Learning Transferable Representation for Bilingual Relation Extraction via Convolutional Neural Networks

Bonan Min[*], **Zhuolin Jiang**[*]
Raytheon BBN Technologies
10 Moulton St
Cambridge, MA 02138
{bonan.min,zhuolin.jiang}@raytheon.com

Marjorie Freedman[†], **Ralph Weischedel**[†]
USC/Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
{mrf,weisched}@isi.edu

Abstract

Typically, relation extraction models are trained to extract instances of a relation ontology using only training data from a single language. However, the concepts represented by the relation ontology (e.g. *ResidesIn*, *EmployeeOf*) are language independent. The numbers of annotated examples available for a given ontology vary between languages. For example, there are far fewer annotated examples in Spanish and Japanese than English and Chinese. Furthermore, using only language-specific training data results in the need to manually annotate equivalently large amounts of training for each new language a system encounters. We propose a deep neural network to learn transferable, discriminative bilingual representation. Experiments on the ACE 2005 multilingual training corpus demonstrate that the joint training process results in significant improvement in relation classification performance over the monolingual counterparts. The learnt representation is discriminative and transferable between languages. When using 10% (25K English words, or 30K Chinese characters) of the training data, our approach results in doubling F1 compared to a monolingual baseline. We achieve comparable performance to the monolingual system trained with 250K English words (or 300K Chinese characters) With 50% of training data.

[*]indicates co-first authors. These two authors made equal contribution.

[†]This work was done while the author was at Raytheon BBN Technologies.

1 Introduction

Semantic relation extraction is critical to many applications including knowledge base population and question answering. The problem is well-studied when relation-specific annotations are available in a single target language. However, the same relations can be represented using a variety of languages. While the evidence of the relation in context is language specific (e.g. *John spent several years living in Beijing* vs 约翰在北京生活了几年), the definition of relation itself is often language independent (e.g. *ResidesIn*) and the meaning should be preserved across languages.

We hypothesize that common, shared representation can be learnt when annotations are available in multiple languages and propose a bilingual relation extraction algorithm for this purpose. Our basic building blocks are Convolutional Neural Networks (CNN) with cross-lingual word embeddings (Ammar et al., 2016). This allows the system to capture lexical similarities across languages as well as phrase-level semantics. Building on CNNs with cross-lingual embeddings, the algorithm is a joint training algorithm which trains a model from annotated datasets in a pair of languages. We require that the annotated classes be consistent across languages, but do not require annotations over parallel (or comparable) text. The base system combines two objectives: an objective that predicts the correct relation labels in each dataset in one of the languages, and another objective to separately learn a shared representation across languages as well as language-specific representations. To further force the learnt representation to be discriminative among classes regardless of language, a discriminative objective for learning the ideal representation (Section 3.3) is added onto the shared, bilingual representation.

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 674–684,
Taipei, Taiwan, November 27 – December 1, 2017

The final combined algorithm essentially learns two types of useful representations: a language-independent relation-specific representation with the shared neurons, and a language-dependent relation-specific representation with the language-specific neurons.

Our contributions are the following:

- Developing a bilingual transfer learning algorithm for relation extraction that can use independent multilingual corpora annotated with the same set of relations. Analysis shows that the representation is discriminative.
- Demonstrating that jointly training from two languages outperforms its monolingual counterparts significantly.
- Showing that knowledge can be transferred from resource-rich language to resource-poor languages: On the ACE multilingual training corpus, we achieve comparable performance with 50% of the target-language training data using our approach and are able to double performance with only 10% (250K words) of target language data. This provides a very cost effective way to develop relation extractors in new languages.

2 Related Work

Relation extraction is typically cast as a multi-class classification problem in which a supervised machine learning model is trained with labeled datasets for classifying relations. Traditional methods (Kambhatla, 2004; Zhou et al., 2005; Zhao and Grishman, 2005; Jiang and Zhai, 2007) either reply on a set of linguistic or semantic features, or use convolution tree kernels (Moschitti, 2006) with syntactic (Zhang et al., 2006), sub-sequence (Bunescu and Mooney, 2005b), or dependency trees (Bunescu and Mooney, 2005a) as means to represent input sentences. Recently, deep neural networks start to show promising results in relation extraction. In particular, Convolutional Neural Networks (Zeng et al., 2014a; dos Santos et al., 2015; Nguyen and Grishman, 2015), Reccurrent/Recursive Neural Networks such as bidirectional LSTMs (Zhang et al., 2015), LSTM along shortest dependency paths (Xu et al., 2015), bidirectional tree-structured LSTM-RNNs (Miwa and Bansal, 2016) are shown to be effective. Attention mechanism (Wang et al., 2016) is also effective in further improving performance. Our baseline monolingual model is similar to (Nguyen and Grishman, 2015) and we does not require parsing or composing multiple models.

