lately proposed such methods, like feature representation transfer [8,9], instance-based transfer learning [10,11], and parameter sharing [12,13]. Cross-resource word embedding [14] has popularized knowledge transfer from high- to low-resource datasets. Ref. [15] overcomes the scarcity of data by training a model on a high-resource dataset and transferring the knowledge features to a low-resource setting in another language. As the data might be gleaned

from different domains, Ref. [16] extended two state-of-the-art transfer learning models as analytical vehicles for multi-label sentiment analysis tasks.

However, the above methods cannot flexibly utilize the features and parameters of different resources when transferring the learning, which leads to a high probability of the occurrence of identification errors. From the perspective of word segmentation, words belonging to an across domain cannot be segmented and recognized well, which degrades recognition.

To solve the above challenges, we examine the settings for transfer learning to improve performance on the target task through training on the source task. we also consider the method that combines characters and words, which can reduce the problem of unclear boundaries in cross-domain word segmentation in Chinese.

As shown in Figure 1, the scheme includes the following steps:

1. Determine the source domain dataset and the target domain dataset. Use the similarity of feature distribution (SFD) and Inter-feature correlation (IFC) to select the label with the highest similarity for feature transfer.

2. Use BiLSTM neural network model to learn and transfer features.

3. In the decoding stage, the characters are first combined with words to search for the maximum path of words in datasets in different domains, and the most complete words are matched so as to map labels.

4. CRF layer is the final prediction layer. Note: All parameters are shared throughout the process.

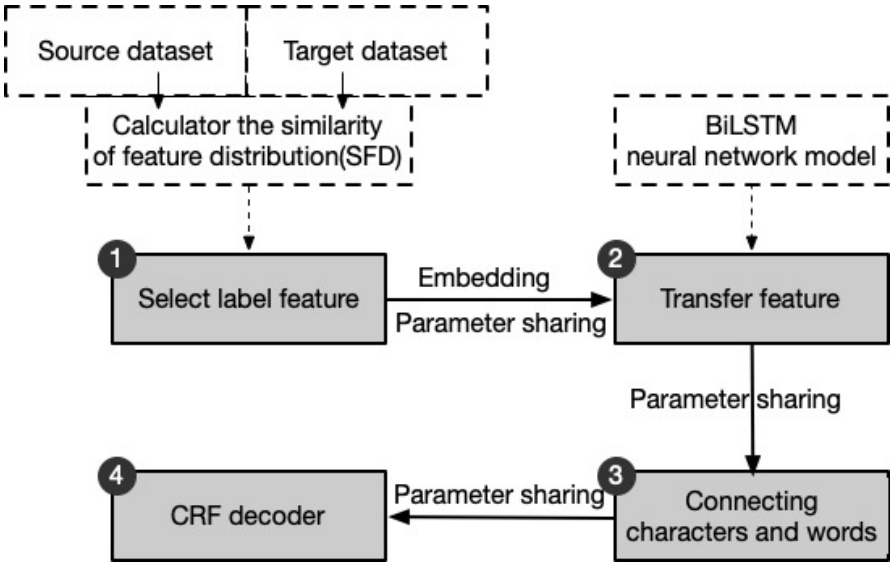

Figure 1. High-level illustration of POISE.

In view of the success of transfer learning, we use it here to solve the problem of cross-domain Chinese entity recognition. The contributions of this study are as follows:

- We develop an symmetry architectures for transfer learning (as shown in Figure 2), respectively. The architecture is extensions of a base model with parameter-sharing and feature transfer, which can hence obtain high-resource features to assist low-resource entity recognition tasks.
- We make full use of words and their order to mine potential lexical information and combine this with character-based tags to match entity vocabularies in different domains according to the path of the processing module. This yields a resolved path to mitigate the degradation in performance caused by unclear boundaries during cross-domain word segmentation in Chinese.
- We build an NER corpus for Chinese tourism, called Zoetral, that collects 8460 posts related to tourism, 5813 sentences related to tourist attractions, and 5377 integrated entities. By using the transfer learning framework, we successfully and automatically identified 2432 tourism entities as extended entities of the Zoetral dataset.

2. Related Work

In recent years, the neural network has been widely applied to research on named entity recognition and yielded impressive results. We briefly review two related directions: named entity recognition and transfer learning.

Named Entity Recognization NER is a popular subject of research in NLP [17]. In the past few years, a lot of related research findings have been reported. Deep learning has recently emerged as the mainstream approach in NER. Methods based on neural networks deliver good performance in terms of identifying entities in English.

Ref. [18] described and evaluated a character-level tagger for language-independent NER. The model consists of stacked bidirectional long short-term memory networks (LSTMs) that take as inputs characters and output tag probabilities for each. Ref. [19] proposed a BiLSTM-supported structure that combines the CRF [20] for NER tasks called the BiLSTM-CRF model. In experiments, this model achieved better performance on NER tasks than the rich features-based CRF model. In [21], for requiring without feature engineering or data pre-processing, it can apply to a wide range of sequence labeling tasks. They proposed a model based on the LSTM-CNN-CRF. The results of experiments showed that their methods outperformed their original counterparts. Ref. [22,23] achieved state-of-the-art performance on NER by using LSTM-CRF models, with characters being integrated into word representations. With the development of Chinese NER task, on the basis of using the neural network model, researchers have also made good achievements in this task [24].

As far as we know, the size of annotated data, which normally limited the performance in the above approaches. In usual cases, it only has the best recognition results when the available annotated data are adequate. Nevertheless, obtaining adequate annotated data is hard for researchers. In recent studies, it has been proved that the knowledge learned from one domain dataset can be applied to another domain dataset in training tasks. The researchers implement this idea by using a technique called transfer learning [1]. In traditional supervised machine learning, an adequate amount of annotated data are required to train an excellent model. On the contrary, these requirements can be relaxed in transfer learning.

Transfer Learning The above idea has been used by researchers in many domains, such as speech recognition [25], text classification [26], and machine translation [27]. It also has a rapid development in named entity recognition [28,29]. In the cross-language NER tasks [30] or cross-domain dataset NER tasks [31], the method based on transfer learning has been well proved to be usable.

