A Multi-task Learning Approach to Adapting Bilingual Word Embeddings for Cross-lingual Named Entity Recognition

Dingquan Wang[1] **Nanyun Peng**[3]* **Kevin Duh**[12]
[1] Center for Language and Speech Processing, Johns Hopkins University
[2] Human Language Technology Center of Excellence, Johns Hopkins University
[3] Information Sciences Institute, University of Southern California
wdd@cs.jhu.edu, npeng@isi.edu, kevinduh@cs.jhu.edu

Abstract

We show how to adapt bilingual word embeddings (BWE's) to bootstrap a cross-lingual name-entity recognition (NER) system in a language with no labeled data. We assume a setting where we are given a comparable corpus with NER labels for the source language only; our goal is to build a NER model for the target language. The proposed multi-task model jointly trains bilingual word embeddings while optimizing a NER objective. This creates word embeddings that are both shared between languages and fine-tuned for the NER task. As a proof of concept, we demonstrate this model on English-to-Chinese transfer using Wikipedia.

1 Introduction

Cross-lingual transfer is an important technique for building natural language processing (NLP) systems for *low-resource* languages, where labeled examples are scarce. The main idea is to transfer labels or models from *high-resource* languages. Representative techniques include (a) projecting labels (or information derived from labels) across parallel corpora (Yarowsky et al., 2011; Das and Petrov, 2011; Che et al., 2013; Zhang et al., 2016), and (b) training universal models using unlexicalized features (McDonald et al., 2011; Täckström et al., 2012; Zirikly and Hagiwara, 2015) or bilingual word embeddings (Xiao and Guo, 2014; Gouws and Søgaard, 2015).

Here, we focus on the bilingual word embedding (BWE) approach. In particular, we are interested in leveraging recent advances in learning BWE from comparable corpora (Hermann and Blunsom, 2014; Vulic and Moens, 2015; Gouws and Søgaard, 2015). A comparable corpus is a collection of document pairs written in different languages but talking about the same topic (e.g. interconnected Wikipedia articles). The advantage of comparable corpora is that they may be more easily acquired in the language and domain of interest. However, cross-lingual transfer on comparable corpora is more difficult than on parallel corpora, due to the difficulty in finding high-quality word translation equivalences.

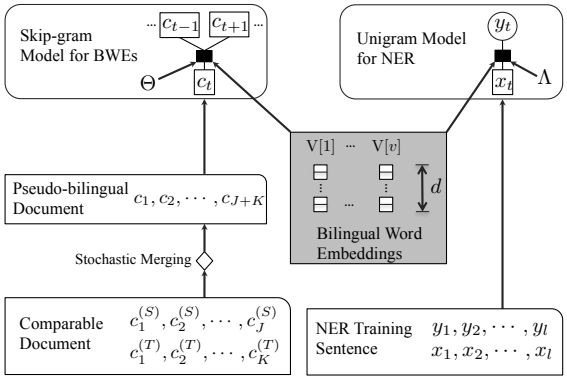

Figure 1: Our multi-task framework, which trains bilingual word embeddings from comparable corpora while optimizing an NER objective on the high-resource language. The NER part of the model is then tested on a low-resource language.

Our contributions are two-fold: First, we investigate cross-lingual transfer on an NER task, and found that pre-trained BWE's do not necessarily help out-of-the-box. This corroborates results in the monolingual setting, where it is widely recognized that training task-specific embeddings is helpful for the downstream tasks like NER (Peng and Dredze, 2015; Ma and Hovy, 2016).

Second, we propose a multi-task learning framework that utilizes comparable corpora to jointly train BWE's and the downstream NER task (Figure 1). We experimented with a Wikipedia

* This research was majorly conducted when the author was at Johns Hopkins University.

corpus, training a NER model from labeled English articles (high-resource) and testing it on Chinese articles (low-resource)[1]. The challenge with training task-specific embeddings in cross-lingual transfer is that the task in which we have labels (English NER) is not equivalent to the task we care about (Chinese NER). Despite this, we demonstrate improvements on NER F-scores with our multi-task model.

2 The Multi-task Framework

Assumptions : We assume two resources: First is a comparable corpus where S ("source") refers to the high-resource language and T ("target") refers to the low-resource language. The comparable corpus is denoted as $\mathcal{C} = \{(\mathbf{c}_i^{(S)}, \mathbf{c}_i^{(T)}) \mid i \in [1, M]\}$, where each $(\mathbf{c}_i^{(S)}, \mathbf{c}_i^{(T)})$ is a tuple of comparable documents written in S and T, and M is the size (total number of tuples) of $\mathcal{C}$.

