Noise Reduction for Distant Supervision in Relation Extraction

胡伟龙

武汉大学计算机学院

2017 年 12 月 25 日

内容目录

Distant Supervision

- Distant Supervision:

  heuristically aligns entities in texts to a give knowledge base.

- Wrong label problem:

  A sentence that mentions two entities may not express the relation which links them in a KB.

Freebase /location/location/contains (Nevada, Las Vegas)

S1. **[Nevada] then sanctioned the sport , and the U.F.C. held its first show in [Las Vegas] in September 2001.**

S2. Pinnacle owns casinos in **[**Nevada**]**, Louisiana , Indiana , Argentina and the Bahamas , but not in the top two American casino cities , Atlantic City and **[**Las Vegas**]**.

S3. **He has retained two of [Nevada] 's most prominent criminal defense lawyers , Scott Freeman of Reno and David Chesnoff of [Las Vegas].**

S4. The state 's population is growing , but not skyrocketing the way it is in Arizona and **[**Nevada**]** , and with no city larger than 100,000 residents , Montana essentially does not have suburbs or exurbs like those spreading around Phoenix, **[**Las Vegas**]** and Denver.

Descriptions

[Nevada**]**: Nevada is a state in the Western, Mountain West, and Southwestern regions of the United States.

[Las Vegas**]**: officially the City of Las Vegas and often known as simply Vegas, is a city in the United States, the most populous city in the state of Nevada, the county seat of Clark County, and the city proper of the Las Vegas Valley.

Related Work

- Mintz et al.(2009):

  Ignored the problem.Single-instance,Single-label.

- Riedel,Yao,and McCallum,(2010):

  At-least-one assumption.Multi-instance,Single-label.

- Hoffmann et al.(2011) and Surdeanu et al.,(2012):

  Multi-instance Multi-label Learning.

- Zeng et al.,(2015):

  Combined MIL and piecewise convolutional neural networks(PCNNs).

内容目录

Distant Supervision via Prototype-Based Global Representation Learning, AAAI-2017

Prototype of "Founder-of" :

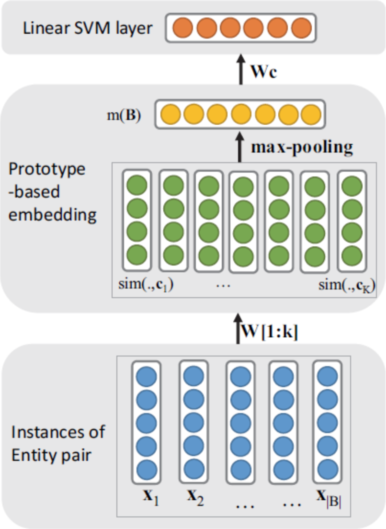
"X is the founder of Y", "X co-found Y" and "X launch Y in..."

Weighted Rejection Sampling Algorithm

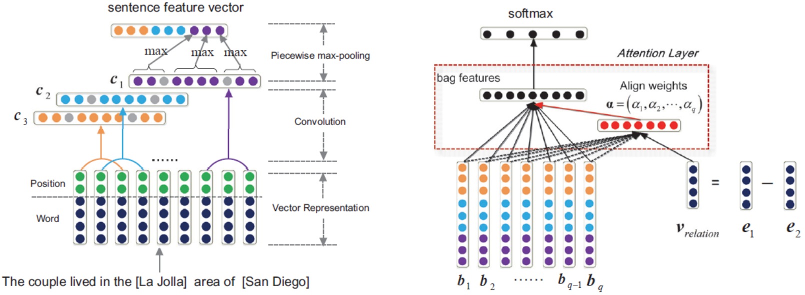
Input:
- The wrongly classified instances $\mathbf{X} = \{\mathbf{x}_1, \ldots \mathbf{x}_m\}$
- The number of sampled prototypes K
- The similarity threshold σ

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

```
For x_i in X:
  Compute σ-NN(x_i)
End for
C ← {}
While Size(C) < K:
  Sample x from X with probability ∝ exp(σ-NN(x))
  If max_k sim(x, c_k) < σ:
    Add x to C
End while
```



Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Description, AAAI-2017



(a) PCNNs Module

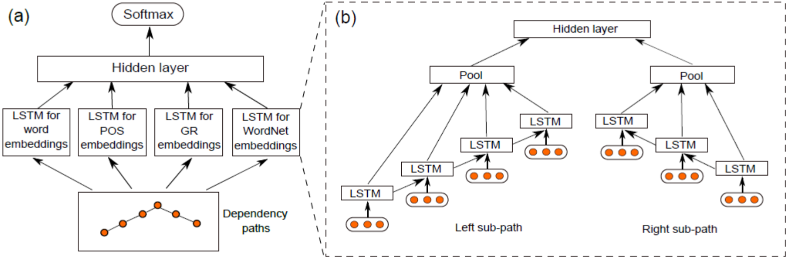
(b) Sentence-level Attention Module

Training Objective:

$$\min \mathcal{L}_A = \sum_{i=1}^{N} logp(r_i|B_i, \theta) \qquad \mathcal{L}_e = \sum_{i=1}^{|\mathcal{D}|} ||e_i - d_i||_2^2$$

$$\min \mathcal{L} = \mathcal{L}_A + \lambda \mathcal{L}_e$$

Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths, ACL-2017



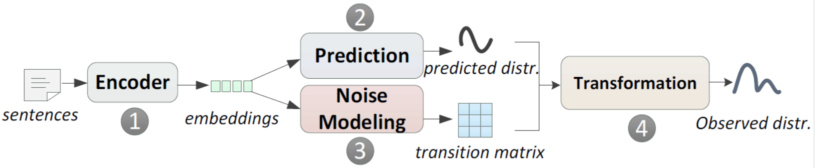
Inspired by the following observations:

- Shortest dependency paths are informative.
- Direction matters.
- Linguistic information helps.

Traning Objective:

$$\mathcal{J} = -\sum_{i=1}^{n_c} t_i \log y_i + \lambda\left(\sum_{i=1}^{\omega} ||W_i||_F^2 + \sum_{i=1}^{\nu} ||U_i||_F^2\right)$$

Learning with Noise:Enhance Distantly Supervised Relation Extraction with Dynamic Transition Matrix, ACL-2017



Transition matrix T for each sentence:

$$T_{ij} = \frac{exp(w_{ij}^T x_n + b)}{\sum_{j=1}^{|\mathbb{C}|} exp(w_{ij}^T x_n + b)}$$

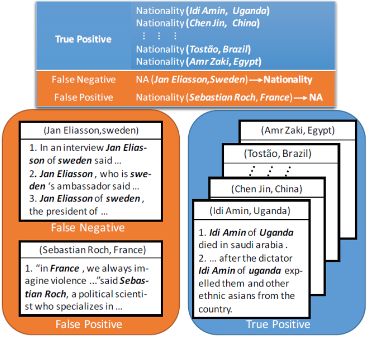
Observed Distribution:

$$\mathbf{o} = \mathbf{T}^T \cdot \mathbf{p}$$

Loss function:

$$L = -\sum_{i=1}^{N}((1-\alpha)\log(o_{iy_i}) + \alpha \log(p_{iy_i})) - \beta trace(\mathbf{T}^i)$$

A Soft-label Method for Noise-tolerant Distantly Supervised Relation Extraction, EMNLP-2017



Soft label r_i for entity pair $< h_i, t_i >$:

$$r_i = \arg\max(o + max(o) A \odot L_i)$$

o_t is calculated as follow:

$$o_t = \frac{exp(Ms_t + b)}{\sum_k exp(Ms_k + b)}$$

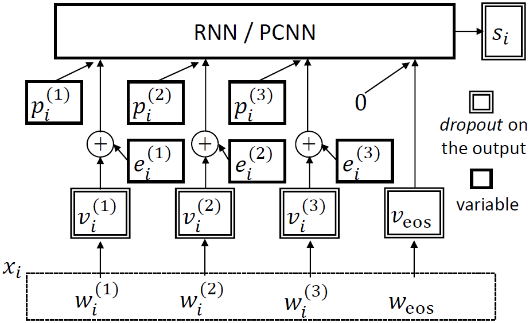
Loss function while training:

