Adversarial Feature Adaptation for Cross-lingual Relation Classification

Bowei Zou, Zengzhuang Xu, Yu Hong, Guodong Zhou*
School of Computer Science and Technology, Soochow University,
Suzhou, 215006, China
zoubowei@suda.edu.cn, nedxuwork@gmail.com,
{yhong, gdzou}@suda.edu.cn

Abstract

Relation Classification aims to classify the semantic relationship between two marked entities in a given sentence. It plays a vital role in a variety of natural language processing applications. Most existing methods focus on exploiting mono-lingual data, e.g., in English, due to the lack of annotated data in other languages. In this paper, we come up with a feature adaptation approach for cross-lingual relation classification, which employs a generative adversarial network (GAN) to transfer feature representations from one language with rich annotated data to another language with scarce annotated data. Such a feature adaptation approach enables feature imitation via the competition between a relation classification network and a rival discriminator. Experimental results on the ACE 2005 multilingual training corpus, treating English as the source language and Chinese the target, demonstrate the effectiveness of our proposed approach, yielding an improvement of 5.7% over the state-of-the-art.

1 Introduction

Relation classification aims to identify the semantic relationship between two nominals labeled in a given sentence. It is critical to many natural language processing (NLP) applications, such as question answering and knowledge base population. For example, the following sentence contains an instance of the *Content-Container*(e_2,e_1) relation between two labeled entity mentions "e_1=*cartridge*" and "e_2=*ink*".

The [***cartridge***]$_{e_1}$ *was marked as empty, even with* [***ink***]$_{e_2}$ *in both chambers.*

An open challenge is how to train a model which is suitable for languages with insufficient available data of relation classification, since manual annotation is time-consuming and human-intensive. This makes it difficult to transferrably use the existing well-trained classification models in other languages.

To tackle this problem, we propose an adversarial feature adaptation approach to transfer latent feature representations from the source language with rich labeled data to the target language with only unlabeled data. Such an approach are both discriminative for relation classification and invariant across languages. This is largely motivated by the adversarial mechanism which has been effectively applied to measure the similarity between distributions in a variety of scenarios, such as domain adaptation (Bousmalis et al., 2016) and multi-modal representation learning (Park and Im, 2016).

In particular, we build two counterpart networks, using convolutional neural networks (CNNs), for source language and target language, to generate the latent feature representations respectively. Then, we use a rival discriminator to identify the correct source of feature representations. At the training step, the network of the source language is trained to maximize the performance on the annotated dataset, while the network of the target language is trained to imitate the feature representations of the source language by rival confusing the discriminator.

We perform the proposed approach on ACE 2005 multilingual training corpus by regarding English as the richly-labeled language (source) and Chinese as the poorly-labeled language (target). Our approach

*corresponding author

Proceedings of the 27th International Conference on Computational Linguistics, pages 437–448
Santa Fe, New Mexico, USA, August 20-26, 2018.

achieves an F1-score of 70.50% with a significant improvement of 5.7%, compared to the state-of-the-art method.

The main contributions of this study are as follows.

- We present a novel neural feature adaptation framework by leveraging a generative adversarial network to transfer feature representations from a richly-labeled language to a poorly-labeled language. To the best of our knowledge, this is the first study on feature adaptation for cross-lingual relation classification.
- Experimental results show that the latent feature representations can be effectively transferred from the source language to the target language. This enables the adaptation of the existing manually annotated resource in one language to a new language.
- In a slightly better scenario that there are a small-scale annotated data available in the target language, our adversarial feature adaptation approach can also be effectively cooperated with the supervised model to further improve the overall performance.

The rest of this paper is organized as follows. In Section 2, we overview the related work. In Section 3, we introduce details of the proposed adversarial feature adaptation approach. We show our experimental results and discussions in Section 4. Finally, we conclude the paper in Section 5.

2 Related Work

In this section, we briefly review the recent progress in cross-lingual relation classification and existing studies on adversarial adaptation.

2.1 Cross-lingual Relation Classification

Labeled data on relation classification are not evenly distributed among languages. While there are various of annotated datasets in English, such as the SemEval'10 Task-8 dataset (Hendrickx et al., 2010) and the SemEval'18 task-7 dataset (Gábor et al., 2018), annotated datasets in other languages are few.

