Low-Resource Sequence Labeling via Unsupervised Multilingual Contextualized Representations

Zuyi Bao[‡], **Rui Huang**[◇], **Chen Li**[‡*] and **Kenny Q. Zhu**[†]
[‡]Alibaba Group
[◇]Zhejiang University, China
[†]Shanghai Jiao Tong University, Shanghai, China
[‡]{zuyi.bzy,puji.lc}@alibaba-inc.com
[◇]iurgnauh@zju.edu.cn, [†]kzhu@cs.sjtu.edu.cn

Abstract

Previous work on cross-lingual sequence labeling tasks either requires parallel data or bridges the two languages through word-by-word matching. Such requirements and assumptions are infeasible for most languages, especially for languages with large linguistic distances, e.g., English and Chinese. In this work, we propose a Multilingual Language Model with deep semantic Alignment (MLMA) to generate language-independent representations for cross-lingual sequence labeling. Our methods require only monolingual corpora with no bilingual resources at all and take advantage of deep contextualized representations. Experimental results show that our approach achieves new state-of-the-art NER and POS performance across European languages, and is also effective on distant language pairs such as English and Chinese.[1]

1 Introduction

Sequence labeling tasks such as named entity recognition (NER) and part-of-speech (POS) tagging are fundamental problems in natural language processing (NLP). Recent sequence labeling models achieve state-of-the-art performance by combining both character-level and word-level information (Chiu and Nichols, 2016; Ma and Hovy, 2016; Lample et al., 2016). However, these models heavily rely on large-scale annotated training data, which may not be available in most languages. Cross-lingual transfer learning is proposed to address the label scarcity problem by transferring annotations from high-resource languages (source languages) to low-resource languages (target languages). In this scenario, a major challenge is how to bridge interlingual gaps with modest resource requirements.

There is a large body of work exploring cross-lingual transfer through language-independent features, such as morphological features and universal POS tags for cross-lingual NER (Tsai et al., 2016) and dependency parsers (McDonald et al., 2011). However, these approaches require linguistic knowledge for language-independent feature engineering, which is expensive in low-resource settings. Other work relies on bilingual resources to transfer knowledge from source languages to target languages. Parallel corpora are widely used to project annotations from the source to the target side (Yarowsky et al., 2001; Ehrmann et al., 2011; Kim et al., 2012; Wang and Manning, 2014). These methods could achieve strong performance with a large amount of bilingual data, which is scarce in low-resource settings.

Recent research leverages cross-lingual word embeddings (CLWEs) to establish inter-lingual connections and reduce the requirements of parallel data to a small lexicon or even no bilingual resource (Ni et al., 2017; Fang and Cohn, 2017; Xie et al., 2018). However, word embedding spaces may not be completely isomorphic due to language-specific linguistic properties, and therefore cannot be perfectly aligned. For example, different from English, Chinese nouns do not distinguish singular and plural forms, while Spanish nouns distinguish masculine and feminine.

On the other hand, NER tags such as person names, organizations, and locations are shared across different languages. Language-independent frameworks such as universal conceptual cognitive annotation (Abend and Rappoport, 2013), universal POS (Petrov et al., 2011a), and universal dependencies (Nivre et al., 2016) are defined to represent different languages in a unified formation. These work serves as our motivation to assume

*Corresponding Author.
† Kenny Q. Zhu was partially supported by NSFC grant 91646205 and Alibaba visiting scholar program.
[1]The code is released at https://github.com/baozuyi/MLMA.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1028–1039, Hong Kong, China, November 3–7, 2019.

that the semantic meanings of words from different languages can be roughly aligned at a conceptual level and it is more reasonable to align deep semantic representations instead of shallow word embeddings. Meanwhile, monolingual contextualized embeddings derived from language models have shown to be effective for extracting semantic information and have achieved significant improvement on several NLP tasks (Peters et al., 2018).

In this paper, we propose a Multilingual Language Model with deep semantic Alignment (MLMA). We train MLMA on monolingual corpora from each language and align its internal states across different languages. Then MLMA is utilized to generate language-independent representations and to bridge the gaps between high-resource and low-resource languages. For evaluation, we conduct extensive experiments on the NER and POS benchmark datasets under cross-lingual settings. The experiment results show that our methods achieve substantial improvements comparing to previous state-of-the-art methods in European languages. We also validate our approaches on a distant language pair, English-Chinese, and the results are competitive with previous methods which use large-scale parallel corpora. Our contributions are as follows:

1. Instead of word-level alignment, we propose MLMA that uses contextualized representations to bridge the inter-lingual gaps.
2. We propose three methods to align contextualized representations without any bilingual resource.
3. Our methods achieve new state-of-the-art performance on cross-lingual NER and POS tasks in European languages, and very competitive results for English-Chinese NER, where previous work uses large parallel data.

