Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge Transfer

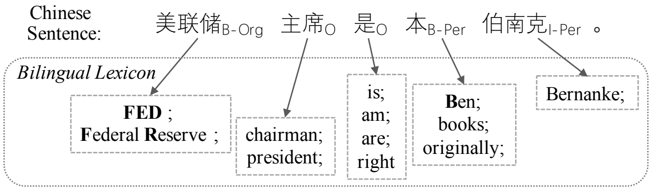
Xiaocheng Feng, Xiachong Feng, Bing Qin*, Zhangyin Feng, Ting Liu
Harbin Institute of Technology, China
{xcfeng, xiachongfeng, bqin, zyfeng, tliu}@ir.hit.edu.cn

Abstract

Neural networks have been widely used for high resource language (e.g. English) named entity recognition (NER) and have shown state-of-the-art results. However, for low resource languages, such as Dutch and Spanish, due to the limitation of resources and lack of annotated data, NER models tend to have lower performances. To narrow this gap, we investigate cross-lingual knowledge to enrich the semantic representations of low resource languages. We first develop neural networks to improve low resource word representations via knowledge transfer from high resource language using bilingual lexicons. Further, a lexicon extension strategy is designed to address out-of lexicon problem by automatically learning semantic projections. Finally, we regard word-level entity type distribution features as an external language-independent knowledge and incorporate them into our neural architecture. Experiments on two low resource languages (Dutch and Spanish) demonstrate the effectiveness of these additional semantic representations (average 4.8% improvement). Moreover, on Chinese OntoNotes 4.0 dataset, our approach achieves an F-score of 83.07% with 2.91% absolute gain compared to the state-of-the-art systems.

1 Introduction

Named entity recognition (NER) is defined as the extraction of a contiguous sequence of textual tokens, which represents the name of an object of a specified class, such as person, location or organization. It plays a vital role in the overall task of Information Extraction (IE) and serves as an intermediate step for subsequent IE tasks, like Relation Extraction and Entity Linking. Current state-of-the-art methods for English NER usually use deep learning algorithms, e.g., Feedforward neural network (FNN) or Recurrent neural network (RNN) [Huang *et al.*, 2015], and build name taggers from annotated data with accompanying entity labels. Such models generalize well on new entities based on features automatically learned from the context. However, a neural-based NER system could still get lower performance if its hidden feature representations cannot be learned adequately, which happens frequently when the annotated data is not enough, especially in low resource scenario [Zhang *et al.*, 2016]. In this paper, we regard English as a high resource language, and other languages, such as Dutch, Spanish, and even Chinese, as low resource languages.



English Translation: Federal Reserve Chairman is Ben Bernanke.

Figure 1: Example of NER labels with bilingual lexicon.

To improve the performance of low resource NER, we present a neural-based sequential tagger which incorporates additional word representations learned from semantic projections based on cross-lingual knowledge. This approach is partly inspired by the previous empirical success of feature-based sequence labeling models with bilingual constraints/inferences for Chinese and English NER [Che *et al.*, 2013; Wang *et al.*, 2013]. Our approach is built on the state-of-the-art LSTM-CRF framework [Lample *et al.*, 2016], which models each word of the input sentence with a contextual embedding based on Bi-LSTM, and then assigns an entity label for each word using CRF. Compared with previous work which are only based on general word embeddings, we design three strategies to enrich the semantic representations, which embodies our main contributions:

(1) We build neural networks to model the external semantic representation of each low resource language word based on the translations from high resource languages. The intuition is that different languages usually contain complementary cues about entities and these cues can be further transferred through bilingual lexicons. Figure 1 shows a simple example for Chinese name tagging. The word “本” is common in Chinese but rarely appears as a name. However, based on a Chinese-English dictionary, one of the English translation candidates of “本” is “Ben”, which provides a strong semantic clue that the word is a person name in English.

*Corresponding author.

(2) The lexicon is usually limited and cannot cover all low resource language words. Thus, we further design a new strategy to extend the bilingual mappings with a linear transformation function. After generating the external semantic representations for the low resource language words based on high resource translations (the output of the previous network), we learn a linear projection function between the low resource word embedding space and the high resource language translation semantic space, accordingly, the out-of-lexicon low resource language words can also be estimated with a new semantic representation.

(3) For each word, we calculate its distributional probabilities over all entity types and add them as additional features to the original word representation for both low resource and high resource languages. The rationale is that the entity type distribution can be regarded as a language-independent knowledge and it is helpful for low resource NER. For example, both the English and Chinese are describing the same entity, even probably with different spelling (e.g., “United States” in English vs. “美国” in Chinese), the entity type of that entity does not change from one language to another [Ni and Florian, 2017]. And if we know that “United States” is a location in English space, then naturally we can predict that the entity type of Chinese word “美国” prefers location.