Using transfer learning, the label features and parameters trained in the source domain can be directly applied to the target domain. This decreases the number of parameters as well as the model size. Therefore, this method was soon introduced into NER tasks in the domain of Chinese low-resource datasets. Many researchers have also performed entity recognition tasks based on transfer learning [32,33] in their work of previous for low-resource entity datasets in Chinese NER task.

However, due to the huge difference between different domain datasets, they do not resolve the problem in segmenting the words nor predict a word label well. Although the work of [33] provides a network to reduce the difference between different datasets, and [32] integrated the language prediction model and algorithm, these do not work in the limited domain. Besides, these models also need high-quality datasets to train between source and target datasets, which can obtain a higher score.

This paper takes advantage of transfer learning to accomplish a Chinese NER task, which can effectively solve the problems of the Chinese NER task and transfer features between several different domain datasets. In the first step, we assume a large amount of training data in a news domain but no data, or a small amount of data, in a target domain. We then transfer NER knowledge from the news domain (here we set news domain as the source domain, and it can be changed to other datasets) to the target domain to improve the training performance of low-resource datasets through cross-domain NNM training.

Comparing with other NER model based transfer learning, we do so using a novel scheme for transfer learning based on a deep hierarchical recurrent neural network that selectively shares hidden

feature representations and part of the parameters between the source domain and the target domain tasks, rather than transfer all of the features in source domain datasets. Furthermore, it is not enough that we only transfer the features and parameters. Considering the effective affection of combining the characters and words in segmenting the words, we introduce this method into our scheme. Combining characters and words can improve the performance in cross-domain entity recognization in the Chinese ENR task. We use gradient-based methods for efficient training. The results of experiments show that our NER model guarantees good performance even when the set of training data are small.

3. Proposed Model

3.1. Model Architecture

We propose a cross-domain named entity recognition frame called POISE. The frame contains label feature embedding, character embedding, word embedding, a BiLSTM module for feature learning and representation, a unit that connecting characters to be a word, and a CRF prediction layer.

The goal of this work is to transfer knowledge and feature from the source to the target domain, which can implement the cross-domain transfer. At first, we assume that there are few labels in the target domain dataset. Next, we have to select the height similarity labels by using the statistic method. Cause there are usually two cases of cross-domain transfer. The two domains have labels that can be mapped to each other or can have disparate label sets. For example, tags in corpus A (e.g., "LOC", "ORG") can be mapped to those in corpus B (e.g., "LOC", "ORG"), whereas some other tags in corpus A (e.g., URL, "WEB") cannot be mapped to those in corpus B (e.g., "LOC", "ORG"). When the two domains have labels that can be mapped to each other, all the model parameters and feature representations are shared in the neural network. We execute a label mapping step on top of the CRF layer. Finally, the two domains have some disparate label sets, the same label dataset can be found from other domain datasets for training. The frame is shown in Figure 2.

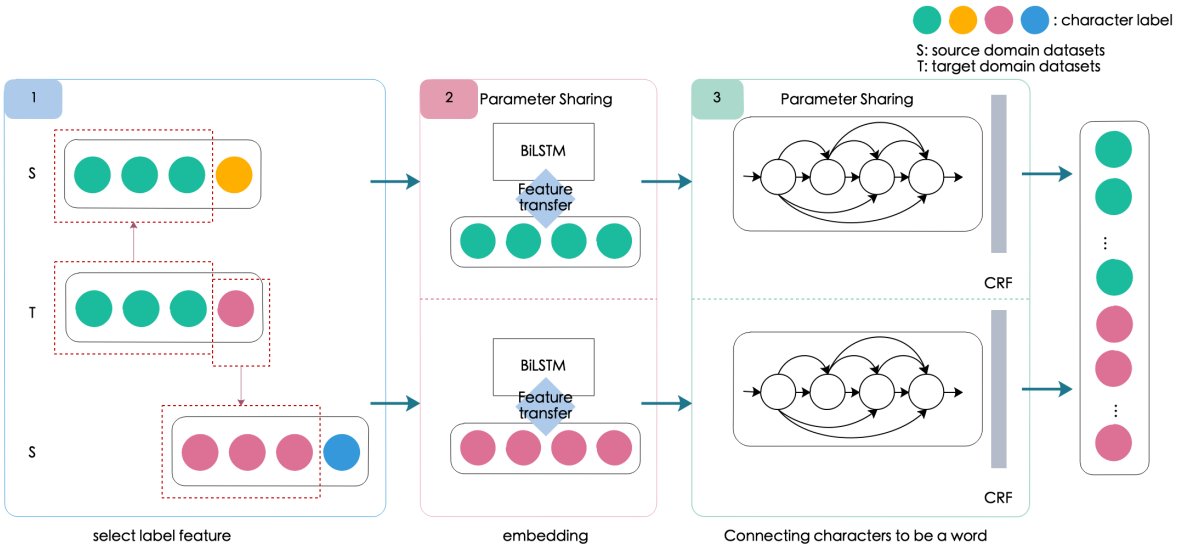

Figure 2. Each two fields have an appropriate set of tags. All parameters and label feature representations in the model are thus shared to complete the mapping from high- to low-resource data.

3.2. Training

In the source and the target tasks, W_s and W_t denote the model parameters, they are shared parameters and task-specific parameters respectively. W_shared are jointly optimized by the two tasks, whereas $W_source, spec$ and $W_target, spec$ are individual trained for each task separately. In the experiment, because we selected the label dataset first, the whole process used shared parameters for

training. If there is no label dataset with high similarity, then the task will directly train the target domain dataset with task-specific parameters.

