Zero-Shot Entity Recognition via Multi-Source Projection and Unlabeled Data

Huixiong Yi* and Jin Cheng

Department of Computer Science and Technology, University of Science and Technology of China, Hefei, China
*Corresponding Author's Email: huixiongyi@gmail.com

Abstract. With the widespread of transfer learning, great success has been achieved in named entity recognition on languages without labeled data. Prior works can be mainly classified into two categories: direct model transfer and annotation projection. However, these works only utilize part of available data, including monolingual word embeddings, labeled data in multiple source languages, and unlabeled data in the target language. To make full use of these data and extract more information, we propose to train multiple teacher models on multiple source languages via the annotation projection method. Then a student model is trained on the unlabeled data in the target language with supervision from all labels obtained from multiple teacher models. Extensive experiments are conducted on three benchmark datasets, and the experimental results prove the effectiveness of the proposed method.

1. Introduction

Named entity recognition (NER) is a tagging task that detects and classifies named entities from text into specified classes, such as people, organizations, and locations. Recent works mainly use deep learning models on NER, such as Recurrent Neural Network (RNN), and these models are trained on labeled datasets. These models only work well when there exists a large amount of labeled training data. These models get poor performance in many languages with limited amounts of labeled data. The problem can be mitigated by adopting cross-lingual transfer. The cross-lingual transfer utilizes knowledge in high-resource source languages with rich labeled data to improve the performance of low-resource target languages with a few or no labeled data. Specifically, in this paper, we focus on zero-resource cross-lingual transfer, where there is no labeled data in the target languages.

Prior works on zero-resource cross-language transfer NER can be mainly classified into two categories: annotation projection and direct model transfer. Annotation projection methods usually translate source-language labeled data into the target language and project labels to construct the pseudo target-language labeled data, on which a NER model is trained. For example, source-language labeled data are translated into target language word by word via finding its nearest neighbor in the shared embedding space, and the entity label of each source word is directly copied to be the label of the corresponding target word [1]. Direct model transfer methods train a NER model on the source language labeled data using delexicalized and language-independent features. Then the NER model is directly applied to the target language.

Both annotation projection methods and direct model transfer methods suffer from several intractable problems. Annotation projection methods depend on the translation system and alignment algorithm, and direct model transfer methods cannot utilize task-specific information in the target language. Furthermore, both methods do not leverage target-language unlabeled data, which is easy
and cheap to obtain and contains much language-specific information. Meanwhile, in practice, there are usually multiple available source languages, which can provide task-specific information.

To tackle the above problems, a new zero-resource cross-lingual transfer method is proposed in this paper. First, the annotation projection methods can be improved via adopting multilingual BERT (mBERT) [2] as a base model, since the mBERT model can produce language-independent features. Then a target-language pseudo-labeled training set is constructed, on which a student NER model is trained. Finally, $K$ teacher models are trained from $K$ source-language datasets. One of the $K$ teacher models is selected to provide the soft labels, and all teacher models vote to get the pseudo hard labels. The student NER model is trained with supervision from both soft labels and pseudo hard labels, and it is the final model to make predictions for the sentences in the target language.

2. Model

In this section, we introduce the proposed zero-resource cross-lingual transfer method in three parts, i.e., basic NER module, annotation projection module, and teacher-student module. Figure 1 shows the overall framework of the proposed method.

![Figure 1. The framework of the proposed method. (a) Training a teacher model on source annotated data. (b) Training a student model on unlabelled target data via multiple teacher models.](image)

2.1. Basic NER Module

In this paper, we use mBERT as the language-independent feature encoder to build a basic NER model. mBERT is a pre-trained language model based on deep bidirectional transformers. mBERT is pre-trained on two auxiliary tasks: masked language model and next sentence prediction. Furthermore, mBERT is pre-trained on Wikipedia with 104 languages so that it can provide language-independent features.

mBERT takes in an input sentence $\mathbf{x} = (x_1, \cdots, x_L)$ and then outputs context hidden vectors $\mathbf{h} = (h_1, \cdots, h_L)$. Then we add a linear classification layer with softmax to compute the probability distribution of labels. The above procedure can be formulated as:

$$\mathbf{h} = \text{mBERT}(\mathbf{x}),$$

$$\hat{y}_i = \text{softmax}(W h_i + b),$$

where $\hat{y}_i$ denotes the predicted probability distribution for the token $x_i$. $\theta = \{\theta_{\text{mBERT}}, W, b\}$ denotes the parameters of the basic NER model.

