Cross-lingual Multi-Level Adversarial Transfer to Enhance Low-Resource Name Tagging

Lifu Huang[1], **Heng Ji**[1], **Jonathan May**[2]
[1] Computer Science Department, Rensselaer Polytechnic Institute
{huangl7,jih}@rpi.edu
[2] Information Sciences Institute, University of Southern California
jonmay@isi.edu

Abstract

We focus on improving name tagging for low-resource languages using annotations from related languages. Previous studies either directly project annotations from a source language to a target language using cross-lingual representations or use a shared encoder in a multitask network to transfer knowledge. These approaches inevitably introduce noise to the target language annotation due to mismatched source-target sentence structures. To effectively transfer the resources, we develop a new neural architecture that leverages multi-level adversarial transfer: (1) word-level adversarial training, which projects source language words into the same semantic space as those of the target language without using any parallel corpora or bilingual gazetteers, and (2) sentence-level adversarial training, which yields language-agnostic sequential features. Our neural architecture outperforms previous approaches on CoNLL data sets. Moreover, on 10 low-resource languages, our approach achieves up to 16% absolute F-score gain over all high-performing baselines on cross-lingual transfer without using any target-language resources.[1]

1 Introduction

Low-resource language name tagging is an important but challenging task. An effective solution is to perform cross-lingual transfer, by leveraging the annotations from high-resource languages. Most of these efforts achieve cross-lingual annotation projection based on bilingual parallel corpora combining with automatic word alignment (Yarowsky et al., 2001; Wang et al., 2013; Fang and Cohn, 2016; Ehrmann et al., 2011; Ni et al., 2017), bilingual gazetteers (Feng et al., 2017; Zirikly and Hagiwara, 2015), cross-lingual word embedding (Fang and Cohn, 2017; Wang et al., 2017; Huang et al., 2018), or cross-lingual Wikification (Kim et al., 2012; Nothman et al., 2013; Tsai et al., 2016; Pan et al., 2017), but these resources are still only available for dozens of languages. Recent efforts on multi-task learning model each language as one single task while all the tasks share the same encoding layer (Yang et al., 2016, 2017; Lin et al., 2018). These methods can transfer knowledge via the shared encoder without using bilingual resources. However, different languages usually have different underlying sequence structures, as shown in Figure 1. Without an explicit constraint, the encoder is not guaranteed to extract language-independent sequential features. Moreover, when the size of annotated resources is not balanced, the encoder is likely to be biased toward the resource-dominant language.

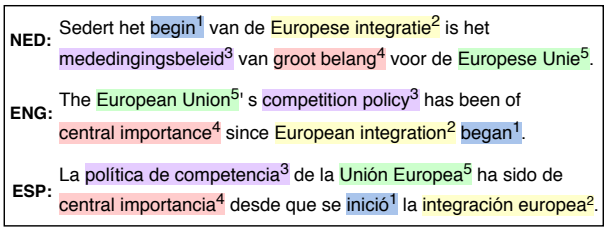

Figure 1: Example of parallel sentences between English (ENG), Spanish (ESP) and Dutch (NED) from Europarl Parallel Corpus (Koehn, 2005). The information units with the same color and superscript are aligned.

Considering these challenges, we develop a new neural architecture which can effectively transfer resources from source languages to improve target language name tagging. Our neural architecture is built upon a state-of-the-art sequence tagger: bi-directional long short-term memory as input to conditional random fields (Bi-LSTM-CRF) (Lample et al., 2016; Huang et al., 2015; Ma and

[1]Our programs will be released at https://github.com/wilburOne/AdversarialNameTagger

Proceedings of NAACL-HLT 2019, pages 3823–3833
Minneapolis, Minnesota, June 2 - June 7, 2019.

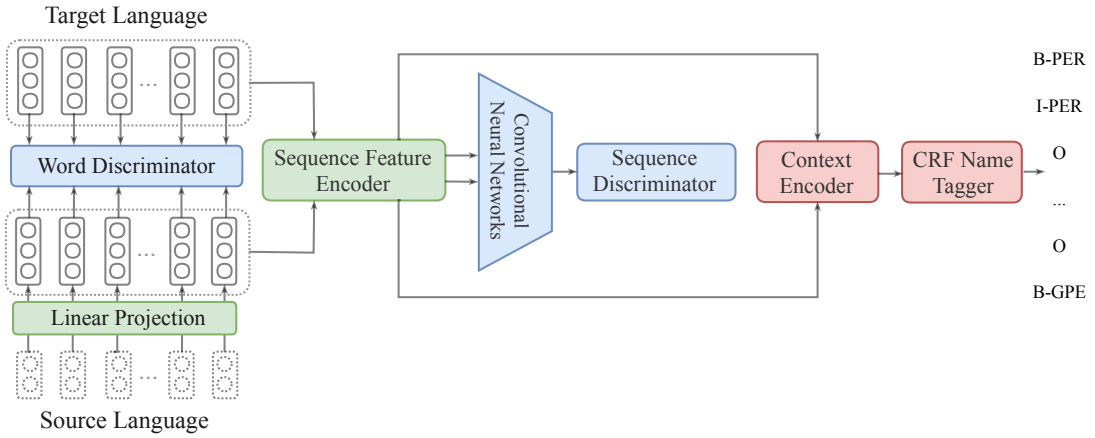

Figure 2: Architecture overview.