We also assume a labeled NER corpus on the high-resource language, which may be disjoint from $\mathcal{C}$. Let $X^{(S)} = \{\mathbf{x}_i^{(S)} \mid i \in [1, N^{(S)}]\}$ and $Y^{(S)} = \{\mathbf{y}_i^{(S)} \mid i \in [1, N^{(S)}]\}$ together form the NER training examples of S, where each $\mathbf{y}_i^{(S)}$ is the gold tag sequence of sentence $\mathbf{x}_i^{(S)}$, and $N^{(S)}$ is the number of training examples.

Training : Given $X^{(S)}$ and $Y^{(S)}$ and $\mathcal{C}$ the training objective (loss) L is:

$$\alpha \underset{X^{(S)},Y^{(S)}}{\mathrm{Ln}}(\mathrm{V}, \Lambda) + (1-\alpha)\underset{\mathcal{C}}{\mathrm{Lm}}(\mathrm{V}, \Theta) \quad (1)$$

Ln is the loss for training the NER tagger in S, Lm is the loss for training the BWE's, and $\alpha \in [0, 1]$ is coefficient for balancing these two losses. Λ, V and Θ are the model parameters, where Λ is Ln-specific parameter, Θ is the Lm-specific parameter and V is the $d \times v$-shape BWE's that shared are by both Ln and Lm. v is the size of the joint vocabulary $\mathcal{V}$ and $\mathcal{V}$ is formed by *concatenating*[2] the vocabulary of S and T, d is the dimension of the word embedding. Figure 1 gives a visualization of the framework.

Evaluation : At test time, given $X^{(T)}$ – the raw sentences of T, we evaluate the F1 score of $\bar{Y}^{(T)}$ predicted by the trained model $\{\mathrm{V}^*, \Lambda^*\}$ against the true label $Y^{(T)}$. Note that this model is trained on NER labels in S only, so it is imperative for the learned BWE's to map S and T words with the same NER label into nearby spaces.

Our framework (Equation 1) is flexible to different definitions of Ln and Lm objectives. Below, we describe a specific instantiation that fits well with cross-lingual NER.

2.1 Design of Ln

Given the labeled training data $X^{(S)}, Y^{(S)}$ of S, we optimize the conditional log-likelihood as Ln (superscripts ^{S} are suppressed for readability):

$$\underset{X,Y}{\mathrm{Ln}}(\mathrm{V}, \Lambda) = \frac{1}{N}\sum_{i=1}^{N} \log p_{\mathrm{V},\Lambda}(\mathbf{y}_i \mid \mathbf{x}_i) \quad (2)$$

$p_{\mathrm{V},\Lambda}(\mathbf{y} \mid \mathbf{x})$ is the conditional probability of $\mathbf{y}$ given $\mathbf{x}$ parameterized by V and Λ such that

$$p_{\mathrm{V},\Lambda}(\mathbf{y} \mid \mathbf{x}) = \frac{\exp(s_{\mathrm{V},\Lambda}(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y}' \in \mathcal{Y}} \exp(s_{\mathrm{V},\Lambda}(\mathbf{x}, \mathbf{y}'))} \quad (3)$$

. $s_{\mathrm{V},\Lambda}(\dots)$ is a score function of a sentence and its possible NER-tag sequence. $\mathcal{Y}$ is the set of all possible NER-tag sequences.

Unigram Model : Given a sentence $\mathbf{x} \in X$ with length l and its label $\mathbf{y} \in Y$, the score function of unigram model is simply:

$$s_{\mathrm{V},\Lambda}(\mathbf{x}, \mathbf{y}) = \sum_{t \in [1,l]} \mathrm{V}[x_t]^\top \cdot W[y_t] + \mathbf{b}[y_t] \quad (4)$$

where $\Lambda \overset{\text{def}}{=} \{W, b\}$, W is a $d \times n$-shape matrix that maps the word vector $\mathrm{V}[x_t]$ into the NER-tag space, n is the number of NER-tag types, which is 7 as will be explained in §3.1. $\mathbf{b}$ is the bias vector with size n. The reason why we call this model unigram is it only looks at the word itself without its context.[3]

2.2 Design of Lm

Here we adopt the method of Vulic and Moens (2015), which is the only mechanism for training on comparable documents as far as we know: First we transform $\mathcal{C}$ into a *pseudo-bilingual* corpora $\mathcal{C}'$ by a stochastic merging process of two documents as shown in Alg. 1. The idea is to mix together the two documents in different languages into a single document (where S and T words are interspersed), then apply a standard monolingual word embedding algorithm.

