Adversarial Multi-lingual Neural Relation Extraction

Xiaozhi Wang[1*], **Xu Han**[1*], **Yankai Lin**[1], **Zhiyuan Liu**[1†], **Maosong Sun**[1,2]
[1]Department of Computer Science and Technology,
State Key Lab on Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing, China
[2]Beijing Advanced Innovation Center for Imaging Technology,
Capital Normal University, Beijing, China

Abstract

Multi-lingual relation extraction aims to find unknown relational facts from text in various languages. Existing models cannot well capture the consistency and diversity of relation patterns in different languages. To address these issues, we propose an adversarial multi-lingual neural relation extraction (AMNRE) model, which builds both consistent and individual representations for each sentence to consider the consistency and diversity among languages. Further, we adopt an adversarial training strategy to ensure those consistent sentence representations could effectively extract the language-consistent relation patterns. The experimental results on real-world datasets demonstrate that our AMNRE model significantly outperforms the state-of-the-art models. The source code of this paper can be obtained from `https://github.com/thunlp/AMNRE`.

1 Introduction

Relation extraction (RE) is a crucial task in NLP, which aims to extract semantic relations between entity pairs from the sentences containing them. For example, given an entity pair (*Bill Gates*, *Microsoft*) and a sentence "*Bill Gates* is the co-founder and CEO of *Microsoft*", we want to figure out the relation `Founder` between the two entities. RE can potentially benefit many applications, such as knowledge base construction (Zhong et al., 2015; Han et al., 2018) and question answering (Xiang et al., 2017).

Recently, neural models have shown their great abilities in RE. Zeng et al. (2014) introduce a convolutional neural network (CNN) to extract relational facts with automatically learning features from text. To address the issue of lack of data, Zeng et al. (2015) incorporate multi-instance learning with a piece-wise convolutional neural network (PCNN) to extract relations in distantly supervised data. Because distant supervision suffer from wrong labeling problems, Lin et al. (2016) further employ a sentence-level selective attention to filter out those noisy sentences in distantly supervised data and achieve state-of-the-art performance. All these neural relation extraction (NRE) models merely focus on extracting relational facts from mono-lingual data, ignoring the rich information in multi-lingual data.

Lin et al. (2017) propose a multi-lingual attention-based neural relation extraction (MNRE) model, which considers the consistency and complementarity in multi-lingual data. MNRE builds a sentence representation for each sentence in various languages and employs a multi-lingual attention to capture the pattern consistency and complementarity among languages.

Although MNRE achieves great success in multi-lingual RE, it still has some problems. MNRE learns a single representation for each sentence in various languages, which cannot well capture both the consistency and diversity of relation patterns in different languages. Moreover, MNRE simply utilizes a multi-lingual attention mechanism and a global relation predictor to capture the consistent relation patterns among multiple languages. From the experimental data, we find that the sentence representations in different languages are still far from each other and linearly separable. Therefore, it is hard for the multi-

[*] indicates equal contribution
[†] Corresponding author: Zhiyuan Liu (liuzy@tsinghua.edu.cn).

Proceedings of the 27th International Conference on Computational Linguistics, pages 1156–1166
Santa Fe, New Mexico, USA, August 20-26, 2018.

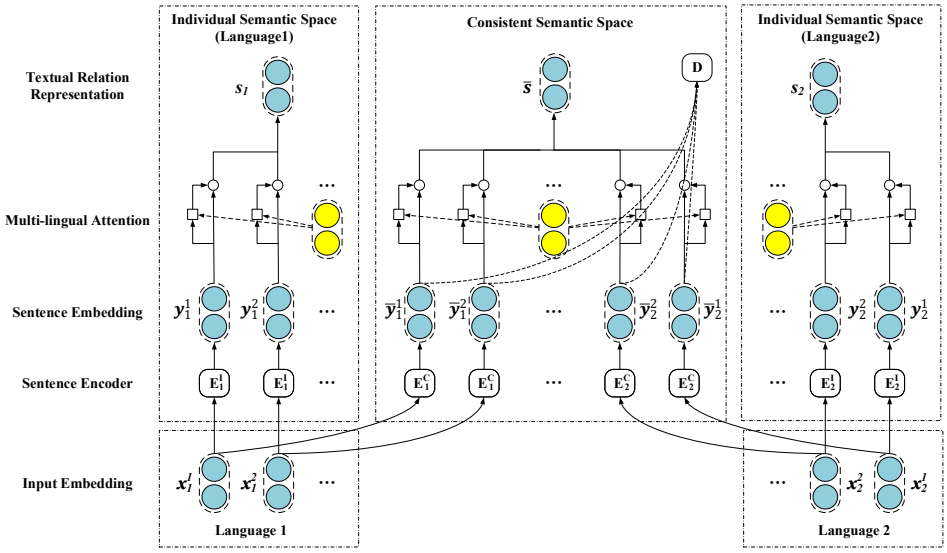


Figure 1: Overall architecture of our adversarial multi-lingual neural relation extraction (AMNRE) which contains two languages.

lingual attention mechanism and global relation predictor to extract relation consistency from distinct sentence representations.