Traditional studies for relation classification usually perform supervised machine learning models trained on mono-lingual labeled datasets, which either rely on a set of linguistic or semantic features (Kambhatla, 2004; Suchanek et al., 2006), or apply tree kernel-based features to represent the input sentences (Bunescu and Mooney, 2005; Qian et al., 2008). Recently, deep neural networks (Zeng et al., 2015; dos Santos et al., 2015) and attention mechanism (Wang et al., 2016) show the effectiveness in relation classification. However, the training of neural network relies on large-scale labeled instances. This makes it difficult to re-construct such classification models for the poorly-labeled language.

Most of existing studies have attempted to leverage parallel data or a knowledge-based system to transfer effective information from the richly-labeled language to the poorly-labeled language. Qian et al. (2014) proposed a bilingual active learning paradigm for Chinese and English relation classification with pseudo parallel corpora and entity alignment. Kim et al., (2014) proposed a cross-lingual annotation projection strategy by employing parallel corpora for relation detection. Faruqui and Kumar (2015) also present a cross-lingual annotation projection method by using machine translation results, rather than parallel data. Verga et al. (2016) performs multi-lingual relation classification by a knowledge base. Min et al. (2017) drive a classifier to learn discriminative representations by joint supervision of classification (softmax) loss and ideal representation loss.

Instead of exploiting external resources and manually selecting a closeness metric, we come up with an adversarial mechanism to provide an adaptive metric for feature adaptation from the richly-labeled language to the poorly-labeled language.

2.2 Adversarial Adaptation

Recently, the generative adversarial networks (GAN) have become increasingly popular, especially in the area of deep generative unsupervised modeling (Goodfellow et al., 2014; Makhzani et al., 2016). For adversarial adaptation, Ganin et al. (2017) proposed the domain adversarial neural networks (DANN) to

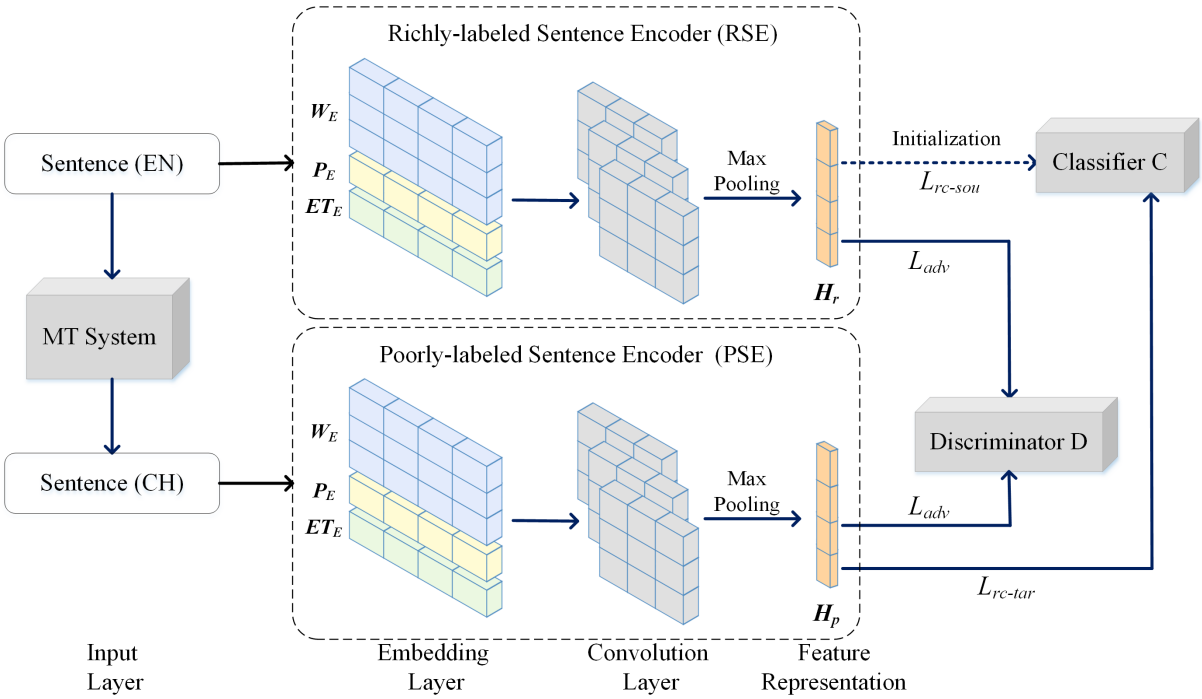

Figure 1: Architecture of the adversarial feature adaptation framework for cross-lingual relation classification.

learn discriminative but domain-invariant representations, transferring the information from the source domain to the target domain. Different from their study, our approach aims to find sharable language-independent latent feature representations for cross-lingual relation classification.