[1] Chinese can be considered a high-resource language for NER, but we use it as a proof-of-concept and do not use any existing Chinese resources.

[2] Same word in different languages is treated separately.

[3] Besides unigram, we also tried LSTM+CRF (Lample et al., 2016) for longer context. Despite good results in monolingual NER, it did poorly in our cross-lingual experiments.

Algorithm 1 Stochastic merging of two documents, where $\text{len}(\mathbf{c})$ returns the number of tokens of document $\mathbf{c}$, $\mathbf{c}[i]$ is the i^{th} token of document $\mathbf{c}$.

Input: Comparable Document: $\mathbf{c}^{(S)}, \mathbf{c}^{(T)}$
Output: Pseudo-bilingual Document: $\mathbf{c}$

$\mathbf{c} \leftarrow []$; $i_1 \leftarrow 0$; $i_2 \leftarrow 0$
$r \leftarrow \frac{\text{len}(\mathbf{c}^{(S)})}{\text{len}(\mathbf{c}^{(S)})+\text{len}(\mathbf{c}^{(T)})}$
while $i_1 < \text{len}(\mathbf{c}^{(S)})$ **and** $i_2 < \text{len}(\mathbf{c}^{(T)})$ **do**
  $p \sim \text{Uniform}[0, 1]$
  if $i_2 == \text{len}(\mathbf{c}^{(T)})$ **or** $p < r$ **then**
    $\mathbf{c}.\text{append}(\mathbf{c}^{(S)}[i_1])$
    $i_1 \leftarrow i_1 + 1$
  else
    $\mathbf{c}.\text{append}(\mathbf{c}^{(T)}[i_2])$
    $i_2 \leftarrow i_2 + 1$
return $\mathbf{c}$

We use the standard skip-gram objective (Mikolov et al., 2013) by considering each pseudo-bilingual document as a single sentence. $\underset{\mathcal{C}}{\text{Lm}}(\text{V}, \Theta)$ is given by:

$$\underset{\mathbf{c}\in\mathcal{C}'}{\text{mean}} \underset{t\in[1,\text{len}(\mathbf{c})]}{\text{mean}} \sum_{\substack{-w\le j\le w \\ j\neq 0}} \log p_{\text{V},\Theta}(c_{t+j}|c_t) \quad (5)$$

, where w is the word window, c_t is the t^{th} token of $\mathbf{c}$. $p_{\text{V},\Theta}(\dots)$ is the standard context probability parameterized by V and Θ such that

$$p_{\text{V},\Theta}(c_O|c_I) = \frac{\text{V}[c_I]^\top \cdot \text{V}'[c_O]}{\sum_{c'_O\in\mathcal{V}} \text{V}[c_I]^\top \cdot \text{V}'[c'_O]} \quad (6)$$

$\Theta \overset{\text{def}}{=} \{\text{V}'\}$, where V' is the context embedding with size $d \times v$. In the implementation, we use negative sampling to save the computation, since we desire using a large vocabulary to handle as many words as possible for cross-lingual transfer.

2.3 Optimization

Full Joint (FJ) Training : First optimize Ln by updating V and Λ. Then optimize Lm by updating V, Θ. Repeat.

Half-fixed Joint (HFJ) Training : Same as FJ Training, except in the Lm optimization step, the English word embeddings are fixed and only the Chinese word embeddings are updated. The motivation is to anchor the English embeddings to fit the NER objective Ln, and encourage the Chinese embedding (of comparable documents) to move towards this anchor.

Inspired by Lample et al. (2016), for both approaches, words with frequency 1 in the NER data are replaced by OOV with probability 0.5 to so that embedding OOV could be optimized.