To address these issues, we propose an adversarial multi-lingual NRE (AMNRE) model. As shown in Figure 1, for an entity pair, we encode its corresponding sentences in various languages through neural sentence encoders. For each sentence, we build an individual representation to grasp its individual language features and a consistent representation to encode its substantially consistent features among languages. Further, we adopt an adversarial training strategy to ensure AMNRE can extract the language-consistent relation patterns from the consistent representations. Orthogonality constraints are also adopted to enhance differences between individual representations and consistent representations for each language.

In experiments, we take Chinese and English to show the effectiveness of AMNRE. The experimental results show that AMNRE outperforms all baseline models significantly by explicitly encoding the consistency and diversity among languages. And we further give a case study and an ablation study to demonstrate the adversarial training strategy could help AMNRE to capture language-consistent relation patterns.

2 Related Works

2.1 Relation Extraction

Traditional supervised RE models (Zelenko et al., 2003; Socher et al., 2012; Santos et al., 2015) heavily rely on abundant amounts of high-quality annotated data. Hence, Mintz et al. (2009) propose a distantly supervised model for RE. Distant supervision aligns knowledge bases (KBs) and text to automatically annotate data, and thus distantly supervised models inevitably suffer from wrong labeling problems.

To alleviate the noise issue, Riedel et al. (2010) and Hoffmann et al. (2011) propose multi-instance learning (MIL) mechanisms for single-label and multi-label problems respectively. Then, Zeng et al. (2015) attempt to integrate neural models into distant supervision. Lin et al. (2016) further propose a sentence-level attention to jointly consider all sentences containing same entity pairs for RE. The attention-based neural relation extraction (NRE) model has become a foundation for some recent works (Ji et al., 2017; Zeng et al., 2017; Liu et al., 2017b; Wu et al., 2017; Feng et al., 2018; Zeng et al., 2018).

Most existing RE models are devoted to extracting relations from mono-lingual data and ignore information lying in text of multiple languages. Faruqui and Kumar (2015) and Verga et al. (2016) first attempt to adopt multi-lingual transfer learning for RE. However, both of these works learn predictive

models on a new language for existing KBs, without fully leveraging semantic information in text. Then, Lin et al. (2017) construct a multi-lingual NRE (MNRE) model to jointly represent text of multiple languages to enhance RE. In this paper, we propose a novel multi-lingual NRE framework to explicitly encode language consistency and diversity into different semantic spaces, which can achieve more effective representations for RE.

2.2 Adversarial Training

Goodfellow et al. (2015) propose adversarial training for image classification tasks. Afterwards, Goodfellow et al. (2014) propose a mature adversarial training framework and use the framework to train generative models. Adversarial networks have recently been used as methods to narrow probability distributions and proven effective in some tasks. In domain adaptation, Ganin et al. (2016) and Bousmalis et al. (2016) adopt adverarial training strategies to transfer the features of one source domain to its corresponding target domain.

Inspired by Ganin et al. (2016), adversarial training has also been explored in some typical NLP tasks for multi-feature fusion. Park and Im (2016) propose a multi-modal representation learning model based on adversarial training. Then, Liu et al. (2017a) employ adversarial training to construct a multi-task learning model for text classification by extending the original binary adversarial training to the multi-class version. And a similar adversarial framework is also adapted by Chen et al. (2017) to learn features from different datasets for chinese word segmentation. In this paper, we adopt adversarial training to boost feature fusion to grasp the consistency among different languages.

3 Methodology

In this section, we introduce the overall framework of our proposed AMNRE in detail. As shown in Figure 1, for each entity pair, AMNRE encodes its corresponding sentences in different languages into several semantic spaces to grasp their individual language patterns. Meanwhile, a unified space is also set up to encode consistent features among languages. By explicitly encoding the consistency and diversity among languages, AMNRE can achieve better extraction results in the multi-lingual scenario.