2 Approach

Our approach belongs to the model transfer (Section 5.2) and mainly consists of three steps:

1. Training a multilingual language model with alignment (MLMA) using monolingual corpora of the source and the target languages. (Section 2.1, 2.2 and 2.3)
2. Building a cross-lingual sequence labeling model based on the language-independent representations from the MLMA. (Section 2.4 and 2.5)
3. Learning the cross-lingual sequence labeling model (with MLMA fixed) on the annotated data of source languages and directly applying it to the target languages.

The architecture of MLMA is shown in Figure 1. In the following sections, we focus on introducing the Step 1 and 2. We first present the architecture of MLMA and describe how to build the unsupervised multilingual alignment. Next, we propose effective methods for collapsing the multi-layer hidden states from MLMA into a single representation. Finally, we introduce the sequence labeling model used in the experiments.

2.1 Language Model Architecture

MLMA is a language model with multi-head self-attention mechanism (Vaswani et al., 2017). The architecture is similar to Radford et al. (2018), except that we combine both a forward and a backward Transformer decoder to build a bidirectional language model. Take the forward direction as an example, given a sentence with N tokens $W = [w_1, w_2, \cdots, w_N]^T$ as input, we first map the sequence of tokens W to token embeddings $\overrightarrow{H}_0 \in \mathbb{R}^{N \times d}$:

$$\overrightarrow{H}_0 = W E_e + E_p \tag{1}$$

where E_e and E_p are the embedding matrix and the positional encoding matrix, and d is the dimension of embeddings and hidden states.

Then n blocks of transformer layers are stacked above the token embeddings. Each block contains a masked multi-head self-attention and a position-wise feedforward layer. The detailed implementation is the same as Vaswani et al. (2017).

$$\overrightarrow{H}_l = \text{TransformerLayer}(\overrightarrow{H}_{l-1}) \tag{2}$$

where $\overrightarrow{H}_l$ refers to the output of the l-th transformer block. Finally, the output distribution over the next tokens is calculated through a softmax function with tied embedding matrix.

$$\overrightarrow{P} = \text{softmax}(\overrightarrow{H}_n E_e^T) \tag{3}$$

For the backward direction, we calculate $\overleftarrow{H}_l$ and $\overleftarrow{P}$ in an analogous way. Finally, we jointly minimize the negative log likelihood of the forward and backward directions:

$$\begin{aligned} NLL = -\sum_{t=1}^{N} (& \log p(w_t | w_1, \ldots, w_{t-1}) \\ & + \log p(w_t | w_{t+1}, \ldots, w_N)) \end{aligned} \tag{4}$$

In a multilingual setting, we share all param-

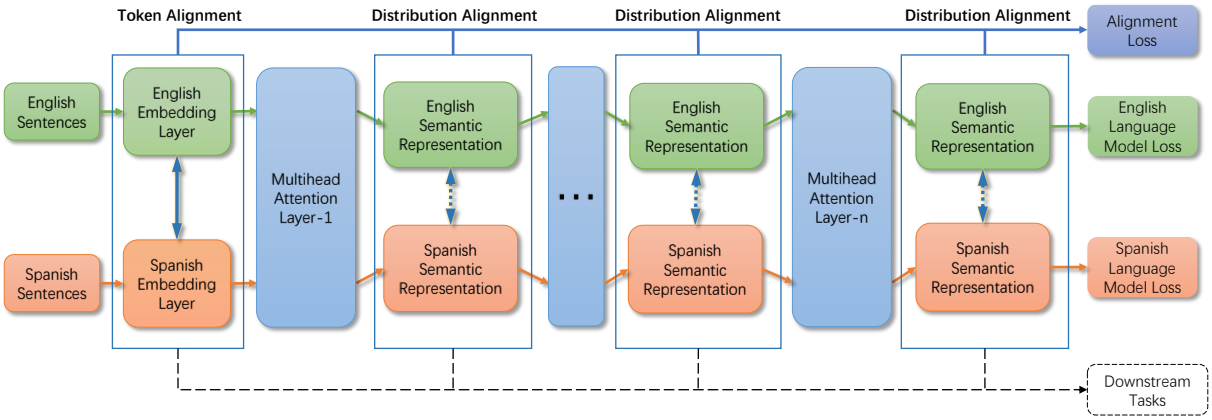

Figure 1: The architecture of MLMA consists of language-specific embedding layers and language-agnostic Transformer layers. MLMA is jointly learned through language modeling loss and alignment loss, and its internal representations are utilized to bridge the gap between source and target languages.

eters in Transformer layers across different languages to facilitate language-agnostic representations, except that we adopt an individual embedding matrix E_e for each language.

2.2 Unsupervised Distribution Alignment

We find that only sharing Transformer layers is not enough for forcing hidden representations from different languages into a common space, as suggested in experiments (Section 3.4). Therefore, we propose three methods to build cross-lingual representations based on identical strings, mean/variance, and average linkage.