There is very little work on multilingual relation extraction. (Qian et al., 2014) proposed an active learning approach for bilingual relation extraction with pseudo parallel corpora. (Kim et al., 2010) and (Kim et al., 2014) proposed cross-lingual annotation projection approach for relation detection with parallel corpora. In contrast, our work don't require parallel corpora nor Machine Translation. More recently, (Faruqui and Kumar, 2015) applied cross-lingual projection for open-domain relation extraction in languages other than English. (Blessing and Schütze, 2012) and Compositional Universal Schema (Verga et al., 2016) performs cross-lingual relation extraction with distant supervision (Mintz et al., 2009; Riedel et al., 2010; Surdeanu et al., 2012; Hoffmann et al., 2011; Ritter et al., 2013). These works are significantly different from ours in that they either operates in the open-domain(Faruqui and Kumar, 2015) without a pre-defined relation schema, or in a distant supervision setting with a KB as source of supervision. POLY (Nakashole et al., 2012) mines relational paraphrases from multilingual sentences which can be useful for relation extraction.

Besides relation extraction, (Huang et al., 2013) performs cross-language knowledge transfer with deep neural networks for speech recognition. (Guo et al., 2016) proposed a distributed representation-based framework for cross-lingual transfer learning for dependency parsers.

3 Bilingual Relation Extraction

Given a pair of monolingual corpora in two different languages and each corpus having been annotated with sentence-level relations [1] of pre-defined types, the goal of bilingual relation extraction is to learn *discriminative* representations to identify the relation between a pair of mentions, regardless of which language the mention pair comes from (*transferable across languages*). We achieve the goal of learning discriminative representation by joint supervision of classification (softmax) loss and ideal representation loss. We achieve the additional cross-lingual transferring goal by learning shared representation across languages.

As shown in Figure 1, our CNN-based bilingual relation extraction model consists of 4 main parts: (1) an embedding layer to encode words (in bilingual space), word positions, entity types and

[1] A sentence with a pair of mentions will be annotated with a relation, if the relation holds between the pair of mentions.

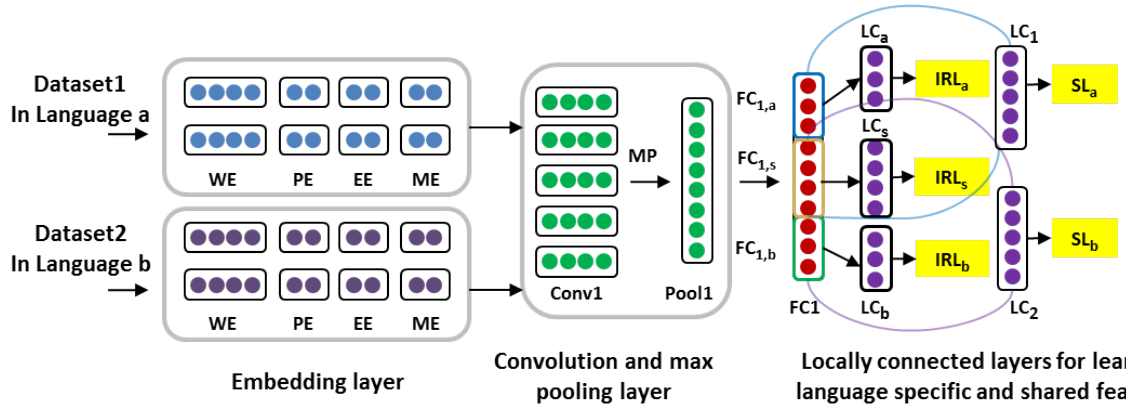


Figure 1: Bilingual relation extraction model trained with both softmax classification loss (SL) and ideal representation loss (IRL). IRL_a and IRL_b are language-specific ideal representation losses for language a and b respectively. IRL_s is the shared representation loss across 2 languages. $FC_{1,a}$ and $FC_{1,b}$ are 2 sets of language-specific neurons for language a and b, respectively. $FC_{1,s}$ is a set of shared neurons. SL_a and SL_b are the softmax classification losses for language a and b respectively.

mention levels by real-valued vectors; (2) a convolution and max pooling layer to generate a fixed-size feature vector for an input sentence; (3) 3 locally connected (LC) layers for learning language-specific and shared representations. These layers are learned to predict discriminative representation using proposed ideal representation loss (IRL) during training; (4) 2 LC layers for learning 2 relation classifiers with the softmax loss (SL). The 5 LC layers and the 5 prediction losses (3 IRLs and 2 SLs) are illustrated in Figure 1.

3.1 Embedding Layer

Word embeddings (WE) The inputs are sentences marked with pairs of mentions of interest. Given an input sentence $\mathbf{x}$ of length t, we firstly transform each word into a real-valued vector of dimension d_1 by looking up a word embedding matrix $W^1 \in \mathbb{R}^{d_1 \times |V|}$, where V is a fixed-sized vocabulary. To project similar words in a pair of languages into close proximity, we use cross-lingual embedding trained with multiCCA (Ammar et al., 2016) to initialize W^1. W^1 will also be finetuned during training.