For each given entity pair, we define its corresponding sentences in n different languages as $\mathcal{T} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$, where $\mathcal{S}_j = \{x_j^1, \ldots, x_j^{|\mathcal{S}_j|}\}$ denotes the sentence set in the j-th language. All these sentences are labeled with the relation $r \in \mathcal{R}$ by heuristical labeling algorithms in distant supervision (Mintz et al., 2009). Our model aims to learn a relation extractor by maximizing the conditional probability $p(r|\mathcal{T})$ with the following three components:

Sentence Encoder. Given a sentence and its target entity pair, we employ neural networks to encode the sentence into a embedding. In this paper, we implement the sentence encoder with both convolutional (CNN) and recurrent (RNN) architectures. Specifically, we set the encoders E_j^I and E_j^C to encode each sentence in the j-th language into its individual and consistent embeddings respectively, and expect these embeddings to capture the diversity and consistency among languages.

Multi-lingual Attention. Since not all sentences are labeled correctly in distant supervision, we adopt multi-lingual attention mechanisms to capture those informative sentences. In practice, we apply language-individual and language-consistent attentions to compute local and global textual relation representations respectively for final prediction.

Adversarial Training. Under the framework of AMNRE, we encode the sentences in various languages into a unified consistent semantic space. We further adopt adversarial training to ensure these sentences are well fused in the unified space after encoding so that our model can effectively extract the language-consistent relation patterns.

We will introduce the three components in detail as follows.

3.1 Sentence Encoder

Given a sentence $x = \{w_1, w_2, \ldots\}$ containing two entities, we apply neural architectures including both CNN and RNN to encode the sentence into a continuous low-dimensional space to capture its implicit semantics.

3.1.1 Input Layer

The input layer transforms all input words in the sentence into corresponding input embeddings by concatenating their word embeddings and position embeddings. The word embeddings are pre-trained by Skip-Gram (Mikolov et al., 2013). The position embeddings are a widely-used technique in RE proposed by Zeng et al. (2014), representing each word's relative distances to the two entities into two k_p-dimensional vectors. The input layer represents the input sentence as a k_i-dimensional embedding sequence $\mathbf{x} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots\}$, where $k_i = k_w + k_p \times 2$, k_w and k_p are the dimensions of word embeddings and position embeddings respectively.

3.1.2 Encoding Layer

After representing the input sentence as a k_i-dimensional embedding sequence, we select both CNN (Zeng et al., 2014) and RNN (Zhang and Wang, 2015) to encode the input embedding sequence $\mathbf{x} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots\}$ to its sentence embedding.

CNN slides a convolution kernel with the window size m to extract the k_h-dimensional local features,

$$\mathbf{h}_i = \text{CNN}\big(\mathbf{w}_{i-\frac{m-1}{2}}, \ldots, \mathbf{w}_{i+\frac{m-1}{2}}\big). \tag{1}$$

A max-pooling is then adopted to obtain the final sentence embedding $\mathbf{y}$ as follows,

$$[\mathbf{y}]_j = \max\{[\mathbf{h}_1]_j, \ldots, [\mathbf{h}_n]_j\}. \tag{2}$$

RNN is mainly designed for modeling sequential data. In this paper, we adopt bidirectional RNN (Bi-RNN) to incorporate information from both sides of the sentence sequence as follows,

$$\overrightarrow{\mathbf{h}}_i = \text{RNN}_f(\mathbf{x}_i, \overrightarrow{\mathbf{h}}_{i-1}), \quad \overleftarrow{\mathbf{h}}_i = \text{RNN}_b(\mathbf{x}_i, \overleftarrow{\mathbf{h}}_{i+1}), \tag{3}$$

where $\overrightarrow{\mathbf{h}}_i$ and $\overleftarrow{\mathbf{h}}_i$ are the k_h-dimensional hidden states at the position i of the forward and backward RNN respectively. $\text{RNN}(\cdot)$ is the recurrent unit and we select gated recurrent unit (GRU) (Cho et al., 2014) as the recurrent unit in this paper. We concatenate both the forward and backward hidden states as the sentence embedding $\mathbf{y}$,

$$\mathbf{y} = [\overrightarrow{\mathbf{h}}_n; \overleftarrow{\mathbf{h}}_1]. \tag{4}$$

For simplicity, we denote such a sentence encoding operation as the following equation,

$$\mathbf{y} = \text{E}(x). \tag{5}$$

For each sentence $x_j^i \in \mathcal{S}_j$, we adopt the individual sentence encoder E_j^I and the consistent sentence encoder E_j^C to embed the sentence into its individual and consistent representations respectively,

$$\{\mathbf{y}_j^1, \mathbf{y}_j^2, \ldots\} = \{\text{E}_j^I(x_j^1), \text{E}_j^I(x_j^2), \ldots\}, \quad \{\bar{\mathbf{y}}_j^1, \bar{\mathbf{y}}_j^2, \ldots\} = \{\text{E}_j^C(x_j^1), \text{E}_j^C(x_j^2), \ldots\}. \tag{6}$$

3.2 Multi-lingual Selective Attention

For each given entity pair, AMNRE adopts multi-lingual selective attention mechanisms to exploit informative sentences in $\mathcal{T}$. We explicitly encode languages' consistency and diversity into individual and consistent representations, thus our attentions are more simple than those proposed in Lin et al. (2017).