To simplify the description, we take the alignment between two languages s and t as an example, but our methods can be directly extended to a scenario with multiple languages by adding the alignment between each pair of languages.

Notation

For the language model, given a sentence with N tokens, the forward internal representation $\overrightarrow{H}_l$ in Eq (2) can be expanded as $\overrightarrow{H}_l = [\overrightarrow{h}_{l,1}, \cdots, \overrightarrow{h}_{l,k}, \cdots, \overrightarrow{h}_{l,N}]^T$, where $\overrightarrow{h}_{l,k}$ refers to the forward hidden representation of the k-th token in the sentence. Then we concatenate the forward and backward hidden representations for each token, $h_{l,k} = \overrightarrow{h}_{l,k} \oplus \overleftarrow{h}_{l,k}$.

We denote the collection of the token representations $h_{l,k}$ at layer l from the whole corpora of language s as C_l^s, which can be regarded as a sampling from the deep semantic space of language s. Similarly, C_l^t is used for language t.

Identical Strings

Similar language pairs such as English and Spanish have a large number of identical strings shared between their vocabularies, which are utilized as the seed dictionary for embedding alignment in previous work (Smith et al., 2017). Similarly, we treat identical strings as explicit supervision signals and align the embeddings of identical strings between different languages. The matching of the embeddings from different languages will lead to an implicit alignment of internal representations. In the experiments, we directly minimize the Euclidean distance between the embeddings of each identical string across different languages:

$$L_{iden} = \frac{\lambda^{iden}}{|W_{iden}^{(s,t)}|} \sum_{w \in W_{iden}^{(s,t)}} ||e_w^s - e_w^t||$$

where $W_{iden}^{(s,t)}$ is the set of identical strings between the vocabulary of language s and language t, and $|W_{iden}^{(s,t)}|$ refers to the number of members in $W_{iden}^{(s,t)}$. λ^{id} is a scaling weight, and e_w^s (e_w^t) is the embedding of word w from embedding matrix E_e^s (E_e^t) of language s (t).

Mean and Variance

In this section, we propose another approach to directly align the distributions of internal representations between different languages. In particular, we leverage the mean and variance of internal distributions for alignment. We denote the mean and variance of C_l^s as m_l^s and v_l^s. Similarly, m_l^t and v_l^t refer to the mean and variance of C_l^t. We minimize the Euclidean distance between the mean and variance of language s and language t for all layers:

$$L_{mv} = \sum_{l=0}^{n} (\lambda_l^m \cdot \frac{||m_l^s - m_l^t||}{|m_l^s| + |m_l^t|} + \lambda_l^v \cdot \frac{||v_l^s - v_l^t||}{|v_l^s| + |v_l^t|})$$

where λ_l is a scaling weight, and $|\cdot|$ is the L1 norm of a vector. Without the denominators, the

model could escape this regularization by learning a mean and variance with low absolute values.

In practice, rather than calculating the mean and variance over the whole source and target corpora, we use the mean and variance of the source and target inner states $h_{l,k}$ in the current mini-batch as an approximation.

Average Linkage

In this method, we employ another metric, average linkage, to perform a more precise point-wise matching. The average linkage is a widely used metric for calculating the similarity of clusters and networks (Yim and Ramdeen, 2015; Seifoddini, 1989; Newman, 2012; Moseley and Wang, 2017). It is sensitive to the shape, thus serves as a better choice than mean and variance. The average linkage measures the similarity of two sets X and Y by calculating the averaged distance between all members of each set:

$$avl(X, Y) = \frac{1}{n_X \cdot n_Y} \sum_{x \in X} \sum_{y \in Y} f(x, y)$$

where n_X (n_Y) is the number of members in X (Y), and f is a distance function. We take Euclidean distance as the distance function f and minimize the average linkage between C_l^s and C_l^t:

$$L_{avl} = \sum_{l=0}^{n} \lambda_l^{avl} \cdot [2 \cdot avl(C_l^s, C_l^t) - avl(C_l^s, C_l^s) - avl(C_l^t, C_l^t)]$$

Similarly, the terms $avl(C_l^s, C_l^s)$ and $avl(C_l^t, C_l^t)$ are used to prevent the model from escaping this regularization. In practice, we calculate L_{avl} between the source and target inner states $h_{l,k}$ inside the mini-batch as an approximation.

The regularization L_{avl} is similar to the maximum mean discrepancy (MMD), which is often employed in domain adaptation (Tzeng et al., 2014; Long et al., 2015) and style transfer (Li et al., 2017) for images. However, different from MMD, our method directly uses Euclidean distance instead of the kernel function.