Hovy, 2016), integrated with multi-level adversarial transfer: (1) word level adversarial transfer, similar to Conneau et al. (2017), applying a projection function on the source language and a discriminator to distinguish each word of the target language from that of the source language, resulting in a bilingual shared semantic space; (2) sentence-level adversarial transfer, where a discriminator is trained to distinguish each sentence of the target language from that of the source language,[2] and a sequence encoder is applied to each sentence of both languages to prevent the discriminator from correctly predicting the source of each sentence, yielding language-agnostic sequential features. These features can better facilitate the resource transfer from the source language to the target language.

Our contributions are twofold: (1) without requiring any parallel corpora or bilingual gazetteers, the multi-level adversarial approach can efficiently transfer annotated resources from the source language to the target language and improve target language name tagging; (2) In addition to outperforming previous high-performing baselines on CoNLL data sets, we also evaluate cross-lingual name tagging on 10 low-resource languages and achieve up to 16% absolute F-score gain over all baselines when there is no annotated resource for the target language.

2 Approach

2.1 Approach Overview

Figure 2 shows the overview of our neural architecture. It consists of three components:

[2]For the name tagging task, 'sequence' always means 'sentence.'

Cross-lingual word embedding learning with adversarial training: Given pre-trained monolingual word embeddings for a target language t and a source language s, we first apply a mapping function to each word representation from s, then feed both the projected source word representations and the target word representations to a word discriminator to predict the language of each word. If the discriminator cannot distinguish the language of t from the projection of s, then we consider t and the projection of s to be in a shared space.

Language-agnostic sequential feature extraction: For each sentence of t and s, we apply a sequence encoder to extract sequential features, and a Convolutional Neural Network (CNN) (Krizhevsky et al., 2012) based sequence discriminator to predict the language source of each sentence. The sequence encoder is trained to prevent the sequence discriminator from correctly predicting the language of each sentence, such that it finally extracts language-agnostic sequential features.

Language-independent name tagger The language-agnostic sequential features from both t and s are further fed into a context encoder to better capture and refine contextual information and a conditional random field (CRF) (Lafferty et al., 2001) based name tagger.

Next we show the details of each component in our architecture.

2.2 Word-level Adversarial Transfer

To better leverage the resources from the source language, our first step is to construct a shared se-

mantic space where the words from the source and target languages are semantically aligned. Without requiring any bilingual gazetteers, recent efforts (Zhang et al., 2017b; Conneau et al., 2017; Chen and Cardie, 2018) explore unsupervised approaches to learn cross-lingual word embeddings and achieve comparable performance to supervised methods. Following these studies, we perform word-level adversarial training to automatically align word representations from s and t.

Formally, assume we are given pre-trained monolingual word embeddings $\mathbf{V}_t = \{\mathbf{v}_1^t, \mathbf{v}_2^t, ..., \mathbf{v}_N^t\} \in \mathbb{R}^{N \times d_t}$ for t, and $\mathbf{V}_s = \{\mathbf{v}_1^s, \mathbf{v}_2^s, ..., \mathbf{v}_M^s\} \in \mathbb{R}^{M \times d_s}$ for s, where $\mathbf{v}_i^t$ and $\mathbf{v}_j^s$ are the vector representations of words w_i^t and w_i^s from t and s, N and M denote the vocabulary sizes, d_t and d_s denote the embedding dimensionality of t and s respectively. We then apply a mapping function f to project s into the same semantic space as t:

$$\widetilde{\mathbf{V}}_s = f(\mathbf{V}_s) = \mathbf{V}_s\mathbf{U} \quad (1)$$

where $\mathbf{U} \in \mathbb{R}^{d_s \times d_t}$ is the transformation matrix. $\widetilde{\mathbf{V}}_s \in \mathbb{R}^{M \times d_t}$ are the projected word embeddings for s, and $\Theta_f = \{\theta_f\}$ denotes the set of parameters to be optimized for f.

Similar to Xing et al. (2015), Conneau et al. (2017), and Chen and Cardie (2018), we constrain the transformation matrix $\mathbf{U}$ to be orthogonal with singular value decomposition (SVD) to reduce the parameter search space:

$$\mathbf{U} = \mathbf{A}\mathbf{B}^\top, \text{with } \mathbf{A}\Sigma\mathbf{B}^\top = \text{SVD}(\widetilde{\mathbf{V}}_s\mathbf{V}_s^\top) \quad (2)$$

To automatically optimize the mapping function f without using extra bilingual signals, we introduce a multi-layer perceptron D as a word discriminator, which takes word embeddings of t and projected word embeddings of s as input features and outputs a single scalar. $D(w_i^*)$ represents the probability of w_i^* coming from t. The word discriminator is trained by minimizing the binary cross-entropy loss:

$$L_{dis}^w = -\frac{1}{I_{t;s}} \cdot \sum_{i=0}^{I_{t;s}} \Big(y_i \cdot \log(D(w_i^*)) + (1 - y_i) \cdot \log(1 - D(w_i^*)) \Big) ,$$

$$y_i = \delta_i(1 - 2\epsilon) + \epsilon ,$$

where $\delta_i = 1$ when w_i^* is from t and $\delta_i = 0$ otherwise. $I_{t;s}$ represents the number of words sampled from the vocabulary of t and s together. ϵ is a smoothed value added to the positive and negative labels. $\Theta_{dis} = \{\theta_D\}$ is the parameter set.