Experiments on two low resource languages (Dutch and Spanish) demonstrate that the additional semantic representations can bring in an average 4.8% F-score gain compared with general word embeddings. Moreover, we also conduct experiments on the Chinese portion of the OntoNotes 4.0 corpus. The results show that we achieve a 2.91% improvement compared to the state-of-the-art system.

2 Methodology

In this section, we first describe the background on LSTM-CRF model, which is the backbone of our approach. Afterwards, we present two neural networks to learn the cross-lingual semantic representation of each low resource language word based on high resource language translations. Lastly, we introduce a lexicon extension strategy to alleviate the out-of-lexicon problem and describe how to learn the entity type distribution based on original word representations and entity type representations in both languages.

2.1 Basic Model: LSTM-CRF

LSTM-CRF model is originally introduced by [Huang *et al.*, 2015], which takes a sequence of elements as the input and outputs a sequence of category labels corresponding to the input sequence[1]. The idea has been successfully applied in POS Tagging, Chunking and NER [Liu *et al.*, 2017a].

The approach of [Lample *et al.*, 2016] is based on LSTM and CRF Tagging models. An illustration of this network is given in the left of Figure 2. The LSTM is a special form of recurrent neural networks (RNNs) with three gated units, *input, output* and *forget*, which could control the passing of information along the sequence and thus improves the modeling of long-range dependencies. Following [Lample *et al.*, 2016], they take a sequence of vectors $\mathtt{X} = \{\mathtt{x}_1, ..., \mathtt{x}_i, ..., \mathtt{x}_n\}$ as input and return another sequence $\mathtt{H} = \{\mathtt{h}_1, ..., \mathtt{h}_i, ..., \mathtt{h}_n\}$ that represents some information about the sequence at every step in the input. For brevity, the details of LSTM equations are given in [Gers *et al.*, 1999]. The conditional random field (CRF) [Jurafsky and Martin, 2000] is a probabilistic graphical model, which works in a sequential way and predict a label sequence $\mathtt{y} = \{y_1, ..., y_i, ..., y_n\}$ corresponding to the input sequence $\mathtt{X}$. They define a score function as follows:

$$s(\mathtt{X}, \mathtt{y}) = \sum_{i=0}^{n} \mathtt{A}_{y_i, y_{i+1}} + \sum_{i=1}^{n} \mathtt{H}_{i, y_i} \tag{1}$$

where $\mathtt{A}$ is a matrix of transition scores such that $\mathtt{A}_{i,j}$ represents the score of a transition from the tag i to tag j. y_0 and y_n are the start and end tags of a sentence. Matrix $\mathtt{A}$ is therefore a square matrix of size $k+2$, k is the number of tags. A softmax over all possible tag sequences yields a probability for the sequence $\mathtt{y}$:

$$p(\mathtt{y}|\mathtt{X}) = \frac{\exp^{s(\mathtt{X}, \mathtt{y})}}{\sum_{\tilde{\mathtt{y}} \in \mathtt{Y}_\mathtt{X}} \exp^{s(\mathtt{X}, \tilde{\mathtt{y}})}} \tag{2}$$

where $\mathtt{Y}_\mathtt{X}$ represents all possible tag sequences for the input sequence $\mathtt{X}$.

Furthermore, [Lample *et al.*, 2016] incorporated character-level structure into word representation. Each input vector $\mathtt{x}_i$ consists of two parts, pre-trained word-level representation $\mathtt{w}_i$ [Mikolov *et al.*, 2013] and task-related character-level representation $\mathtt{c}_i$. They adopted a bidirectional LSTM to capture information in both forward and backward directions and concatenate the outputs of these two LSTMs as $\mathtt{c}_i$.

2.2 Improved with Bilingual Lexicon

We present an overview of the developed networks for modeling bilingual lexicons, as illustrated in the right of Figure 2. Following the same setting in Section 2.1, given a low resource language sentence $\boldsymbol{X} = \{x_1, x_2, ..., x_i, ..., x_n\}$, we assume that each word x_i has a corresponding high resource language translation T_i based on the bilingual lexicon [2]. The translation T_i can be viewed as a combination of multiple translation items and each translation item consists of multiple high resource language words. To make better understanding of the high resource language translation of a low resource word, all translation items should be encoded into the encoder. One simple way is to take the concatenation of all translation words as the input for the vanilla RNN unit [Liu *et al.*, 2017b]. We also map each high resource language translation word into its embedding vector. Therefore, translation word vectors $\{\mathtt{t}_1, ..., \mathtt{t}_i, ..., \mathtt{t}_l\}$ are stacked and regarded as the translation memory unit $\mathtt{T} \in \mathbb{R}^{d \times l}$, where l is the number of all translation words. An example is given in Figure 1. The Chinese word “美联储” has two translation items in English, namely “FED” and “Federal Reserve”. We can get its translation sequence as [“FED”, “Federal”, “Reserve”], of which size is 3.