2.3 Training of MLMA

During the training stage of MLMA, we sample equivalent number of sentences from the monolingual corpora of each language for each mini-batch. Then MLMA is optimized through a combination of the language modeling loss L_{lm} and the alignment regularization loss L_{reg}. For each alignment method, we use its corresponding alignment loss:

$$\begin{aligned} L =& L_{lm} + L_{reg} \\ \text{where } L_{lm} =& \sum_{i \in \{s,t\}} \lambda_i^{lm} \cdot NLL_i, \\ & L_{reg} \in \{L_{id}, L_{mv}, L_{avl}\} \end{aligned}$$

where λ_i^{lm} is used for balancing the convergence speed of different languages. NLL_i is the negative log likelihood of language i in Eq (4).

2.4 Cross-lingual Representations

After the MLMA is trained, we fix its parameters and extract the hidden states as cross-lingual contextualized representations (CLCRs). In this section, we propose two effective strategies for integrating these multi-layer high-dimensional representations into downstream models.

Self-Weighted Sum For each token, we concatenate all layers of hidden states and feed them into a multi-layer perceptron (MLP) to calculate a $(n+1)$-dimensional weight vector, $s = \text{softmax}(\text{MLP}(h_{0,k} \oplus \cdots \oplus h_{n,k}))$. Then we calculate a weighted sum of these layers according to the weight vector, $\mathbf{CLCR}_k = \sum_{l=0}^{n} s_l \cdot h_{l,k}$.

Fully-Weighted Sum We introduce a weight matrix, $F \in \mathbb{R}^{(n+1) \times 2d}$, with separate weights for each hidden dimension. The weight matrix F is softmaxed by column and used to calculate a weighted sum of all layers for each hidden dimension, $\mathbf{CLCR}_k = \sum_{l=0}^{n} F_l \odot h_{l,k}$, where $\odot$ is the element-wise product.

The parameters of the MLP and F are trained during the learning of sequence labeling model.

2.5 Sequence Labeling Model

The sequence labeling model is then built on the CLCRs. For both NER and POS tasks, we use an LSTM-CRF model following Lample et al. (2016), which consists of a character-level LSTM, a word-level LSTM, and a linear-chain CRF.

More specifically, given a sequence of words as $[w_1, w_2, \ldots, w_N]$, where w_k is composed of a sequence of characters $[c_{k,1}, c_{k,2}, \ldots, c_{k,m}]$. First, for each word w_k, the character-level LSTM takes its character sequence $[c_{k,1}, c_{k,2}, \ldots, c_{k,m}]$ as input and outputs a vector e_k to represent this word. Then the pre-trained $\mathbf{CLCR}_k$ is concatenated with e_k to form a word-level embedding x_k. Finally, the sequence of word-level embeddings $[x_1, x_2, \ldots, x_N]$ are fed into the word-level LSTM, and the linear-chain CRF are employed to

predict the probability distribution for all possible output label sequences.

3 Experiments

We first introduce the datasets used in the experiment and then the implementation details of our models, before presenting the results on NER and POS tasks.

3.1 Datasets

For cross-lingual NER, we evaluate the proposed approaches on CoNLL 2002/2003 datasets (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), which contain four European languages, English (en), Spanish (es), Dutch (nl), German (de) and four entity types (person, location, organization, and MISC). We also evaluate a distant language pair, English-Chinese, on OntoNotes(v4.0) dataset (Hovy et al., 2006). We adopt the same dataset split and four valid entity types (person, location, organization, and GPE) as described in (Wang and Manning, 2014).

For cross-lingual POS, we use the Danish (da), Dutch (nl), German (de), Greek (el), Italian (it), Portuguese (pt), Spanish (es) and Swedish (sv) portion from CoNLL 2006/2007 dataset (Buchholz and Marsi, 2006; Nivre et al., 2007). Following previous work (Fang and Cohn, 2017), we train the sequence labeling model on Penn Treebank data and adopt the universal POS tagset (Petrov et al., 2011b).

In all cases, the sequence labeling model is trained on the source language (English) training data and is tested on the target language test data.

3.2 Details of MLMA

We adopt a 6-layer bi-directional Transformer decoder with 8 attention heads. The dimension size of hidden states and inner states are 512 and 2048, respectively. The dropout rates after attention and residual connection are both 0.1. We use the Adam optimization scheme (Kingma and Ba, 2014) with a learning rate of 0.0001 and a gradient clip norm of 5.0. The vocabulary size of each language is 200,000, and we train the model with a sampled softmax (Jean et al., 2015) of 8192 samples. We only keep the sentences containing less than 200 tokens for training and group them into batches by length. Each batch contains around 4096 tokens for each language. The language modeling weight λ_i^{lm} is set to be 1.0 for each language. For alignment, λ_l^m, λ_l^v, λ_l^{al} are set to be 0.1, 0.01 and 1.0 for every layer l, and λ^{iden} is set to be 100.