MultiCCA only requires two monolingual corpora and a bilingual parallel dictionary. It first trains monolingual embeddings for each language independently from each monolingual corpus, capturing semantic similarity within each language. Then given the dictionary, it applies canonical correlation analysis (CCA) to estimate linear projections from the monolingual embeddings to bilingual embeddings. This makes translationally equivalent words in different languages to be embedded nearby each other.

Position embeddings (PE) The words close to the argument mentions are more informative to determine the relationship. Similar to (Santos et al., 2015), for each word, we map its relative distances to two argument mentions to two real-valued vectors of dimension d_2 by a embedding matrix $W^2 \in \mathbb{R}^{d_2 \times |D|}$, where D is the set of relative distances in a dataset. We obtain two vectors for each word with respect to the first and the second argument of the relation mention.

Entity type embeddings (EE) and mention level embeddings (ME) For each word, we map its entity type and mention level into real-valued vectors using embedding matrix $W^3 \in \mathbb{R}^{d_3 \times |E|}$ and $W^4 \in \mathbb{R}^{d_4 \times |M|}$ respectively. E is the set of entity types while M is the set of mention levels.

The final embedding dimension for each token is $n_1 = (d_1 + 2d_2 + d_3 + d_4)$. This layer will produce an embedding representation $\mathbf{x}^{(1)} \in \mathbb{R}^{n_1 \times t}$ when fed with an input sentence $\mathbf{x}^{(0)} = \mathbf{x}$. W^1, W^2, W^3, W^4 are parameters to be learnt via the end-to-end model.

3.2 Convolution and Max Pooling Layer

Relations can be expressed by words or their combinations. The model should utilize all local features extracted around each word in the sentence and predict the relation globally. Convolution operation is a natural approach to achieve this goal (Zeng et al., 2014b; Santos et al., 2015). Given a convolution filter i of window size k, the convolution operation on an input sentence $\mathbf{x}$ will produce a score vector $\mathbf{z} = (z_1, ..., z_{(t-k+1)})$, where $z_j = g_1(\mathbf{w}_i \mathbf{x}_j + b_i)$. $\mathbf{w}_i \in \mathbb{R}^{kn_1}$ are the linear transform parameters for filter i, $\mathbf{x}_j$ denotes the

j-th context window in $\mathbf{x}$, b_i is a bias scalar and g_1 is a non-linear function such as the rectified linear unit (ReLU). Then the max operation is applied to identify the most informative n-gram feature from this score vector: $m_i = \max(\mathbf{z})$. We replicate this process for a set of filters with different window sizes to capture important n-gram features from an input sentence. The matrix $W^{(2)} = [\mathbf{w}_1, ..., \mathbf{w}_{n_2}]$, where n_2 is the total number of filters, and vector $\mathbf{b}^{(2)} = [b_1, ..., b_{n_2}]$ are parameters to be learnt in this convolution layer. Finally we obtain a fixed-sized feature vector $\mathbf{x}^{(2)} = \mathbf{m} = [m_1, ..., m_{n_2}] \in \mathbb{R}^{n_2}$. The representation $\mathbf{x}^{(2)}$ is generated by taking max pooling over entire sentence with filters of multiple window sizes. To prevent these neurons that generate $\mathbf{m}$ from co-adapting and force them to learn individual useful features, a dropout layer is added after the pooling layer for regularization.

We added a fully connected layer to combine information captured by these filters: $\mathbf{x}^{(3)} = g_2(W^{(3)}\mathbf{x}^{(2)} + \mathbf{b}^{(3)}) \in \mathbb{R}^{n_3}$. $W^{(3)}$ and $\mathbf{b}^{(3)}$ are parameters learnt in this layer, and g_2 is a non-linear function.

3.3 Learning Transferable, Discriminative Bilingual Representation

Given two sets of training examples X^a and X^b in two languages a and b, we aim at learning transferable, discriminative bilingual representations. We achieve this by weight sharing at a high-level layer. We further improve the discriminative power of learnt representations using ideal representation loss. The bilingual representation learning model is shown in Figure 1.

Shared and Language-specific Neurons We aim at not only learning representation shared by both languages, but also learning representation specific to each language. We partition the neurons in the $FC1$ layer into three disjoint sets: two language-specific sets ($FC_{1,a}$ and $FC_{1,b}$) and a shared bilingual set $FC_{1,s}$ to represent shared features across languages. We expect that $FC_{1,a}$ and $FC_{1,b}$ can model language-specific features, while the common set $FC_{1,s}$ can model the share features.

Discriminative Representation Learning As described in (Zeiler and Fergus, 2014; Krizhevsky et al., 2012), the neurons from high-level layers of a CNN tend to extract more abstract and class-specific features. If each neuron in a high-level layer of our CNN activates only when a specific relation is presented, this will lead to a discriminative representation for relations. Such representation, when learnt across language, would be extremely useful for transferring useful information across languages for relation extraction.