There have been some previous work applying adversarial adaptation technique to NLP tasks, such as that for sentiment analysis (Chen et al., 2016) and parsing (Sato et al., 2017). These studies learn the domain-invariant or domain-specific features by a shared network. Our work differs from them, since we force a second network with identical structure to learn the latent feature representations from the supervised network. Qin et al. (2017) propose a feature imitation approach. An adversarial mechanism is used between explicit and implicit discourse relation samples. Different from their study, we migrates the feature representations from one language to another (non-parallel). To the best of our knowledge, this is the first work to employ the adversarial feature adaptation for cross-lingual relation classification.

3 Adversarial Feature Adaptation for Cross-lingual Relation Classification

The common semantic information between different language motivates our adversarial feature adaptation approach. In this section, let us take a glance at the framework first, and then the relation classification networks (sentence encoders) and the adversarial training procedure.

Figure 1 illustrates the schematic overview of the framework which consists of four key components: 1) a Richly-labeled Sentence Encoder (RSE) with English sentences as the inputs, 2) a Poorly-labeled Sentence Encoder (PSE) which takes translated Chinese sentences as the input, 3) a language discriminator D distinguishing between the feature representations from the above two encoders, and 4) a relation classifier C to predict the relation label. In general, a GAN consists of a generative network G and a discriminator D, in which G generates instances by a distribution $P_{G(x)}$, and D aims to determining whether a instance is from $P_{G(x)}$ or the real data distribution $P_{data(x)}$. In our approach, the PSE is taken as the generative network which generates the feature representations $\boldsymbol{H_p}$ to confuse the discriminator.

In training step, the relation classifier aims to predict the labels, while the language discriminator attempts to distinguish between the feature representations extracted by the two sentence encoders ($\boldsymbol{H_p}$ or $\boldsymbol{H_r}$). In test step, we utilize the PSE to encode the input sentences in target language, and apply the same classifier to predict to relation labels.

3.1 Components

As a common neural network model that yields good performance for monolingual relation classification, we employ CNN to transform a sentence with pairs of entity mentions into a distributed representation $\boldsymbol{H}$. Note that, this plug-in architecture can also be implemented with the other networks, e.g., a long short-term memory network (LSTM)[1].

Embedding Layer

Following Zeng et al. (2014)'s work, we build an embedding layer to encode words, word positions, and entity types by real-valued vectors.

Given an input sentence $S = (w_1, w_2, ..., w_n)$, we first transform every word into a real-valued vector of dimension d_w using a word embedding matrix $\boldsymbol{W_E} \in \mathbb{R}^{d_w \times |V|}$, where V is the input vocabulary. Since the structures of RSE and PSE are the same, it is necessary if the word representations for both languages have a shared vocabulary. Therefore, bilingual word embeddings (Shi et al., 2015) are employed to map words from different languages into the same feature space.

To capture the informative features of the relationship between words and the entity mentions, we map the relative distances to entity mentions of each word to two real-valued vectors of dimension d_p using a position embedding matrix $\boldsymbol{P_E} \in \mathbb{R}^{d_p \times |D|}$, where D is the set of relative distances which are mapped to a vector initialized randomly (dos Santos et al., 2015). For each word, we obtain two position vectors with respect to the two entity mentions.

For each word, we also incorporate its entity type embedding to reflect the relationship between the entity type and the relation type. Each word is mapped to a real-valued vector using embedding matrix $\boldsymbol{ET_E} \in \mathbb{R}^{d_{et} \times |E|}$, where E is the set of entity types.

Finally, we represent a input sentence as a vector sequence $\mathbf{w}=\{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n\}$ with the embedding dimension $d = (d_w + 2d_p + d_{et})$.

Convolution Layer

After encoding the input sentence, a convolution layer extracts local features by sliding a window of length w over the sentence and perform a convolution within each sliding window. The output for the ith sliding window is

$$\mathbf{p}_i = \mathbf{W_c}\mathbf{w}_{i-w+1:i} + \mathbf{b}, \tag{1}$$

where $\mathbf{w}_{i-w+1:i}$ denotes the concatenation of w word embeddings within the ith window, $\mathbf{W_c} \in \mathbb{R}^{d_c \times (w \times d)}$ is the convolution matrix and $\mathbf{b} \in \mathbb{R}^{d_c}$ is the bias vector (d_c is the dimension of output of the convolution layer).