3.2.1. Input

A set of inputs $C = [c_1, c_2, c_3 ... c_j]$ and $W = [W_1, W_2, W_3 ... W_j]$ are given from the NER training set of the source $S_n er = (x_i, y_i)_{i=1}^{m}$ or the target domain $\tau_n er = (x_i, y_i)_{i=1}^{n}$. The original text set $S_{lm} = (x_i)_{i=1}^{P}$ of a source field or the original text set $\tau_n er = (x_i)_{i=1}^{q}$ of a target domain, the text is automatically divided from the source domain and the target domain, and the trained word embedding is constructed as $\mathbb{D}$.

Before feature transfer, in order to migrate effective feature information, we merge the data of the source domain and the target domain, and then cluster the merged data set based on features, thus gathering highly correlated features into the same cluster. Then, the similarity of the distribution of each feature between the source domain data and the target domain data is calculated as the sorting basis, and the features in each cluster are sorted in descending order. Finally, the top-ranked features are selected from each cluster as the features that need to be migrated finally (as shown in Figure 3).

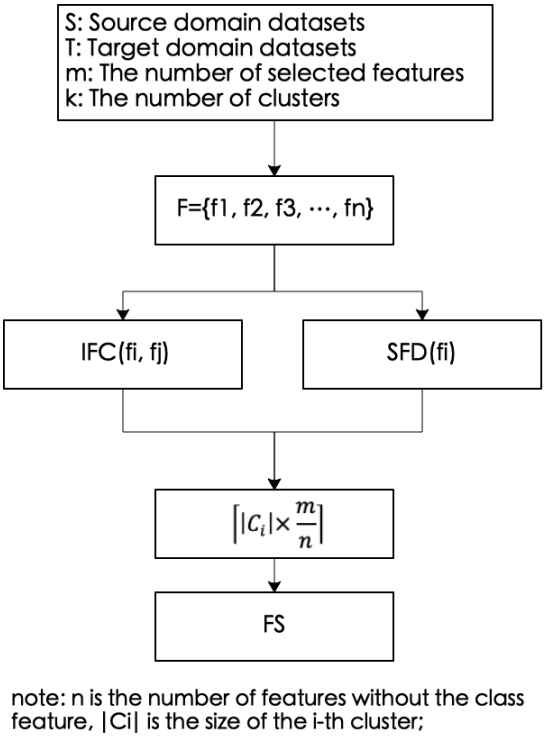

Figure 3. Feature selection.

Inter-feature correlation (IFC) is used to measure the correlation between two features f_i and f_j. In this paper, symmetric uncertainty (SU) is used to calculate the correlation between the two features in feature selection.

The similarity of feature distribution (SFD) is used to measure the similarity of the distribution of a specific feature f_i on two different data sets. We use Kolmogorov–Smirnov (K-S) test to verify whether the distribution of the two sets of data is similar.

In order to verify the effectiveness of the experiment, we selected three datasets in different domains and selected three entity label types for training and testing.

3.2.2. Character-Based BiLSTM Model

Long short-term memory networks (LSTMs) can capture long-range dependencies. They do so by using several gates that control the ratio of the input to assign to the memory cell, and by setting the ratio from the previous state to forget. We use the following implementation:

$$i_t = \sigma(W_{x_i} X_t + W_{h_i} h_{t-1} + W_{c_i} C_{t-1} + b_i) \quad (1)$$

$$c_t = (1 - i_t) \odot c_{t-1} + i_t \odot \tanh(W_{x_c} x_t + W_{h_c} h_{t-1} + b_c) \quad (2)$$

$$o_t = \sigma(W_{x_o} X_t + W_{h_o} h_{t-1} + W_{c_o} c_t + b_o) \quad (3)$$

$$h_t = o_t \odot \tanh(c_t) \quad (4)$$

where σ is the sigmoid function, and $\odot$ is the element-wise product.

Previous work has shown that the character-based approach is more suitable for Chinese NER tasks than the word-based approach. We thus use the character-based approach to improve the performance of the network caused by unclear boundaries of word partitioning across domains. In the processing, each encoded character is represented by an embedded character (e^c representing a character-embedded lookup table)

$$X_j^c = e^c(c_j) \quad (5)$$

A bidirectional LSTM is applied to $x_1, x_2, \ldots, x_m$ to obtain $\overrightarrow{h}^c_1, \overrightarrow{h}^c_2, \ldots, \overrightarrow{h}^c_m$ and $\overleftarrow{h}^c_1, \overleftarrow{h}^c_2, \ldots, \overleftarrow{h}^c_m$, we use two sets of parameters in the left-to-right and right-to-left directions, respectively. The representation of each character in hidden vector is:

$$h_j^c = [\overrightarrow{h}^c_j; \overleftarrow{h}^c_j] \quad (6)$$

3.2.3. Connecting Characters into Words

For word embedding, the calculation of c_e considers the lexical subsequence $W^{\mathbb{D}}_{b,e}$ in the sentence, that is, the character refers to the original text to divide the word list $\mathbb{D}$ and form the word through path matching [24]. In this way, while collecting contextual information in the module, the problem of ambiguity in cross-domain recognition can be mitigated by combining the feature of the character with that of the word. The word embedding $W^{\mathbb{D}}_{b,e}$ is used to represent each word, where each subsequence $W^{\mathbb{D}}_{b,e}$ is represented as (e^w in the word embedding lookup table):

$$X^w_{b,e} = e^w(W^d_{b,e}) \quad (7)$$

In addition, a word $C^W_{b,e}$ is used to indicate the state of repetition $X^w_{b,e}$ from the beginning of a sentence. $C^W_{b,e}$ is calculated as follows:

$$\begin{bmatrix} i^c_j \\ o^c_j \\ f^c_j \\ \widetilde{c}^c_j \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(W^{c\intercal} \begin{bmatrix} x^w_{b,e} \\ h^c_b \end{bmatrix} + b^w \right) \quad (8)$$

$$c^w_{b,e} = f^w_{b,e} \bigodot c^c_b + i^w_{b,e} \bigodot \widetilde{c}^w_{b,e} \quad (9)$$

$i^w_{b,e}$, $f^w_{b,e}$ is a set consisting of an input gate and a forgetting gate. Because the label works only at the character level, the output gate is not available for word units.