3 Experiments

3.1 Data

We use the EN-ZH portion of the Wikipedia Comparable Corpora[4]. For experiment purposes, we sampled 19K document pairs[5] as our comparable corpora $\mathcal{C}$. The NER labeled data on English (S) is obtained by collecting the first **paragraph** of each English document in $\mathcal{C}$ as $X^{(S)}$, and labeling it with Stanford NER tagger (Finkel et al., 2005) to generate $Y^{(S)}$.[6]

For the NER test data in Chinese (T), we separately sampled 1K documents and collected the first sentence as $X^{(T)}$. We ran automatic word segmentation[7] and manually labeled $X^{(T)}$ to generate $Y^{(T)}$. The English side of these 1K tuple is treated as held-out data for tuning NER hyperparameters, and is labeled with the same Stanford NER tagger. We use the BIO tagging scheme for 3 basic named-entity types ("LOC" for location, "ORG" for organization and "PER" for person), so the output space is 7 tags. The data statistics are shown in Table 1. The size of BWE's is about 1M with 514K being Chinese words.

	$\mathcal{C}$	$X^{(S)}$	$X^{(T)}$
#Document Tuple	19K	-	-
#Sentence	1.8M	45K	1K
#Token	24M	994K	20K

Table 1: Data statistics.

3.2 Results

We compare our multi-task model with FJ and HFJ alternate training[8] against a baseline where BWE's

[4] http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

[5] We started from 20K pairs in totall, then we first sampled out 1K from the Chinese side for the final evaluation and left 19K for training. We consider it as a reasonable number compared to Vulic and Moens (2015), which used 14K pairs for Spanish-English and 19K for Italian-English.

[6] We treat these automatic NER tags as "gold" labels to simulate a scenario where we want cross-lingual transfer on T to do at least as well as on S. But naturally our model can also use human annotations if available.

[7] https://github.com/fxsjy/jieba

[8] For training the BWE's, we borrow the same hyperparameters as Vulic and Moens (2015) – learning rate of **0.025**,

d	Baseline	HFJ	FJ
64	6.54	**13.5**	7.01
128	14.95	**20.27**	17.14
256	13.69	**24.29**	16.53
512	**21.23**	17.7	20.14

Table 2: The F1 scores on the Chinese held-out data averaged over multiple restarts. d is the dimension of the BWE's. All the BWE's are initialized by the output of Vulic and Moens (2015) trained on $\mathcal{C}$. "Baseline" fixes the BWE's and only trains Λ, "HFJ" and "FJ" are proposed joint training methods described in §2.3

are pre-trained on $\mathcal{C}$, then held fixed when training a NER tagger. This baseline corresponds to a two-step procedure where word embeddings pre-trained on comparable corpora is used as features when training an NER.[9]

Table 2 shows the F1 scores. We observe the joint training methods (HFJ & FJ) outperform the baseline method with embedding size 64, 128, and 256. For example, for $d = 256$, HFJ achieves an F1 score of 24%, compared to the baseline of 13%, implying that jointly tuning the embedding on both comparable corpora and NER objectives (HFJ) is better than fine-tuning only NER objective after training on comparable corpora (baseline). Further, HFJ results are better than FJ, implying when optimizing Lm, it is better to tune the Chinese embedding toward an anchored English embedding, rather than allow both to be updated.

We note that the trend is different for $d = 512$: it might be that as the NER model grows larger, there is a risk of overfitting[10] the English NER data and losing generality on Chinese NER. We observe similar trends when we replaced the unigram model with a LSTM+CRF (see §2.1) in Ln. This degraded results for all systems, e.g. with $d = 256$, "Baseline" F1 dropped from 13.69 to 6.07, and "FJ" dropped from 16.53 to 8.62.

It is interesting to see the impact of joint training on BWE's. In Table 3, given a Chinese query, we show the most similar words in English returned by computing the Euclidean distance between BWE's ($d = 256$). While both pre-trained and jointly-trained BWE's retrieve correct English translations, jointly-trained BWE's retrieves more words with the same NER type. For example, "Spanish", and "Greek" – the adjective forms of "Spain" and "Greece", rank highly with the pre-trained BWE's due to semantic similarity. But these may degrade NER since these adjectives are not labeled "LOC". Our multi-task model mitigates this confusion.

4 Conclusion & Future Work

We show how a multi-task learning approach can help adapt bilingual word embeddings (BWE's) to improve cross-lingual transfer. Joint training of BWE's encourages the BWE's to be task-specific, and outperforms the baseline of using pre-trained BWE's. We showed promising results on the challenging task of cross-lingual NER on comparable corpora, where the target language has no labels. Future work will aim to improve the absolute F1 scores by combining limited labels in the low-resource languages, via exploiting document structure in Wikipedia (Richman and Schone, 2008; Steinberger et al., 2011; Tsai et al., 2016; Ni and Florian, 2016; Pan et al., 2017). While we only focus on the most difficult case where the source language and target languages are not in the same family, and a bilingual dictionary is not available in this paper, it is interesting to study how this technique could be applied when the different levels of supervision are available on various language pairs in the future.