For languages except English, the latest dump of Wikipedia is used as monolingual corpora. For English, we use 1B Word Benchmark (Chelba et al., 2013) to reduce the effects of potential internal alignment in Wikipedia (Zirikly and Hagiwara, 2015; Tsai et al., 2016).

All characters are preprocessed to lowercase, and Chinese text are converted into the simplified version through OpenCC[2]. The corpora of European languages are tokenized by nltk (Loper and Bird, 2002) and Chinese text is segmented using Ltp[3].

3.3 Details of Sequence Labeling Model

In our experiments, we set the hidden size of word-level LSTM and character-level LSTM to be 300 and 100, respectively. The character embedding size is set to be 100. We apply dropout at both the input and the output of word-level LSTM to prevent overfitting. The dropout rate is set to be 0.5. We train the sequence labeling model for 20 epochs using Adam optimizer with a batch size of 20 and perform an early stopping when there is no improvement for 3 epochs. We set the initial learning rate to be 0.001 and decay the learning rate by 0.1 for each epoch. We do not update the pre-trained cross-lingual deep representations from MLMA during training. For each model, we run it five times and report the mean and standard deviation. We disable the character-level LSTM in English-German and English-Chinese NER as they have a different character pattern from English. For POS, we disable the character-level LSTM following Fang and Cohn (2017).

3.4 Results for NER

We first train a multilingual language model without alignment (MLM) and report its performance of cross-lingual NER in Table 1. The poor performance demonstrates that only sharing part of the parameters in a language model is far from enough for cross-lingual transfer.

As shown in Table 1, the mean/variance alignment strategy (MLMA-Mv) is competitive with previous work which utilizes extra bilingual resources (Section 5.1). The average linkage strategy (MLMA-Avl) performs a more precise align-

[2] https://github.com/BYVoid/OpenCC
[3] https://github.com/HIT-SCIR/pyltp

Model	es	nl	de	Extra Resources
MLM (w/o alignment) + s.w.s.	21.16 ± 1.40	33.97 ± 1.49	15.46 ± 1.21	None
MLM (w/o alignment) + f.w.s.	23.61 ± 2.33	32.94 ± 1.62	16.38 ± 1.09	None
Täckström et al. (2012)	59.30	58.40	40.40	parallel corpus
Nothman et al. (2013)	61.00	64.00	55.80	Wikipedia
Wang and Manning (2014)	-	-	60.00	parallel corpus
Tsai et al. (2016)	60.55	61.60	48.10	Wikipedia
Ni et al. (2017)	65.10	65.40	58.50	Wikipedia, parallel corpus, 5K dict.
Mayhew et al. (2017)	65.95	66.50	59.11	Wikipedia, 1M dict.
Xie et al. (2018)	72.37	71.25	57.76	None
MUSE*	66.17 ± 1.15	65.52 ± 0.78	55.46 ± 0.59	None
Multilingual BERT*	66.42 ± 1.15	69.21 ± 0.48	**70.78** ± 0.36	None
MLMA-Iden + s.w.s.	69.45 ± 0.91	68.82 ± 0.82	55.75 ± 1.64	None
MLMA-Iden + f.w.s.	67.10 ± 0.78	68.15 ± 0.67	55.25 ± 1.29	None
MLMA-Mv + s.w.s.	73.81 ± 0.83	70.61 ± 1.79	57.70 ± 0.71	None
MLMA-Mv + f.w.s.	74.12 ± 1.00	71.72 ± 0.70	57.84 ± 0.80	None
MLMA-Avl + s.w.s.	75.01 ± 0.79	76.22 ± 0.42	60.98 ± 1.00	None
MLMA-Avl + f.w.s.	74.43 ± 0.50	76.02 ± 0.55	60.50 ± 0.43	None
MLMA-Avl (init) + s.w.s.	75.72 ± 0.80	**76.90** ± 0.30	63.01 ± 0.83	None
MLMA-Avl (init) + f.w.s.	76.30 ± 0.76	76.85 ± 0.43	62.85 ± 0.47	None
MLMA-Avl (multi) + s.w.s.	**79.36** ± 0.57	74.89 ± 0.28	65.93 ± 0.32	None
MLMA-Avl (multi) + f.w.s.	79.34 ± 0.35	74.74 ± 0.40	66.53 ± 0.35	None

Table 1: NER F1 scores on test sets of European languages. For previous work which reports multiple results, we only list their best performance on each language. Results of methods with mark * are obtained by running their released source code or models. The results of MUSE embeddings are produced by using them for direct model transfer. "MLM" denotes our multilingual language model without alignment. The three alignment methods, **Iden** = identical strings, **Mv** = mean and variance, **Avl** = average linkage, respectively. "s.w.s" and "f.w.s." are self-weighted sum and fully-weighted sum. "init" represents using MUSE to initialize the embedding matrices in the MLMA. "multi" refer to the multi-source transfer.