Given the value of c^w, more looped paths are available for information to flow to each c_e. We connect all c^w with $B \in b|w^{\mathbb{D}}_{b,e}$ in the c_e cell (as shown in Figure 4). We use an additional gate $i_{B,e}$ for each unit of the subsequence $c_{b,e}$ to control its contribution to $c^w_{B,e}$:

$$i^c_{B,e} = \sigma(W^{l\intercal} \begin{bmatrix} emb(c_e) \\ c^w_{b,e} \end{bmatrix} + b^l) \tag{10}$$

Therefore, the unit value is:

$$e^c_e = \sum_{B \in b|w_{b,e} \in \mathbb{D}} \alpha^c_{b,e} \bigodot c^w_{b,e} + \alpha^c_e \odot \widetilde{c}^c_e \tag{11}$$

In Equation (11), $i^c_{B,e}$ and i^c_e are normalized to $\alpha^c_{b,e}$ and α^c_e, respectively by setting the sum to one:

$$\alpha^c_{b,e} = \frac{\exp(i^c_{B,e})}{\exp(i^c_e) + \sum_{B \in b|w_{b,e} \in \mathbb{D}} \exp(i^c_{B,e})} \tag{12}$$

$$\alpha^c_e = \frac{\exp(i^c_e)}{\exp(i^c_e) + \sum_{B \in b|w_{b,e} \in \mathbb{D}} \exp(i^c_{B,e})} \tag{13}$$

The final hidden vector h^c_j is still calculated according to Equation (12). In training for named entity recognition, the loss is propagated back to the parameters W^c, b^c, W^w, b^w, W^l, and b^l focus on relevant words during NER.

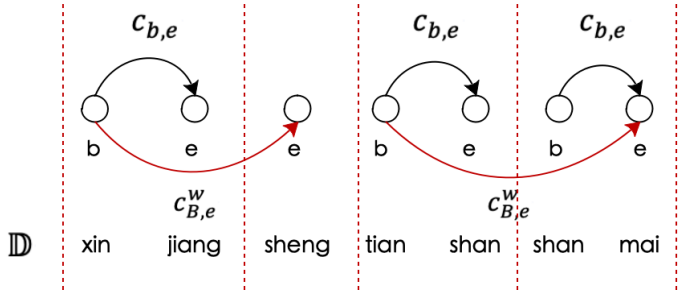

Figure 4. Connecting characters into words.

3.2.4. CRF Layer

As described above, the standard CRF layer is $h = [\overrightarrow{h}_1 \oplus \overleftarrow{h}_1, \ldots, \overrightarrow{h}_n \oplus \overleftarrow{h}_n]$. The mark sequence $y = l_1, l_2, \ldots, l_i$ is obtained by dividing the output probability $p(y|x)$ by the input sentence x:

$$p(y|x) = \frac{\exp \sum_i (w^{l_i}_{CRF} \bullet h_i + b^{(l_{i-1}, l_i)}_{CRF})}{\sum_{y'} \exp \sum_i (w^{l'_i}_{CRF} \bullet h_i + b^{(l'_{i-1}, l'_i)}_{CRF})} \tag{14}$$

where y' represents any tag sequence, $w^{l_i}_{CRF}$ is the model parameter specific to l_i and $b^{(l_{i-1}, l_i)}_{CRF}$, and is the bias specific to l_{i-1} and l_i. We find the tag sequences with the highest score in the word- or character-based input sequences by using the first-order Viterbi algorithm to. Given a set of training

data $\{(s_i, y_i)\}|_{i=1}^{N}$, the training model of sentence-level logarithmic likelihood loss is normalized to L_2. (λ is the regularization parameter of L_2 and Θ represents the parameter set):

$$L = \sum_{i=1}^{N} \log\left(P\left(y_i|s_i\right)\right) + \frac{\lambda}{2}\Theta 2 \tag{15}$$

4. Datasets

Previous work has contributed to Chinese named entity recognition data corpora, most of the content of which from the news and social media networks. For sports, finance, and tourism, while work needed to collect data has been carried out, the corpora are not adequate in general and need to be continually supplemented. Therefore, by relying on the rich tourism information database for China, we collected 5813 sentences and labeled them according to NER.

Corpus Statistics

We introduce detailed information about `Zoetral`, the Chinese tourism named entity recognition dataset that we built. The primary work involved collecting all the information about famous travel destinations in the cities, counties, and villages in Xinjiang Province and other provinces in China. We collected the dataset of sentences on tourism and used the `BIOE` method to annotate the NER travel domain datasets. A sample of the annotations is shown in Figure 5. We crawled 8460 posts from Qunar (https://www.qunar.com/), Meituan (https://www.meituan.com/), Ctrip (https://www.ctrip.com/), and other commonly used travel websites in China by using crawler tools. We manually selected 5813 sentences from them containing information on scenic spots, where each sentence contained 10 to 50 characters to ensure good performance of BiLSTM processing. Moreover, 5377 entities were extracted manually featuring three types. Specifically, 4674 items on "LOC" (location), 283 items on "PER" (personal), and 420 on "ORG" (organization).

	Label		Label		Label
tian	B-LOC	yao	B-LOC	de	O
shan	I-LOC	chi	E-LOC	bei	O
tian	I-LOC	"	O	ce	O
chi	E-LOC	,	O	,	O
,	O	wei	O	ju	O
you	O	yu	O	wu	B-LOC
cheng	O	chang	B-LOC	lu	I-LOC
tian	B-LOC	ji	I-LOC	mu	I-LOC
chi	I-LOC	zhou	E-LOC	qi	I-LOC
guo	I-LOC	fu	B-LOC	shi	E-LOC
jia	I-LOC	kang	I-LOC	yue	O
di	I-LOC	shi	E-LOC	1	O
zhi	I-LOC	jing	O	1	O
gong	I-LOC	nei	O	0	O
yuan	E-LOC	,	O	gong	O
,	O	bo	B-LOC	li	O
gu	O	ge	I-LOC	,	O
cheng	O	da	I-LOC	shi	O
"	O	feng	E-LOC	xin	B-LOC
...	...	...	...	...	...