Max-pooling Layer

We merge all local features via a max-pooling layer and apply a hyperbolic tangent function to obtain a fixed-sized final representations. The ith element of the output vector $\mathbf{x} \in \mathbb{R}^{d^c}$ is

$$[\mathbf{x}]_j = \tanh \max_i \mathbf{p}_{ij}. \tag{2}$$

Classifier and Discriminator

While the Classifier C is a fully-connected layer followed by a softmax classifier, the Discriminator D is a binary classifier which is implemented as a fully-connected neural network with a sigmoid activation function. Discriminator D takes the feature representations as input, to discriminate whether the feature representation comes from RSE or from PSE.

3.2 Adversarial Training

Although the aforementioned architecture could be applied to cross-lingual relation classification by leveraging the RSE module to train on source language and the MT module to translate the instances from

[1]As an alternative, a bidirectional LSTM (BiLSTM) is tried as the basic network. The experimental results are shown in Subsection 4.2.

Algorithm 1 Adversarial Training Procedure

Input: Training dataset

Output: RSE with classifier C

1: Initialize $\boldsymbol{\theta}_h$ and $\boldsymbol{\theta}_C$ by minimizing Eq.(3).
2: **repeat**
3: Train the discriminator D through Eq.(4)
4: Train PSE and classifier C through Eq.(7)
5: **until** convergence

the target language to the source language, there is no guarantee that the latent feature representations of the source language exist in the target language, or vice versa. On the other hand, the error propagation of MT module also should be considered. Therefore, we introduce adversarial training into our cross-lingual relation classification framework.

Algorithm 1 illustrates the adversarial training procedure. First, we pre-train the RSE and the classifier C by minimizing the relation classification (RC) loss function on source language (step line 1). Then we interleave the optimization of the adversarial loss function and the RC loss function on the target language at each iteration (step line 2-5). Finally, if the discriminator cannot tell the language of a input sentence using the adversarially trained features, then those features from PSE are effectively language-invariant. Upon successful training, the feature representations ($\boldsymbol{H_p}$) are thus encouraged to be both discriminative for relation classification and invariant across languages. Referring to the expression of Qin et al. (2017), the three loss function of our adversarial training are as follows.

RC Loss for Training on Source Language

We denote the parameters of the RSE and classifier C as $\boldsymbol{\theta}_r$ and $\boldsymbol{\theta}_C$, respectively, and the objective can be learned by minimizing the cross-entropy loss as

$$\mathcal{L}_{rc\text{-}sou}(\boldsymbol{\theta}_r, \boldsymbol{\theta}_C) = \mathbb{E}_{(\boldsymbol{x_e}, y)\sim data}\left[\mathcal{J}(C(H_r(\boldsymbol{x_e}; \boldsymbol{\theta}_r); \boldsymbol{\theta}_C), y)\right], \tag{3}$$

where $\mathbb{E}_{(\boldsymbol{x_e}, y)\sim data}[\cdot]$ denotes the expectation in terms of the data distribution, $\mathcal{J}(\boldsymbol{p}, y) = -\sum_k \mathrm{II}(y = k)\log p_k$ is the cross-entropy loss between predictive distribution $\boldsymbol{p}$ and ground-truth label y, $C(\boldsymbol{H_r}(\boldsymbol{x}))$ is the final prediction of classifier C when the input is the feature representation of RSE ($\boldsymbol{H_r}(\boldsymbol{x})$), $(\boldsymbol{x_e}, y)$ is the pair of input and output of relation classification model, where $\boldsymbol{x_e}$ is an English instance, and y is the relation label.

Adversarial Loss

The adversarial loss $\mathcal{L}_{adv}$ is used to train the discriminator D to make a correct estimation that where the feature representation comes from. Formally, the parameters of the discriminator D is denoted as $\boldsymbol{\theta}_D$. The training objective of D is to distinguish the input source of feature representation as far as possible:

$$\min_{\boldsymbol{\theta}_D} \mathcal{L}_{adv} = \mathbb{E}_{(\boldsymbol{x_e}, \boldsymbol{x_c}, y)\sim data}\left[\log(1 - D(\boldsymbol{H_r}(\boldsymbol{x_e}; \boldsymbol{\theta}_D))) + \log D(\boldsymbol{H_p}(\boldsymbol{x_c}; \boldsymbol{\theta}_D))\right], \tag{4}$$

where $D(\boldsymbol{H})$ denotes the output of discriminator D to estimate the probability that $\boldsymbol{H}$ comes from the RSE rather than the PSE, $C(\boldsymbol{H_l}(\boldsymbol{x}))$ is the final prediction of classifier C when the input is the feature representation of PSE ($\boldsymbol{H_l}(\boldsymbol{x})$), and $(\boldsymbol{x_c}, y)$ is the pair of input and output of relation classification model, where $\boldsymbol{x_c}$ is a translated Chinese instance.