To achieve this, we partition the neurons in each high-level layer into subsets, and encourage each subset to only represent sentences of one of the relation types. This results in an explicit correspondence between blocks of neurons and relation types. Specifically, we partition the neurons in each sub-layer (*i.e.*, LC_a, LC_b and LC_s) into disjoint subsets and associate each subset with one specific relation label. For sentences of different relation types, we represent them using disjoint subsets of neurons. To do so, we integrate the ideal representation loss introduced in (Jiang et al., 2011) into our objective function during training. Let $(\mathbf{x}_i; y_i)$ denote training example $\mathbf{x}_i$. The ideal representation loss can be defined as:

$$L_r = \|\mathbf{q}_i - \mathbf{c}_i\| \tag{1}$$

where $\mathbf{c}_i = W^{(l_d)}\mathbf{x}_i^{(l_d-1)} + \mathbf{b}^{(l_d)}$ is the *predicted representation* for example $\mathbf{x}_i$ from locally connected layer $l_d = \{LC_a, LC_b, LC_s\}$. $\mathbf{x}_i^{(l_d-1)}$ is the representation of example $\mathbf{x}_i$ from $FC_{1,a}$, $FC_{1,b}$ or $FC_{1,s}$. $W^{(l_d)}$ and $\mathbf{b}^{(l_d)}$ are linear transform parameters and bias parameters learnt from these locally connected layers. We define $\mathbf{q}_i$ as the ideal representation corresponding to training sample $\mathbf{x}_i$ from locally connect layer l_d. The non-zero values of $\mathbf{q}_i$ occur at the indices where the training example $\mathbf{x}_i$ and neurons from layer l_d shared the same relation label. For example, suppose we have six training samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_6\}$ and their relation labels $\mathbf{y} = [y_1, \ldots, y_6] = [1, 1, 2, 2, 3, 3]$. Further assume layer l_d has six neurons $\{p_1, \ldots, p_6\}$ with $\{p_1, p_2\}$ associated with label 1, $\{p_3, p_4\}$ label 2, and $\{p_6, p_7\}$ label 3. Then the ideal representations for these six samples are given by:

$$[\mathbf{q}_1, \ldots, \mathbf{q}_6] = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \tag{2}$$

where each column is an ideal discriminative representation corresponding to a training sample. Minimizing the ideal representation loss term ensure that the input sentence from the same type

have similar representations while those from different types have dissimilar representations.

Relation Classification We trained a pair of softmax classifiers for relation classification for language a and b. The softmax classification loss $L_s^*, * = \{a, b\}$ can be defined as:

$$L_s^* = -\log(\frac{e^{o_{y_i}}}{\sum_j e^{o_j}}) \quad (3)$$

where o_j is the j-th element of the predicted relation scoring vector from LC_1 or LC_2 layer in Figure 1. Let l_s denote these two locally connected layers, *i.e.*, $l_s = \{LC_1, LC_2\}$. We have $\mathbf{o} = W^{(l_s)}\mathbf{x}_i^{(l_s-1)} + \mathbf{b}^{(l_s)}$, where $W^{(l_s)}$ and $\mathbf{b}^{(l_s)}$ are learnt parameters from layer l_s. $\mathbf{x}_i^{(l_s-1)}$ is the representation of example $\mathbf{x}_i$ by concatenating the activations from $FC_{1,a}$ and $FC_{1,s}$, or $FC_{1,b}$ and $FC_{1,s}$.

We combine the ideal representation loss and softmax classification loss to obtain the final loss function for a training sample $(\mathbf{x}_i; y_i)$:

$$L(\mathbf{x}_i, y_i) = \begin{cases} L_s^a + \lambda(L_r^a + L_r^s) & \text{if } \mathbf{x}_i \text{ in lang. } a \\ L_s^b + \lambda(L_r^b + L_r^s) & \text{if } \mathbf{x}_i \text{ in lang. } b \end{cases} \quad (4)$$

where L_s^a and L_s^b are the softmax loss for samples in language a and b, respectively. They are computed via equation 3. The terms L_r^a, L_r^b are the ideal representation losses for language a and b respectively, while L_r^s is shared by both languages. The term $L_r^*, * = \{a, b, s\}$ can be computed by equation 1. We minimized equation 4 using stochastic gradient descent.

4 Experiments

Parameter setting In the embedding layer, we used the pretrained 100-dimension bilingual word embeddings in (Ammar et al., 2016) to initialize W^1. We set the dimension of the other three embedding matrices W^2, W^3 and W^4 to 50 and initialize them randomly. In the convolution layer, we set the filter widths to $[2, 3, 4, 5]$, and use 150 filters per width. The number of neurons in the $FC1$ layer ($FC_{1,a} \cup FC_{1,s} \cup FC_{1,b}$) is 300. We use $tanh$ for g_1, g_2. All parameters are tuned with the ACE development set (described in next subsection). In bilingual experiments, we assigned 56 neurons [2] for each of $FC_{1,a}$ and $FC_{1,b}$ to learn language-specific features and the remaining neurons for language-shared features. λ is fixed to be 0.5. Training is done via stochastic gradient descent with Adam (Kingma and Ba, 2014) optimizer with a learning rate of 0.001.