3.2.1 Language-individual Attention

Since it is intuitive that each language has its own characteristic, we set language-individual attention mechanisms for different languages. In the individual semantic space of the j-th language, we assign a query vector $\mathbf{r}_j$ to each relation $r \in \mathcal{R}$. The attention score for each sentence in $\mathcal{S}_j = \{x_j^1, x_j^2, \ldots\}$ is defined as follows,

$$\alpha_j^i = \frac{\exp(\mathbf{r}_j^\top \mathbf{y}_j^i)}{\sum_{k=1}^{|\mathcal{S}_j|} \exp(\mathbf{r}_j^\top \mathbf{y}_j^k)}. \tag{7}$$

The attention scores can be used to compute language-individual textual relation representations,

$$\mathbf{s}_j = \sum_{k=1}^{|\mathcal{S}_j|} \alpha_j^k \mathbf{y}_j^k. \tag{8}$$

3.2.2 Language-consistent Attention

Besides language-individual attention mechanisms, we also adopt a language-consistent attention to take all sentences in all languages into consideration. In the consistent semantic space, we also assign a query vector $\bar{\mathbf{r}}$ to each relation $r \in \mathcal{R}$ and the attention score for each sentence is defined as follows,

$$\beta_j^i = \frac{\exp(\bar{\mathbf{r}}^\top \bar{\mathbf{y}}_j^i)}{\sum_{l=1}^{n} \sum_{k=1}^{|\mathcal{S}_l|} \exp(\bar{\mathbf{r}}^\top \bar{\mathbf{y}}_l^k)}. \tag{9}$$

The attention scores can be used to compute language-consistent textual relation representations,

$$\bar{\mathbf{s}} = \sum_{l=1}^{n} \sum_{k=1}^{|\mathcal{S}_l|} \beta_l^k \bar{\mathbf{y}}_l^k. \tag{10}$$

3.3 Relation Prediction

With the language-individual textual relation representations $\{\mathbf{s}_1, \mathbf{s}_2, \ldots\}$ and the language-consistent textual relation representation $\bar{\mathbf{s}}$, we can estimate the probability $p(r|\mathcal{T})$ over each relation $r \in \mathcal{R}$,

$$p(r|\mathcal{T}) = p(r|\bar{\mathbf{s}}) \prod_{j=1}^{n} p(r|\mathbf{s}_j). \tag{11}$$

$p(r|\bar{\mathbf{s}})$ and $p(r|\mathbf{s}_j)$ can be defined as follows,

$$p(r|\mathbf{s}_j) = \text{softmax}[\mathbf{R}_j \mathbf{s}_j + \mathbf{d}_j], \quad p(r|\bar{\mathbf{s}}) = \text{softmax}[\bar{\mathbf{R}}\bar{\mathbf{s}} + \bar{\mathbf{d}}], \tag{12}$$

where $\mathbf{d}_j$ and $\bar{\mathbf{d}}$ are bias vectors, $\mathbf{R}_j$ is the specific relation matrix of the j-th language, and $\bar{\mathbf{R}}$ is the consistent relation matrix. We define the objective function to train the relation extractor as follows,

$$\min_{\theta} \mathcal{L}_{nre}(\theta) = -\sum_{l} \log p(r_l|\mathcal{T}_l), \tag{13}$$

where θ is all parameters in the framework. In the training phase, $p(r|\mathcal{T})$ is computed using the labeled relations as the attention queries. In the test phase, we need to use each possible relation as attention queries to compute $p(r|\mathcal{T})$ for relation prediction since the relations are unknown in advance.

3.4 Adversarial Training

In our framework, we encode sentences of various languages into a consistent semantic space to grasp the consistency among languages. One possible situation is that sentences of different languages are aggregated in different places of the space and linearly separable. In this case, our purpose of mining substantially consistent relation patterns in different languages is difficult to be reached. Inspired by Ganin et al. (2016), we adopt adversarial training into our framework to address this problem.

In the adversarial training, we define a discriminator to estimate which kind of languages the sentences from. The probability distributions over these sentences are formalized as follows,

$$\text{D}(\bar{\mathbf{s}}_j^i) = \text{softmax}(\text{MLP}(\bar{\mathbf{s}}_j^i)), \tag{14}$$

where MLP is a two-layer multilayer perceptron network.