The mapping function f and word discriminator D are two adversarial players, thus we flip the word labels and optimize f by minimizing the following loss:

$$L_f^w = -\frac{1}{I_{t;s}} \cdot \sum_{i=0}^{I_{t;s}} \Big((1 - y_i) \cdot \log(D(w_i^*)) + y_i \cdot \log(1 - D(w_i^*)) \Big) ,$$

$$y_i = \delta_i(1 - 2\epsilon) + \epsilon$$

Following the standard training procedures of deep adversarial networks (Goodfellow et al., 2014), we train the word discriminator and the mapping function successively with stochastic gradient descent (SGD) (Bottou, 2010) to minimize L_{dis}^w and L_f^w. Similar to Conneau et al. (2017), after word-level adversarial training, we also adopt a refinement step to construct a bilingual dictionary for the top-k most frequent words in the source language[3] based on $\widetilde{\mathbf{V}}_s$ and $\mathbf{V}_t$, and further optimize $\mathbf{U}$ with Equation 2 in a supervised way.

2.3 Sentence-level Adversarial Transfer

Once s is projected into the same semantic space as t, we can regard both sentences as coming from one unified language and directly project annotations from s to t. However, name tagging not only relies on word level features, but also on sequential contextual features for entity type classification. Without constraints, the sequence encoder can only extract sequential features for both t and s based on their final training signals while these features are not necessarily beneficial to the target language. Thus, we further design sentence level adversarial transfer to encourage the encoder to extract language-agnostic sequential features.

Given a sentence $x^t = \{w_1^t, w_2^t, ...\}$ from t and a sentence $x^s = \{w_1^s, w_2^s, ...\}$ from s, we first use $\mathbf{V}_t$ and $\widetilde{\mathbf{V}}_s$ to initialize a vector representation for each w_i^t and w_i^s. We also apply a character-based CNN (denoted as CharCNN) (Kim et al., 2016) for each language to compose a word representation from its characters. For each word, we

[3] We set k=15,000 in our experiment.

concatenate its word representation and character based representation. Then we feed the sequence of vector representations into a weight sharing Bi-LSTM encoder E to obtain sequential features $\mathbf{H}_t = \{\mathbf{h}_1^t, \mathbf{h}_2^t, ...\}$ and $\mathbf{H}_s = \{\mathbf{h}_1^s, \mathbf{h}_2^s, ...\}$ for x^t and x^s respectively. The parameter set of optimizing both language-dependent CharCNN and the sequence encoder can be denoted as $\Theta_e = \{\theta_{\text{CharCNN}_t}, \theta_{\text{CharCNN}_s}, \theta_E\}$.

Based on these sequential features, we use a sequence discriminator to predict the language source of each sentence. Given a sentence x^* and its sequential features $\mathbf{H} = \{\mathbf{h}_1^*, \mathbf{h}_2^*, ...\}$ from E, we first apply a language-independent CNN with max-pooling to get an overall vector representation for x^*, then feed it into another multi-layer perceptron, $\tilde{D}$, to predict the probability that x^* comes from language t. The sequence discriminator is trained by minimizing the following binary cross-entropy loss:

$$L_{dis}^x = -\frac{1}{\tilde{I}_{t;s}} \cdot \sum_{i=0}^{\tilde{I}_{t;s}} \Big(\tilde{y}_i \cdot \log(\tilde{D}(x_i^*)) + (1 - \tilde{y}_i) \cdot \log(1 - \tilde{D}(x_i^*)) \Big) ,$$

$$\tilde{y}_i = \tilde{\delta}_i(1 - 2\eta) + \eta ,$$

where $\tilde{\delta}_i = 1$ if the sentence x_i^* is from t and $\tilde{\delta}_i = 0$ otherwise. $\tilde{I}_{t;s}$ represents the number of sentences sampled from the whole data set of t and s. η is another smoothed value for sequence labels. $\Theta_{\widetilde{dis}} = \{\theta_{\text{CNN}}, \theta_{\tilde{D}}\}$ denotes the parameter set for optimizing the sequence discriminator.

The sequence encoder E and the sequence discriminator $\tilde{D}$ are two adversarial players and E is optimized by trying to fool $\tilde{D}$ to correctly predict the language source of each sentence. Thus we flip the sequence labels and optimize E by minimizing the following loss:

$$L_e^x = -\frac{1}{\tilde{I}_{t;s}} \cdot \sum_{i=0}^{\tilde{I}_{t;s}} \Big((1 - \tilde{y}_i) \cdot \log(\tilde{D}(x_i^*)) + \tilde{y}_i \cdot \log(1 - \tilde{D}(x_i^*)) \Big) ,$$

$$\tilde{y}_i = \tilde{\delta}_i(1 - 2\eta) + \eta$$

2.4 Name Tagger Training

With the language-agnostic sequential features from E, we can directly combine all annotated training data from both t and s to train the name tagger for t. To do so, we feed the sequential features from E to another Bi-LSTM encoder E_c to refine the context information for each token, and use a CRF output layer to render predictions for each token, which can effectively capture dependencies among name tags (e.g., an "inside-organization" token cannot follow a "beginning-person" token).