The loss function $\mathcal{L}$ for the NER model is defined as the cross-entropy between the predicted label distribution and the ground-truth one for each token:

$$\mathcal{L}(\theta) = -\frac{1}{L} \sum_{i=1}^{L} \text{CrossEntropy}(\hat{y}_i, y_i),$$

where $y_i$ denotes a one-hot vector of the ground-truth label corresponding to $x_i$. Note that all NER models share the same architecture with the basic NER model here.
2.2. Annotation Projection Module

Following [3], source-language labeled data is translated into target language word-by-word, and we directly copy the label of each source word to the corresponding target word. Then we directly use the translated data to train a NER model. The details of this method are introduced below.

First, word embedding matrices $S$ and $T$ in the source and target languages are obtained via FastText embeddings. Next, a linear mapping matrix $M$ is learned between $S$ and $T$, as formulated below:

$$
M = \arg\min_M \|MS - T\|^2 \text{ s.t. } MM^T = I,
$$

where $M$ is a square parameter matrix. Finally, we select the nearest-neighbor in the target-language space as the translation result of each word in the source language. Here we adopt cross-domain similarity local scaling (CSLS) metric to measure the distance, as formulated below:

$$
\text{CSLS}(s_i, t_j) = 2 \cos(s_i, t_j) - \frac{1}{N} \sum_{t_n \in \mathcal{N}_T(s_i)} \cos(s_i, t_n) - \frac{1}{N} \sum_{s_n \in \mathcal{N}_S(t_j)} \cos(s_n, t_j),
$$

where $s_i$ is the embedding of a source-language word and $t_j$ is the embedding of a target-language word. $\mathcal{N}_T(s_i)$ denotes $K$ target-language nearest neighbors of the source word. $\mathcal{N}_S(t_j)$ denotes $K$ source-language nearest neighbors of the target word.

2.3. Teacher-Student Module

As illustrated in Figure 1(a), the source-language labeled data is first translated into the target language to get the pseudo labeled data on the target language. Then, a teacher model is trained on the pseudo labeled data. As illustrated in Figure 1(b), $K$ teacher models are first trained on $K$ source-language labeled datasets via the procedure described above. Then, a student NER model is trained on the unlabeled target-language data with supervision from the $K$ teacher models. Specifically, we select one of the $K$ teacher models to get the soft label and use all $K$ teacher models to get the hard label.

2.3.1. Soft Label

The unlabelled target-language data is denoted as $\{\overline{x}\}$, where $\overline{x} = (\overline{x}_1, \cdots, \overline{x}_L)$ denotes an unlabeled sentence. The probability distribution predicted from the teacher model is regarded as the soft label of the unlabeled target-language data. The mean squared error is adopted to measure the distance between output distributions of the teacher network and the student network. The loss function $L_{\text{soft}}$ for the soft label is formulated below:

$$
L_{\text{soft}} = -\frac{1}{L} \sum_{l=1}^{L} \text{MSE}(\hat{y}_l^{\text{Teacher}}, \hat{y}_l^{\text{Student}}),
$$

where $\hat{y}_l^{\text{Teacher}}$ and $\hat{y}_l^{\text{Student}}$ are the output distribution predicted by the teacher model and the student model, respectively.

2.3.2. Hard Label

In teacher-student learning, the knowledge from the teacher may be inaccurate, i.e. the result predicted from the teacher model is incorrect. Therefore, the student model can be improved by using correct hard labels. However, there is no hard label of the unlabeled target-language data. To deal with this, we propose to obtain the pseudo hard labels by integrating all predictions from $K$ teacher models. The $K$ teacher models are trained from $K$ pseudo target-language training data.

Take the $k$-th source language for example. The $k$-th source language is translated into target language to get the $k$-th pseudo target-language training data, on which the $k$-th teacher model is trained. The label for $\overline{x}_i$ predicted from the $k$-th teacher model is denoted as $\hat{y}_i^{\text{Teacher}}$, as formulated below:

$$
\hat{y}_i^{\text{Teacher}} = \arg\max_j \hat{y}_i^{\text{Teacher}}_j,
$$

where $\hat{y}_i^{\text{Teacher}}_j$ is the probability distribution of the $i$-th label predicted from the $k$-th teacher model. Then we adopt a voting scheme to obtain the final pseudo hard label $\hat{y}_i^{\text{Teacher}}$, as formulated below:
\[ \hat{y}_i^{\text{Teacher}} = \arg\max_k \sum_{k=1}^K f_{\text{onehot}}(\hat{y}_k^{\text{Teacher}}), \]  

where \( f_{\text{onehot}} \) is a one-hot encoding function. The loss function \( \mathcal{L}_{\text{hard}} \) for the pseudo hard label is defined as the cross-entropy between the pseudo hard label obtained from \( K \) teacher models and the output probability distribution from the student model, as formulated below:

\[
\mathcal{L}_{\text{hard}} = -\frac{1}{L} \sum_{i=1}^L \text{CrossEntropy}(\hat{y}_i^{\text{Teacher}}, \hat{y}_i^{\text{Student}}).
\]

