Transductive Non-linear Learning for Chinese Hypernym Prediction

Chengyu Wang[1], Junchi Yan[1,2], Aoying Zhou[1], Xiaofeng He[1]*

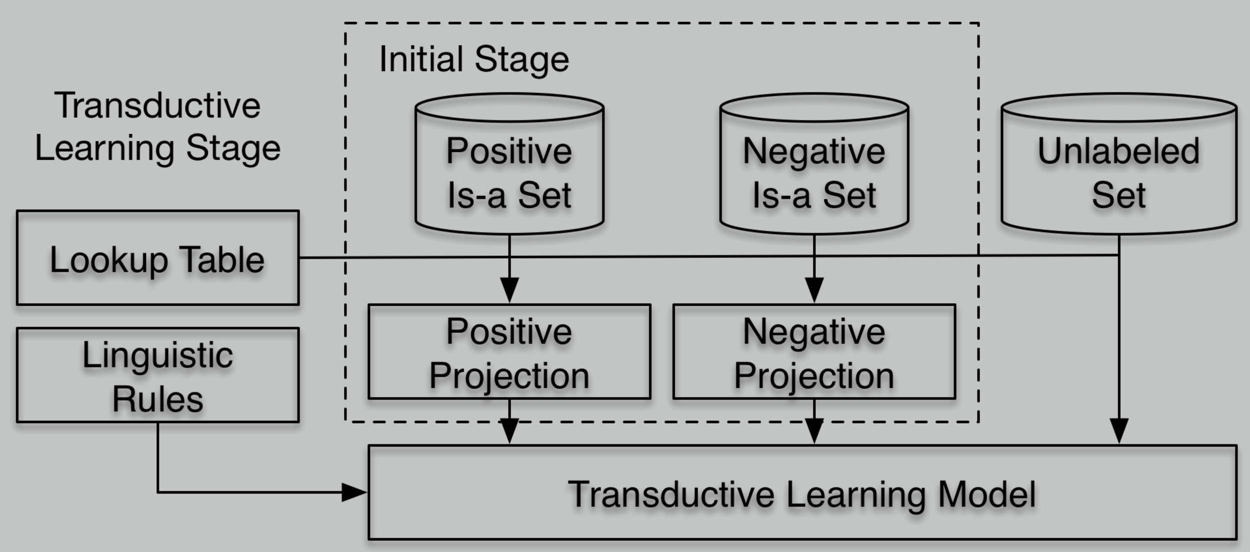
[1] Shanghai Key Laboratory of Trustworthy Computing, East China Normal University
[2] IBM Research — China

Introduction

- Learning hypernymy relations is essential for taxonomy construction, fine-grained entity categorization, knowledge base population, etc.
- Extracting hypernyms for entities is still challenging for Chinese due to flexible language expressions.
- Our work maps Chinese hyponyms to hypernyms in the embedding space by transductive non-linear learning.

General Framework

- Initial Stage
  - Train two linear projection models to capture the semantics of is-a and not-is-a relations based on the training set (i.e., D^+ and D^-).
  - Estimate the prediction and confidence scores for unlabeled data D^U.
- Transductive Learning Stage
  - Learn the final prediction score for each pair $(x_i, y_i) \in D^U$ based on initial prediction, linguistic rules and non-linear regularization.



Initial Model Training

- Train skip-gram models to map each word or concept with multiple words x_i to its embedding vector $\mathbf{x}_i$.
- Train two linear projection models with Tikhonov regularizers based on word embeddings. One for is-a relations, the other for not-is-a relations.

$$J(\mathbf{M}^+) = \frac{1}{2} \sum_{(x_i,y_i)\in D^+} \|\mathbf{M}^+\mathbf{x}_i - \mathbf{y}_i\|_2^2 + \frac{\lambda}{2}\|\mathbf{M}^+\|_F^2$$

$$J(\mathbf{M}^-) = \frac{1}{2} \sum_{(x_i,y_i)\in D^-} \|\mathbf{M}^-\mathbf{x}_i - \mathbf{y}_i\|_2^2 + \frac{\lambda}{2}\|\mathbf{M}^-\|_F^2$$

- Estimate the prediction score $score(x_i, y_i)$ and the confidence score $conf(x_i, y_i)$ for each $(x_i, y_i) \in D^U$.

$$score(x_i, y_i) = \tanh(\|\mathbf{M}^-\mathbf{x}_i - \mathbf{y}_i\|_2 - \|\mathbf{M}^+\mathbf{x}_i - \mathbf{y}_i\|_2)$$

$$conf(x_i, y_i) = \frac{|\|\mathbf{M}^+\mathbf{x}_i - \mathbf{y}_i\|_2 - \|\mathbf{M}^-\mathbf{x}_i - \mathbf{y}_i\|_2|}{\max\{\|\mathbf{M}^+\mathbf{x}_i - \mathbf{y}_i\|_2, \|\mathbf{M}^-\mathbf{x}_i - \mathbf{y}_i\|_2\}}$$

High prediction score: large probability of is-a relation between x_i and y_i. High confidence score: large probability that the models can predict the existence of is-a relations correctly.