Specifically, given an input sentence $x = \{w_1, w_2, ...w_n\}$, we extract language-agnostic sequential features with E, and further obtain a new sequence of contextual features $\widetilde{\mathbf{H}} = \{\tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, ..., \tilde{\mathbf{h}}_n\}$ with E_c. Then we a apply a linear layer ℓ to further convert each $\tilde{\mathbf{h}}_i$ to a score vector $\mathbf{y}_i$, in which each dimension denotes the predicted score for a tag (the starting, inside or outside of a name mention with a pre-defined entity type). Then we feed the sequence of score vectors $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ into the CRF layer. The score of a sequence of tags $\mathbf{Z} = \{z_1, z_2, ..., z_n\}$ is defined as:

$$Score(x, \mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{n} (R_{z_{i-1}, z_i} + Y_{i, z_i})$$

where R is a transition matrix and $R_{p,q}$ denotes the binary score of transitioning from tag p to tag q.

Algorithm 1 Multi-level Adversarial Training for Improving Target Language Name Tagging

Input: Monolingual pre-trained word embeddings $\mathbf{V}_t$ for target language t, and $\mathbf{V}_s$ for source language s. Annotated sentence set Δ_t for t and Δ_s for related language s.

for $iter = 1$ to *word_epoch* **do**
  for $a = 1$ to $word\_dis\_steps$ **do**
    sample a batch of words $\mathbf{b}_t \sim \mathbf{V}_t$, $\mathbf{b}_s \sim \mathbf{V}_s$
    $loss = L_{dis}^w([\mathbf{b}_t, f(\mathbf{b}_s)])$
    update Θ_{dis} to minimize $loss$
  sample a batch of words $\mathbf{b}_t^{'} \sim \mathbf{V}_t$, $\mathbf{b}_s^{'} \sim \mathbf{V}_s$
  $loss^{'} = L_f^w([\mathbf{b}_t^{'}, f(\mathbf{b}_s^{'})])$
  update Θ_f to minimize $loss^{'}$
build a parallel dictionary with $\mathbf{V}_t$ and $f(\mathbf{V}_s)$ and refine projected word embeddings $\widetilde{\mathbf{V}}_s = f(V_s)$
for $iter = 1$ to seq_epoch **do**
  sample a batch of sentences $\tilde{\mathbf{b}}_t \sim \Delta_t$, $\tilde{\mathbf{b}}_s \sim \Delta_s$
  extract sequential features from $\tilde{\mathbf{b}}_t$, $\tilde{\mathbf{b}}_s$ with E
  $loss = L_{dis}^x([E(\tilde{\mathbf{b}}_t), E(\tilde{\mathbf{b}}_s)])$
  update Θ_e, $\Theta_{\widetilde{dis}}$ to minimize $loss$
  for $g = 1$ to seq_tagger_steps do
    sample a batch of sequences $\tilde{\mathbf{b}}_t^{'} \sim \Delta_t$, $\tilde{\mathbf{b}}_s^{'} \sim \Delta_s$
    $loss^{'} = L_e^x([E(\tilde{\mathbf{b}}_t^{'}), E(\tilde{\mathbf{b}}_s^{'})]) + L_{crf}([\tilde{\mathbf{b}}_t^{'}, \tilde{\mathbf{b}}_s^{'}])$
    update Θ_e, Θ_c to minimize $loss^{'}$

$Y_{i,z}$ represents the unary score of assigning tag z to the i-th word.

Given the annotated sequence of tags $\mathbf{Z}$, the CRF loss is:

$$L_{crf} = \log \sum_{\mathbf{Z}' \in \tilde{\mathbf{Z}}} e^{Score(x, \mathbf{Y}, \mathbf{Z}')} - Score(x, \mathbf{Y}, \mathbf{Z})$$

where $\tilde{\mathbf{Z}}$ is the set of all possible tagging paths. The parameter set for optimizing the name tagger can be denoted as $\Theta_c = \{\theta_{E_c}, \theta_\ell, \theta_{\text{CRF}}\}$.

We jointly optimize the sequence encoder E, the context encoder E_c and the CRF together by minimizing the loss $L^{'} = L_e^x + L_{crf}$, and successively minimize L_{dis}^x and $L^{'}$ with SGD. The end-to-end training for our neural architecture is described in Algorithm 1.

3 Experiment

3.1 Data and Experimental Setup

We evaluate our methods from multiple settings. We first evaluate our architecture on 10 low-resource languages from the DARPA LORELEI project. The annotations are released by the Linguistic Data Consortium (LDC).[4] Each dataset has four predefined name types: person (PER), organization (ORG), location (LOC) and geo-political entity (GPE). For each target low-resource language, we choose a source language if they are from the same language family or use the same script. To show the impact of resource transfer between distinct languages, we also use English as a source language for each target low-resource language. We create the English annotated resource by combining the TAC-KBP 2015 English Entity Discovery and Linking (Ji et al., 2015) data set and the Automatic Content Extraction (ACE2005) data set.[5] To avoid the impact of parameter initialization, we perform 5-fold cross validation. For each experiment, we run twice and get the averaged F-score. Table 1 shows the statistics of each data set.

We also evaluate our approach on high-resource languages. We use Dutch (nl) and Spanish (es) data sets from the CoNLL 2002 (Tjong Kim Sang, 2002) shared task as target languages, and use English (en) data from the CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) shared task as the source language. All the data sets have four pre-defined name types: PER, ORG, LOC and miscellaneous (MISC). Table 2 shows the statistics of these data sets.