$$J(\theta) = \sum_{i=1}^{n} \log p(r_i | s_i; \theta)$$

loss function in testing stage:

$$G(\theta) = \sum_{i=1}^{n} \log p(l_i | s_i; \theta)$$

Adversarial Training for Relation Extraction, EMNLP-2017



Loss Function:

$$L(X;\theta) = -\sum_{i=1}^{K} \log P(r_i|X;\theta)$$

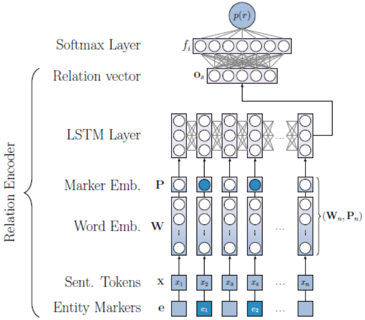
Adversarial Training:

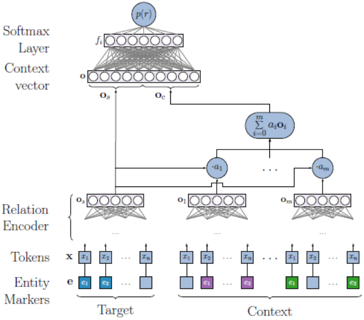
$$L_{adv}(X;\theta) = L(X + e_{adv};\theta) \quad \text{where} \quad e_{adv} = \arg\max_{||e|| \leq \varepsilon} L(X + e;\hat{\theta})$$

Approximately:

$$e_{adv} = \varepsilon g/||g||, \quad where \quad g = \bigtriangledown_V L(X;\hat{\theta})$$

Context-Aware Representations for Knowledge Base Relation Extraction, EMNLP-2017





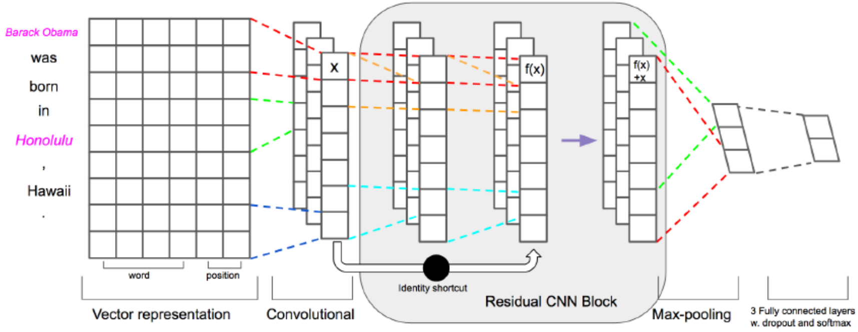
LSTM baseline:

$$p(r| < e_1, e_2 >, \mathbf{x}; \theta) = \frac{exp(f_r)}{\sum_{i=1}^{n_r} exp(f_i)} \quad f_i = \mathbf{y}_i \cdot \mathbf{o}_s + b_i$$

ContextAtt:

$$\mathbf{o}_c = \sum_{i=0}^{m} a_i \mathbf{o}_i \quad a_i = \frac{exp(o_i A o_s)}{\sum_{j=0}^{m} exp(o_j A o_s)}$$

Deep Residual Learning for Weakly-Supervised Relation Extraction, EMNLP-2017



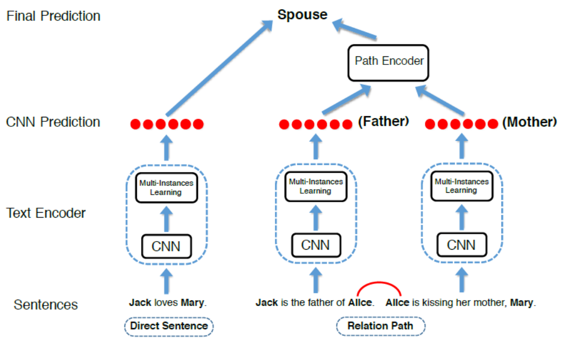
Motivation:

Previous neural relation extraction models are relatively shallow CNNs.