Benchmarking baseline monolingual models As we are not aware of prior work on bilingual relation extraction in similar settings, we first benchmark our baseline monolingual model on two popular monolingual datasets. We also use the ACE development dataset (described below) for tuning the parameters mentioned previously. The baseline monolingual model is similar to (Nguyen and Grishman, 2015) and only takes a English dataset as input. It is simplified from the model in Figure 1 by replacing the FC and LC layers with a fully-connected layer followed by a softmax loss. We evaluate it on two English datasets: the SemEval-2010 Task 8 dataset (Hendrickx et al., 2010) and the ACE 2005 dataset. The SemEval dataset contains 10,717 annotated examples (8,000 for training and 2,717 for testing). For ACE, to be comparable to state-of-the-art Neural Network models, we use the split in (Gormley et al., 2015; Nguyen and Grishman, 2015): find the ACE articles from news domains: broadcast conversation (*bc*), broadcast news (*bn*), newswire (*nw*), and uses news (*bn* & *nw*) as the training set, half of *bc* as the development set, the other half of *bc* as the test set. Table 1 and Table 2 [3] show that the performance of our monolingual baseline and various other systems. The monolingual model [4] achieves higher performance than state-the-art CNN methods with similar structure and no additional semantic features.

For the rest of the experiments, we focus on bilingual experiments. We evaluated our model on the ACE 2005 multilingual training corpus [5] which contains 596 English and 633 Chinese documents [6] . The ACE corpus consist of articles from weblogs, broadcast news, newsgroups, broadcast conversation, and is annotated exhaustively with

[2] We choose 56 to leave sufficient number of neurons for sharing across languages. We try 56, 63, 70 and no significant difference was observed.

[3] The results reported in (Nguyen and Grishman, 2016) used a rich feature set (e.g., dependency parses). For fair comparison, we reported their results by running their code without those features.

[4] For fair comparison, we use the pre-trained word2vec word embeddings (Mikolov et al., 2013) with 300 dimensions for the monolingual experiments. For all other experiments, we use cross-lingual embeddings (Ammar et al., 2016).

[5] https://catalog.ldc.upenn.edu/LDC2006T06

[6] The English and Chinese documents are not translation of each other.

Classifier	F1
SVM (Rink and Harabagiu, 2010)	77.6 (82.2)
CNN (Zeng et al., 2014a)	78.9 (82.7)
CNN (Nguyen and Grishman, 2015)	82.8
RNN (Socher et al., 2012)	74.8 (77.6)
MVRNN (Socher et al., 2012)	79.1 (82.4)
FCM (Yu et al., 2014)	80.6 (83.0)
Our English baseline	**83.23**

Table 1: Performance on SemEval-2010 Task 8 dataset (Hendrickx et al., 2010). The numbers inside parentheses are the systems using features such as WordNet (all systems), dependency parses (Yu et al., 2014) and Google n-grams (Rink and Harabagiu, 2010). Our approach does not use these features but still achieves the best result.

Classifier	P	R	F1
CNN (Nguyen and Grishman, 2016)	63.3	58.2	60.6
Our English baseline	62.1	60.1	**61.1**

Table 2: Performance on English ACE 2005 with the data split setting in (Gormley et al., 2015).

relations. For documents in English or Chinese, we collected all annotated relation mentions, and generated relation mentions for the *Other* class by sampling pairs of entity mentions within a sentence but is not annotated as having a relation. We refer to the resulted English dataset as *ACE05.ENG*, and refer to the Chinese ACE dataset as *ACE05.CHN* in the rest of this section. We divided relation mentions into 5 folds, performed cross-validation and averaged the results. As is standard (e.g. (Grishman et al., 2005)), we use the mention boundaries and types provided by the annotated data as input.

4.1 Comparing Bilingual Models to Alternative Approaches

To verify the effectiveness of our bilingual model, we compare four approaches:

- **Baseline**: We train two monolingual models on *ACE05.ENG* and *ACE05.CHN* respectively. These two models used bilingual word embedding and only used the softmax classification loss during training.
- **Bilingual-FT**: To see if having access to additional training data from another language helps, we pre-train each monolingual model with the other language(source)'s training dataset and then finetune it using the target language's training dataset [7]
- **Bilingual-Joint** We train a bilingual model on the Chinese and English datasets jointly with the softmax loss. The model is similar to the bilingual model in Figure 1, but without the ideal representation losses [8].
- **Bilingual-Joint-IRL** We train the bilingual model on *ACE05.ENG* and *ACE05.CHN* jointly using the softmax and ideal representation loss (the complete model in Figure 1).