[1] Entity types are usually represented in BIOES format (which stand for *Begin, Inside, Outside, End*, and *Single*, indicating the position of the token in the entity) as this scheme has been reported to outperform others such as BIO [Ratinov and Roth, 2009].

[2] We construct bilingual lexicons from online translators such as Bing Dict and FAIR (Facebook AI Research) dictionary.

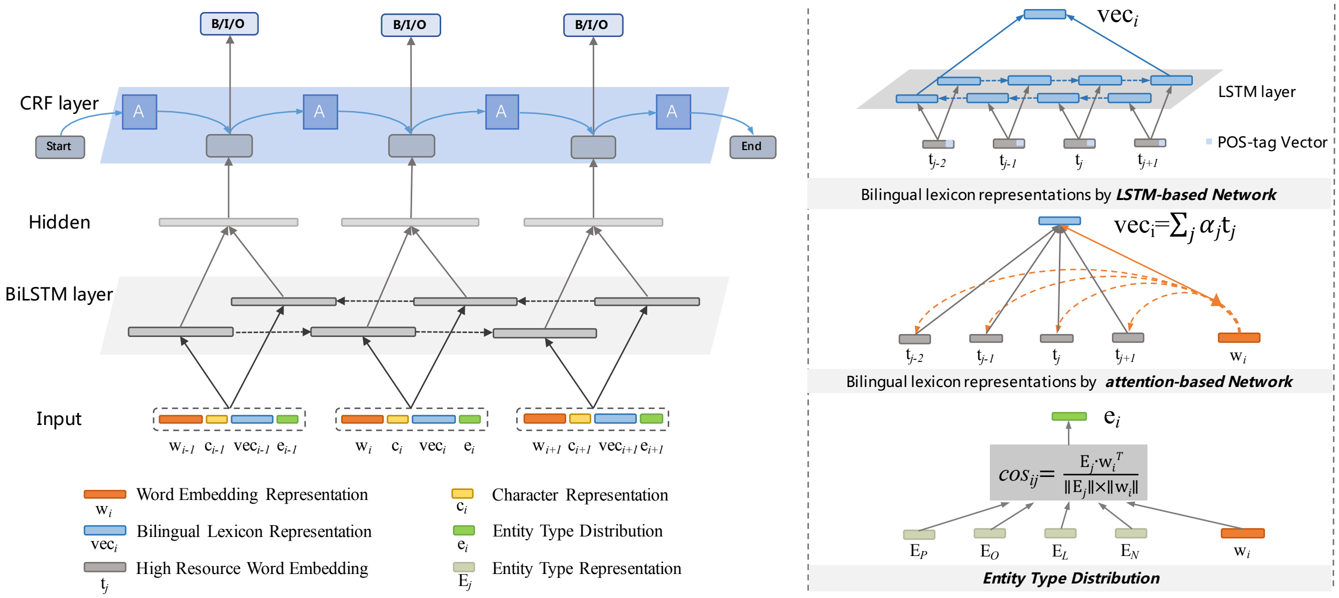

B/I/O
B/I/O
B/I/O
CRF layer
A
A
A
A
Start
End
Hidden
BiLSTM layer
Input
w_{i-1} c_{i-1} vec_{i-1} e_{i-1}
w_i c_i vec_i e_i
w_{i+1} c_{i+1} vec_{i+1} e_{i+1}
w_i Word Embedding Representation
vec_i Bilingual Lexicon Representation
t_j High Resource Word Embedding
c_i Character Representation
e_i Entity Type Distribution
E_j Entity Type Representation
vec_i
LSTM layer
POS-tag Vector
t_{j-2} t_{j-1} t_j t_{j+1}
Bilingual lexicon representations by LSTM-based Network
vec_i=Σ_j α_j t_j
t_{j-2} t_{j-1} t_j t_{j+1} w_i
Bilingual lexicon representations by attention-based Network
e_i
cos_ij= E_j·w_i^T / ‖E_j‖×‖w_i‖
E_P E_O E_L E_N w_i
Entity Type Distribution

Figure 2: Main architecture of our model. In this framework, the translation representation of low resource word $\texttt{vec}_i$ can be modeled by two approaches, LSTM-based network or attention-based model.

Following [Liu *et al.*, 2017b], we first present a LSTM-based network for modeling bilingual lexicons. To better encode the structural information of different translation items, we incorporate the POS-tag information of each translation item into their corresponding translation words. In our method, each POS-tag label is also mapped to a d_p (a hyperparameter) dimensional vector, which is randomly initialized and optimized by the model. Then we combine the embedding of each translation word t_i and its corresponding POS-tag vector $\texttt{p}_i$ as $[\texttt{t}_i, \texttt{p}_i]$, and then feed it to the Bi-LSTM unit. Finally, we concatenate the outputs of forward LSTM and backward LSTM as translation representations vec. We name this model as *LSTM-CRF+BL*^{LSTM}, as illustrated in the upper-right of Figure 2.

