A Unified Model for Cross-Domain and Semi-Supervised NER in Chinese Social Media

Hangfeng He and Xu Sun

MOE Key Laboratory of Computational Linguistics, Peking University
School of Electronics Engineering and Computer Science, Peking University
{*hangfenghe, xusun*} *@pku.edu.cn*

Overview

Motivation

Named entity recognition (NER) in Chinese social media is important but difficult because of its informality and strong noise. Previous methods only focus on in-domain supervised learning which is limited by the rare annotated data. However, there are enough corpora in formal domains and massive in-domain unannotated texts which can be used to improve the task.

Framework

We propose a unified model which can learn from out-of-domain corpora and in-domain unannotated texts. The unified model contains two major functions.
i) **Cross-domain learning function** can learn out-of-domain information based on domain similarity.
ii) **Semi-Supervised learning function** can learn in-domain unannotated information by self-training.

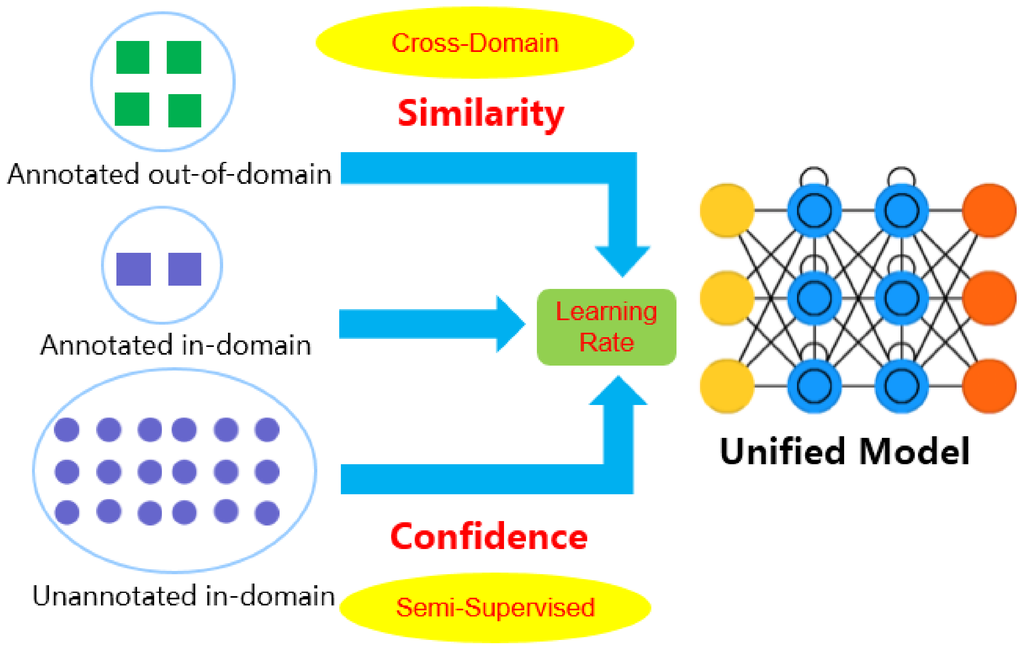


Figure 1: Illustration of the framework.

Proposal

Cross Domain Learning

For cross-domain learning, learning rate is adjusted by similarity function automatically. The learning rate for sentence x is computed as follow:

$$\alpha(x) = \alpha_0 * func(x, IN) \tag{1}$$

where α_0 is the fixed learning rate for in-domain sentences, $func(x, IN)$ indicates the similarity between sentence x and in-domain corpus IN, which is from 0 to 1.
In our model, we consider three different similarity functions: Cross Entropy, Gaussian RBF Kernel and Polynomial Kernel. Finally, we choose the polynomial kernel with $d = 1$ for our cross-domain learning.
We consider polynomial kernel as follow:

$$func(x, IN) = \frac{1}{C} \frac{< v_x, v_{IN} >^d}{\|v_x\|^d . \|v_{IN}\|^d} \tag{2}$$

The definition of C, v_x and v_{IN} are the same as Gaussian RBF kernel. If $d = 1$, the normalized kernel has the form $\frac{1}{C} cos\theta$,where θ is the angle between v_x and v_{IN} in the Euclidean space, which is exactly the cos kernel.