Model	zh
Wang and Manning (2014)◇	**64.40**
MUSE*	35.35 ± 0.84
Xie et al. (2018)*	44.13 ± 1.49
Our methods	
MLMA-Iden + s.w.s.	11.08 ± 0.89
MLMA-Iden + f.w.s.	11.17 ± 0.69
MLMA-Avl + s.w.s.	50.11 ± 1.51
MLMA-Avl + f.w.s.	45.88 ± 2.49
MLMA-Avl (init) + s.w.s	60.33 ± 1.39
MLMA-Avl (init) + f.w.s	58.92 ± 1.22

Table 2: NER F1 scores on test sets for Chinese. The notations are the same as Table 1. Methods with mark ◇ require parallel corpora.

ment and gains a further improvement. We conducted experiments of using all three alignments together, and results show no significant improvement over average linkage alone. These results agree with our statements that average linkage performs a more precise matching, and thus, carries the benefits brought by the other methods.

To demonstrate the strengths of the proposed cross-lingual contextualized representations (CLCRs) over cross-lingual word embeddings (CLWEs), we also report the results of using CLWEs for direct model transfer in Table 1. Specifically, we compare with the unsupervised method MUSE from Conneau et al. (2017). The experiment results demonstrate its effectiveness for cross-lingual sequence labeling. The alignment method using identical strings (MLMA-Iden) outperforms MUSE, suggesting that the contextual-level representations are more effective than the word-level ones. The other proposed methods (MLMA-Mv and MLMA-Avl) achieve significant improvement over MUSE and MLMA-Iden, which shows the benefit of directly aligning the contextualized representations.

Combination with CLWEs We further demonstrate that CLWEs are compatible with our methods by using MUSE embeddings to initialize the embedding layer of our multilingual language model. The results of MLMA-Avl (init) shown in Table 1 indicate that the CLWEs lead to a better initialization and improved performance.

Multi-source Transfer We conduct experiments of multi-source transfer based on method MLMA-Avl and report the performance as MLMA-Avl (multi) in Table 1. The experiment settings largely follow the previous work (Mayhew et al., 2017). They employ two source languages for each target language and use syntactic features to choose the related source languages. For Spanish and German, we use English and Dutch as source languages. English and Spanish are adopted for

Model	es	nl	de	da	el	it	pt	sv	Avg.
Das and Petrov (2011)◇	**84.2**	79.5	82.8	83.2	**82.5**	86.8	**87.9**	80.5	83.31
Fang and Cohn (2017)†	68.40	64.50	65.90	73.5	65.5	64.8	67.8	66.0	67.05
Fang and Cohn (2017)†‡	81.20	82.30	78.90	81.9	80.1	81.9	82.1	78.1	80.81
Xie et al. (2018)*	73.25	75.46	80.72	29.75	71.65	71.19	76.48	64.36	67.86
MUSE*	78.30	80.84	81.10	73.99	63.16	80.63	82.79	66.38	75.90
Multilingual BERT*	83.86	84.79	**87.16**	**83.77**	82.27	**88.39**	87.86	**81.07**	**84.90**
Our methods									
MLMA-Avl + s.w.s.	81.60	85.10	84.10	83.38	77.04	84.46	86.93	80.78	82.92
MLMA-Avl + f.w.s.	81.20	85.54	84.92	83.45	77.48	84.80	87.43	80.72	83.19
MLMA-Avl (init) + s.w.s.	82.73	85.79	85.76	82.44	80.55	86.76	85.99	79.75	83.72
MLMA-Avl (init) + f.w.s.	82.27	**85.97**	86.37	82.25	81.31	86.28	86.53	80.00	83.87

Table 3: POS accuracy on test sets of European languages. The notations are consistent with Table 1. Fang and Cohn (2017) report different results according to different resource requirements. We only list their best results in each setting. Methods with mark ◇, †, ‡ require parallel corpora, bilingual lexicons, and training data respectively.

Dutch. The multi-source transfer leads to a significant improvement for Spanish and German, but a slight decline for Dutch. In the follow-up experiment, we find that the Spanish training set achieves a poor cross-lingual performance on Dutch. Similar results are observed in the experiments of Spanish to English and Dutch to English. These results suggest that the cross-lingual transfer may be directional, and we leave this issue for future work.

Comparison with BERT We also compare the performance of our MLMA with the released multilingual BERT (Devlin et al., 2018). As shown in Table 1, our MLMA-Avl achieves a better performance on Spanish and Dutch. For German, BERT achieves a high performance as it employs effective subword information through BPE. The architecture of BERT also performs better than LSTM.

It is worth mentioning that, in previous work and this work, the corpora used in the experiments are limited to the source and the target language. In contrast, the multilingual BERT is jointly learned on Wikipedia of 102 languages and may benefit from a multi-hop transfer. BERT employs a shared BPE vocabulary for different languages, which implicitly performs a subword alignment similar to MLMA-Iden. Meanwhile, the proposed MLMA-Mv and MLMA-Avl methods are compatible with BERT and can be used to align the inner states of BERT.

