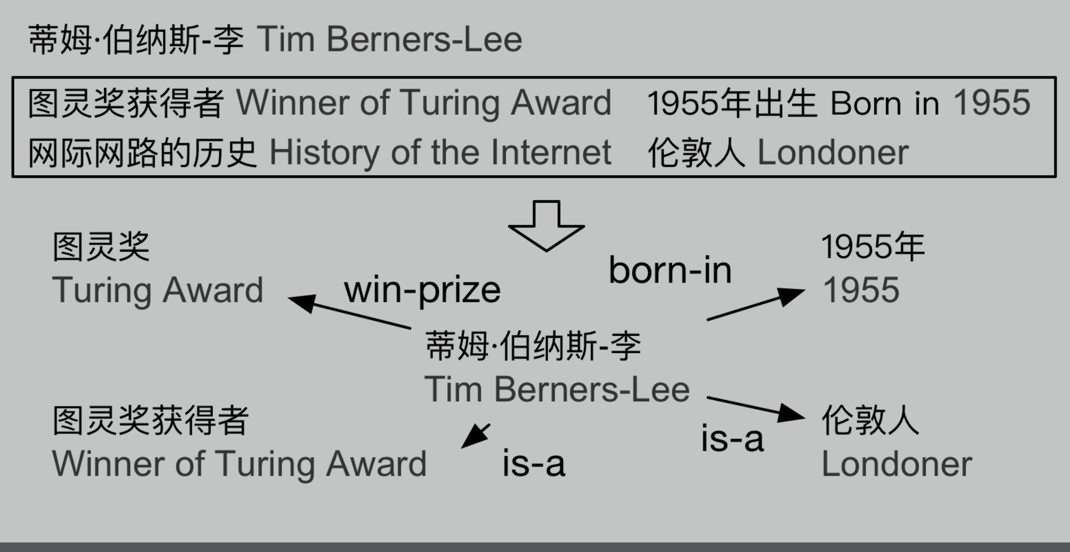
Learning Fine-grained Relations from Chinese User Generated Categories

Chengyu Wang, Yan Fan, Xiaofeng He*, Aoying Zhou
School of Computer Science and Software Engineering, East China Normal University

Introduction

- User generated categories (UGCs) express rich semantic relations implicitly.
- While most methods use pattern matching for English, learning relations from Chinese UGCs poses challenges due to the flexibile expressions.
- Our work uses weakly supervised methods to extract relations from Chinese UGCs based on projection learning and graph mining.



Mining Is-a Relations

Initial model training

- Use existing labeled sets and heuristic rules to generate training data automatically (i.e., is-a and not-is-a relation pairs).
- Train a skip-gram model to map each word 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.

$$J(\mathbf{M}^+, \mathbf{B}^+) = \frac{1}{2} \sum_{(e,c_h)\in D^+} \|\mathbf{M}^+\mathbf{e} + \mathbf{B}^+ - \mathbf{c}_h\|_2^2 + \frac{\lambda}{2}\|\mathbf{M}^+\|_F^2 + \frac{\lambda}{2}\|\mathbf{B}^+\|_F^2$$

$$J(\mathbf{M}^-, \mathbf{B}^-) = \frac{1}{2} \sum_{(e,c_h)\in D^-} \|\mathbf{M}^-\mathbf{e} + \mathbf{B}^- - \mathbf{c}_h\|_2^2 + \frac{\lambda}{2}\|\mathbf{M}^-\|_F^2 + \frac{\lambda}{2}\|\mathbf{B}^-\|_F^2$$

where e is a Wikipedia concept and c_h is the head word of a UGC of entity e in its corresponding Wikipedia page.

- Estimate the prediction score $s(e, c)$ for each unlabeled (e, c) pair.

$$s(e, c) = \tanh(\|\mathbf{M}^+\mathbf{e} + \mathbf{B}^+ - \mathbf{c}_h\|_2 - \|\mathbf{M}^-\mathbf{e} + \mathbf{B}^- - \mathbf{c}_h\|_2)$$

High prediction score means there is a large probability of is-a relation between e and c.

Score refinement by collective inference

- Denote $\tilde{g}(h)$ as the un-normalized global prediction score for head word h of UGCs:

$$\tilde{g}(h) = \ln(1 + |D_h| + |D_h^+|)\frac{|D_h^+| + \sum_{(e,c)\in D_h} s(e, c)}{|D_h| + |D_h^+|}$$

where H is the collection of head words of UGCs.