Acknowledgment

We thank Mark Dredze and Mo Yu for the useful discussion and Jason Eisner for the suggestion of some relevant previous work. Finally, we thank the anonymous reviewers for the high-quality suggestions.

negative sampling with **25** samples and subsampling rate of value **1e-4**. The dropout rate of **0.3** is decided by the best F1 score of the language-specific NER tagger on the English held-out data. The coefficient α for balancing two Ln and Lm should presumably be chosen by tuning on the labelled data for Chinese, which is not available in our setting. So we heuristically set it to **0.5** by assuming they are equally helpful. Our fully unsupervised setting has no NER training data available on the Chinese side for tuning. To prevent the training from overfitting to the English data, we heuristically early stop after **10000** pairs of alternating updates of Lm and Ln.

[9] Another possible baseline is using only Ln for training, this will not work because Ln only consists of English data, so the Chinese embeddings will stay random when English embedding are optimized, resulting random outputs on the Chinese side.

[10] When training with $d = 512$, the F1 on training data is consistently > 60 for the last epochs, which is about 50 with $d = 256$.

Query	Type	Top 8 results in English
NBA	ORG	**NBA**, rebounds, **Knicks**, **Lakers**, Lewiston-Porter, **76ers**, guard-forward, **Celtics**
NBA		**NBA**, **Lakers**, Gervin, rebounds, **Celtics**, **Cavaliers**, **Knicks**, **All-Defensive**
西班牙	LOC	**Spain**, Spanish, **Nogueruelas**, Rosanes, **Mazarete**, **Ólvega**, Marquesado, **Montija**
Spain		**Spain**, Rosanes, **Cenicientos**, **Madrid**, Sorita, **Alcahozo**, **Nogueruelas**, **Villaralto**
希腊	LOC	**Greece**, Greek, **Achaia**, annalistic, heroized, Gigantomachy, Hecabe, river-god
Greece		Hachadoor, **Greece**, Demoorjian, **Safranbolu**, **Scicli**, **Holasovice**, **Sighisoara**, **Litomysl**
卡卡	PER	**Kakashi**, **Moure**, Uzumaki, cosplayed, **Uchiha**, humanizing, **Yens**, hilarious
Kaka		**Kakashi**, **Kaka**, **Moure**, **Nedved**, **Suazo**, **Batistuta**, Uzumaki, **Quagliarella**

Table 3: Top similar words in the English given a Chinese query. "Type" is the gold named-entity type of the query. For each query, the upper row is calculated with the baseline BWE's, and the lower row is calculated with HFJ BWE's. The words with the same type as the query are bold-faced, and we observe more of these cases with HFJ.

References

Wanxiang Che, Mengqiu Wang, Christopher D. Manning, and Ting Liu. 2013. Named entity recognition with bilingual constraints. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 52–62, Atlanta, Georgia. Association for Computational Linguistics.

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.

Rose Jenny Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370. Association for Computational Linguistics.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, Denver, Colorado. Association for Computational Linguistics.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Distributed Representations without Word Alignment. In *Proceedings of ICLR*.

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.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *CoRR*, abs/1603.01354.

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, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Jian Ni and Radu Florian. 2016. Improving multilingual named entity recognition with wikipedia entity type mapping. pages 1275—1284.

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 the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alexander E Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *ACL*, pages 1–9.

Ralf Steinberger, Bruno Pouliquen, Mijail Kabadjov, Jenya Belyaeva, and Erik van der Goot. 2011. Jrc-names: A freely available, highly multilingual named entity resource. *RECENT ADVANCES IN*, page 104.

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.

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, Berlin, Germany. Association for Computational Linguistics.

Ivan Vulic and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*. ACL.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129, Ann Arbor, Michigan. Association for Computational Linguistics.

D. Yarowsky, G. Ngai, and R. Wicentowski. 2011. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Dongxu Zhang, Boliang Zhang, Xiaoman Pan, Xiaocheng Feng, Heng Ji, and Weiran Xu. 2016. Bitext name tagging for cross-lingual entity annotation projection. page 461–470.

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, Beijing, China. Association for Computational Linguistics.