Language	# of Sents	# of Tokens	# of Names
Amharic (am)	4,770	71,399	3,891
Tigrinya (ti)	5,023	95,364	6,201
Arabic (ar)	4,781	80,715	4,937
Farsi (fa)	3,855	72,629	3,966
Oromo (om)	2,987	52,876	4,985
Somali (so)	3,453	78,400	5,571
Swahili (sw)	4,155	96,902	6,044
Yoruba (yo)	1,599	46,084	2,016
Uyghur (ug)	3,961	60,999	2,575
Uzbek (uz)	11,135	177,816	10,937
English (en)	17,936	388,120	23,938

Table 1: Data set statistics for each low-resource language.

For fair comparison, we use the same pre-trained word embeddings of English, Dutch and Spanish as Lin et al. (2018), while for each low-resource language we train their word embeddings using the documents from their LDC packages with FastText.[6] Table 3 lists the key hyper-parameters we used in our experiments.

3.2 Baselines

We compare our methods with three categories of baseline methods:[7]

- **Monolingual Name Tagging** Using monolingual annotations only, the current state-of-the-art name tagging model is the Bi-LSTM-CRF network (Huang et al., 2015; Lample et al., 2016; Ma and Hovy, 2016).[8]

- **Multi-task Learning** Lin et al. (2018) apply multi-task learning to boost name tagging performance by introducing additional annotations from source languages using a weight sharing context encoder across multiple languages.

- **Language Universal Representations** We apply word adversarial transfer only to project the source language into the same semantic space as the target language, then train the name tagger on the annotations of source and target languages. Word-Adv[1] refers to the approach which is directly trained on the combination of the anno-

[4] The annotations are from: am (LDC2016E87), ti (LDC2017E39), ar (LDC2016E89), fa (LDC2016E93), om (LDC2017E27), so (LDC2016E91), sw (LDC2017E64), yo (LDC2016E105), ug (LDC2016E70), uz (LDC2016E29)

[5] The data sets are LDC2015E103 and LDC2006T06

[6] https://fasttext.cc/

[7] All the baselines are trained for 100 epochs

[8] For each word, we also combine its word embedding with a CharCNN based representation.

Language	Resource	Train	Dev	Test
English (en)	source language	204,567 (23,499)	51,578 (5,942)	46,666 (5,648)
Dutch (nl)	target language	202,931 (13,344)	37,761 (2,616)	68,994 (3,941)
Spanish (es)	target language	264,715 (18,797)	52,923 (4,351)	51,533 (3,558)

Table 2: CoNLL data set statistics: # of tokens and # of names (between parentheses).

Parameter Name	Value
Monolingual Embedding Size	100
CharCNN Filter Size	25
CharCNN Filter Widths	[2, 3]
LSTM Hidden Size	100
Droupout Rate	0.5
Smoothing Value ϵ for Word Discriminator	0.1
Word Adversarial Training Epochs	5
Smoothing Value η for Sequence Discriminator	0.3
Sequence Adversarial & Name Tagging Training Epochs	60
# of Steps for Sequence Tagging Training	5
Batch Size	20
Initial Learning Rate	0.01
Optimizer	SGD

Table 3: Hyper-parameters.

target (source)	Cross-lingual Word-Adv[1]	Multitask Learning	Our Approach
am (ti)	15.19	19.72	**26.86**
ti (am)	16.20	9.06	**29.36**
ar (fa)	1.53	3.52	**13.83**
fa (ar)	2.59	0.91	**11.14**
om (so)	4.66	3.40	**14.14**
so (om)	4.12	2.98	**20.02**
sw (yo)	7.20	5.60	**18.25**
yo (sw)	13.07	6.14	**23.73**

Table 4: Cross-lingual transfer when the target language has no resources (F-score %).

tations, while Word-Adv[2] refers to the baseline that is first trained on the target language annotations and then further tuned on the related language annotations.

3.3 Cross-lingual Transfer with Zero Target Language Annotated Resource

We first evaluate our approach on a cross-lingual transfer setting without using any annotated training data from the target language. We conduct experiments on 8 low-resource languages. Among those, some pairs, such as Amharic (am) and Tigrinya (ti), Oromo (om) and Somali (so), or Yoruba (yo) and Swahili (sw), are from the same language family and are closely related, while some are not, such as Arabic (ar) and Farsi (fa). Since our approach requires some unlabeled sentences from the target language to train the sentence-level discriminator, we entirely remove the annotations from the annotated data set of the target language. Table 4 presents the results.

Our approach significantly outperforms the previous methods on all languages. Specifically, compared with the Word-Adv[1] baseline, which only performs word-level adversarial transfer, our approach achieves 10% absolute F-score gain on average, which demonstrates the effectiveness of the sentence-level adversarial transfer. In addition, compared with Lin et al. (2018), who only apply a shared context-encoder to transfer the knowledge, our approach not only includes a language-sharing encoder, but also performs multi-level adversarial training to encourage the semantic alignment of words from both languages and a sequence encoder to extract language-agnostic sequential features.

