Adversarial Multi-lingual Neural Relation Extraction

Xiaozhi Wang, Xu Han, Yankai Lin, Zhiyuan Liu, Maosong Sun

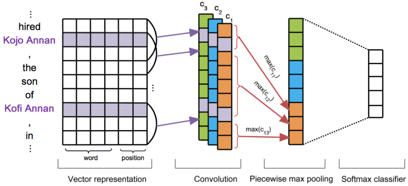
Tsinghua University

August 21, 2018

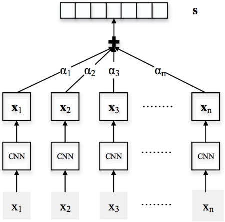
Introduction

- Relation Extraction: Extract semantic relations between entity pairs from plain text.
- E.g.
  Entity pair: (*Bill Gates*, *Microsoft*)
  Sentence: Bill Gates is the co-founder and CEO of Microsoft.
  Relation: `Founder`
- Application: Automatically construct large-scale knowledge graph from plain text.

Related Work: Distant Supervision



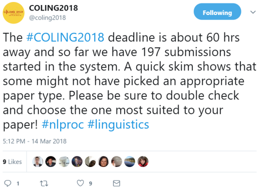
(a) Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks (Zeng et al., 2015)

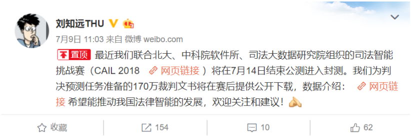


(b) Neural Relation Extraction with Selective Attention over Instances (Lin et al., 2016)

Multi-lingual Relation Extraction

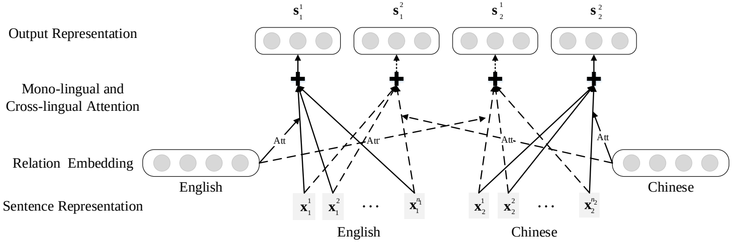
Web data are typically of various languages:





Rich information in multi-lingual data could benefit relation extraction.

Previous work: MNRE



Neural relation extraction with multi-lingual attention (Lin et al., 2017)

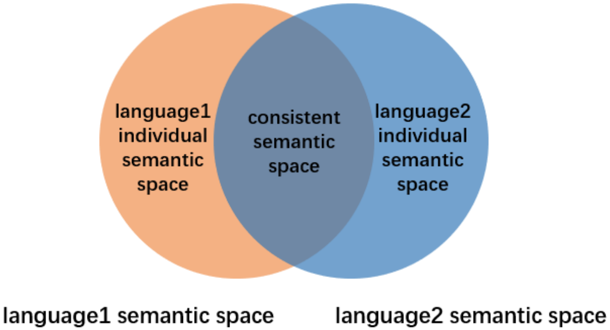
Problems in MNRE

- A single representation for each sentence: cannot well capture both consistency and diversity.
- Cross attention: time-consuming.

Motivation

Each sentence contains two parts of information:

- language-consistent information: semantics...
- language-individual information: syntax, special meaning of some words...