Figure 5. Examples of types of BIOE annotations of Zoetral.

5. Experiment

5.1. Datasets

We carried out training and testing on three cross-domain datasets and compared the results with those of four other models. We used the MSRA (NER) Chinese named entity recognition dataset as data for the source domain. Three datasets of target domain data were used (see Table 1). Each dataset consisted of three parts: the training set, the test set, and the verification set. The MSRA (NER) datasets (https://download.csdn.net/detail/weixin_43098787/10992902) contained three types of entities—PER (person), LOC (location), and ORG (organization)—for a total of 46,365 items. The training set contained 27,819 items, and the validation and testing sets contained 9273 and 9273 items, respectively. The NER dataset containing items extracted from social media consisted of 8821 items from Sina Weibo (https://github.com/quincyliang/nlp-dataset/tree/master/ner-data/weibo) (a Chinese social media website). It was divided into 3307 items for the training set, 1103 for the verification set, and 1103 for the test set. The news datasets contained 23,061 items from the People's Daily (https://download.csdn.net/detail/weixin_43098787/10992902). We also collected 3820 articles on tourism in the dataset on tourism and manually marked them as a test set.

Table 1. Statistics of the datasets.

Corpus	Type	Total	Train	Dev	Test
MSRA(NER)	–	46,365	27,819	9273	9273
Weibo NER	Social media	5514	3307	1103	1103
People's Daily NER	News	23,061	13,837	4612	4612
Zoetral NER	Tourism	5377	3226	1075	1075

5.2. Baseline

We compared four prevalent models—transfer model T-B, neural adaptation layers, unified model, and the Lattice LSTM [24,34–36]—with the proposed POISE model through experiments. Ref. [35] proposed a lightweight yet effective method for domain adaptation for neural models. They introduced adaptation layers on top of neural architectures such that no re-training was required using data from the source domain.

Ref. [36] proposed a unified model that can learn from out-of-domain corpora and in-domain unannotated text. The unified model contains two major functions: one for cross-domain learning and the other for semi-supervised learning. There are two major functions in the unified model: one is cross-domain learning and the other for semi-supervised learning. The cross-domain learning method can learn out-of-domain label based on feature similarity, and the semi-supervised learning method can learn in-domain unannotated label through self-training.

Ref. [34] explored the problem of transfer learning for sequence tag tasks, in which a source task with a number of annotations was used to improve performance on a task with fewer annotations.

Ref. [24] investigated a lattice LSTM model for a Chinese NER task. It can encode a sequence of input characters and words, finding potential words that match a lexicon. The model explicitly leverages word and word sequence information, and not suffer from segmentation errors.

5.3. Experiment Setting

The MSRA NER datasets were pre-trained into a 300-dimensional word embedding. We also trained the NER data that had been selected from other cross-domain datasets into a 300-dimensional pre-training word embedding. Thirty-dimensional initialized character embeddings were adopted for all datasets. The number of dimensions of hidden states of the word-based BiLSTM was set to 200, and that of the source-shared and targeted BiLSTM was set to 100 for model-1 and model-2. The parameters were optimized using the Adam optimizer by adopting a gradient clipping and a learning rate decay of five. We also established an initial learning rate β of 0.002 in all experiments.

The learning rate β_t in each iteration was updated with $\beta_t = \beta 0/(1 + \rho \times t)$, and the decay rate ρ to 0.03. To avoid overfitting, we applied dropout to the output of the pre-processing layer and the BiLSTM layer. In the experiment, we used all data from the source domain (MSRA NER) and the target domain (Sina Weibo news, People's Daily NER, and tourism NER) for training, repeatedly adjusting the parameters for experimental comparison, and chose the best results. Tables 2–4 show these after 100 epochs. We run these typical models on 4 Tesla K80 GPUs and save the best model on the validation set for testing. We are interested in a model that performs robustly across a diverse set of tasks. To this end, for baseline, we use the same hyperparameters as those in the original paper.

5.4. Results and Analysis

As shown in Tables 2–4, our POISE outperformed the other methods. Transfer learning was the most beneficial for tasks with low-resource datasets, and enabled generalization even with few labeled examples.

Table 2. The results of Weibo NER. The best results in each metric are represented in bold.

	Prec	**Recall**	**F1**
Transfer model T-B [34]	83.29	69.34	75.68
Neural Adaptation Layers [35]	88.36	73.85	80.46
Unified Model [36]	93.75	76.20	**85.86**
Lattice Model [24]	92.84	**78.29**	84.95
POISE	**93.80**	77.62	84.96

Table 3. The results of news NER. The best results in each metric are represented in bold.

	Prec	**Recall**	**F1**
Transfer model T-B [34]	87.12	73.94	79.99
Neural Adaptation Layers [35]	95.38	79.30	86.60
Unified Model [36]	97.64	81.73	88.98
Lattice Model [24]	96.32	80.79	87.87
POISE	**98.32**	**84.37**	**90.81**

Table 4. The results of Zoetral NER. The best results in each metric are represented in bold.

	Prec	**Recall**	**F1**
Transfer model T-B [34]	82.48	68.47	74.82
Neural Adaptation Layers [35]	94.07	74.93	83.42
Unified Model [36]	93.64	**76.41**	84.15
Lattice Model [24]	92.38	74.29	82.35
POISE	**96.72**	74.83	**84.38**

Impact of character-level and word-level approaches

By using the Zoetral datasets as target domain in an example, along with the same data for the source domain, the results show that the accuracy of the POISE model was 8.92% higher than the adaptive layer method proposed in [35]. When we used the transfer model T-B model to train the Chinese NER datasets, the [35] model worked well for English entity recognition but its performance degraded when applied to Chinese text. Although it rendered the characteristics of the input and the model parameters isomorphic between the domains by adding a word adaptation layer, the isomorphism was forced in the process of migration of Chinese text, resulting in reduced accuracy (see especially Tables 2 and 4, where there were significant differences between the domains of the target and the source datasets). We used the BiLSTM module based on the feature transfer in transfer learning for the Chinese NER task. We matched words with the input characters and training the character feature of the source domain to transfer to the target domain. In this way, we obtained better word and character features, where this significantly mitigated the degradation in performance caused by the unclear boundary of cross-domain word segmentation in Chinese.