The results are summarized in Table 3. The pre-training approach (bilingual-FT), which pre-trains a model on the additional language's dataset and then finetune on the target language dataset, achieves better results than training the monolingual baseline. The bilingual-Joint approach can simultaneously learn language-specific and shared bilingual representations, therefore it is able to generalize across languages while still making use of language-specific information. Its performance exceeds that of baseline and bilingual-FT. The bilingual-Joint-IRL method achieves the best result: it encourages the learnt representation to be discriminative among relation types. In particular, the learnt cross-lingual representation is encouraged to differentiate relation types regardless of language. It captures both language-specific and cross-lingual relation semantics, and thus has the best of both worlds.

4.2 Transfer representations to lower-resource languages

For just a handful of languages, (e.g. English, Chinese) corpora of relation annotation are readily available. However, for most languages few (if any) such resources exist. We show that our approach yields large gains in F1 when a small amount of *target* training is combined with large amounts of existing *source* annotation.

We demonstrate our model's ability to use large amounts of *source* language data to supplement limited in-domain *target* language data with either Chinese and English as *target*. In each setting, we down-sample the *target* language dataset to 10%, 20%, 30%, 40% and 50% of its full training size and we report F1 in the test dataset in the *target* language. Figure 2 compares performance of *bilingual-Joint-IRL* with the performance of *baseline*. With access to existing, additional resources in the *source* through the bilingual model, performance on *target* doubles its F1 scores when only

[7] For example, we pre-train the Chinese model with ACE05.ENG, fine-tune with ACE05.CHN, then test with Chinese.

[8] This performs bilingual representation learning by only sharing a subset of neurons in $FC1$

Approach	ACE05.ENG			ACE05.CHN		
	Precision	Recall	F1	Precision	Recall	F1
baseline	72.5	68.3	70.4	75.6	75	75.3
bilingual-FT	70.8	72.1	71.4	78.4	75.6	77.0
bilingual-Joint	72.6	73.3	72.9	77.7	76.4	77.1
bilingual-Joint-IRL	74.3	75.6	74.9	80.9	77.1	78.9

Table 3: Performances of the bilingual models on the ACE 2005 multilingual training corpus.

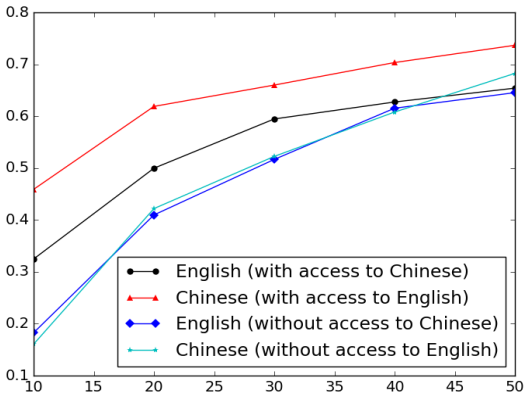


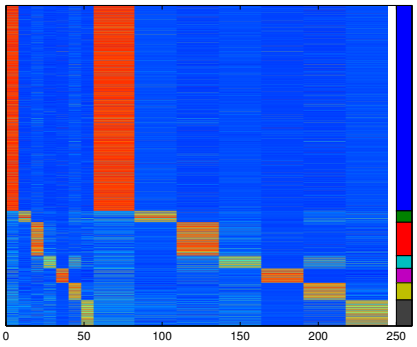
Figure 2: Performance of Chinese and English relation extraction by varying the size of in-domain training data. For the model with access to another language (*source*), we use the full dataset from the *source* but only a fraction of in-domain data in the *target* language. We sample each dataset from 10% to 50% with 10% as the incremental step size. x-axis is the percentage of training data in *target* is used, and y-axis shows F1 scores.

10% training data is available, and gains 30% to 66% relative improvement in F1 with 20% training data. The bilingual extractor trained with only 50% of the *target* training data achieves performance nearing ($< 5\%$ difference) that of the baseline approach using all of the *target* resources. This results provides a new and cost-effective way to perform relation extraction when resources are limited for a new language.

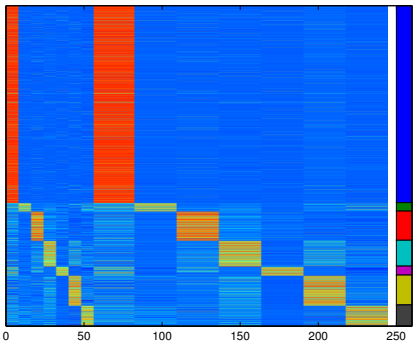
4.3 Analysis on learnt representation

Learning discriminative representations We use LC_a and LC_s to represent a sentence from language a, and use LC_b and LC_s to represent a sentence from language b. We visualize the predicted representations for testing examples in both English and Chinese in Figure 3. X-axis shows the 244 neurons [9] from locally connect layers LC_a-LC_s, and LC_b-LC_s. Y-axis shows all testing examples (each row represents an example). Each

[9]The number of neurons in layers LC_a and LC_b are both 56, while layer LC_s contains 188 neurons.