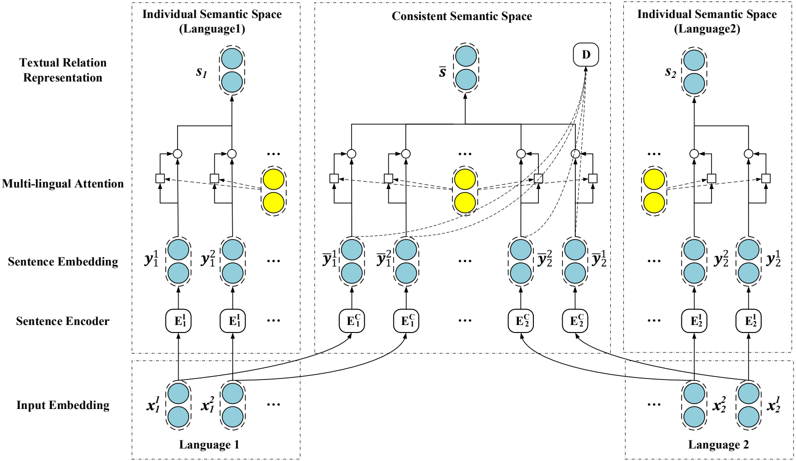


Our model

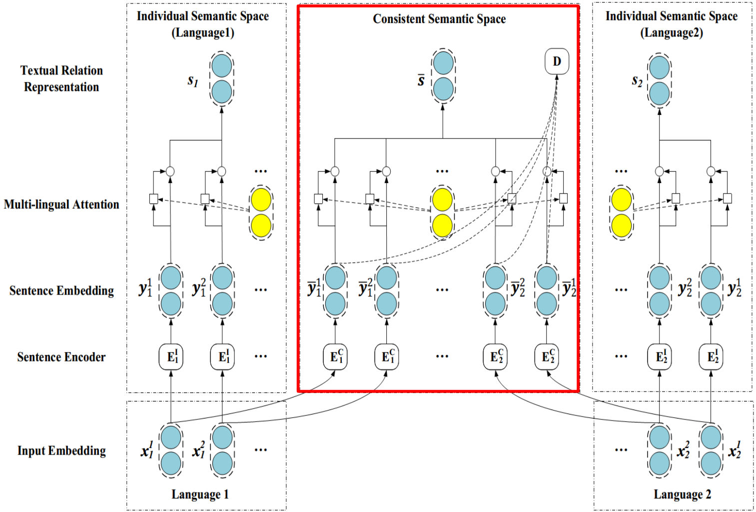
Explicitly encoder language consistency and diversity into different semantic spaces.

- **Adversarial Training**: extract language-consistent information.
- **Orthogonality Penalty**: separate individual and consistent space.
- **Attention Mechanism** in each space: denoising.

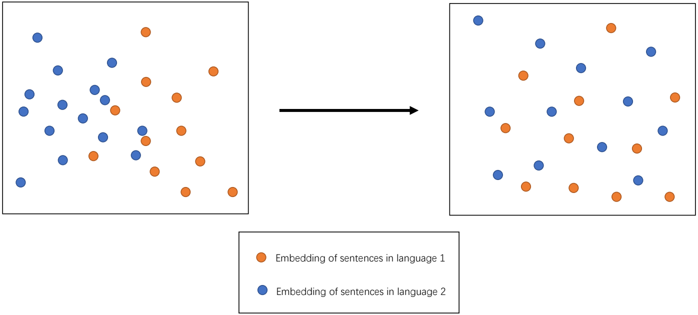
Overall architecture



Adversarial Training



Adversarial Training



Adversarial Training

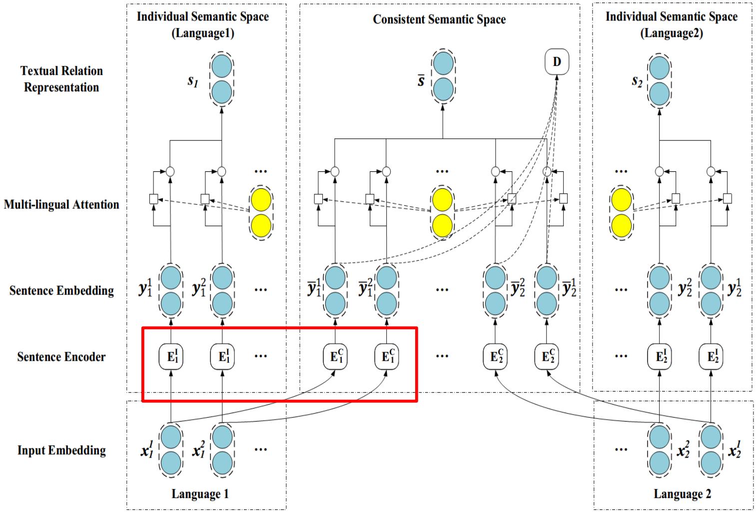
Discriminator estimates which kind of languages:

$$\mathsf{D}(\mathbf{s}_j^i) = \mathsf{softmax}(\mathsf{MLP}(\mathbf{s}_j^i)), \tag{1}$$

where MLP is a two-layer perceptron network.
Train the encoder adversarially:

$$\min_{\theta_{\mathsf{E}}^C} \max_{\theta_{\mathsf{D}}} \sum_{j=1}^{n} \sum_{i=1}^{|\mathcal{S}_j|} \log[\mathsf{D}(\mathsf{E}_j^C(x_j^i))]_j, \tag{2}$$