Contrary to the discriminator, the consistent sentence encoders are expected to produce sentence embeddings that cannot be reliably predicted by the discriminator. Hence, the adversarial training process is a min-max game and can be formalized as follows,

$$\min_{\theta_{\text{E}}^C} \max_{\theta_{\text{D}}} \sum_{j=1}^{n} \sum_{i=1}^{|\mathcal{S}_j|} \log[\text{D}(\text{E}_j^C(x_j^i))]_j, \tag{15}$$

where $[\cdot]_j$ is the j-th value of the vector.

The formula means that given a sentence of any language, the corresponding sentence encoder of its language generates the sentence embedding to confuse the discriminator. Meanwhile, the discriminator

tries its best to predict the language of the sentence according to the sentence embedding. After sufficient training, the encoders and the discriminator reach a balance, and sentences of different languages containing similar semantic information can be well encoded into adjacent places of the space. In training, we optimize the following loss functions instead of Eq. 15,

$$\min_{\theta_{\mathrm{E}}^{C}} \mathcal{L}_{adv}^{E}(\theta_{\mathrm{E}}^{C}) = \sum_{l} \sum_{\mathcal{S}_j \in \mathcal{T}_l} \sum_{x_j^i \in \mathcal{S}_j} \log[\mathrm{D}(\mathrm{E}_j^C(x_j^i))]_j, \quad \min_{\theta_{\mathrm{D}}} \mathcal{L}_{adv}^{D}(\theta_{\mathrm{D}}) = -\sum_{l} \sum_{\mathcal{S}_j \in \mathcal{T}_l} \sum_{x_j^i \in \mathcal{S}_j} \log[\mathrm{D}(\mathrm{E}_j^C(x_j^i))]_j, \tag{16}$$

where θ_{E}^{C} and θ_{D} are all parameters of the consistent sentence encoders and the discriminator.

We notice that language-individual semantics could be wrongly encoded into the consistent semantic space, and may have negative effects on extracting language-consistent features. Inspired by Bousmalis et al. (2016), we adopt orthogonality constraints to alleviate this issue. We minimize the following penalty function:

$$\min_{\theta_{\mathrm{E}}} \mathcal{L}_{penalty}(\theta_{\mathrm{E}}) = \sum_{j=1}^{n} \left\| \mathbf{I}_j^T \mathbf{C}_j \right\|_F, \tag{17}$$

where $\mathbf{I}_j$ and $\mathbf{C}_j$ are two matrices whose row vectors are the embeddings of sentences in the j-th language encoded by E_j^I and E_j^C respectively. θ_{E} is parameters of the all encoders. And $\|\cdot\|_F$ is the squared Frobenius norm.

3.5 Implementation Details

During training process, we combine the extraction and adversarial objective functions as follows,

$$\mathcal{L} = \mathcal{L}_{nre}(\theta) + \lambda_1 \mathcal{L}_{adv}^{D}(\theta_{\mathrm{D}}) + \lambda_2 \mathcal{L}_{adv}^{E}(\theta_{\mathrm{E}}^{C}) + \lambda_3 \mathcal{L}_{penalty}(\theta_{\mathrm{E}}), \tag{18}$$

where λ_1, λ_2, and λ_3 are harmonic factors. All models are optimized using stochastic gradient descent (SGD). In practice, we integrate λ_1 and λ_2 into the alternating ratio among the loss functions, and we calibrate a 1:1:5 ratio among $\mathcal{L}_{nre}(\theta) + \lambda_3 \mathcal{L}_{penalty}(\theta_{\mathrm{E}})$, $\mathcal{L}_{adv}^{D}(\theta_{\mathrm{D}})$ and $\mathcal{L}_{adv}^{E}(\theta_{\mathrm{E}}^{C})$. λ_3 is set as 0.02.

4 Experiments

4.1 Datasets and Evaluation

We evaluate our models on a multi-lingual relation extraction dataset developed by Lin et al. (2017). The dataset consists of English and Chinese data, and has 176 relations including a special relation `NA` indicating that there is no relation between entities. The whole dataset is divided into three parts for training, validation and test. The statistics of the dataset are listed in Table 1.

Dataset		#Rel	#Sent	#Fact	Dataset		#Rel	#Sent	#Fact
English	Training	176	1,022,239	47,638	Chinese	Training	176	940,595	42,536
	Validation	176	80,191	2,192		Validation	176	82,699	2,192
	Test	176	162,018	4,326		Test	176	167,224	4,326

Table 1: Statistics of the dataset

We evaluate all models by the held-out evaluation following previous works (Mintz et al., 2009; Lin et al., 2017). In experiments, we report precision-recall curves of recall under 0.3 since we focus more on the performance of those top-ranked results. To give a complete view of the performance, we also report the area under the curve (AUC).