Transductive Non-linear Learning (I)

- Initialize the m-dimensional final prediction vector $\mathbf{F}$ where $m = |D^+| + |D^-| + |D^U|$.

$$F_i = \begin{cases} 1 & (x_i, y_i) \in D^+ \\ -1 & (x_i, y_i) \in D^- \\ u_i & (x_i, y_i) \in D^U, u_i \sim Uniform(-1, 1) \end{cases}$$

- Define the objective considering results of initial prediction: $O_s = \|\mathbf{W}(\mathbf{F} - \mathbf{S})\|_2^2$.
  - $\mathbf{S}$ is the initial prediction vector. $\mathbf{W}$ is set as follows:

$$W_{i,j} = \begin{cases} conf(x_i, y_i) & i = j, (x_i, y_i) \in D^U \\ 1 & i = j, (x_i, y_i) \in D^+ \cup D^- \\ 0 & \text{Otherwise} \end{cases}$$

Transductive Non-linear Learning (II)

- Define the objective considering linguistic rules: $O_r = \|\mathbf{F} - \mathbf{R}\|_2^2$.
  - Compute the TP/TN rate γ_i for each positive/negative rule.
  - If (x_i, y_i) matches a collection of positive rules $C_{(x_i,y_i)}$, define R_i as: $R_i = \max\{F_i, \max_{c\in C_{(x_i,y_i)}} \gamma\}$.
  - If (x_i, y_i) matches a collection of negative rules $C_{(x_i,y_i)}$, define R_i as: $R_i = -\max\{-F_i, \max_{c\in C_{(x_i,y_i)}} \gamma\}$.
- Define the non-linear regularizer based on the *TransLP* framework: $O_n = \mathbf{F}^T\mathbf{\Sigma}^{-1}\mathbf{F}$.

$$\mathbf{\Sigma}(i, j) = \begin{cases} \cos(\mathbf{x}_i, \mathbf{x}_j) & y_i = y_j \\ 0 & \text{Otherwise} \end{cases}$$

  It assumes F_i and F_j w.r.t. (x_i, y_i) and (x_j, y_j) is similar if the candidate hypernyms y_i and y_j are the same and the candidate hyponyms x_i and x_j are similar in semantics.
- Optimize the combined objective function via blockwise gradient descent.

$$J(\mathbf{F}) = O_s + O_r + \frac{\mu_1}{2}O_n + \frac{\mu_2}{2}\|\mathbf{F}\|_2^2$$

- Predict y_i is a hypernym of x_i if $F_i > \theta$ ($\theta \in (-1, 1)$).

Experiments

- Datasets: Two public Chinese datasets (i.e., FD and BK), consisting of Chinese entity pairs with labeled positive/negative is-a relations.
- Metrics: Precision, Recall and F-Measure.
- Results: Our approach outperforms all baselines for Chinese.

Dataset	FD			BK		
Method	P	R	F	P	R	F
Taxonomy Matching	54.3	38.4	45.0	61.2	47.5	53.5
Linear Projection	64.1	56.0	59.8	71.4	64.8	67.9
Piecewise Linear Projection	66.4	59.3	62.6	72.7	67.5	70.0
Iterative Linear Projection	69.3	64.5	66.9	73.9	69.8	71.8
Vector Concatenation Model	67.7	**75.2**	69.7	80.3	75.9	78.0
Vector Addition Model	65.3	60.7	62.9	72.7	65.6	68.9
Vector Subtraction Model	71.9	60.6	65.7	78.4	60.7	68.4
Ours (Initial)	70.7	69.2	69.9	81.7	78.5	80.0
Ours	**72.8**	70.5	**71.6**	**83.6**	**80.6**	**82.1**

  - Examples of model prediction.

Candidate Hypernym	P	T	Candidate Hypernym	P	T
Entity: 乙烯(Ethylene)			**Entity**: 孙燕姿(Stefanie Sun)		
化学品(Chemical)	√	√	歌手(Singer)	√	√
有机化学(Organic Chemistry)	×	×	明星(Star)	√	√
有机物(Organics)	√	√	人物(Person)	√	√
气体(Gas)	√	√	金曲奖 (Golden Melody Award)	√	×
自然科学(Natural Science)	×	×	音乐人(Musician)	√	√

- Supplementary experiments: Our approach is comparable to many existing methods in the English environment. (please refer to the paper for details).

Conclusion and Future Work

- We propose a transductive non-linear learning approach for Chinese hypernym prediction. It has high accuracy and does not require parsing Chinese texts and training deep classification models.
- Future work: constructing a complete taxonomy from texts in Chinese.

References

[1] Fu et al. Learning semantic hierarchies via word embeddings. *ACL* 2014. pages 1199–1209.

[2] Liu and Yang. Bipartite edge prediction via transductive learning over product graphs. *ICML* 2015. pages 1880–1888.

[3] Wang and He. Chinese hypernym-hyponym extraction from user generated categories. *COLING* 2016. pages 1350–1361.

[4] Mirza and Tonelli. On the contribution of word embeddings to temporal relation classification. *COLING* 2016. pages 2818–2828.