- Re-normalize the prediction score $s(e, c)$ based on the initial prediction score and global prediction score.

$$f(e, c) = \beta s(e, c) + (1 - \beta)g(h)$$

where $\beta \in (0, 1)$ is the tuning parameter and $g(h)$ is the normalized version of $\tilde{g}(h)$:

$$g(h) = \frac{\tilde{g}(h)}{\max_{h'\in H} |\tilde{g}(h')|}$$

- Expand the number of hypernyms by the following heuristic rule: Finally, we regard c_h as a valid hypernym of e if c is predicted as a hypernym of e and c_h is also a Wikipedia concept.

Mining Non-taxonomic Relations (I)

Single-pass category pattern mining

- Extract category patterns by replacing entity placeholders with specific entity names in UGCs. For example, the pattern is "[E]获得者'(Winner of [E])" for "图灵奖获得者(Winner of Turing Award)". The pair "(蒂姆·伯纳斯-李, 图灵奖)(Tim Berners-Lee, Turing Award)" can be extracted as a candidate relation instance.
- Calculate the pattern support score $supp(p)$ of pattern p and filter out low-support patterns by:

$$supp(p) = |R_p| \cdot \ln(1 + L_p)$$

where R_p is the collection of extracted pairs for pattern p and L_p is the pattern length.

Mining Non-taxonomic Relations (II)

Graph-based raw relation extractor

- For each pattern p, construct a graph G where nodes are extracted candidate relation pairs based on p and weighted edges are the semantic similarities between the pairs.
- Detect a Maximum Edge Weight Clique (MEWC) C^* in G and treat pairs in C^* as seed relation instances that p may represent. We propose a Monte Carlo based method to extract the MEWC from the graph approximately. Please refer to the paper for details.
- Extract relation instances for the underline relation that p may present by finding pairs that are similar enough to the seed relation instances.

Relation mapping

- Map extracted pairs to relation triples by defining the relation predicates through i) direct verbal mapping, ii) direct non-verbal mapping and iii) indirect mapping.

Experiments

Experiments on is-a relation extraction

- Dataset: 1,788 labeled entity-UGC pairs extracted from Chinese Wikipedia.
- Metrics: Precision, Recall and F-Measure.
- Results: Our approach outperforms all competitive baselines.

Method	Precision (%)	Recall (%)	F-Measure (%)
Concat Model	79.5	64.2	67.2
Sum Model	80.9	70.1	72.6
Diff Model	78.3	69.0	71.5
Piecewise Projection	78.9	72.3	75.5
Our Method (w/o Exp)	89.2	88.1	88.7
Our Method	**89.8**	**88.3**	**89.0**

Experiments on non-taxonomic relation extraction

- Dataset: All entity-UGC pairs in Chinese Wikipedia.
- Metrics: Size (#extractions for a certain relation type), Accuracy and Coverage (whether the extracted relations are covered by a large existing Chinese KB).
- Results: Our approach can extract a large amount of novel relations with high accuracy.

Relation	Size	Accuracy (%)	Coverage (%)
毕业(graduated-from)	44,118	98.0	22.9
位于(located-in)	29,460	97.2	8.5
建立(established-in)	20,154	95.0	31.5
出生(born-in)	11,671	98.3	41.4
成员(member-of)	8,445	96.0	4.2
启用(open-in)	8,956	98.2	21.6

- Please refer to more supplementary experiments in the paper.

Conclusion and Future Work

- We propose a weakly supervised framework to extract relations from Chinese UGCs. It requires very little human intervention and has high accuracy for the Chinese language.
- Future work includes:
  - Improving our work for short text knowledge extraction;
  - Designing a general framework for cross-lingual UGC relation extraction.

Key References

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

[2] Wang et al. Transductive Non-linear Learning for Chinese Hypernym Prediction. *ACL* 2017. pages 1394–1404.

[3] Nastase and Strube. Decoding Wikipedia Categories for Knowledge Acquisition. *AAAI* 2008. pages 1219–1224.

[4] Pasca. German Typographers vs. German Grammar: Decomposition of Wikipedia Category Labels into Attribute-Value Pairs. *WSDM* 2017.

Contact info: chywang2013@gmail.com
School of Computer Science and Software Engineering, East China Normal University