4.2 Experiment Settings

Following the settings of previous works, we use the pre-trained word embeddings learned by Skip-Gram as the initial word embeddings. We implement the MNRE framework proposed by Lin et al. (2017) by ourselves. For fair comparision, we set most of the hyperparameters following Lin et al. (2017). We list the best setting of hyperparameters in Table 2.

Batch Size B	160	Convolution Kernel Size m	3
Learning Rate α	0.002	Dropout Probability p for CNN and RNN	0.5
Hidden Layer Dimension k_h for CNN	230	Dropout Probability p_d for the Discriminator	0.1
Hidden Layer Dimension k_h for RNN	200	Word Dimension k_w	50
Hidden Layer Dimension k_d for the Discriminator	2048	Position Dimension k_p	5

Table 2: Parameter settings.

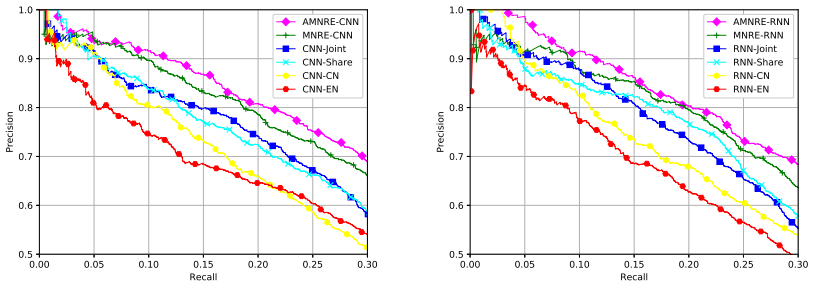


Figure 2: The aggregated precision-recall curves for proposed models and various baseline models. Left: models with CNN as sentence encoders. Right: models with RNN as sentence encoders.

4.3 Overall Evaluation Results

To evaluate the effectiveness of our proposed models **AMNRE-CNN** and **AMNRE-RNN**, we compare the proposed models with various neural methods: **MNRE-CNN** and **MNRE-RNN** are multi-lingual attention-based NRE models with CNN and RNN sentence encoders respectively (Lin et al., 2017); **CNN-EN** and **RNN-EN** are vanilla selective-attention NRE models trained with English data, which are the state-of-the-art models in mono-lingual RE (Lin et al., 2016); **CNN-CN** and **RNN-CN** are trained with Chinese data; **CNN-Joint** and **RNN-Joint** are naive joint models which predict relations by directly summing up ranking scores of both English and Chinese; **CNN-Share** and **RNN-Share** are another naive joint models which train English and Chinese models with shared relation embeddings. The results of precision-recall curves are shown in Figure 2 and the results of AUC are shown in Table 3.

From the results, we have the following observations:

(1) Both for CNN and RNN, the models jointly utilizing English and Chinese sentences outperform the models only using mono-lingual sentences. This demonstrates that the rich information in multi-lingual data is useful and can significantly enhance existing NRE models.

(2) The **-Joint** models achieve similar performance with the **-Share** models, and both of them underperform the **MNRE** and **AMNRE** models. They all benefit from the multi-lingual information, but the models with multi-lingual attentions can better take advantage of multi-lingual data. It indicates that designing targeted schemes to extract rich multi-lingual information is crucial.

(3) **AMNRE** achieves the best results among all the baseline models over the entire range of recall in Figure 2, even as compared with **MNRE**. **AMNRE** also outperforms **MNRE** with 3 percentage points increasing in the AUC results. It indicates our proposed framework which explicitly encodes language-consistent and language-individual semantics better extract multi-lingual information, and therefore lead to the significant improvement in RE performance.

Models	CNN-EN	CNN-CN	CNN-Joint	CNN-Share	MNRE-CNN	AMNRE-CNN
AUC	36.6	33.2	37.1	37.0	43.4	**46.2**
Models	RNN-EN	RNN-CN	RNN-Joint	RNN-Share	MNRE-RNN	AMNRE-RNN
AUC	34.5	34.4	36.5	37.6	44.2	**47.3**

Table 3: The AUC results of different models (%).

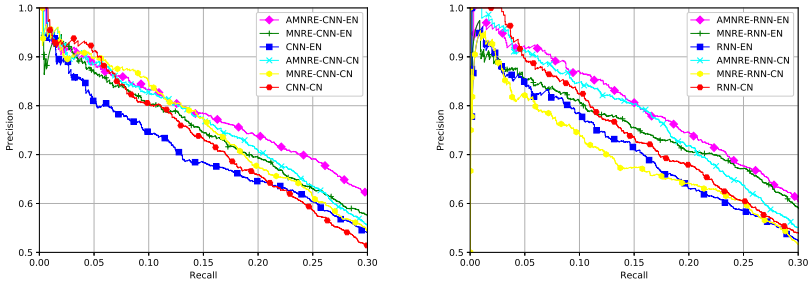


Figure 3: The aggregated precision-recall curves for proposed models and various baseline models in the mono-lingual scenario. Left: models with CNN as sentence encoders. Right: models with RNN as sentence encoders.