Considering that each word in the translation does not contribute equally to the semantic meaning of the original low resource word, we further introduce an attention-based network to model the bilingual lexicons, which is similar as the attention-based memory network in question answering [Sukhbaatar *et al.*, 2015]. In detail, taking an external translation unit $\texttt{T} \in \mathbb{R}^{d\times l}$ and a low resource word vector $\texttt{x}_i \in \mathbb{R}^d$ as input, the attention model outputs a continuous vector $vec \in \mathbb{R}^d$, which is a weighted sum of each piece of memory in $\texttt{T}$:

$$vec = \sum_{j=1}^{l} \alpha_j \texttt{t}_j \tag{3}$$

where l is the memory unit size, $\alpha_j \in [0, 1]$ is the weight of t_j and $\sum_j \alpha_j = 1$. We implement a neural network based attention model based on previous work [Bahdanau *et al.*, 2014]. For each piece of translation memory $\texttt{t}_j$, we use a feed forward neural network to compute its semantic relatedness with the low resource word. The scoring function is calculated as follows:

$$g_j = \tanh(\texttt{W}_{att}\texttt{x}_i + \texttt{U}_{att}\texttt{t}_j + b_{att}) \tag{4}$$

where $\texttt{W}_{att} \in \mathbb{R}^d$, $\texttt{U}_{att} \in \mathbb{R}^d$ and $b_{tt} \in \mathbb{R}^{1\times 1}$.

After obtaining $g_1, g_2, ...g_l$, we feed them to a $softmax$ function to calculate the final importance distribution $\alpha_1, \alpha_2, ...\alpha_l$. We name this model as *LSTM-CRF+BL*^{ATT}, as illustrated in the right-middle of Figure. 2

$$\alpha_j = \frac{\exp(g_j)}{\sum_{z=1}^{l} \exp(g_z)} \tag{5}$$

2.3 Improved with Mapping based Lexicon Extension Strategy

In the actual situation, the bilingual lexicons can not cover all low resource language words. To overcome this challenge, we design a lexicon extension strategy to estimate the translation representations of out-of-lexicon word.

Suppose there is a low resource language word set $\texttt{W} = \{w_1, ..., w_i, ..., w_f\}$, each word has a low resource word vector $\texttt{w}_i$ and a high resource language translation vector $\texttt{vec}_i$. We learn a linear projection function as the transformation between the two semantic space, as follows:

$$\texttt{vec}_i = \texttt{Mw}_i \tag{6}$$

where $\texttt{M}$ is the mapping matrix. We minimize the following objective to optimize $\texttt{M}$:

$$loss_M = \sum_{i=1}^{f} ||\texttt{vec}_i - \texttt{Mw}_i||_2 \tag{7}$$

After obtain $\texttt{M}$, for each out-of-lexicon word o_i, we can estimate the translation representation veo_i as follows:

$$\texttt{veo}_i = \texttt{Mo}_i \tag{8}$$

2.4 Improved with Language-Independent Entity Type Distribution

In this section, we introduce the entity type-based distributional features, which denotes the probabilities of each word to be tagged as each entity type. Word embeddings have been empirically shown to preserve linguistic regularities, such as similar words tend to be close to each other in the same space [Mikolov *et al.*, 2013]. We observe that the same property also applies to the words with the same entity type. For example, the distance between the word "Microsoft" and "Bill Gates" is larger than that between "Microsoft" and "IBM". Therefore, we can learn an approximate representation of each entity type, and use the similarities between each entity type representation and each word embedding as the entity type-based distributional feature.

In this work, we focus on three most common named entity types, i.e., P (Person), L (Location), O (Organization), and discard the others. Taking low resource language as an example, we randomly select 10 entities from each entity type and average their embeddings as the entity type representation. At the same time, we randomly generate one vector representing non-entity; Therefore, four entity type vectors $\{\mathtt{E}_P, \mathtt{E}_O, \mathtt{E}_L, \mathtt{E}_N\}$ are constructed, each $\mathtt{E}_j \in \mathbb{R}^d$. Afterwards, we use standard $cosine$ function to calculate the semantic relatedness between the low resource word embedding $\mathtt{w}_i$ and the entity type representation $\mathtt{E}_j$.

$$e_{ij} = \frac{\mathtt{w}_i^T \cdot \mathtt{E}_j}{||\mathtt{w}_i|| \times ||\mathtt{E}_j||} \tag{9}$$

For high resource language, we also calculate the entity distribution of each word. In the end, each low resource language word and each high resource language word are assigned with an entity distributional feature vector with dimensionality 4, $\mathtt{e}_i = \{e_P, e_O, e_L, e_N\}$, as illustrated in the bottom-right of Figure 2.

