Distant Supervision for Relation Extraction Combined with Sentence-Level Attention and Pattern Filtering

Weilong Hu

Computer School of Wuhan University

16,March,2018

Relation extraction(RE):

extract triple $< h, r, t >$ from unstructured text.

Example:

'Donald Trump was born in New York.'

$< Donald\ Trump, Born - in, New\ York >$

Supervised models:

labour-intensive and even harder for open domains.

Distant supervision:

- assums that a sentence related to a entity pair may express the same relation which they express in KBs.
- given x sentences, y triples, $x \times y$ instances were generated.

Relation Instance	Knowledge Base	Labelled relation
S1: Obama was born in 1961 in Honolulu Hawaii.	<Obama, Born-in, Honolulu>	Born-in
S2: Obama made an unannounced return to Honolulu Monday	<Obama, Born-in, Honolulu>	Born-in

图: Examples of wrong label problem

To alleviate this problem:

- (Riedel et al., 2010; Hoffmann et al., 2011) proposed the *at-least-one* assumption and adopt *multi-instance* learning.
- (Surdeanu et al., 2012) introduce *multi-instance multi-label* learning.
- (Zeng et al., 2014; Lin et al., 2016; Zhou et al., 2016) adopted neural networks to overcome previous methods' weakness.

Previous methods:

- (Zeng et al., 2015) only utilize one sentence that is most likely to be valid.
- (Lin et al., 2016; Zhou et al., 2016) design attention strategy to automatically learn weights over multiple instances.

Our model:

- We develop an algorithm for clustering positive patterns for each relation.Compared with other methods, our algorithm obtains semantic patterns instead of lexcial.
- We combine sentence-level attention with pattern filtering, which is significant for selecting valid instances.
- Experimental results shows that our model combined sentence-level attention with patterns filtering makes full use of correct instances in relation extraction.

Prototype of "Founder-of" :

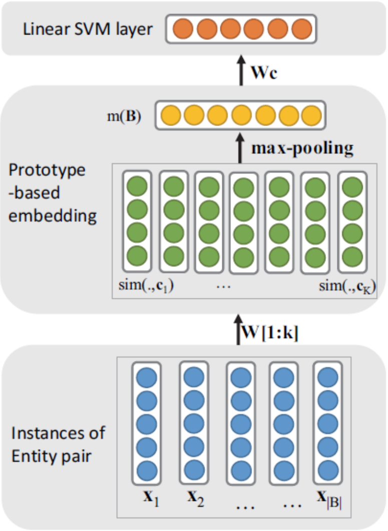
"X is the founder of Y", "X co-found Y" and "X launch Y in..."

Weighted Rejection Sampling Algorithm

Input:
- The wrongly classified instances $\mathbf{X} = \{\mathbf{x}_1, \ldots \mathbf{x}_m\}$
- The number of sampled prototypes K
- The similarity threshold σ

Output: The new prototypes $\mathbf{C}=\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$

For $\mathbf{x}_i$ in $\mathbf{X}$:
  Compute σ-NN($\mathbf{x}_i$)
End for
$\mathbf{C} \leftarrow \{\}$
While Size($\mathbf{C}$) < K:
  Sample $\mathbf{x}$ from $\mathbf{X}$ with probability $\propto \exp(\sigma\text{-NN}(\mathbf{x}))$
  If $\max_k \text{sim}(\mathbf{x}, \mathbf{c}_k) < \sigma$:
    Add $\mathbf{x}$ to $\mathbf{C}$
End while



Related work Neural Relation Extraction with Selective Attention over Instances

Average:

$$s = \sum_i \frac{1}{n} x_i$$

Selective Attention:

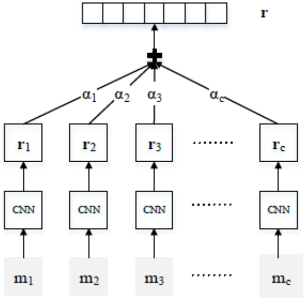
$$\alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)}$$