RC Loss for Training on Target Language

We denote the parameters of the PSE as $\boldsymbol{\theta}_p$. The training objective is to minimize the discriminator's chance of correctly telling apart the features:

$$\mathcal{L}_p(\boldsymbol{\theta}_p) = \mathbb{E}_{\boldsymbol{x_c}\sim data}\left[\log D(\boldsymbol{H_p}(\boldsymbol{x_c}; \boldsymbol{\theta}_p))\right]. \tag{5}$$

The parameters of classifier C is denoted as $\boldsymbol{\theta}_C$. The training objective of C is to correctly classify

Relation Type	Source (EN)	Target (CH)		
	train	train	dev.	test
PHYS	1,958	1,094	170	314
PART-WHOLE	1,299	1,596	228	454
ART	869	441	63	125
ORG-AFF	2,511	1,528	218	436
PER-SOC	1,087	462	66	132
GEN-AFF	954	1,354	193	386
Other	1,446	1,081	154	308
Total	10,124	7,556	1,092	2,055

Table 1: Discription of our datasets. The data in this tabel denote the number of samples of corresponding sets. The relation type "Other" means that the relation of entity mentions is not among the aforementioned six types.

relations. The objective can be learned by minimizing the cross-entropy loss:

$$\mathcal{L}_C(\boldsymbol{\theta}_C) = \mathbb{E}_{(\boldsymbol{x_c},y)\sim data}\left[\mathcal{J}(C(\boldsymbol{H_p}(\boldsymbol{x_c};\boldsymbol{\theta}_C), y)\right]. \tag{6}$$

Finally, we combine the above objectives Eq.(5) and (6) of the relation classifiers, and minimize the joint loss:

$$\min_{\boldsymbol{\theta}_p,\boldsymbol{\theta}_C} \mathcal{L}_{rc\text{-}tar} = \lambda\mathcal{L}_p(\boldsymbol{\theta}_p) + \mathcal{L}_C(\boldsymbol{\theta}_C) \tag{7}$$

where λ is a balancing parameters calibrating the weights of the classification loss and the feature-regulating loss.

4 Experimentation

In this section, we first describe our datasets, detailed settings, and evaluation metrics used in the experiments. Then we show the effectiveness of our adversarial feature adaptation framework for cross-lingual relation classification. Finally, we further investigate the semi-supervised settings, where a small amount of labeled data of the target language exists.

4.1 Experimental Settings

In this paper we regard English as the richly-labeled (source) language and Chinese as the poorly-labeled (target) language. Note that, in fact, our model could generalize to any pair of source and target languages in principle. We conduct our experiments on the commonly used ACE 2005 multilingual training corpus (Walker et al., 2006)[2] dataset. Table 1 shows the detailed descriptions of the datasets. We utilize all of the seven types of labeled relation mentions, and evaluate all of the systems by using the micro- and macro-F1 scores over the six types of relations excluding "Other". The English and Chinese datasets are not translation of each other.

We pre-train bilingual word embeddings by CLSim[3] (Shi et al., 2015), which provides 50-dimensional embeddings for 800k parallel sentence pairs. We set the dimensions of the position embeddings and the entity type embeddings to 20 and 30, respectively, with random initialization following a continuous uniform distribution. To obtain the translated sentence pairs automatically and easily, we employ the commercial Google Translate engine[4], which is a highly engineered machine translation system. The mention boundaries and the entity type tags are provided by the ACE 2005 multilingual training corpus.

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

[3]http://nlp.csai.tsinghua.edu.cn/∼lzy/src/acl2015_bilingual .html

[4]https://translate.google.com

Hyper-parameter	value
Window size w	3
Convolutional filters f	100
Batch size B	100
Maximum iteration I	100
Dropout probability p	0.5

Table 2: Parameter settings.

Model (EN⇒CH)	Micro			Macro		
	P	R	F1	P	R	F1
BI-AL	62.73	64.69	63.69	64.28	65.33	64.80
CNN-MT-Source	**67.11**	66.27	66.68	**70.69**	66.82	68.70
CNN-MT-Target	65.34	66.87	66.10	70.46	66.44	68.39
BiLSTM-MT-Source	65.60	64.54	65.07	68.25	65.97	67.09
BiLSTM-MT-Target	65.52	64.92	65.22	68.96	66.88	67.90
CNN-GAN	66.79	**70.11**	**68.41**	68.73	**72.35**	**70.50**
BiLSTM-GAN	66.22	67.41	66.81	68.02	68.07	68.05

Table 3: Performance of systems on the Chinese test sets of ACE 2005 multilingual training corpus for cross-lingual relation classification. CNN-GAN: our adversarail feature adaptation approach with a CNN sentence encoder; BiLSTM-GAN: similar as CNN-GAN, with a BiLSTM sentence encoder.