2.5 Low Resource NER

Now, the vectors of each word in the input low resource sentence is made up of four parts: a word embedding $\mathtt{w}_i$, a character-level representation $\mathtt{c}_i$, a high resource translation vector $\mathtt{vec}_i$ or $\mathtt{veo}_i$, and an entity type distributional representation $\mathtt{e}_i$. we regard the concatenation vectors of these four representations as word representation $\mathtt{x}_i = [\mathtt{w}_i, \mathtt{c}_i, \mathtt{vec}_i, \mathtt{e}_i]$, and feed them into the previous LSTM-CRF model for NER (Section 2.1). The model is trained in a supervised manner by minimizing the cross entropy error of sequence labeling and L2 loss:

$$loss = -\sum_{\mathtt{X} \in \mathtt{C}} \sum_{\mathtt{y} \in \tilde{\mathtt{Y}}_{\mathtt{X}}} p^g(\tilde{\mathtt{y}}|\mathtt{X}) \mathtt{log}(p(\tilde{\mathtt{y}}|\mathtt{X})) + loss_M \tag{10}$$

where $p(\mathtt{y}|\mathtt{X})$ is the probability of predicting sequence $\mathtt{X}$ as tag sequence $\tilde{\mathtt{y}}$. $\mathtt{C}$ denotes all training sentences. $\mathtt{X}$ is the input sentence representations. $\mathtt{Y}_{\mathtt{X}}$ represents all possible tag sequences for the input sequence. $p^g(\tilde{\mathtt{y}}|\mathtt{X})$ is 1 or 0, indicating whether the correct sequence tag is $\tilde{\mathtt{y}}$. We use back propagation to calculate the gradients of all the parameters, and update them with stochastic gradient descent. We randomize other parameters with uniform distribution $U(0.01, 0.01)$, and set the learning rate as 0.01. Table 1 illustrates the word embedding parameters used in our experiments[3]. For brevity, the details of other parameters are given in our codes [4].

Languages	Embedding Corpus	Embedding dimension	Vocabulary size
Dutch	FAIR	300	33315
Spanish	FAIR	300	31673
Chinese	Gigword V5	300	66785
English	Gigword V5	300	97473

Table 1: Embedding parameters used in our experiments on four languages.

Languages	Dataset	Train	Dev	Test
Dutch	CoNLL-2002	15520	2822	5077
Spanish	CoNLL-2002	8323	1914	1517
Chinese	Ontonotes 4.0	22761	3903	2730

Table 2: # of sentences.

3 Experiment

We apply our neural architecture for NER on various datasets and evaluate the effectiveness separately. In this section, we will describe the detailed experimental settings and discuss the results.

3.1 Dataset

We evaluate the proposed approach on two low resource languages (including Spanish and Dutch[5]), and Chinese[6] , which is distinct from Latin-based languages. In this paper, we regard English as high resource language and all other languages as low resource languages. Table 2 shows the detailed description of the data sets used in our experiments. In this paper, we focus on four entity types (Person, Location, Organization, None), which are commonly adopted in previous NER studies [Che *et al.*, 2013; Wang *et al.*, 2013].

3.2 Low Resource NER

We compare with the following baseline methods on the two languages.

- *LSTM-CRF* [Lample *et al.*, 2016] is introduced in section 2.1. Compared with standard Bi-LSTM, it adds a CRF layer to impose several hard constraints of the "grammar".
- *LM-LSTM-CRF* [Liu *et al.*, 2017a] is also a LSTM-CRF-based sequence labeling framework and incorporates residual network and language model to extract character-level knowledge from the self-contained order information.
- *CLNER* (Cross-Lingual Named Entity Recognition) [Ni *et al.*, 2017], a weakly supervised method, which creates automatically labeled NER data for a target language via

[3] FAIR:https://github.com/facebookresearch/MUSE

[4] Our code is available at: https://github.com/scir-code/lrner.

[5] CoNLL: https://github.com/synalp/NER/tree/master/corpus/

[6] Ontonotes: https://catalog.ldc.upenn.edu/ldc2011t03

annotation projection on comparable corpora. The recognizer is a prototype-based neural model.

Our model has several variations, which are detailed below.

- *LSTM-CRF+BL*^{LSTM} extends *LSTM-CRF* by taking into account of the bilingual lexicon, and uses a LSTM-based network towards the translations.
- *LSTM-CRF+BL+M*^{LSTM}: an extension of *LSTM-CRF+BL*^{LSTM} by further incorporating the lexicon extension strategy.
- *LSTM-CRF+BL+M+E*^{LSTM}: an extension of *LSTM-CRF+BL+M*^{LSTM} by further concatenating the semantic representation of each word with entity type distribution features in both languages.
- *LSTM-CRF+BL*^{ATT} extends *LSTM-CRF* by taking into account of the bilingual lexicon, and uses an attention-based network towards the translations.
- *LSTM-CRF+BL+M*^{ATT}: an extension of *LSTM-CRF+BL*$^{+ATT}$ by further incorporating the lexicon extension strategy.
- *LSTM-CRF+BL+M+E*^{ATT}: an extension of *LSTM-CRF+BL+M*$^{+ATT}$ by further concatenating the semantic representation of each word with entity type distribution features in both languages.