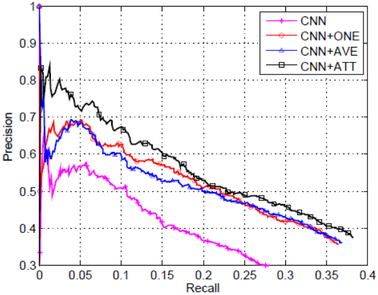
e_i scores how well x_i matches relation r:

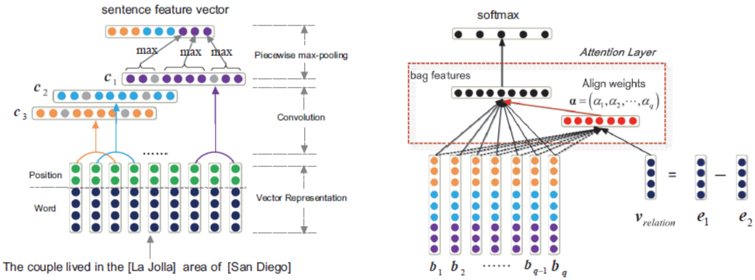
$$e_i = x_i \mathbf{A} \mathbf{r}$$

set vector s is:

$$s = \sum_i \alpha_i x_i$$







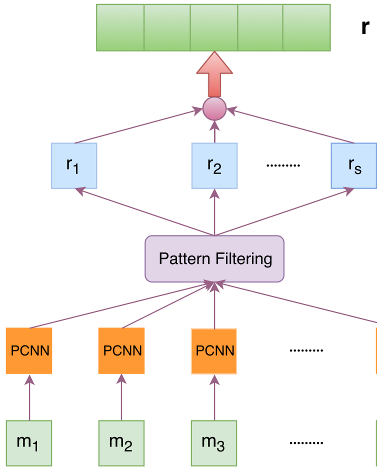
(a) PCNNs Module

(b) Sentence-level Attention Module

If an instance expresses the relation r, its feature vector should has higher similarity with $v_{relation}$, otherwise lower similarity.

$$v_{relation} = e_1 - e_2$$

$$\alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^{q} \exp(w_j)} \qquad w_i = \mathbf{W}_a^T(tanh[\mathbf{b}_i; v_{relation}]) + b_a$$



where m_i indicates the original sentences for an entity pair, r_j indicates filering sentence features of the original sentences, α_i is the weight given by sentence-level attention.

S_0	**Donald Trump** *was born in* **New York**.
S_1	**Obama** *was born in* **New York**.
S_2	**Donald Trump** *was born in* **Obama**.
S_3	**Donald Trump** *is working in* **New York**.

表: Observations of relation patterns.

Algorithm 1 Positive patterns sampling algorithm

Input: The relation instances set $\mathbf{O}$

Output: The positive pattern matrix $\mathbf{M}$

```
for instance o in O do
    Ecode o and store its pattern into P
end for
for subset P_r in P do
    Conduct cluster on P_r
    Store top-K_1 groups into M_r
end for
return M
```

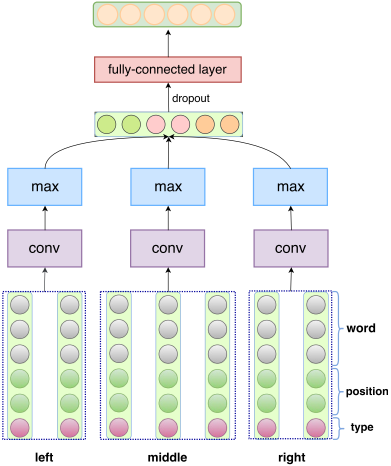


图: Piece-wise CNN for instance pattern encoder

given a sentence $x = \{w_1, w_2, \cdots, w_L\}$ where w_i is a word and L is the length of x. Instance pattern encoder encode x into a vector p.

