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 flexible 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 $x_i$.
- Train two linear projection models with Tikhonov regularizers based on word embeddings. One for is-a relations. The other for not-is-a relations.

$$J(M^+, B^+) = \frac{1}{2} \sum_{(c,h) \in O} \|M^+e + B^+ - c_h\|^2 + \frac{\lambda}{2} \|M^+\|^2 + \frac{\lambda}{2} \|B^+\|^2$$

$$J(M^-, B^-) = \frac{1}{2} \sum_{(c,h) \in O} \|M^-e - B^- + c_h\|^2 + \frac{\lambda}{2} \|M^-\|^2 + \frac{\lambda}{2} \|B^-\|^2$$

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.

$$(e, c) \rightarrow s(e, c) = \text{tanh} (\|M^+e + B^- - c_h\| - \|M^-e - B^+ + c_h\|)$$

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

Score refinement by collective inference

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

$$\hat{g}(h) = \ln(1 + |D_h| + |D^+_h|)^{\beta} + \sum_{e \in E_h} s(e, c)$$

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.

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

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

$$g(h) = \frac{\hat{g}(h)}{\max_{c \in H} \hat{g}(h)}$$

- 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 “(获得者, 2015) (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:

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

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) 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-Measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concat Model</td>
<td>79.5</td>
<td>74.2</td>
<td>78.3</td>
</tr>
<tr>
<td>Sum Model</td>
<td>76.9</td>
<td>70.1</td>
<td>73.6</td>
</tr>
<tr>
<td>Diff Model</td>
<td>78.3</td>
<td>69.9</td>
<td>71.5</td>
</tr>
<tr>
<td>Piecewise Projection</td>
<td>79.8</td>
<td>72.3</td>
<td>75.5</td>
</tr>
<tr>
<td>Our Method (w/o Exp)</td>
<td>80.3</td>
<td>86.1</td>
<td>86.7</td>
</tr>
<tr>
<td>Our Method</td>
<td>80.9</td>
<td>80.3</td>
<td>80.9</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Size</th>
<th>Accuracy (%)</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[E]2015 graduated-from</td>
<td>44.119</td>
<td>98.1</td>
<td>97.2</td>
</tr>
<tr>
<td>[E]2015 located-in</td>
<td>29.499</td>
<td>97.2</td>
<td>8.5</td>
</tr>
<tr>
<td>[E]2015 established-in</td>
<td>20.154</td>
<td>95.0</td>
<td>31.5</td>
</tr>
<tr>
<td>[E]2015 born-in</td>
<td>10.071</td>
<td>98.3</td>
<td>41.4</td>
</tr>
<tr>
<td>[E]2015 member-of</td>
<td>8.645</td>
<td>95.0</td>
<td>4.0</td>
</tr>
<tr>
<td>[E]2015 open-in</td>
<td>8.926</td>
<td>98.2</td>
<td>21.1</td>
</tr>
</tbody>
</table>

- 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


Contact info: chywang2013@gmail.com

School of Computer Science and Software Engineering, East China Normal University