3.5 A Case Study of Chinese NER

We conduct experiments and evaluate our approaches on a distant language pair, English-Chinese. The experiment results are shown in Table 2. Wang and Manning (2014) utilize 80K parallel sentences for annotation projection and report a strong performance. As Chinese and English do not share the alphabet, the number of identical strings is significantly smaller than similar languages pairs such as English-Spanish. Therefore, the MLMA-Iden achieves a lower result comparing to MUSE which uses adversarial training. The MLMA-Avl method performs a direct alignment of internal representations and achieves a significant improvement over the word-level methods. The initialization from CLWEs also proves its effectiveness for distant language pairs by gaining further improvement and reaching a comparable result with Wang and Manning (2014). This experiment suggests that cross-lingual transfer is still challenging between distant language pairs.

3.6 Results for POS

We evaluate our methods on another sequence labeling task POS, and the results are shown in Table 3. We compare with previous studies using unsupervised cross-lingual clustering (Fang and Cohn, 2017) and large-scale parallel corpora (Das and Petrov, 2011). As shown in Table 3, our models with deep semantic alignment outperform previous lexicon-based cross-lingual clustering by a large margin. When comparing to the previous method with a small amount of training data, the MLMA-Avl method obtains an improved accuracy without training data in the target languages. For further comparison, We also list the performance of applying the method from Xie et al. (2018) and

MUSE
brown: oliváceo (olive), negruzcas (blackish), negruzco (blackish), marrón (brown), ocráceo (ochraceous)
chair: vicepresidenta (vice president), vicedecano (vice dean), cátedra (chair), vicedecana (vice dean), catedrático (professor)

MLMA-Avl	
[Brown]'s office told news outlets of his visit to Afghanistan ...	**[Neira]** escapó meses después rumbo a Miami para ... (**Neira** escaped months later heading to Miami to ...)
Wearing a **[brown]** suit with matching hat, ...	La corona y vientre del macho son de un **[verde]** esmeralda brillante iridiscente, ... (The crown and belly of the male are of an iridescent bright emerald **green**, ...)
Sweden currently holds the EU **[chair]**.	Tras tomar posesión de su **[asiento]** , Lois decide limpiar el lago para empezar, ... (After taking possession of her **seat**, Lois decides to clean the lake to begin, ...)
It's an honor to be asked to **[chair]** the Man Booker Prize, ...	..., y fue la primera mujer en **[presidir]** un sindicato AFL-CIO. (..., and was the first woman to **preside** over an AFL-CIO union.)

Table 4: English words and their nearest Spanish words according to MUSE and MLMA-Avl.

multilingual BERT to the POS task.[4]

POS mainly relies on the information of each single word, and parallel corpora providing word alignment are effective for cross-lingual POS. Thus, previous annotation projection methods through parallel corpora are strong approaches for cross-lingual POS and often achieve a significantly better performance against previous unsupervised methods. The experimental results show that the proposed CLCRs are competitive and even achieve better average accuracy.

3.7 Self-Weighted v.s. Fully-Weighted Sum

As shown in Table 1, 2 and 3, we observe that Self-Weighted Sum (SWS) generally outperforms Fully-Weighted Sum (FWS) in NER tasks, while the opposite is true for POS tasks. SWS allows weights to vary at each position in a sequence, while FWS imposes adaptive weights for each hidden dimension. We hypothesize that NER is more context-sensitive and requires models to adapt to different context information, which makes SWS a better option. On the other hand, the POS of words is more independent across different context, but certain feature dimensions in contextualized representations may be critical for making a judgment. Therefore, FWS has the edge over SWS for its ability to select out these dimensions.

4 What is Connected during Alignment?

In this section, we dive into the MLMA and investigate the question of what is connected between different languages during the alignment. From English 1B and Spanish Wikipedia, we randomly select 1,000 sentences for each language and extract their cross-lingual contextual representations using our MLMA-Avl model. We calculate the nearest neighbors in cosine distance for each word, and some of them are listed in Table 4.

In these cases, the MLMA can disambiguate word senses according to context information. For example, for the word *brown* in English, the MLMA groups color *brown* with *verde* (green), and name *Brown* with *Neira* (a person name in Spanish) in the Spanish corpus. The proposed method is different from unsupervised translation in that, instead of learning a precise matching between English and Spanish words, the CLCRs establishes a high-level semantic connection between the source and the target language. The next example demonstrates that the MLMA is able to distinguish the part-of-speech of words. It connects an English verb *chair* with a Spanish verb *presidir* (preside), while a noun *chair* with a noun *asiento* (seat) in Spanish. To compare with unsupervised cross-lingual word embeddings, we list the top 5 similar words calculated using MUSE. As shown in Table 4, MUSE successfully groups the English word *brown* with Spanish words that are related to colors. However, without the help of contextual information, its ability of word sense disambiguation is limited.