Models	CNN-EN	MNRE-EN	AMNRE-EN	RNN-EN	MNRE-EN	AMNRE-EN
AUC	36.6	39.6	**42.7**	34.5	42.2	**43.2**
Models	CNN-CN	MNRE-CN	AMNRE-CN	RNN-CN	MNRE-CN	AMNRE-CN
AUC	33.2	34.6	**37.9**	33.5	34.8	**36.4**

Table 4: The AUC results of different models in the mono-lingual scenario (%).

4.4 Mono-lingual Evaluation Results

To further verify that every mono-lingual RE models can benefit from our proposed framework, which explicitly consider language-consistent relation patterns, we train models with multi-lingual data and evaluate the performance of these models in the mono-lingual RE scenario. To show the results clearly, we report the precision-recall curves in Figure 3 and the AUC results in Table 4.

From the results, we can observe that:

(1) As compared with the models directly learned with the mono-lingual data, the models exploiting the multi-lingual information perform better in the mono-lingual scenario. This demonstrates that there is latent consistency among languages, and grasping this consistency from multi-lingual data can provide additional information for models in each language to enhance their results in the mono-lingual scenario.

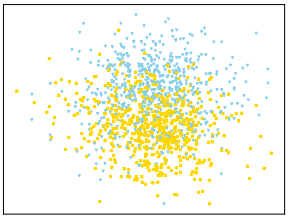
(2) Our proposed models achieve the best precision over the entire range of recall and also significantly improve the AUC results as compared with both MNRE and mono-lingual RE models. It indicates that due to the consistent semantic space in our framework, language-consistent information lying in the multi-lingual data is better mined and serve the mono-lingual scenario.

4.5 Effectiveness of Adversarial Training and Orthogonality Constraints

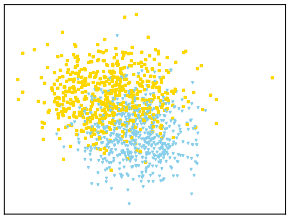
We adopt an adversarial training strategy to fuse the features from different languages to extract consistent relation patterns. Orthogonality constraints are also adopted to separate the consistent and individual feature spaces. To measure the effectiveness of them, we conduct an ablation study which compares the proposed models with the similar models but without adversarial training strategy (**AMNRE-noA**), without orthogonality constraints (**AMNRE-noO**), and without both of them (**AMNRE-noBoth**). The AUC results are shown in Table 5.

We can observe that both the adversarial training strategy and orthogonality constraints have significant influence on the performance of our proposed model. This demonstrates the effectiveness of adversarial training strategy and orthogonality constraints for multi-lingual RE. To give a more intuitive picture of the effect of these two mechanisms, we visualize the distribution of sentence feature embeddings encoded by the individual and consistent encoders using t-SNE (Maaten and Hinton, 2008). The results are shown in Figure 4.

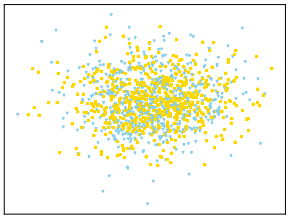
Figure 4(a) shows that there are obvious differences between the feature embeddings encoded from the same sentences by individual and consistent encoders. It indicates the orthogonality constraints are effective to separate the individual and consistent latent spaces. From the comparison between Figure



(a) The same English sentences encoded by the consistent encoder (yellow) and individual encoder (blue).



(b) The English sentences (yellow) and the Chinese sentences (blue) encoded by their own consistent encoders without adversarial training.



(c) The English sentences (yellow) and the Chinese sentences (blue) encoded by their own consistent encoders with adversarial training.

Figure 4: The visualization of sentence feature embeddings with different mechanisms.

Models	AMNRE-CNN	AMNRE-CNN-noA	AMNRE-CNN-noO	AMNRE-CNN-noBoth
AUC	**46.2**	44.1	43.9	41.3
Models	AMNRE-RNN	AMNRE-RNN-noA	AMNRE-RNN-noO	AMNRE-RNN-noBoth
AUC	**47.3**	43.5	43.5	42.2

Table 5: The AUC results of the proposed models and ablated models.(%)

4(b) and Figure 4(c), we can observe that the feature embeddings from different languages are well-mixed due to the adversarial training strategy. We can more easily to grasp latent consistency among languages after multi-feature fusion.