Effect of transfer learning

The parameter-sharing scheme proposed in [34] has contributed significantly to the transfer learning from high- to low-resource data, but methods of learning within the scheme itself have not been the subject of much research and experimentation. Considering that differences in resource representation between domains may occur in the process of migration and learning, and may in turn lead to poor identification performance, the POISE model was able to transfer the character label features between the source and the target domains. Even though they did not share words, they had common characters. The BiLSTM layer calculated a better score than traditional method by learning the feature-related information in the characters. When sharing parameters, tasks were differentiated based on the similarity of labels of the domains of the data. Compared to the lattice LSTM model [24], lattice LSTM does not use the transfer learning, so that the model depends on the quality of the dataset. We know through experiments that the effect of the dataset from other domains is obviously lower than the effect of the dataset provided by the original Lattice LSTM work. There is no doubt about the contribution of the work from Lattice LSTM, but when the data in different domains are applied, there is still a big problem of wrong identification and the general effect has a low effect.

Impact of combining characters to be a word

We also compared our method with the unified model proposed in [36] that also uses out-of-domain and unlabeled data in case of unsupervised training as well as segmented words based on characters to obtain their positions in words when analyzing text. Although it has an excellent recognition effect in some cross-domains, it is not suitable for recognition tasks involving domains with large differences. When the source domain consisted of the MSRA entity datasets, the target domain used data from different fields for the experiments. The results show that because this was similar to the field of training data, the results after learning were better than the test results for data from other fields, as shown in Table 3. Moreover, the testing effect of this model for microblogs was better than that for tourism. We see that the unified model was better than the parameter-sharing scheme proposed by [34]. However, in terms of cross-domain learning, its performance was poorer than that of the adaptive layer method proposed by [35]. Therefore, we made full use of the words, characters, and character features to mine potential lexical information, and combined them to match entity vocabularies in different fields according to calculations in the processing module. In this way, we effectively mapped labels from the high-resource to the low-resource data, improved the accuracy of label prediction, and reduced errors.

5.5. Ablation Study

We used the POISE model for a comparison of ablation experiments for training and testing on datasets of the three target areas. The methods with and without parameter sharing were trained. A comparison of the results is shown in Table 5.

The data to be predicted in the target field should be trained and learned using only data of the source domain without any reference annotation. In each set of the domain data, the result using transfer learning was better than that without it. As shown in Figure 6, by learning from the data of the source domain and applying the method to predict those of the target domain, the recognition effect was improved.

Table 5. Results of transfer learning. The best results in each metric are represented in bold.

		Prec	Recall	F1
News	− parameter sharing	76.83	59.19	66.87
	+ parameter sharing	**98.32**	**84.37**	**90.81**
Weibo	− parameter sharing	78.42	60.83	68.51
	+ parameter sharing	**93.80**	**77.62**	**84.96**
Zoetral	− parameter sharing	65.73	48.22	55.63
	+ parameter sharing	**96.72**	**74.83**	**84.38**

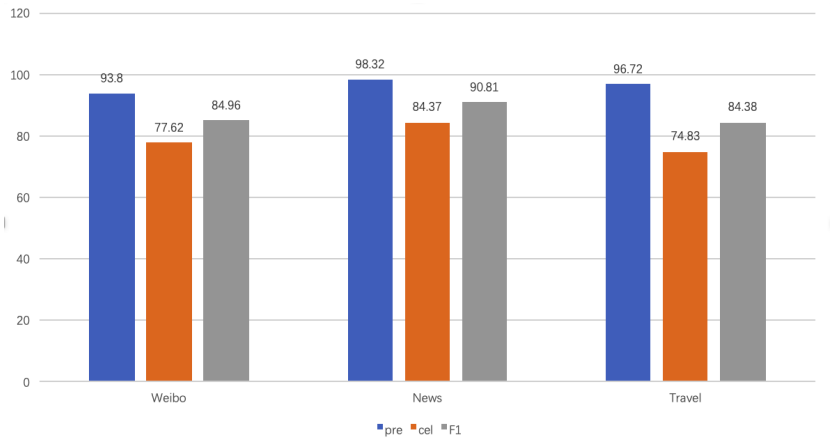

Figure 6. A comparison of the results of the POISE model on datasets of different domains.

5.6. Case Study

As shown in Figure 7, the following problems occurred in a comparison of experimental cases between the proposed model and the other four models:

(1) The parameter-sharing scheme proposed by [34] does not adequately divide the vocabulary in the data of the target domain, which affected performance in terms of tag learning.

(2) Ref. [35] could not identify "xi zheng jun dian jiang tai" as the LOC tag in the Zoetral datasets, but did identify "xi zheng jun" as a PER tag. Its efficiency of recognition decreased, owing to the large number of characters in the name field, and "yi ji" was forcibly mapped to the LOC label.

(3) According to the data analysis in Section 5.4, the experiment comparing the unified model [36] with the POISE model showed that the former performed well on datasets of similar types to that of the source domain, but yielded poorer performance when the target and source domains were dissimilar. As shown in Figure 7, this model did not identify the LOC label "nan hu gong yuan" in the Zoetral datasets. The comparison shows that the proposed model adequately partitioned the vocabulary of the target domain, which provided accurate labels for prediction according to its characteristics.

(4) In the comparative experiment on the lattice LSTM, we found that the system could not transfer knowledge from a high-resource to a low-resource one. However, when the data were sufficient in quantity, the model yielded good identification results.