All the models are optimized using ADADELTA (Zeiler, 2012). We pick the parameters showing the best performance on the development set (in Column 4, Table 1) via early stopping, and report the scores on the test set (in Column 5, Table 1). Table 2 shows the best settings of model parameters in our experiments.

We compare with the following baselines for cross-lingual relation classification.

- **CNN-MT-Source** All the instances in the English training set (the "Source" Column in Table 1) are translated into Chinese (target language) by Google Translator. A CNN model with the same structure of PSE is trained by leveraging this new translated training data on Chinese.
- **CNN-MT-Target** Contrary to the **CNN-MT-Source**, all the instances in the Chinese test set (the "Target-test" Column in Table 1) are directly translated into English (source language) by Google Translator. A CNN model with the same structure of RSE is trained by leveraging the English training set.
- **BiLSTM-MT-Source** and **BiLSTM-MT-Target** They are similar to the settings of the **CNN-MT-Source** and the **CNN-MT-Target**, respectively. For further examining the robustness of our adversarial feature adaptation framework, we replace the sentence encoder networks (CNNs) with the BiLSTM networks for both RSE and PSE. The BiLSTM model is proposed by Zhang et al. (2015), which is one of the state-of-the-art mono-lingual relation classification system. For fair comparison, we only retain the word embeddings, the position embeddings, and the entity type embeddings.
- **BI-AL** This model is proposed by Qian et al. (2014), which is a bilingual active learning system for Chinese and English relation classification with pseudo parallel corpora and entity alignment.

4.2 Experimental Results

Adversarial Feature Adaptation

Table 3 shows a performance comparison of our adversarial feature adaptation models (*-GAN) with baselines. With the CNN-GAN system, we achieve a micro-F1 score of 68.41% and a macro-F1 score

Model (CH⇒EN)	Micro			Macro		
	P	R	F1	P	R	F1
CNN-MT-Source	64.86	73.53	68.92	64.18	72.13	67.93
CNN-MT-Target	68.08	72.60	70.27	66.60	71.11	68.78
BiLSTM-MT-Source	68.30	73.24	70.68	67.06	70.79	68.88
BiLSTM-MT-Target	66.85	71.56	69.12	66.30	70.74	68.45
CNN-GAN	71.65	**77.53**	**74.47**	69.51	**73.74**	**71.56**
BiLSTM-GAN	**72.20**	75.66	73.89	**69.69**	72.03	70.84

Table 4: Performance of systems on the opposite direction for cross-lingual relation classification, in which Chinese is treated as the source language and English is treated as the target language (contrary to Table 3). The partition of English dataset is the same as that in Table 1 (i.e. the training set 70%, the development set 10%, and the test set 20%).

of 70.50%, which outperform the active learning based system (BI-AL) with a relative improvement of about 5%. It indicates that our adversarial feature adaptation framework substantially outperforms the state-of-the-art model for cross-lingual relation classification without any annotated data on the target language.

Moreover, we pay more attention to evaluate the proposed adversarial feature adaptation model against bidirectional MT-based baselines, including 1) translate the Chinses text into English, then leveraging the English relation classification model (trained on English dataset) to identify the entity type (CNN-MT-Target and BiLSTM-MT-Target), and 2) translate the English training set into Chinese, then train an Chinese relation classification model to classify the entity types (CNN-MT-Source and BiLSTM-MT-Source). Our approach improves about 2% of macro-F1 score over the machine translation based systems. Besides, the performances of CNN-MT-Source and CNN-MT-Target are comparative, which indicates that the translated direction may be insignificant for cross-lingual relation classification via MT.

We can also see that all of the CNN-based adversarial feature adaptation models achieve better results than the corresponding BiLSTM-based ones within the same settings. The reason might be that the BiLSTM is fitter for encoding order information and long-range context dependency for sequence labeling problem, while the CNN is suited to extracting local and position-invariant features. For relation classification, the essential features are always distributed around the entity mentions, which would be better utilized by CNN[5].