The total loss for the teacher-student model is \( \mathcal{L} \), as formulated below:

\[
\mathcal{L} = \mathcal{L}_{\text{soft}} + \mathcal{L}_{\text{hard}}.
\]

The student model is trained to minimize \( \mathcal{L} \), and it is the only model to be used to predict the entity labels for the input sentence of the target language during the inference stage.

### Table 1. Statistics of the benchmark dataset.

| Language   | Type     | Train | Dev  | Test  |
|------------|----------|-------|------|-------|
| Spanish-es | Sentence | 8323  | 1915 | 1517  |
| (ConLL-2002) | Entity  | 18798 | 4351 | 3558  |
| Dutch-nl   | Train    | 15806 | 2895 | 5195  |
| (ConLL-2002) | Test    | 13344 | 2616 | 3941  |
| English-en | Train    | 14987 | 3466 | 3684  |
| (ConLL-2003) | Test    | 23499 | 5942 | 5648  |
| German-de  | Train    | 12705 | 3068 | 3160  |
| (ConLL-2003) | Test    | 11851 | 4833 | 3673  |

### 3. Experiments

#### 3.1. Datasets

The experiments are conducted on several benchmark datasets, including CoNLL-2002 Spanish and Dutch NER, CoNLL-2003 English and German NER. Table 1 shows the statistics of all datasets. For all experiments, when we use one language as the target language, the other as all source languages. The teacher model trained on English is selected to produce soft labels for the student model.

#### 3.2. Experiment Setup

The FastText embeddings and the MUSE model are used in the translation procedure. The cased multilingual BERT\textsubscript{BASE} is used as the language-independent feature encoder, which has 12 Transformer blocks, 12 self-attention heads, 768 hidden units. Specifically, dropout is applied to the mBERT with a rate of 0.1, and the parameters of the bottom three layers of mBERT are frozen. During the training process, we train all teacher and student models for 3 epochs and choose minibatchsize as 32. AdamW is used as the optimizer of all models, while we set the learning rate to 5e-5 for the teacher model and 1e-4 for the student model.

#### 3.3. Results and Analysis

To evaluate the performance of the proposed model, experiments are conducted on three target languages, including Spanish, Dutch, and German. The evaluation metric is F1-score. Table 2 presents the results of the proposed model and other prior state-of-the-art methods. The results show that our method achieves improvements compared with other methods. Specifically, the improvement of F1-score is 2.70 for Spanish, and the average improvement is 1.27 for all languages. The results confirm the effectiveness of our proposed model.
The improvement is mainly credited with the following elements. First, the proposed model combines the advantages of the two mainstream methods. Next, the target-language unlabeled data is used to obtain target language-specific information. Finally, multiple source languages are used in the proposed model to get task-specific information of NER.

Table 2. Results of multi-source cross-lingual NER.

| Method | es  | nl  | de  | avg |
|--------|-----|-----|-----|-----|
| Täckström, McDonald, and Uszkoreit (2012) [4] | 61.90 | 59.90 | 36.40 | 52.73 |
| Chen et al. (2019) [5] | 73.50 | 72.40 | 56.00 | 67.30 |
| Moon et al (2019) [6] | 76.53 | 83.45 | 72.44 | 77.47 |
| Wu et al (2020) [7] | 78.00 | 81.33 | 75.33 | 78.22 |
| Ours | **80.70** | 83.11 | 74.65 | **79.49** |

4. Related Work

4.1. Named Entity Recognition

Machine learning and deep learning have been widely used to solve problems in feature selection [8][9], computer vision [10], natural language processing [11][12]. Various models based on machine learning algorithms, such as SVM [13][14], MPCVM[15][16], have achieved decent performance in NER. With the widespread use of deep learning, many neural network models have been proposed in NER, such as LSTM, LSTM-CRF. CNN has also been introduced into NER to capture the character information, which contains much morphological and orthographic information. A lattice-structured LSTM model [17] was proposed to use word and character information at the same time and made great improvements.

4.2. Cross-lingual Learning

Cross-lingual learning approaches are powerful to address low resource NER tasks. At present, there exist two types of approaches: annotation projection based on the translation and the direct model transfer.