(a) ACE05.ENG set



(b) ACE05.CHN set

Figure 3: Visualization of predicted presentation for test examples in both languages. X-axis indicates the 244 neurons from layers LC_a-LC_s (or LC_b-LC_s). Y-axis corresponds to the indices of test examples. The color shows strength of pairwise association between examples and neurons (the brighter the color, the stronger association). Each color in the color bar located at the right most of each subfigure represents one relation type for a subset of testing examples.

graph shows strong associations between testing examples from 7 types and 7 blocks of neurons in the hidden layer in 1) language-specific representations (the first 56 neurons as shows in x-axis [0, 55] in Figure 3 (a) and (b)) as well as 2) bilingual representation (the remaining 188 neurons as shown in x-axis [56,243] in each figure). The visualization is strongly block-diagonal. This shows that the learnt representation is discriminative in

Relation	English	Chinese
ORG-AFF	... while japanese officials ...	...应新加坡 总理吴作栋 的邀请...
PHYS	...which is ... outside the center of baghdad...	...前日在广州 被公安 寻回...
PART-WHOLE	...this is that city hall in orlando...	...亚齐省首府班达亚齐
PER-SOC	...you go to your grandma 's house	该名家长 起初以为孩子 撒谎...
GEN-AFF	joseph britt of kennesaw , ga , recently...	日本 历史上有过女天皇
ART	i 've decided i 'll take the train home later	我们 的军舰 在广里...

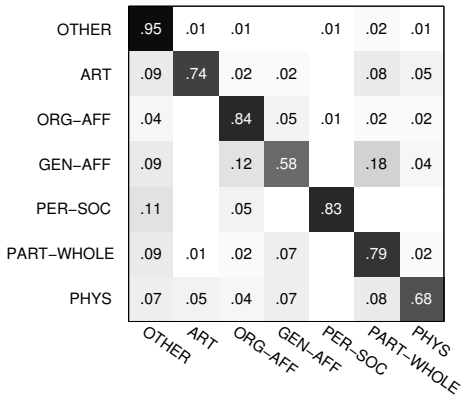
Table 4: Examples of relation mentions that are found to be similar to its cross-lingual counterparts using our model. The two examples in each row is similar to each other.

predicting relation types.

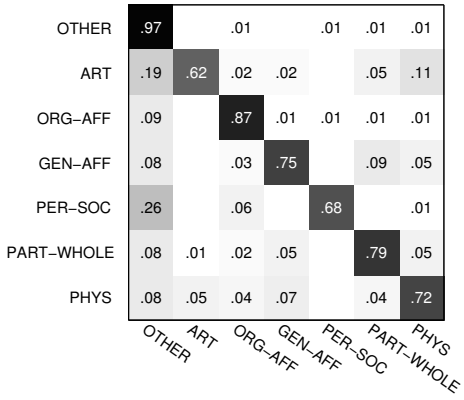
Confusion matrix on bilingual relation extraction To further understand the discriminative power of the bilingual model with regard to predicting relation types, we plot the confusion matrix using the bilingual model over all examples in the ACE datasets. Figure 4 shows that our model performs well on differentiating the classes. It also shows that our algorithm performs slightly worse in differentiating *PHYS*, *PART-WHOLE* and *GEN-AFF* relations, in both English and Chinese. This is caused by the trumping rules in ACE relation definition. ACE annotation guideline defines a relation *Org-Location-Origin* (a subtype of *GEN-AFF*) for locations of ORG's, and defines another relation *Geographical* (a subtype of *PART-WHOLE*) for locations of facilities/locations/GPEs, and further defines a relation *Located* (as a subtype of *PHY*) to capture the physical location of a person. These relations share the same high-level semantics but are defined as different ACE types.

Examples of similar relation instances across languages We took the last-layer activations as a "relation embeddings" representation for each relation instance. To see how well the cross-lingual representation learning did, we calculated pairwise similarities [10] between all relation instances in Chinese and English. Table 4 shows examples in which each pair of instances in the same row in the two languages are very similar [11] to each other.

The model learns lexical similarity such as *japanese officials* and 新加坡 总理 *(Singapore prime minister)*. It also knows *city hall in orlando* is similar to 亚齐省首府班达亚齐 *(Aceh capital of Banda Aceh)*. It further learns complicated correspondence such as *you go to your grandma 's house* and 该名家长 起初以为孩子 撒谎*(The parent previously thought the child was lying*) The diverse range of examples in Table 4 shows that the model not only captures lexical similarity, but also syntactic and long-range semantic similarities across the pair of languages.



(a) ACE05.ENG set



(b) ACE05.CHN set

Figure 4: Confusion matrices of *bilingual-Joint-IRL* on the ACE datasets.