To validate the language independence of our feature adaptation framework, we also implement our adversarial feature adaptation systems and the MT-based systems in Table 1 on the same corpus from the opposite direction. Specifically, we regard Chinese as the source language and English as the target, then learn the feature representations from the Chinese dataset to predict the relation labels of the English dataset. Table 4 indicates that our approach also outperforms the baselines on learning the feature representations from Chinese to English. Besides, the results further validate the conclusions mentioned above: 1) Our system outperforms the MT-based systems for cross-lingual relation classification, and 2) the CNN-based systems achieve better performances than the BiLSTM-based systems in the same settings.

To further provide an empirical insight into the relationship between 1) the data size of labeled training set for supervised relation classification (the blue curve in Figure 2), and 2) the data size required by our adversarial feature adaptation system with only unlabeled sentences (the orange dashed curve in Figure 2), we simulate a supervised scenario by adding labeled Chinese instances for training a CNN-based relation classification system (CNN-CH in Table 5). We start from adding 100 labeled sentences and keep adding 100 sentences each time until 900. As shown in Figure 2, when adding the same number of labeled sentences, the CNN-CH system can better utilize the extra supervision. The margin is naturally decreasing as more supervision is incorporated, until the training set contains more than 700 instances.

[5]Yin et al. (2017) also demonstrated this conclusion for relation classification task.

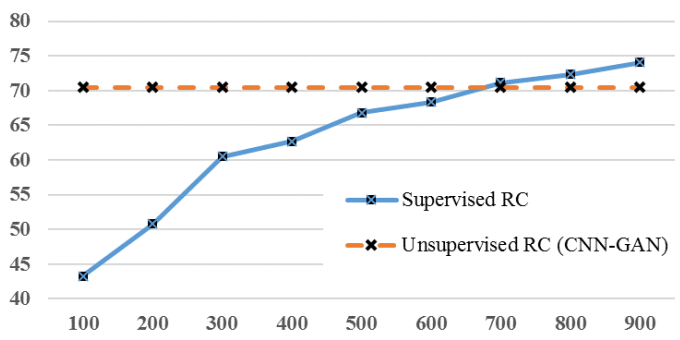

Figure 2: Comparison with a supervised relation classification system.

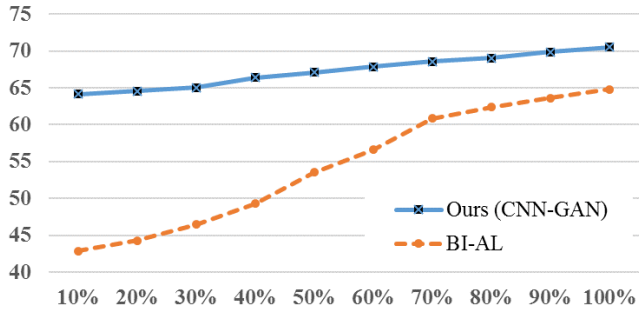

Figure 3: Comparison of the performances (macro-F1) when adding different sizes if labeled data of the source language from 10% to 100%.

It indicates that our adversarial feature adaptation system can achieve comparable performance to the supervised system trained on a small labeled dataset.

Another interesting find is that it seems a very small amount of supervision (e.g., 500 labeled instances) could significantly help the supervised relation classification system. However, it is worth noting that the manual annotation on such amount of dataset is still time-consuming and human-intensive, since the people should annotate not only the entity mentions and their relation, but also the external information such as the lexical or syntactic features if necessary.

Figure 3 compares the performances of our CNN-GAN system (the blue curve in Figure 3) and a bilingual active learning system BI-AL (Qian et al., 2014) (the orange dashed curve in Figure 3) when training with different sizes of labeled data from the source language. As we can see, the margin of our approach is not significant when the size of the source-language instances is relatively small. When using 10% of the training data, our system only declines 6.35% of performance (from 70.50% to 64.15%), while for the bilingual active learning system, the gap widens to 20.91% (from 64.80% to 42.89%). It indicates that our feature adaptation approach can efficiently utilize the translated supervision from the source language.

Semi-supervised Scenario

In this paper, we mainly focus on the the scenario of unsupervised cross-lingual relation classification, i.e., without any labeled dataset. However, for broader comparisons, we also test our framework when a few labeled training instances of the target language are available. Actually, our approach can be easily generalized to a semi-supervised setting. We employ a simple way that directly combine the softmax layers of the CNN-CH system and the CNN-GAN system, to integrate the supervision from target language into relation classification.