Pattern Filtering:

whether a instance pattern is valid depends on its similarity to positive patterns.

$$sim(s) = \sum_{i=1}^{R} \sum_{j=1}^{K_1} d(s, \mathbf{M}_{ij})$$

For training:

$$sim(s) = \sum_{j=1}^{K_1} d(s, \mathbf{M}_{rj})$$

Sentence-level Attention:

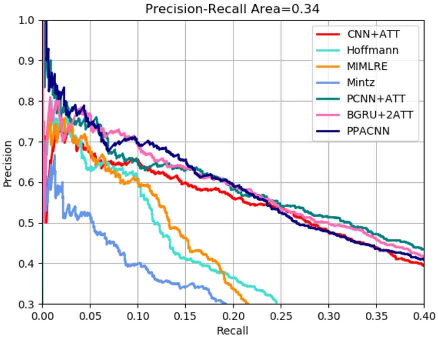
$$\alpha_i = \frac{exp(sim(x_i))}{\sum_k exp(sim(x_k))}$$

bag feature of $B = \{x_1, x_2, \cdots, x_m\}$ can be calculated as:

$$\mathbf{b} = \sum_{i=1}^{m} \alpha_i x_i$$

Test Settings	One				Two				All			
P@N(%)	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
CNN+AVE	75.2	67.2	53.8	60.9	70.3	62.7	55.8	62.9	67.3	64.7	58.1	63.4
CNN+ATT	76.2	65.2	60.8	67.4	76.2	65.7	62.1	68.0	76.2	68.6	59.8	68.2
PCNN+AVE	71.3	63.7	57.8	64.3	73.3	65.2	62.1	66.9	73.3	66.7	62.8	67.6
PCNN+ATT	73.3	69.2	60.8	67.8	77.2	71.6	66.1	71.6	76.2	73.1	67.4	72.2
BGRU+2ATT	73.0	64.0	59.3	65.4	77.0	70.0	64.0	70.3	**81.0**	73.5	68.0	74.2
PPACNN	**83.0**	**75.0**	**68.7**	**75.6**	**82.0**	**76.5**	**70.3**	**76.3**	80.0	**76.0**	**70.3**	**75.4**

Table 3: P@N results for relation extraction in entity pairs with different number of sentences



- Conclusion:

  We introduce a novel model combined setencelevel atterntion with pattern filtering for relation extraction under distant supervision. The pattern filtering can select more valid instances in a bag by matching with positive patterns. The sentencelevel attention can make full use of all selected instances by assigning different weights for different instance according to its similarity. We conduct experiments on a widely used dataset and our model achives competitive performance.

- Feature Works:

  In the future, we will explore how to extract patterns for relation from outer text.

Our model:

- We use coarse-grained entity type to assist in relation classification.
- We combine sentence-level attention with pattern filtering, which is significant for selecting valid instances.
- Experimental results shows that our model combined sentence-level attention with patterns filtering makes full use of correct instances in relation extraction.

Test Settings	One				Two				All			
P@N(%)	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
CNN+AVE	75.2	67.2	53.8	60.9	70.3	62.7	55.8	62.9	67.3	64.7	58.1	63.4
CNN+ATT	76.2	65.2	60.8	67.4	76.2	65.7	62.1	68.0	76.2	68.6	59.8	68.2
PCNN+AVE	71.3	63.7	57.8	64.3	73.3	65.2	62.1	66.9	73.3	66.7	62.8	67.6
PCNN+ATT	73.3	69.2	60.8	67.8	77.2	71.6	66.1	71.6	76.2	73.1	67.4	72.2
BGRU+2ATT	73.0	64.0	59.3	65.4	77.0	70.0	64.0	70.3	**81.0**	73.5	68.0	74.2
PPACNN	79.0	74.5	66.0	73.2	81.0	75.0	69.3	75.1	81.0	73.5	69.0	74.5
PPACNN+T	**83.0**	**75.0**	**68.7**	**75.6**	**82.0**	**76.5**	**70.3**	**76.3**	80.0	**76.0**	**70.3**	**75.4**

Table 3: P@N results for relation extraction in entity pairs with different number of sentences