Orthogonality Penalty



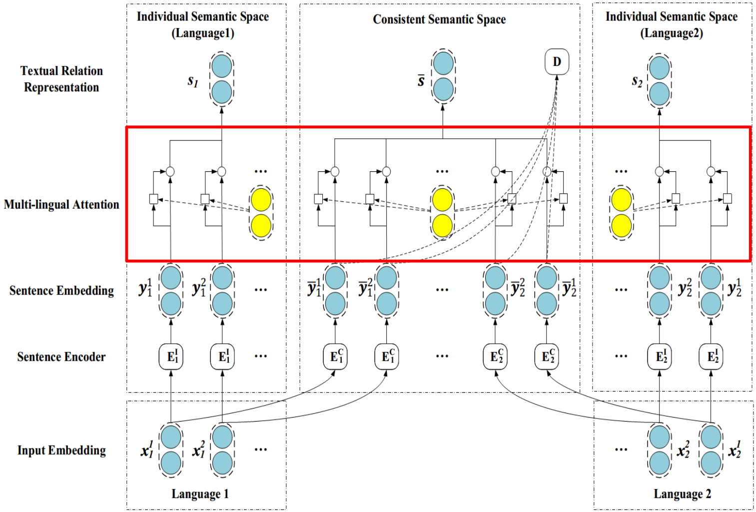
Orthogonality Penalty

Language-individual semantics could also appear in consistent semantic space.
We minimize the following penalty function to avoid it:

$$\min_{\theta_{\mathsf{E}}} \mathcal{L}_{penalty}(\theta_{\mathsf{E}}) = \sum_{j=1}^{n} \left\| \mathbf{I}_j^T \mathbf{C}_j \right\|_F, \tag{3}$$

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.

Orthogonality Penalty



Multi-lingual Selective Attention

Language-individual Attention

$$\alpha_j^i = \frac{\exp(\mathbf{r}_j^\top \mathbf{y}_j^i)}{\sum_{k=1}^{|\mathcal{S}_j|} \exp(\mathbf{r}_j^\top \mathbf{y}_j^k)}. \tag{4}$$

$$\mathbf{s}_j = \sum_{i=1}^{|\mathcal{S}_j|} \alpha_j^i \mathbf{y}_j^i. \tag{5}$$

Language-consistent Attention

$$\beta_j^i = \frac{\exp(\mathbf{r}^\top \mathbf{y}_j^i)}{\sum_{l=1}^{n} \sum_{k=1}^{|\mathcal{S}_l|} \exp(\mathbf{r}^\top \mathbf{y}_l^k)}. \tag{6}$$

$$\mathbf{s} = \sum_{j=1}^{n} \sum_{i=1}^{|\mathcal{S}_j|} \beta_j^i \mathbf{y}_j^i. \tag{7}$$

Experiments

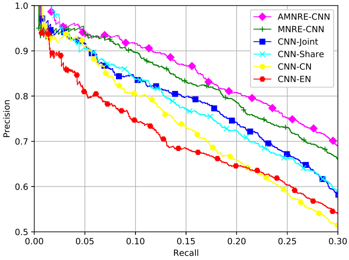
- Dataset: bilingual dataset from Wikipedia[1] and Baidu Baike[2]
- Held-out Evaluation

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

[1] www.wikipedia.org
[2] baike.baidu.com

Overall Evaluation Results



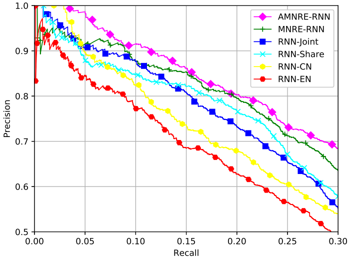


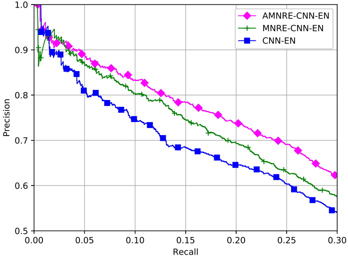
Figure 1: Precision-recall curves. Left: CNN models. Right: RNN models.

Overall Evaluation Results

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 2: The AUC results of different models (%).

Mono-lingual Evaluation Results



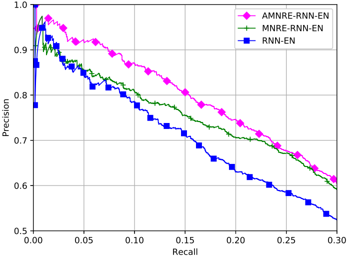


Figure 2: Precision-recall curves in English scenario. Left: CNN models. Right: RNN models.

Visualization

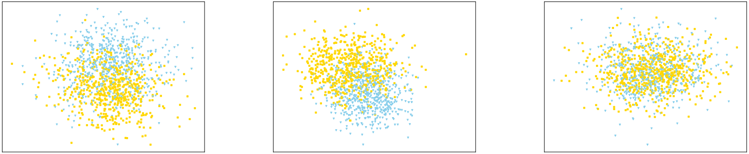


Figure 3: The visualization of sentence feature embeddings with different mechanisms using t-SNE.

Case Study

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 3: 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.

Conclusion

- We propose a novel adversarial multi-lingual neural relation extraction model.
- Experiments on real-world dataset demonstrate that our model achieves state-of-the-art results.

Future Work

- More languages.
- Combine Machine Translation.
- Transfer to another task.

The End

Thanks for listening.
Questions are welcome.

(a) Code

(b) Paper