5 Case Study

To further show the effectiveness of our proposed model to extract the language-consistent semantic information, we give an example in Table 6. We adopt the cosine similarity to measure the similarity between sentence embeddings encoded by consistent encoders. The first sentence in the middle column is the standard Chinese translation of the left sentence, thus they share the same semantic information. We observe that in our proposed model, the feature embedding similarity between these two sentences are significantly higher than the other English sentences sharing entity pair and relational fact but differing in semantics. It indicates that sentences in different languages containing similar semantics can be indeed encoded into adjacent places of the consistent space in our framework.

Relation : `Located in`		Cosine Similarity
There are eighteen small glaciers in the **North Island** on **Mount Ruapehu**.	北岛的**鲁阿佩胡山**上有十八个小冰川。	**0.584**
	...the bottom of the **North Island** of New Zealand up to the area of **Mount Ruapehu**.	0.3538
	It is located on the south-eastern **North Island** volcanic plateau, ...south-east of **Mount Ruapehu**.	0.342

Table 6: The example highlighting entities for the case study by measuring the cosine similarities between the sentence in the left column and each sentence in the middle column.

6 Conclusion and Future Work

In this paper, we introduce a novel adversarial multi-lingual neural relation extraction model (AMNRE). AMNRE builds both individual and consistent representations for each sentence to consider the consistency and diversity of relation patterns among languages. It also employs an adversarial training strategy and orthogonality constraints to ensure the consistent representations could extract the language-consistent features to extract relations. The experimental results on real-world datasets demonstrate that

our AMNRE could effectively encode the consistency and diversity among languages, and achieves state-of-the-art performance in relation extraction.

We will explore the following directions as our future work: (1) AMNRE can be also implemented in the scenario of multiple languages, and this paper shows the effectiveness of AMNRE on the dataset with two languages (English and Chinese). In the future, we will explore AMNRE in much more other languages such as French, Spanish, and so on. (2) AMNRE simply aligns the sentences with similar semantics in different languages with an adversarial training strategy. In fact, machine translation is a typical approach to align sentences in various languages. In the future, we will combine machine translation with our model to further improve the extraction performance.

Acknowledgments

We thank Jiacheng Zhang for his help. This work is supported by the National Natural Science Foundation of China (NSFC No. 61621136008, 61772302) and Tsinghua University Initiative Scientific Research Program (20151080406). This research is part of the NExT++ project, supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@Singapore Funding Initiative.

References

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Proceedings of NIPS*, pages 343–351.

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-criteria learning for chinese word segmentation. In *Proceedings of ACL*, pages 1193–1203.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: encoder-decoder approaches. In *Proceedings of SSST-8*.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual open relation extraction using cross-lingual projection. In *Proceedings of NAACL*, pages 1351–1356.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of AAAI*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *JMLR*, 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 NIPS*, pages 2672–2680.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of ICML*, pages 1–10.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *Proceedings of AAAI*.

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 ACL*, pages 541–550.

Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao, et al. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of AAAI*, pages 3060–3066.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*, pages 2124–2133.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of ACL*, pages 34–43.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017a. Adversarial multi-task learning for text classification. In *Proceedings of ACL*, pages 1–10.

Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017b. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of EMNLP*, pages 1790–1795.

Laurens Van Der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *JMLR*, 9(12):2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, pages 1–12.

Mintz, Mike, Steven, Jurafsky, and Dan. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL*, pages 1003–1011.

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

Sebastian Riedel, Limin Yao, and Andrew Mccallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of ECML-PKDD*, pages 148–163.

Cicero Nogueira Dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of ACL*, pages 626–634.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP*, pages 1201–1211.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew Mccallum. 2016. Multilingual relation extraction using compositional universal schema. In *NAACL*, pages 886–896.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of EMNLP*, pages 1778–1783.

Yang Xiang, Qingcai Chen, Xiaolong Wang, and Yang Qin. 2017. Answer selection in community question answering via attentive neural networks. *IEEE SPL*, 24(4):505–509.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *JMLR*, 3(2):1083–1106.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING*, pages 2335–2344.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*, pages 1753–1762.

Wenyuan Zeng, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Incorporating relation paths in neural relation extraction. In *Proceedings of EMNLP*.

Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Large scaled relation extraction with reinforcement learning. In *Proceedings of AAAI*.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.

Huaping Zhong, Jianwen Zhang, Zhen Wang, Hai Wan, and Zheng Chen. 2015. Aligning knowledge and text embeddings by entity descriptions. In *Proceedings of EMNLP*, pages 267–272.