Experimental results are given in Table 3. Evaluation metric is F measure [Manning and Schütze, 1999]. We can find that our method *LSTM-CRF+BL+M+E*^{LSTM} yields the best performance on two languages compared with many strong baselines. The performance of *CLNER* are relatively low because both of them utilize indirectly acquired features based on linguistic resources and cross-lingual entity mappings. *LSTM-CRF* obtains significant improvement over *CLNER* by integrating semantic representations of low resource language words and learning the constraints between entity labels. Besides, we surprisingly find that all of our variants outperform the strong baseline *LSTM-CRF* on two languages, which demonstrate the effectiveness of each of the additional semantic representations for low resource NER. Among the six variants of our model, LSTM-based models perform better than attention-based models, which indicates that the sequence feature is more important for modeling lexicon structure. In the last, two real examples are given in Table 4 to demonstrate the effectiveness of the additional bilingual lexicon representations for low resource NER.

Model	Dutch	Spanish
LSTM-CRF	81.74	83.41
LM-LSTM-CRF	86.24	85.13
CLNER	69.30	65.50
LSTM-CRF+BL^{LSTM}	86.48	86.02
LSTM-CRF+BL+M^{LSTM}	87.94	86.03
LSTM-CRF+BL+M+E^{LSTM}	**88.39**	**86.42**
LSTM-CRF+BL^{ATT}	82.46	83.91
LSTM-CRF+BL+M^{ATT}	83.83	84.27
LSTM-CRF+BL+M+E^{ATT}	86.07	85.34

Table 3: Comparison of different methods on low resource NER.

Example 1 Dutch	*Some are in the George Grard Foundation.* Een aantal is in de Stichting George Grard.							
Golden	O	O	O	O	O	B-ORG	I-ORG	E-ORG
LSTM-CRF	O	O	O	O	O	B-LOC	I-LOC	E-LOC
Our model	O	O	O	O	O	B-ORG	I-ORG	E-ORG
Example 2 Spanish	*The delegate of the Andalusian Gov in Cádiz,* El delegado del Gobierno andaluz en Cádiz,							
Golden	O	O	O	S-ORG	O	O	S-LOC	
LSTM-CRF	O	O	O	O	O	O	O	
Our model	O	O	O	S-ORG	O	O	S-LOC	

Table 4: Case study for Dutch and Spanish NER (Italic Sentences Show the English Translations for the Dutch and Spanish Examples). Our model is *LSTM-CRF+BL+M*^{LSTM}.

3.3 Chinese NER

Chinese	Precision	Recall	F-score
Soft-Align	77.37	71.13	74.13
LSTM-CRF	82.58	76.92	79.65
LM-LSTM-CRF	81.90	78.50	80.16
LSTM-CRF+BL^{LSTM}	82.01	82.81	82.41
LSTM-CRF+BL+M^{LSTM}	82.05	83.24	82.64
LSTM-CRF+BL+M+E^{LSTM}	**82.84**	**83.32**	**83.07**
LSTM-CRF+BL^{ATT}	81.72	80.68	81.20
LSTM-CRF+BL+M^{ATT}	82.41	80.44	81.42
LSTM-CRF+BL+M+E^{ATT}	82.46	80.97	81.71

Table 5: Comparison of different methods on Chinese NER.

To demonstrate the effectiveness of our models on large-scale corpora, we show the results on Chinese Ontonotes 4.0 NER in Table 5. Additionally, we add a strong baseline for Chinese NER with bilingual constraints, namely Soft-Align[7]. From Table 5, we can still get consistent improvements on Chinese NER over previous state-of-the-art methods. Specifically, we observe that *LSTM-CRF+BL+M+E*^{LSTM} achieves a significant gain in Recall. This is reasonable since the semantic representation of each Chinese word is much richer in *LSTM-CRF+BL+M+E*^{LSTM} than other models.

3.4 Fine-Grained Performance on Different Groups

This subsection studies the effectiveness of our cross-lingual representations. For comparison purposes, we select the baselines: *LSTM-CRF* [Lample *et al.*, 2016] and *LM-LSTM-CRF* [Liu *et al.*, 2017a] and compare them with LSTM-based bilingual lexicon models. Moreover, to prove that cross-lingual representation could capture more valuable semantics, especially for the entities that appear in the testing data but never appear in the training data, we divide the entities in the testing data into two parts (**A**: appearing in both testing and training data with the same entity type, or **B**: appearing in testing data only) and perform evaluations separately. Experimental results are shown in Table 6 and illustrate that for all situations, cross-lingual representation brings in significant improvements compared with word embedding and character-level representation in NER. For situation **B** in three lan-

[7] Che *et al.* proposed a novel Integer Linear Programming-based inference algorithm with bilingual constraints for English and Chinese NER.