Here we use some Arabic (Farsi) examples to further show the effectiveness of each level of adversarial training in our architecture. Without using any annotated training data from Arabic, both our approach and the Word-Adv[1] baseline successfully identify الفرنسية (*French*) as a *GPE* from the Arabic (ar) sentence in Figure 3, since with word-level adversarial training, the semantics of الفرنسية is well aligned with the *GPE* names in Farsi annotated data, such as فرانسه (*France*), روسیه (*Russia*) and آلمان (*Germany*). However, both the Word-Adv[1] and Lin et al. (2018) baselines fail to identify الجزائرية (*Algerian*) as a *GPE* since its top ranked similar words in Farsi include مذاکرات (*negotiations*), دوحه (*Doha*) and توافقنامه (*agreement*). With sentence-level adversarial training, our approach successfully captures language-agnostic sequential features, such as "او (*or*) *usually connects two names with the same type*", thus our approach successfully identifies الجزائرية (*Algerian*) as a *GPE* name.

3.4 Cross-lingual Transfer for Low-Resource Languages

We also investigate the impact of cross-lingual transfer when the target languages have some annotated resources. For each target low-resource language, we explore the use of a related low-resource language vs. using the high-resource En-

target (related)	Monolingual Bi-LSTM-CRF	Cross-lingual Embedding Word-Adv[1]	Cross-lingual Embedding Word-Adv[2]	Multitask Learning	Our Approach Multi-Adversarial
am (ti)	72.23	72.15	72.01	72.35	**73.98**
ti (am)	74.68	74.43	74.83	74.71	**74.93**
ar (fa)	48.92	48.37	47.90	47.53	**49.76**
fa (ar)	64.35	63.93	64.43	63.21	**65.09**
om (so)	76.37	76.43	76.19	76.18	**77.19**
so (om)	77.63	77.31	77.13	77.99	**78.15**
sw (yo)	77.01	77.31	77.85	**77.86**	76.28
yo (sw)	68.97	68.89	69.62	70.12	**70.59**
ug (uz)	68.73	68.53	68.29	68.39	**69.46**
uz (ug)	74.59	74.21	74.74	74.56	**75.37**
am (en)	72.23	72.43	71.63	72.22	**73.35**
ti (en)	74.68	74.61	74.69	74.68	**74.80**
ar (en)	48.92	48.50	47.91	47.40	**50.08**
fa (en)	**64.35**	64.04	64.25	63.44	63.92
om (en)	76..27	76.68	76.53	76.2	**77.29**
so (en)	77.63	76.67	77.88	77.88	**78.21**
sw (en)	77.01	77.52	76.84	**77.89**	77.01
yo (en)	68.97	69.21	69.46	70.43	**70.88**
ug (en)	68.73	68.14	68.79	68.69	**69.06**
uz (en)	74.59	73.95	74.46	74.48	**74.75**

Table 5: Cross-lingual transfer when the target language has resources (F-score %).

AR: ويكون نائب المدعي العام قد اعتبر ان الادلة ضد الموقوفين الذين يحملون الجنسية الفرنسية[3] او[2] الجزائرية[1] في غالبيتهم ، كافية .

EN: The deputy prosecutor has ruled that the evidence against those with French[3] or[2] Algerian[1] nationality is mostly sufficient.

Figure 3: Example of an Arabic (ar) name tagging output with Farsi (fa) annotated training data only.

glish as our source language. Table 5 shows the performance on 10 low-resource languages.

Comparing cross-lingual embedding based baselines to the monolingual baseline, we observe that for most low-resource languages, directly adding the annotations from the source language to the target language slightly hurts the model. This suggests that when the training data for the target language is not enough, the model will be very sensitive to noise. The multitask learning based baseline (Lin et al., 2018) performs better than the monolingual baseline only when the target and source languages are very close, such as Amharic (am) and Tigrinya (ti), or Swahili (sw) and Yoruba (yo).

By introducing annotated training data from English, the performance of all the baselines becomes worse than the monolingual baseline. Since the script and sequence structure of English is very different from these low-resource languages, the addition of English to the limited target language training data yields a considerably noisy corpus. However, by forcing the sequence encoder to extract language-agnostic features, our approach still achieves better performance than the monolingual baseline for most languages. All of these experiments demonstrate that our approach is more effective in leveraging annotations from other languages to improve target language name tagging.

3.5 Cross-lingual Transfer for High Resource Languages

Language	Model	F-score
Dutch	Lample et al. (2016)	81.74
	Yang et al. (2017)	85.19
	Lin et al. (2018)	85.71
	Gillick et al. (2016)	82.84
	Word-Adv[1]	85.87
	Word-Adv[2]	86.43
	Our Model (Bi-LSTM)	**86.87**
Spanish	Lample et al. (2016)	85.75
	Yang et al. (2017)	85.77
	Lin et al. (2018)	85.02
	Gillick et al. (2016)	82.95
	Word-Adv[1]	85.92
	Word-Adv[2]	85.84
	Our Model (Bi-LSTM)	**86.41**

Table 6: Comparison on cross-lingual transfer for Dutch and Spanish with various baselines: monolingual baseline (Lample et al. (2016)), multitask baselines (Yang et al. (2017) and Lin et al. (2018)), language universal representation baselines (Gillick et al. (2016), Word-Adv[1], Word-Adv[2]).

We finally investigate the results when both the source and target languages are all high-resource

languages. Table 6 presents the performance on Dutch and Spanish while using English as the source language. Our approach significantly outperforms all the other approaches even when the size of the annotated training data for the target language is huge. We notice that our approach achieves larger improvement on Dutch than Spanish. The reason may be that, compared with Spanish, Dutch is much closer to English (Cutler and Pasveer, 2006). Both English and Dutch are from the same *West Germanic* branch of the *Indo-European* language family while Spanish is from the *Italic* branch.