Semi-Supervised Learning

For sentence x, our prediction is the tag sequence with highest score as following:

$$y_{max}(x) = \underset{\bar{y} \in Y(x)}{argmax}\, s(x, \bar{y}, \theta)$$

we consider the tag sequence with the second highest score:

$$y_{2nd}(x) = \underset{\bar{y} \in Y(x) \text{ and } \bar{y} \neq y_{max}}{argmax} s(x, \bar{y}, \theta)$$

Then our sentence confidence is defined as follow:

$$confid(x) = \frac{y_{max}(x) - y_{2nd}(x)}{y_{max}(x)} \tag{3}$$

Our semi-supervised learning function is dynamic. The learning rate $\alpha_t(x)$ for unannotated sentence x in epoch t is computed as follow:

$$\alpha^t(x) = \alpha_0^t * confid(x, t) \tag{4}$$

where α_0^t is the learning rate for in-domain sentences at epoch t, $confid(x, t)$ is the confidence of sentence x at epoch t.

Unified Model

In our unified model, learning rate $\alpha^t(x)$ for every sentence x at epoch t is computed as follow:

$$\alpha^t(x) = \alpha_0^t * weight(x, t) \tag{5}$$

where $weight(x, t)$ is used to adjust learning rate for sentence x. The definition of $weight(x, t)$ is as follow:

$$weight(x, t) = \begin{cases} 1.0 & \text{x is in-domain,} \\ func(x, IN) & \text{x is out-of-domain,} \\ confid(x, t) & \text{x is unannotated.} \end{cases}$$

where $func(x, IN)$ is the similarity between sentence x and in-domain corpus IN and $confid(x, t)$ is the confidence of sentence x at epoch t.

Experiments

Results

In our experiments, we use Weibo NER corpus and SIGHAN corpus. Results are shown in Table 1.

Models	Named Entity			Nominal Mention				
	Precision	Recall	F1	Precision	Recall	F1	Overall	OOV
BiLSTM-MMNN	65.74	33.65	44.51	70.42	50.51	58.82	51.44	14.35
+ All Data Merge	43.58	45.02	44.29	28.81	17.17	21.52	32.27	**30.87**
Cross-Domain Learning (proposal)	52.94	**51.18**	52.05	71.63	51.01	59.59	55.70	**30.87**
Semi-Supervised Learning (proposal)	**68.42**	36.97	48.00	73.43	53.03	61.58	54.57	15.65
Unified Model (proposal)	61.68	48.82	**54.50**	**74.13**	**53.54**	**62.17**	**58.23**	28.70

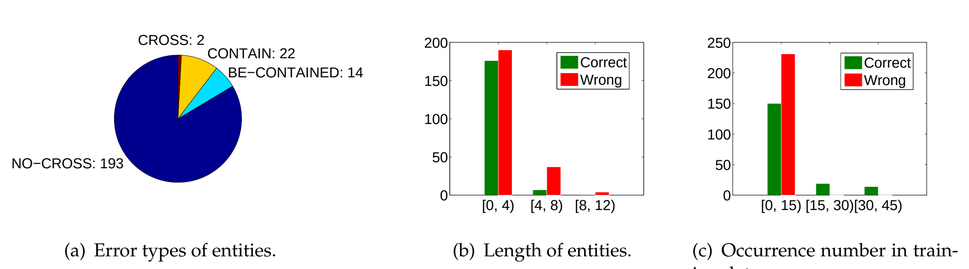
Table 1: NER results for named and nominal mentions on test data. We can see that our cross-domain and semi-supervised learning improve NER. Our unified model outperforms previous work.

Error Analysis

We design six metrics to do error analysis as follows:

- Sentence length.
- Entity length.
- Five error types: CONTAIN, BE-CONTAINED, SPLIT, CROSS, NO-CROSS.
- Occurrence number in training data.
- Unknown word rate of sentence.
- Unknown word rate of entity.

We show some important details of basic model analysis in Figure 2.



(a) Error types of entities. (b) Length of entities. (c) Occurrence number in training data.

Figure 2: Results of basic model. Green bars denote correct predictions and red bars denote wrong predictions. We can know that our main error type is CROSS and basic model is not good at long entities or entities with few occurrences in training data.

Conclusions

We propose a unified model for NER in Chinese social media. The model can learn from out-of-domain corpora and in-domain unannotated texts. In our experiments, our unified model outperforms previous work. Furthermore, our targeted and detailed error analysis not only helps us understand the advantage of our model but also points out the aspects we need to pay more attention to.