Transfer model T-B (Yang et al.,2017)	wo men/yao/jin guo/shi you/zhi/cheng/ke/la/ma/yi
POISE Model	wo men/yao/jin guo/shi you zhi cheng(B-LOC I-LOC I-LOC E-LOC)/ke la ma yi(B-LOC I-LOC I-LOC E-LOC)
Neural Adaptation Layers (Lin and Lu,2018)	you/cheng ji si han(B-PER I-PER I-PER E-PER)/xi zheng jun(B-PER I-PER E-PER)/dian jiang tai(B-LOC I-LOC E-LOC)/deng/li shi/wen hua/yi ji(B-LOC E-LOC)
POISE Model	you/cheng ji si han(B-PER I-PER I-PER E-PER)/xi zheng jun dian jiang tai (B-LOC I-LOC I-LOC I-LOC I-LOC E-LOC)/deng/li shi/wen hua/yi ji(B-LOC E-LOC)
Unified Model (He and Sun,2017)	nan/hu/gong yuan(B-LOC E-LOC)/zuo/luo/zai/shi zheng fu(B-ORG I-ORG E-ORG)/nan ce
POISE Model	nan hu gong yuan(B-LOC I-LOC I-LOC E-LOC)/zuo/luo/zai/shi zheng fu(B-ORG I-ORG E-ORG)/nan ce
Lattice Model (Zhang and Yang,2018)	zhu yao/you/hong shan ta(B-LOC I-LOC E-LOC)/、/da fo si(B-LOC I-LOC E-LOC)/he/yuan tiao lou(B-LOC I-LOC E-LOC)/deng/ji ge/zhong dian/qu yu
POISE Model	zhu yao/you/hong shan ta(B-LOC I-LOC E-LOC)/、/da fo si(B-LOC I-LOC E-LOC)/he/yuan tiao lou(B-LOC I-LOC E-LOC)/deng/ji ge/zhong dian/qu yu

Figure 7. Examples of the results of the POISE and other models on NER tasks. The results in blue represent those for location, those in green represent the recognition results for person, and yellow represents the results for organization.

The comparison shows that the proposed model adequately partitioned the vocabulary of the target domain, which yielded accurate labels for prediction according to its characteristics.

5.7. Discussion

Through experiments, we know that the transfer learning method does perform well in dealing with tasks in the domain of low-resource datasets. At the same time, we also discussed in depth some problems existing in the application of this method in the Chinese NER.

(1) English NER and Chinese NER use the same transfer learning method, but the results are different. Word segmentation problems in Chinese NER will reduce recognition performance, especially in datasets in different domains. Therefore, based on the work proposed by [24], we add the method of combining the characters and words. Through the comparison of three groups of datasets in different domains, the method of word combination is indeed conducive to improving the accuracy of word segmentation.

(2) In the application of the transfer learning method, we have done four groups of comparative experiments. Among them, [34–36] is all NER model based on transfer learning method, while Lattice LSTM [24] is the NER model based on training a large number of datasets. The initial experimental results show that the NER model based on the transfer learning method has an outstanding application effect only in one domain, but not in datasets in others. As shown in Tables 2–4, after the screening of datasets and label features, we can obtain the final experimental results, which shows that this kind of method depends on the similarity of data to a great extent. For the work of [24], the model proposed by them is outstanding in the Chinese NER task, but it is not outstanding in some low-resource datasets. We expanded a large number of datasets and trained them with Lattice LSTM to improve the results. By comparing the two types of NER models, it can be proved that transfer learning does play a role in NER tasks.

(3) When comparing the same type of model based on the transfer learning method, we find that selecting data with high label similarity for mapping between high and low resources is conducive to improving the training results of low-resource datasets. In the initial experiment, we directly used the model proposed in [34] for mapping, completely ignoring the differences in labels and other issues, and the final experimental results were very poor. After the similarity screening of features added to the dataset, the experimental results can be effectively improved. Therefore, before we do the baseline

comparison experiment, we first selected the common features in the homologous datasets of the three groups of target domain datasets to obtain the results in Section 5.4.

To sum up, the steps we added can improve the application of transfer learning in Chinese NER. However, we do not deny that this scheme still depends on the size and quality of the source dataset and the feature similarity between the source and the target domain dataset. The feature similarity between datasets in different domains is still a problem worth exploring in the future.

6. Conclusions and Future Work

In this paper, we proposed a scheme which focuses on improving the performance of named entity recognition task in low-recourse datasets. Specifically, by transferring the label features and parameters between low- and high-resource datasets, it can obtain an effective result of named entity recognition. Besides, we consider differences in domains between the source and the target texts, combines the characters and words to solve the segmentation problems in Chinese NER. Experiments on three target datasets showed that our method allows for zero-sample learning, and can deliver consistently superior performance on different domains, which can provide that this scheme is suitable for Chinese named entity recognition.

Although the NER task performance of low-resource datasets has been improved under the POISE model, there are still some problems to be solved in this model. The method of combining characters and words is used to find the maximum path of words. Although the result obtains better performance, the redundancy is too large, the time complexity is also very high, and there is still room for optimization. In addition, although transfer learning can help feature conversion between different datasets, this feature conversion is limited. Between the two datasets, only features with high similarity can be transferred and learned. If the similarity is low, it will lead to poor transfer learning effects and even identify a large number of wrong labels. Therefore, how to better optimize transfer learning is still a research hotspot.

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

Funding: This work was funded by the Opening Foundation of the Key Laboratory of Xinjiang Uyghur Autonomous Region of China (grant number 2016D03023); the National Natural Science Foundation of China (grant number 61662077); the Scientific Research Program of the State Language Commission of China (grant number ZDI135-54). Supported by Xinjiang Uyghur Autonomous Region Graduate Research and Innovation Project Grant Number XJ2020G071, Dark Web Intelligence Analysis and User Identification Technology Grant Number 2017YFC0820702-3, National Language Commission Research Project Grant Number ZDI135-96, and funded by National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC).

Acknowledgments: We thank the anonymous reviewers for their valuable feedback. These authors contributed equally to this work.