3.6 Impact of Annotation Size from Source and Target Languages

We use Amharic as the target language and Tigrinya as the source language to show the impact of the size of their annotations. Specifically, to explore the impact of the size of target language annotations, we use 0, 10%, 50%, or 100% annotated training data from Amharic. Similarly, to show the effect of the size of source language annotations, for each experiment, we also gradually add 0, 20%, 50%, or 100% annotated training data from Tigrinya. For all experiments, we use the same dev and test set of Amharic. As Figure 4 shows, as we gradually add annotations from the source or target language, the performance can always be improved. When the size of target language annotations is small, such as 400 sentences, we can achieve 5%-30% F-score gain by adding about 4,000 sentences from the source language. When the size of target language annotations is over 2,000 sentences, the improvement is about 2% if we add in about 4,000 sentences from source language annotations.

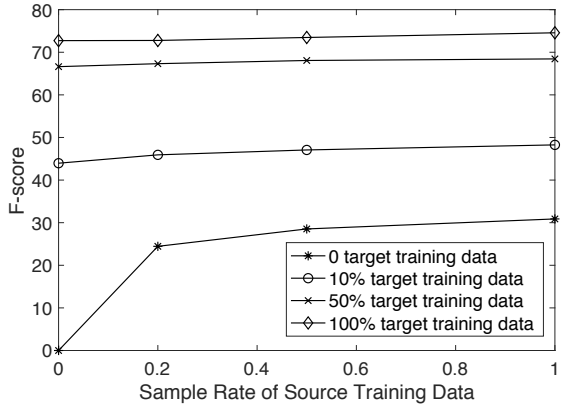

Figure 4: The impact of the size of annotations from source and target languages on Amharic name tagging.

4 Related Work

Name tagging methods based on sequence labeling have been widely studied in recent years. Huang et al. (2015) and Lample et al. (2016) propose an effective Bi-LSTM-CRF architecture; the Bi-LSTM encodes previous and following contexts, and the CRF is used for tag prediction. Other studies incorporate a character-level CNN (Ma and Hovy, 2016), global contexts (Zhang et al., 2018), or language models (Liu et al., 2018; Peters et al., 2017, 2018; Devlin et al., 2018) to improve name tagging. In addition, several approaches (Zhang et al., 2016a, 2017a; Al-Badrashiny et al., 2017) attempt to incorporate hand-crafted linguistic features into a Bi-LSTM-CRF to improve low-resource name tagging performance.

Recent attempts on cross-lingual transfer for name tagging can be divided into two categories: the first projects annotations from a source language to a target language via parallel corpora (Yarowsky et al., 2001; Wang and Manning, 2013; Wang et al., 2013; Zhang et al., 2016b; Fang and Cohn, 2016; Ehrmann et al., 2011; Enghoff et al., 2018; Ni et al., 2017), a bilingual gazetteer (Feng et al., 2017; Zirikly and Hagiwara, 2015), Wikipedia anchor links (Kim et al., 2012; Nothman et al., 2013; Tsai et al., 2016; Pan et al., 2017), and language universal representations, including Unicode bytes (Gillick et al., 2016) and cross-lingual word embeddings (Fang and Cohn, 2017; Wang et al., 2017; Huang et al., 2018; Xie et al., 2018). The second is based on multitask learning via a weight sharing encoder (Yang et al., 2016, 2017; Lin et al., 2018). Compared to these studies, our approach not only automatically learns cross-lingual word embeddings without requiring any parallel resources, but also carefully extracts language-agnostic sequential features, yielding better performance.

Adversarial training has also been extensively studied and applied for cross-lingual and cross-domain transfer. Several studies (Barone, 2016; Zhang et al., 2017c,b; Conneau et al., 2017; Chen and Cardie, 2018) explore adversarial training to automatically induce bilingual and multilingual word representations without using any parallel corpora or bilingual gazetteers. Adversarial training is also applied to extract language-agnostic (Chen et al., 2016; Zou et al., 2018; Wang and Pan, 2018; Kim et al., 2017a; Muis et al.,

2018; Cao et al., 2018) and domain-agnostic features (Kim et al., 2017b; Ganin et al., 2016; Tzeng et al., 2017; Chen et al., 2017; Li et al., 2017; Fu et al., 2017; Bousmalis et al., 2016; Shi et al., 2018) for cross-lingual and cross-domain adaptation. Compared with these methods, our approach combines both word-level and sentence-level adversarial training.

5 Conclusions and Future Work

We design a new neural architecture which integrates multi-level adversarial transfer into a Bi-LSTM-CRF to improve low-resource name tagging. With word-level adversarial training, it can automatically project the source language into a shared semantic space with the target language without requiring any comparable data or bilingual gazetteers. Moreover, considering the different underlying sequential structures among various languages, we further design a sentence-level adversarial transfer to encourage the sequence encoder to extract language-agnostic features. The experiments show that our approach achieves the state-of-the-art on both CoNLL data sets and 10 low-resource languages. In the future, we will further explore selecting the feature-consistent annotations from the source language and add to the target language, and explore unsupervised pretrained cross-lingual language models (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018; Lample and Conneau, 2019) for cross-lingual low resource name tagging.