The key of the annotation projection is to construct a target-language pseudo-labeled dataset. Therefore, the approach depends mainly on the quality of the translation and the alignment of the label. Although there are some problems with the annotation projection, this intuitive approach has been widely used in various NLP tasks under the cross-lingual setting, such as POS tagging, parsing, and NER. For example, Google Translate is utilized twice to translate entities and sentences, and the alignment of the label is completed via distributional statistics [18]. The direct model transfer relies on the delexicalized and language-independent features. Once a model is trained via the direct model transfer, it will be directly applied to target languages. However, this approach is strongly limited by the fact that it requires a generic feature representation across languages. Many previous works have studied the problem. A cross-lingual word cluster [4] was built to generate universal features. The distributed representations of words were projected into a common space as language-independent features. mBERT was used to produce language-independent features [19]. Later, a teacher-student network was applied to enhance the model trained on mBERT [7].

5. Conclusion

In this paper, we propose the method for zero-resource cross-lingual NER. The proposed method synthesizes the advantages of direct model transfer and annotation projection method. Meanwhile, the proposed method utilizes the unlabeled target-language data and multiple available source-language datasets to obtain more information. Experimental results on their benchmark datasets empirically indicate that the proposed model is effective.
6. References

[1] Wu Y, Schuster M, Chen Z, et al. 2020 Google's neural machine translation system: Bridging the gap between human and machine translation arXiv preprint arXiv:1609.08144

[2] Devlin J, Chang M W, Lee K, et al. 2019 BERT: Pre-training of deep bidirectional transformers for language understanding Proc. of the 2019 Conf. of the NAACL-HLT pp 214-24

[3] Xie J, Yang Z, Neubig G, et al. 2018 Neural crosslingual named entity recognition with minimal resources Proceedings of the 2018 Conf. on EMNLP pp 369-79

[4] Täckström O, McDonald R, Uszkoreit J 2012 Cross-lingual word clusters for direct transfer of linguistic structure Proceedings of the 2012 Conf. of the NAACL-HLT

[5] Chen X, Hassan Awadallah A, Hassan H, et al. 2019 Multi-Source Cross-Lingual Model Transfer: Learning What to Share Proc. of the 57th Annual Meeting of ACL

[6] Moon T, Aswathy P, Ni J, et al. 2019 Towards lingua franca named entity recognition with bert arXiv preprint arXiv:1912.01389

[7] Wu Q, Lin Z, Karlsson B F, et al. Single-/Multi-Source 2020 Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language Proc. of the 58th Annual Meeting of ACL

[8] Wu X, Jiang B, Zhong Y, et al. Tolerant Markov Boundary Discovery for Feature Selection Proc. of the 29th CIKM pp 2261-64

[9] Wu X, Jiang B, Yu K, et al. 2020 Multi-Label Causal Feature Selection AAAI pp 6430-37

[10] Zhou X, Chen H and Li J 2018 An Automatic GPR B-scan Image Interpreting Model IEEE Transactions on Geoscience and Remote Sensing

[11] Lyu S, Cheng J, Wu X, et al. 2020 Auxiliary Learning for Relation Extraction IEEE Transactions on Emerging Topics in Computational Intelligence

[12] Lyu S and Liu J 2020 Hybrid Framework of Convolution and Recurrent Neural Networks for Text Classification Proc. of IEEE Int. Conf. on Knowledge Graph pp 313-20

[13] Tang F, Peter T, Pedro A and Chen H 2015 The Benefits of Modelling Slack Variables in SVMs Neural Computation pp 954-81

[14] Yu L, Chen H, Wang S and Lai K 2009 Evolving Least Squares Support Vector Machines for Stock Market Trend Mining IEEE Transactions Evolutionary Computation pp 87-102

[15] Chen H, Peter T and Yao X 2014 Efficient Probabilistic Classification Vector Machine With Incremental Basis Function Selection IEEE Transactions on Neural Networks and Learning Systems pp 356-69

[16] Lyu S, Tian X, Li Y et al. 2019 Multi-class Probabilistic Classification Vector Machines IEEE Transactions on Neural Networks and Learning Systems vol 31 pp 3906-19

[17] Zhang Y, Yang J 2018 Chinese NER Using Lattice LSTM Proc. of the 56th Annual Meeting of ACL pp 1554-64

[18] Jain A, Paranjape B, Lipton Z C 2019 Entity Projection via Machine-Translation for Cross-Lingual NER Proc. of the 2019 Conf. on EMNLP-IJCNLP pp 1080-92

[19] Wu S, Dredze M, Bentz 2019 Becas: The Surprising Cross-Lingual Effectiveness of BERT Proc. of the 2019 Conf. on EMNLP-IJCNLP pp 833-44