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

Abbreviations

The following abbreviations are used in this manuscript:

NER	Named entity recognition
NLP	natural language processing
IFC	Inter-feature correlation
SU	Symmetric uncertainty
SFD	Similarity of feature distribution

References

1. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A comprehensive survey on transfer learning. *Proc. IEEE* **2020**. [CrossRef]
2. Lin, Y.H.; Chen, C.Y.; Lee, J.; Li, Z.; Zhang, Y.; Xia, M.; Rijhwani, S.; He, J.; Zhang, Z.; Ma, X.; et al. Choosing Transfer Languages for Cross-Lingual Learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3125–3135.
3. Peng, Y.; Yan, S.; Lu, Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, 1 August 2019; pp. 58–65.
4. Qi, T.; Xu, Y.; Ling, H. Tourism scene classification based on multi-stage transfer learning model. *Neural Comput. Appl.* **2019**, *31*, 4341–4352. [CrossRef]
5. Chen, Q.; Zheng, Z.; Hu, C.; Wang, D.; Liu, F. Data-driven Task Allocation for Multi-task Transfer Learning on the Edge. In Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas, TX, USA, 7–10 July 2019.
6. Wang, H.; Hu, J. Deep Multi-Task Transfer Network for Cross Domain Person Re-Identification. *IEEE Access* **2019**, *8*, 5339–5348. [CrossRef]
7. Pfeiffer, J.; Vulić, I.; Gurevych, I.; Ruder, S. MAD-X: An Adapter-based Framework for Multi-task Cross-lingual Transfer. *arXiv* **2020**, arXiv:2005.00052.
8. Yin, X.; Yu, X.; Sohn, K.; Liu, X.; Chandraker, M. Feature transfer learning for face recognition with under-represented data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5704–5713.
9. Nahmias, D.; Cohen, A.; Nissim, N.; Elovici, Y. Deep feature transfer learning for trusted and automated malware signature generation in private cloud environments. *Neural Netw.* **2020**, *124*, 243–257. [CrossRef] [PubMed]
10. Wang, T.; Huan, J.; Zhu, M. Instance-based deep transfer learning. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 367–375.
11. Arnold, A.O.; Cohen, W.W. Instance-based Transfer Learning for Multilingual Deep Retrieval. *arXiv* **2019**, arXiv:1911.06111.
12. Ma, J.; Li, Z.Z.J.C.A.; Hong, L. *SNR: Sub-Network Routing for Flexible Parameter Sharing in Multi-Task Learning*; Association for the Advancement of Artificial Intelligence: Menlo Park, CA, USA, 2019.
13. Savarese, P.; Maire, M. Learning Implicitly Recurrent CNNs Through Parameter Sharing. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
14. Jia, C.; Liang, X.; Zhang, Y. Cross-Domain NER using Cross-Domain Language Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 2464–2474.
15. Taslimipoor, S.; Rohanian, O. Cross-lingual transfer learning and multitask learning for capturing multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 155–161.
16. Tao, J.; Fang, X. Toward multi-label sentiment analysis: A transfer learning based approach. *J. Big Data* **2020**, *7*, 1–26. [CrossRef]
17. Li, J.; Sun, A.; Han, J.; Li, C. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.* **2020**. [CrossRef]
18. Kuru, O.; Can, O.A.; Yuret, D. Charner: Character-level named entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*; The COLING 2016 Organizing Committee: Osaka, Japan, 2016; pp. 911–921.
19. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
20. Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Available online: http://portal.acm.org/citation.cfm?id=655813 (accessed on 8 September 2020).

21. 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 (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 1064–1074.
22. Liu, L.; Shang, J.; Ren, X.; Xu, F.F.; Gui, H.; Peng, J.; Han, J. Empower Sequence Labeling with Task-Aware Neural Language Model. *arXiv* **2018**, arXiv:1709.04109v4.
23. Jin, Y.; Xie, J.; Guo, W.; Luo, C.; Wu, D.; Wang, R. LSTM-CRF neural network with gated self attention for Chinese NER. *IEEE Access* **2019**, *7*, 136694–136703. [CrossRef]
24. Zhang, Y.; Yang, J. Chinese NER Using Lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 1554–1564.
25. Shivakumar, P.; Georgiou, P. Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations. *Comput. Speech Lang.* **2020**, *63*, 101077. [CrossRef] [PubMed]
26. Banerjee, S.; Akkaya, C.; Perez-Sorrosal, F.; Tsioutsiouliklis, K. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 6295–6300.
27. Luo, G.; Yang, Y.; Yuan, Y.; Chen, Z.; Ainiwaer, A. Hierarchical transfer learning architecture for low-resource neural machine translation. *IEEE Access* **2019**, *7*, 154157–154166. [CrossRef]
28. Bhatia, P.; Arumae, K.; Celikkaya, B. *Dynamic Transfer Learning for Named Entity Recognition*; Springer: Berlin/Heidelberg, Germany, 2019.
29. Chen, L.; Moschitti, A. Transfer learning for sequence labeling using source model and target data. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Menlo Park, CA, USA, 2019; Volume 33, pp. 6260–6267.
30. Al-Smadi, M.; Al-Zboon, S.; Jararweh, Y.; Juola, P. Transfer Learning for Arabic Named Entity Recognition With Deep Neural Networks. *IEEE Access* **2020**, *8*, 37736–37745. [CrossRef]
31. Francis, S.; Van Landeghem, J.; Moens, M.F. Transfer Learning for Named Entity Recognition in Financial and Biomedical Documents. *Information* **2019**, *10*, 248. [CrossRef]
32. Peng, D.; Wang, Y.; Liu, C.; Chen, Z. TL-NER: A transfer learning model for Chinese named entity recognition. *Inf. Syst. Front.* **2019**, 1–14. [CrossRef]
33. 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.
34. Yang, Z.; Salakhutdinov, R.; Cohen, W.W. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv* **2017**, arXiv:1703.06345.
35. Lin, B.Y.; Lu, W. Neural adaptation layers for cross-domain named entity recognition. *arXiv* **2018**, arXiv:1810.06368.
36. He, H.; Sun, X. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media. In *Thirty-First AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Menlo Park, CA, USA, 2017.