5 Conclusion

We present a bilingual relation extraction algorithm to learn discriminative and transferable representation across languages via a Convolutional Neural Network. Experiments show that it outperforms monolingual algorithms and the baseline algorithms significantly both when large amounts

[10] We use cosine distance as the similarity metric. The smaller cosine distance is, the more similar the pair.

[11] Here we define *very similar* as ranked within top-50 most similar instances in the entire dataset, with regard to the query instance from the other language.

of data are available in both languages and when only limited training data is available in the target language.

Acknowledgments

This work was supported by DARPA/I2O Contract No. FA8750-13-C-0008 under the DEFT program. The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Department of Defense or the U.S. Government. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Andre Blessing and Hinrich Schütze. 2012. Crosslingual distant supervision for extracting relations of different complexity. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1123–1132, New York, NY, USA. ACM.

Razvan Bunescu and Raymond Mooney. 2005a. A Shortest Path Dependency Kernel for Relation Extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 724–731, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Razvan Bunescu and Raymond J Mooney. 2005b. Subsequence kernels for relation extraction. In *NIPS*, pages 171–178.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual Open Relation Extraction Using Cross-lingual Projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1356, Denver, Colorado. Association for Computational Linguistics.

Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1774–1784, Lisbon, Portugal. Association for Computational Linguistics.

Ralph Grishman, David Westbrook, and Adam Meyers. 2005. Nyus english ace 2005 system description. In *ACE 2005 Evaluation Workshop*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A Distributed Representation-based Framework for Cross-lingual Transfer Parsing. *J. Artif. Int. Res.*, 55(1):995–1023.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Seaghdha, Sebastian Pado, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA. Association for Computational Linguistics.

Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. 2013. Cross-language Knowledge Transfer using Multilingual Deep Neural Network with Shared Hidden Layers. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Jing Jiang and ChengXiang Zhai. 2007. A systematic exploration of the feature space for relation extraction. In *Proceedings of NAACL HLT 2007*. Association for Computational Linguistics.

Zhuolin Jiang, Zhe Lin, and Larry Davis. 2011. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *Proceedings of CVPR*.

Nanda Kambhatla. 2004. Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Information Extraction. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 178–181, Barcelona, Spain. Association for Computational Linguistics.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2010. A Cross-lingual Annotation Projection Approach for Relation Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 564–571, Beijing, China. Coling 2010 Organizing Committee.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2014. Cross-lingual annotation projection for weakly-supervised relation extraction. *ACM Transactions on Asian Language Information Processing (TALIP)*, 13(1):3.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Makoto Miwa and Mohit Bansal. 2016. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.

Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *ECML*, volume 4212, pages 318–329. Springer.

Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. Patty: A taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1135–1145, Jeju Island, Korea. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation Extraction: Perspective from Convolutional Neural Networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2016. Combining Neural Networks and Log-linear Models to Improve Relation Extraction. In *Proceedings of IJCAI Workshop on Deep Learning for Artificial Intelligence*, New York, USA.

Longhua Qian, Haotian Hui, Ya'nan Hu, Guodong Zhou, and Qiaoming Zhu. 2014. Bilingual Active Learning for Relation Classification via Pseudo Parallel Corpora. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592, Baltimore, Maryland. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling Relations and Their Mentions without Labeled Text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Barcelona, Spain. Springer, Berlin, Heidelberg.

Bryan Rink and Sanda Harabagiu. 2010. Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. 2013. Modeling missing data in distant supervision. *Transactions of the Association for Computational Linguistics*.

Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying Relations by Ranking with Convolutional Neural Networks. 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 626–634, Beijing, China. Association for Computational Linguistics.

Cicero Nogueira Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural network. In *Proceedings of ACL*.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1201–1211, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance Multi-label Learning for Relation Extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea. Association for Computational Linguistics.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 886–896, San Diego, California. Association for Computational Linguistics.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation Classification via Multi-Level Attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307, Berlin, Germany. Association for Computational Linguistics.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language*

Processing, pages 1785–1794, Lisbon, Portugal. Association for Computational Linguistics.

Mo Yu, Matthew R. Gormley, and Mark Dredze. 2014. Factor-based compositional embedding models. In *Proceedings of the NIPS Learning Semantics Workshop*.

Matthew Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of ECCV*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014a. Relation Classification via Convolutional Deep Neural Network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014b. Relation classification via convolutional deep neural network. In *Proceedings of COLING*.

Min Zhang, Jie Zhang, and Jian Su. 2006. Exploring Syntactic Features for Relation Extraction using a Convolution Tree Kernel. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 288–295, New York City, USA. Association for Computational Linguistics.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional Long Short-Term Memory Networks for Relation Classification. In *The 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78.

Shubin Zhao and Ralph Grishman. 2005. Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 419–426, Stroudsburg, PA, USA. Association for Computational Linguistics.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring Various Knowledge in Relation Extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434, Ann Arbor, Michigan. Association for Computational Linguistics.