Model	Dutch			Spanish			Chinese		
	A	**B**	All	**A**	**B**	All	**A**	**B**	All
	53.56%	46.44%	100%	68.30%	31.70%	100%	72.81%	27.19%	100%
LSTM-CRF	92.31	54.71	81.74	85.72	67.43	83.41	83.48	64.22	79.65
LM-LSTM-CRF	96.50	58.85	86.24	87.06	69.87	85.13	84.09	65.46	80.16
LSTM-CRF+BL^{LSTM}	96.93	58.97	86.48	87.62	70.97	86.02	85.39	68.65	82.41
LSTM-CRF+BL+M^{LSTM}	96.98	62.20	87.94	87.66	71.07	86.03	85.83	68.92	82.64
LSTM-CRF+BL+M+E^{LSTM}	**97.00**	**64.10**	**88.39**	**87.93**	**71.69**	**86.42**	**86.02**	**69.89**	**83.07**

Table 6: Comparison of the results for *LSTM-CRF*, *LM-LSTM-CRF* and our LSTM-based networks. **A** denotes the entities appearing in both training and test datasets, and **B** indicates all other cases. Evaluation metric is F measure.

guages, our model (*LSTM-CRF+BL+M+E*^{LSTM}) yields an average 6.44% improvement, which is 2 times in situation **A**. This demonstrates that the cross-language representation has better ability to model non-covered entities than word-level and character-level representations.

4 Related Work

There exist two threads of related work regarding the topics in this paper, which are Monolingual NER and how to improve it with other languages (Cross-lingual NER).

4.1 Monolingual NER

Named entity recognition is typically regarded as a kind of sequence labeling problem in literature. Therefore, standard feature-based classification approaches such as conditional random fields (CRFs) [Lafferty *et al.*, 2001], hidden markov models (HMMs) [Florian *et al.*, 2003] and maximum entropy classifiers [Chieu and Ng, 2002] can be naturally employed to build a name tagger. Despite the effectiveness of feature engineering, it is labor intensive and unable to discover the discriminative or explanatory factors of data [Bengio *et al.*, 2015]. To handle this problem, some recent studies [Chiu and Nichols, 2015; Santos and Guimarães, 2015; Huang *et al.*, 2015; Yang *et al.*, 2017; Lample *et al.*, 2016; Liu *et al.*, 2017a; Peters *et al.*, 2017; Ma and Hovy, 2016] used neural network methods and got promising results . The representative approaches include BiLSTM-CNN [Chiu and Nichols, 2015; Santos and Guimarães, 2015], CNN-CRF [Huang *et al.*, 2015; Yang *et al.*, 2017], LSTM-CRF [Lample *et al.*, 2016; Liu *et al.*, 2017a; Peters *et al.*, 2017] and LSTM-CNN-CRF [Ma and Hovy, 2016]. Furthermore, character-based representations had been proved to be effective in capturing the orthographic and morphological evidence. Also, most of these models added a CRF layer, and reported significant improvement over pure RNN models. Our architecture is based on the success of LSTM-CRF model and is further modified to enrich the word representation with cross-lingual knowledge information.

4.2 Cross-lingual NER

The idea of utilizing multilingual resources to improve monolingual name tagger systems has been studied extensively. [Li *et al.*, 2012] presented a cyclic CRF model and performed approximate inference using loopy belief propagation. Although, their feature-rich CRF formulation of bilingual edge is powerful, an obvious drawback of this approach is the requirement of manually annotate bilingual NER data. Therefore, [Chen *et al.*, 2010] proposed approaches to extract bilingual named entity pairs from unannotated bitext. The verification was based on bilingual named entity dictionaries. In this regard, one of the most interesting papers is [Burkett *et al.*, 2010], which explored an "up-training" mechanism by using the outputs from a strong monolingual model as groundtruth, and thereby simulated a learning environment, where a bilingual model is trained to help a "weakened" monolingual model recover the results of the strong model. [Kim *et al.*, 2012] proposed a method of labeling bilingual corpora with named entity labels automatically based on Wikipedia. [Che *et al.*, 2013; Wang *et al.*, 2013] tackled the problem of jointly recognizing and aligning bilingual named entities. For low resource NER, [Zhang *et al.*, 2016] proposed an expectation-driven model that designed a large number of language-specific features (rules, patterns, gazetteers, etc.) via consulting and encoding linguistic knowledge from native speakers. [Ni and Florian, 2017; Ni *et al.*, 2017] developed approaches to improve multilingual name tagging performances with Wikipedia entity type mapping and word distribution mapping. However, these methods suffer from error propagation. Moreover, the selection and collection of task related features are time-consuming and labor intensive. Our approach differs in that it does not acquire any hand-craft features and bilingual lexicons are one of the most basic language resources for all languages.