5 Related Work

Previous work in cross-lingual transfer learning can be roughly divided into two main branches: annotation projection and model transfer.

5.1 Annotation Projection

In annotation projection approaches, parallel or comparable corpora are commonly used (Yarowsky et al., 2001; Ehrmann et al., 2011; Das and Petrov, 2011; Li et al., 2012; Täckström et al.,

[4]The poor performance of Xie et al. (2018) on en-da is due to the low quality of word translation pairs generated by their method.

2013; Wang and Manning, 2014; Ni et al., 2017). The source language sentences of parallel corpora are first annotated either manually or by a pre-trained tagger. Then, annotations on the source side are projected to the target side through word alignment to generate distantly supervised training data. Finally, a model of the target language is trained on the generated data. Wikipedia contains multilingual articles for various topics and can thus be used to generate parallel/comparable corpora or even weakly annotated target language sentences (Kim et al., 2012).

However, parallel corpora and Wikipedia can be rare for true low-resource languages. Mayhew et al. (2017) reduce the resource requirement by proposing a cheap translation method, which "translates" the training data from the source to the target language word by word through a bilingual lexicon. While Xie et al. (2018) reduce the requirement of bilingual lexicons by an unsupervised word-by-word translation through CLWEs.

5.2 Model Transfer

Model transfer methods train a model on the source language with language-independent features. Thus, the trained model can be directly applied to the target language.

McDonald et al. (2011) design a cross-lingual parser based on delexicalized features like universal POS tags. Täckström et al. (2012) reveal that cross-lingual word cluster features induced using large parallel corpora are useful. Lexicon and Wikipedia also demonstrate effectiveness for language-independent feature engineering. Zirikly and Hagiwara (2015) generate multilingual gazetteers from the source language gazetteers and comparable corpus. Page categories and linkage information to entries from Wikipedia are extracted as strong language-independent features (wikifier features) (Tsai et al., 2016). Bharadwaj et al. (2016) facilitate the cross-lingual transfer through phonetic features, which work well between languages like Turkish, Uzbek, and Uyghur, but are not strictly language independent. Recently, CLWEs are used as language-invariant representations for direct model transfer in NER (Ni et al., 2017) and POS (Fang and Cohn, 2017).

Some of the previous work also proposes sequence labeling models with shared parameters between languages for performing cross-lingual knowledge transfer (Lin et al., 2018; Cotterell and Duh, 2017; Yang et al., 2017; Ammar et al., 2016; Kim et al., 2017). However, these models are usually obtained through joint learning and require annotated data from the target language.

6 Conclusion

In this paper, we focused on a low-resources cross-lingual setting and proposed transfer learning methods based on the alignment of deep semantic spaces between different languages. The proposed multilingual language model bridges different languages by automatically learning cross-lingual disambiguated representations. Abundant NER and POS experiments are conducted on the benchmark datasets. Experimental results show that our approaches using only monolingual corpora achieve improved performance comparing to previous strong cross-lingual studies with extra resources.

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472. Association for Computational Linguistics.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 149–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96. Asian Federation of Natural Language Processing.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 600–609, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 118–124. Association for Computational Linguistics.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10. Association for Computational Linguistics.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838. Association for Computational Linguistics.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 694–702. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics.

Qi Li, Haibo Li, Heng Ji, Wen Wang, Jing Zheng, and Fei Huang. 2012. Joint bilingual name tagging for parallel corpora. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1727–1731, New York, NY, USA. ACM.

Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. 2017. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A multi-lingual multi-task architecture for low-resource sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 799–809. Association for Computational Linguistics.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. 2015. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. 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)*, pages 1064–1074. Association for Computational Linguistics.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72. Association for Computational Linguistics.

Benjamin Moseley and Joshua Wang. 2017. Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search. In *Advances in Neural Information Processing Systems*, pages 3094–3103.

Mark EJ Newman. 2012. Communities, modules and large-scale structure in networks. *Nature physics*, 8(1):25.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artif. Intell.*, 194:151–175.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011a. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2011b. A universal part-of-speech tagset. *CoRR*, abs/1104.2086.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Hamid K Seifoddini. 1989. Single linkage versus average linkage clustering in machine cells formation applications. *Computers & Industrial Engineering*, 16(3):419–426.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 219–228. Association for Computational Linguistics.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Mengqiu Wang and Christopher D. Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association for Computational Linguistics*, 2:55–66.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. *arXiv preprint arXiv:1808.09861*.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Odilia Yim and Kylee T Ramdeen. 2015. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *The quantitative methods for psychology*, 11(1):8–21.

Ayah Zirikly and Masato Hagiwara. 2015. Cross-lingual transfer of named entity recognizers without parallel corpora. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 390–396. Association for Computational Linguistics.