Residual learning:

Tackle the vanishing gradient problem in deep networks.

Incorporating Relation Paths in Neural Relation Extraction, EMNLP-2017



Global score function:

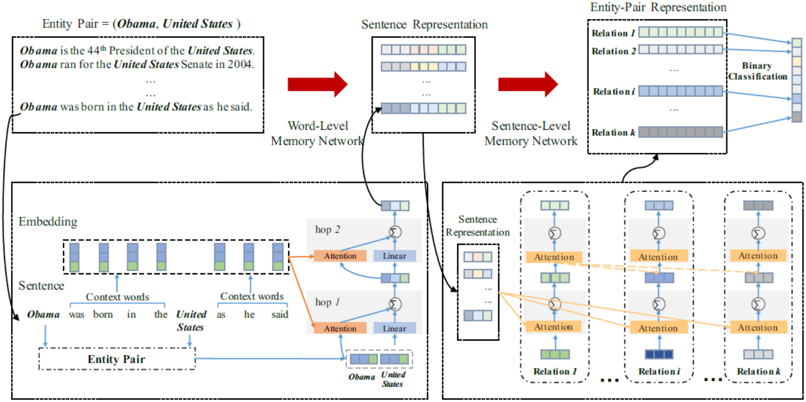
$$E(h, r, t|s) = \max_i p(r|\theta, s_i) \qquad G(h, r, t|p_i) = E(h, r_A, e)E(e, r_B, t)p(r|r_A, r_B)$$

$$G(h, r, t|P) = \max_i G(h, r, t|p_i) \qquad L(h, r, t) = E(h, r, t|S) + \alpha G(h, r, t|P)$$

Traning Objective:

$$J(\theta) = \sum_{(h,r,t)} \log(L(h, r, t))$$

Effective Deep Memory Networks for Distant Supervised Relation Extraction, IJCAI-2017



Memorry Network:

$$Network = < m, I, G, O, R >$$

Two ovservations:

- Not all context words contribute equally to the inference of relation.
- There exists dependencies between different relations.

内容目录

Observation

Example: Obama was born in Honolulu, in 1961.

- Trump was born in Honolulu, in 1961.
- Obama was born in Honolulu, in 1961. —-Pattern
- Beijing was born in Honolulu, in 1961.
- Trump_LOC was born in Honolulu, in 1961. ——Entity type

Heuristically

X_PER was born in Y_LOC, in 1961.

Observation

Example: Obama was born in Honolulu, in 1961.

- Trump was born in Honolulu, in 1961.
- Obama was born in Honolulu, in 1961. —-Pattern
- Beijing was born in Honolulu, in 1961.
- Trump_LOC was born in Honolulu, in 1961. ——Entity type

Heuristically

X_PER was born in Y_LOC, in 1961.

Observation

Example: Obama was born in Honolulu, in 1961.

- Trump was born in Honolulu, in 1961.
- Obama was born in Honolulu, in 1961. —-Pattern
- Beijing was born in Honolulu, in 1961.
- Trump_LOC was born in Honolulu, in 1961. ——Entity type

Heuristically

X_PER was born in Y_LOC, in 1961.

Observation

Example: Obama was born in Honolulu, in 1961.

- Trump was born in Honolulu, in 1961.
- Obama was born in Honolulu, in 1961. —-Pattern
- Beijing was born in Honolulu, in 1961.
- Trump_LOC was born in Honolulu, in 1961. ——Entity type

Heuristically

X_PER was born in Y_LOC, in 1961.

Observation

Example: Obama was born in Honolulu, in 1961.

- Trump was born in Honolulu, in 1961.
- Obama was born in Honolulu, in 1961. —-Pattern
- Beijing was born in Honolulu, in 1961.
- Trump_LOC was born in Honolulu, in 1961. ——Entity type

Heuristically

X_PER was born in Y_LOC, in 1961.

Observation

Example: Obama was born in Honolulu, in 1961.