5 Conclusions

Low resource NER is a very important yet challenging problem in natural language processing. In this paper, we focus on this problem by incorporating cross-lingual knowledge into a neural architecture, which guides low resource name tagging to achieve a better performance. Specifically, we use bilingual lexicons to bridge cross-lingual semantic mapping and design a lexicon extension strategy to alleviate the out-of-lexicon problem. Moreover, we regard entity type distribution as language-independent features and model them in our architecture. Experiments on three languages, namely, Dutch, Spanish and Chinese demonstrate the effectiveness of our model for low resource language NER. In the future, we will incorporate other knowledge resources, such as FrameNet and WordNet, from high resource languages into our neural architecture. We will also extend our architecture to other NLP tasks, such as event extraction, sentiment analysis.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) via grant 61632011, 61772156 and 61472107.

References

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, 2014.

[Bengio *et al.*, 2015] Yoshua Bengio, Ian J Goodfellow, and Aaron Courville. Deep learning. *Nature*, 2015.

[Burkett *et al.*, 2010] David Burkett, Slav Petrov, John Blitzer, and Dan Klein. Learning better monolingual models with unannotated bilingual text. In *CONLL*, pages 46–54, 2010.

[Che *et al.*, 2013] Wanxiang Che, Mengqiu Wang, Christopher D Manning, and Ting Liu. Named entity recognition with bilingual constraints. In *HLT-NAACL*, 2013.

[Chen *et al.*, 2010] Yufeng Chen, Chengqing Zong, and Keh-Yih Su. On jointly recognizing and aligning bilingual named entities. In *ACL*, pages 631–639, 2010.

[Chieu and Ng, 2002] Hai Leong Chieu and Hwee Tou Ng. Named entity recognition: a maximum entropy approach using global information. In *COLING*, pages 1–7, 2002.

[Chiu and Nichols, 2015] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*, 2015.

[Florian *et al.*, 2003] Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. Named entity recognition through classifier combination. In *HLT-NAACL*, pages 168–171, 2003.

[Gers *et al.*, 1999] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

[Huang *et al.*, 2015] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[Jurafsky and Martin, 2000] Dan Jurafsky and Jameás H Martin. *Speech & language processing*. Pearson Education India, 2000.

[Kim *et al.*, 2012] Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *ACL*, 2012.

[Lafferty *et al.*, 2001] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, volume 1, pages 282–289, 2001.

[Lample *et al.*, 2016] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[Li *et al.*, 2012] Qi Li, Haibo Li, Heng Ji, Wen Wang, Jing Zheng, and Fei Huang. Joint bilingual name tagging for parallel corpora. In *CIKM*, pages 1727–1731. ACM, 2012.

[Liu *et al.*, 2017a] Liyuan Liu, Jingbo Shang, Frank Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. Empower sequence labeling with task-aware neural language model. *arXiv preprint arXiv:1709.04109*, 2017.

[Liu *et al.*, 2017b] Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. Table-to-text generation by structure-aware seq2seq learning. *CoRR*, 2017.

[Ma and Hovy, 2016] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[Manning and Schütze, 1999] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[Ni and Florian, 2017] Jian Ni and Radu Florian. Improving multilingual named entity recognition with wikipedia entity type mapping. *EMNLP*, 2017.

[Ni *et al.*, 2017] Jian Ni, Georgiana Dinu, and Radu Florian. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. *ACL*, page 1470–1480, 2017.

[Peters *et al.*, 2017] Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*, 2017.

[Ratinov and Roth, 2009] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

[Santos and Guimarães, 2015] Cicero Nogueira dos Santos and Victor Guimarães. Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv:1505.05008*, 2015.

[Sukhbaatar *et al.*, 2015] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *NIPS*, pages 2440–2448, 2015.

[Wang *et al.*, 2013] Mengqiu Wang, Wanxiang Che, and Christopher D Manning. Effective bilingual constraints for semi-supervised learning of named entity recognizers. In *AAAI*. Citeseer, 2013.

[Yang *et al.*, 2017] Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. Transfer learning for sequence tagging with hierarchical recurrent networks. *CoRR*, 2017.

[Zhang *et al.*, 2016] Boliang Zhang, Xiaoman Pan, Tianlu Wang, Ashish Vaswani, Heng Ji, Kevin Knight, and Daniel Marcu. Name tagging for low-resource incident languages based on expectation-driven learning. In *Proceedings of NAACL-HLT*, pages 249–259, 2016.