Table 5 lists the performances of these semi-supervised relation classification systems. We see that our ensemble model (*-CH-EN) which can employ not only the labeled data from the target language but from the source language, slightly improves over the supervised model (*-CH). It indicates that our

Model	P	R	F1
bilingual-Joint-IRL	**80.9**	77.1	78.9
CNN-CH	79.09	81.15	80.11
BiLSTM-CH	76.23	79.94	78.04
CNN-CH-EN	79.61	**81.65**	**80.62**
BiLSTM-CH-EN	77.55	79.50	78.52

Table 5: Performance of the semi-supervised relation classification systems. bilingual-Joint-IRL system: a bilingual approach by joint supervision of classification loss and ideal representation loss (Min et al., 2017); *-CH systems: A CNN/BiLSTM with the same architecture of our CNN/BiLSTM feature extractor (trained on 7,556 instances of the CH training set); *-CH-EN systems: a simple ensemble way that directly combine the softmax layers of the CNN-CH system (trained on 3,778 instances of the CH training set) and CNN-GAN system (trained on 3,778 instances of the EN training set).

adversarial model can transfer some useful knowledge and information from the source language to the target language for relation classification, by which both the language-specific and language-invariant features could be learned.

Besides, better ensemble methods could be attempted to exploit the information across language for semi-supervised relation classification. In addition, we see that our model has obtained improvement over the previous best-performing system (bilingual-Joint-IRL in Table 5) of semi-supervised relation classification.

5 Conclusion

In this paper we introduce an adversarial feature adaptation approach for cross-lingual relation classification without labeled dataset, which leverages the data on richly-labeled language to help relation classification on the poorly-labeled language. We evaluate our approach on ACE 2005 multilingual training corpus. Experimental results show that this approach can effectively transfer feature representations from a richly-labeled language to another poorly-labeled language, and outperforms several baselines including active learning models and highly competitive MT-based baselines. The code is available at `https://github.com/zoubowei/feature_adaptation4RC`.

Theoretically speaking, our adversarial feature adaptation approach can be flexibly implemented in the scenario of multiple languages, while this paper focuses on two languages of English and Chinese. Thus in future, we will extend this approach to more languages and explore its significance.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 61703293, No. 61751206, and No. 61672368). We would like to thank the anonymous reviewers for their insightful comments and suggestions.

References

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. *Advances in Neural Information Processing Systems*, pages 343–351.

Razvan C. Bunescu and Raymond J. Mooney. 2005. Subsequence kernels for relation extraction. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS'05)*, pages 171–178.

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*.

Cicero Nogueira 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 (ACL'15), pages 626–634.

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 (NAACL'15)*, pages 1351–1356.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, Thierry Charnois. 2018. SemEval-2018 Task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of International Workshop on Semantic Evaluation (SemEval'18)*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Mario Marchand, and Victor Lempitsky. 2017. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030.

Ian J. 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 Advances in Neural Information Processing Systems Conference (NIPS'14)*, pages 2672–2680.

Iris Hendrickx, Nam Kim Su, Zornitsa Kozareva, Preslav Nakov, 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*, pages 33–38.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics on Interactive poster and demonstration sessions*.

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.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. 2016. Adversarial autoencoders. In *Proceedings of the International Confer- ence on Learning Representations (ICLR'16)*.

Bonan Min, Zhuolin Jiang, Marjorie Freedman, and Ralph Weischedel. 2017. Learning transferable representation for bilingual relation extraction via convolutional neural networks. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing (IJCNLP'17)*, pages 674–684.

Gwangbeen Park and Woobin Im. 2016. Image-Text multi-Modal representation learning by adversarial backpropagation. *arXiv preprint arXiv:1612.08354*.

Longhua Qian, Haotian Hui, YaNan 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 (ACL'14)*, pages 582–592.

Longhua Qian, Guodong Zhou, Fang Kong, Qiaoming Zhu, and Peide Qian. 2008. Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING'08)*, pages 697–704.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric P. Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pages 1006–1017.

Motoki Sato, Hitoshi Manabe, Hiroshi Noji, and Yuji Matsumoto. 2017. Adversarial training for cross-domain universal dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (CoNLL'17)*, pages 71–79.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL'15, Short Papers)*, pages 567–572.

Fabian M. Suchanek, Georgiana Ifrim, and Gerhard Weikum. 2006. Combining linguistic and statistical analysis to extract relations from web documents. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'06)*, pages 712–717.

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 (NAACL'16)*, pages 886–896.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. *ACE 2005 multilingual training corpus*. Linguistic Data Consortium.

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 (ACL'16)*, pages 1298–1307.

Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schtze. 2017. Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923*.

Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*, pages 1753–1762.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING'14)*, pages 2335–2344.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78.