Acknowledgments

This research is based upon work supported in part by U.S. DARPA LORELEI Program # HR0011-15-C-0115, and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract # FA8650-17-C-9116, and ARL NS-CTA No. W911NF-09-2-0053. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Mohamed Al-Badrashiny, Jason Bolton, Arun Tejasvi Chaganty, Kevin Clark, Craig Harman, Lifu Huang, Matthew Lamm, Jinhao Lei, Di Lu, Xiaoman Pan, et al. 2017. Tinkerbell: Cross-lingual cold-start knowledge base construction. In *Proceedings of TAC 2017*.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. *Proceedings of ACL 2016*.

Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT 2010*.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Proceedings of NIPS 2016*.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2018. Adversarial transfer learning for chinese named entity recognition with self-attention mechanism. In *Proceedings of EMNLP 2018*, pages 182–192.

Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of ICCV 2017*.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of EMNLP 2018*.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv preprint arXiv:1606.01614*.

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

Anne Cutler and Dennis Pasveer. 2006. Explaining cross-linguistic differences in effects of lexical stress on spoken-word recognition. In *3rd International Conference on Speech Prosody*. TUD press.

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 RANLP 2011*, pages 118–124.

Jan Vium Enghoff, Sren Harrison, and Zeljko Agi. 2018. Low-resource named entity recognition via multi-source projection: Not quite there yet? In *The 4th Workshop on Noisy User-generated Text*.

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource pos tagging using cross-lingual projection. *Proceedings of CoNLL 2016*.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of ACL 2017*.

Xiaocheng Feng, Lifu Huang, Bing Qin, Ying Lin, Heng Ji, and Ting Liu. 2017. Multi-level cross-lingual attentive neural architecture for low resource name tagging. *Tsinghua Science and Technology*, pages 633–645.

Lisheng Fu, Thien Huu Nguyen, Bonan Min, and Ralph Grishman. 2017. Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of IJCNLP 2017*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of NAACL-HLT 2016*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of NIPS 2014*.

Lifu Huang, Kyunghyun Cho, Boliang Zhang, Heng Ji, and Kevin Knight. 2018. Multi-lingual common semantic space construction via cluster-consistent word embedding. *Proceedings of EMNLP 2018*.

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

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of tac-kbp2015 tri-lingual entity discovery and linking. In *Proceedings of TAC 2015*.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017a. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of EMNLP 2017*.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of ACL 2012*, pages 694–702.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017b. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of ICML 2017*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Proceedings of AAAI 2016*, pages 2741–2749.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of NIPS 2012*, pages 1097–1105.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT 2016*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *Proceedings of IJCAI 2017*.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A multi-lingual multi-task architecture for low-resource sequence labeling. In *Proceedings of ACL 2018*, volume 1, pages 799–809.

Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Fangzheng Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *Proceedings of AAAI 2018*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of ACL 2016*.

Aldrian Obaja Muis, Naoki Otani, Nidhi Vyas, Ruochen Xu, Yiming Yang, Teruko Mitamura, and Eduard Hovy. 2018. Low-resource cross-lingual event type detection via distant supervision with minimal effort. In *Proceedings COLING 2018*.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of ACL 2017*.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, pages 151–175.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of ACL 2017*, pages 1946–1958.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of ACL 2017*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL 2018*.

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

Ge Shi, Chong Feng, Lifu Huang, Boliang Zhang, Heng Ji, Lejian Liao, and Heyan Huang. 2018. Genre separation network with adversarial training for cross-genre relation extraction. In *Proceedings of EMNLP 2018*.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL 2002*, pages 1–4.

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 HLT-NAACL 2003*.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of CoNLL 2016*.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of CVPR 2017*.

Dingquan Wang, Nanyun Peng, and Kevin Duh. 2017. A multi-task learning approach to adapting bilingual word embeddings for cross-lingual named entity recognition. In *Proceedings of IJCNLP 2017*.

Mengqiu Wang, Wanxiang Che, and Christopher D Manning. 2013. Joint word alignment and bilingual named entity recognition using dual decomposition. In *Proceedings of ACL 2013*, pages 1073–1082.

Mengqiu Wang and Christopher D Manning. 2013. Cross-lingual pseudo-projected expectation regularization for weakly supervised learning. *arXiv preprint arXiv:1310.1597*.

Wenya Wang and Sinno Jialin Pan. 2018. Transition-based adversarial network for cross-lingual aspect extraction. In *Proceedings of IJCAI 2018*, pages 4475–4481.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of EMNLP 2018*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of NAACL 2015*.

Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *arXiv preprint arXiv:1603.06270*.

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 HLT 2001*.

Boliang Zhang, Di Lu, Xiaoman Pan, Ying Lin, Halidanmu Abudukelimu, Heng Ji, and Kevin Knight. 2017a. Embracing non-traditional linguistic resources for low-resource language name tagging. In *Proceedings of IJCNLP 2017)*, pages 362–372.

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

Boliang Zhang, Spencer Whitehead, Lifu Huang, and Heng Ji. 2018. Global attention for name tagging. In *Proceedings of CoNLL 2018*, pages 86–96.

Dongxu Zhang, Boliang Zhang, Xiaoman Pan, Xiaocheng Feng, Heng Ji, and XU Weiran. 2016b. Bitext name tagging for cross-lingual entity annotation projection. In *Proceedings of COLING 2016*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of ACL 2017*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017c. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of EMNLP 2017*.

Ayah Zirikly and Masato Hagiwara. 2015. Cross-lingual transfer of named entity recognizers without parallel corpora. In *Proceedings of ACL 2015*.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of COLING 2018*.