- Trump was born in Honolulu, in 1961.
- Obama was born in Honolulu, in 1961. —-Pattern
- Beijing was born in Honolulu, in 1961.
- Trump_LOC was born in Honolulu, in 1961. ——Entity type

Heuristically

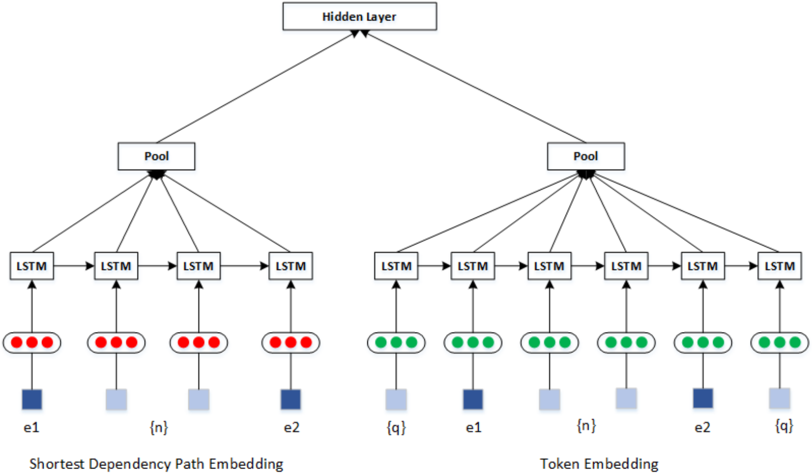
X_PER was born in Y_LOC, in 1961.

Mathmatical motivation

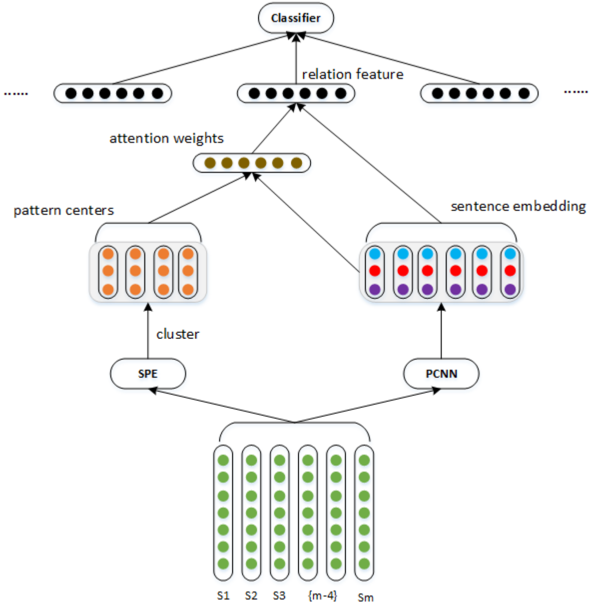
- Learning relation representation to build relation space.
- A sentence may express multiple relations.
- Instances in relation space may be more sparse to classify.

Sentence Pattern Finder

Sentence Pattern Encoder(SPE)



Architecture



Model

Defination:

- $T_{ij}(i \leq r, j \leq k)$: Pattern j related to relatin i.
- $W_{ij}(i \leq r, j \leq k)$: weight of pattern j related to relation i.
- $S_{ij}(i \leq r)$: Sentence j of Relation i.
- $D(S_{ij}, T_{it})(i \leq r, t \leq k)$: Distance between sentence S_{ij} and pattern T_{it}.
- α_{ij}: Attentin weight of sentence S_{ij}.

$$\alpha_{ij} = \frac{exp\{\sum_{t=1}^{k} D(S_{ij}, T_{it}) \cdot W_{it} + b_i\}}{\sum_{j=1}^{m_i} exp\{\sum_{t=1}^{k} D(S_{ij}, T_{it}) \cdot W_{it} + b_i\}}$$

Traning Objective:

$$L = -\sum_{i=1}^{N} \log p(r_i | s_i, \theta)$$

New Relation Class

- The meaning of small Wij:

  Pattern Tij may not express relation i

- Clustering of $\sum Tij$
  - Isolated points may be new relations.
  - $O(N)->O(r \cdot k)$

Thanks for Listening!

huweilong@whu.edu.cn