Learning Fine-grained Relations from Chinese User Generated Categories

Chengyu Wang, Yan Fan, Xiaofeng He*, Aoying Zhou
Shanghai Key Laboratory of Trustworthy Computing,
School of Computer Science and Software Engineering, East China Normal University
{chywang2013,eileen940531}@gmail.com
{xfhe,ayzhou}@sei.ecnu.edu.cn

Abstract

User generated categories (UGCs) are short texts that reflect how people describe and organize entities, expressing rich semantic relations implicitly. While most methods on UGC relation extraction are based on pattern matching in English circumstances, learning relations from Chinese UGCs poses different challenges due to the flexibility of expressions. In this paper, we present a weakly supervised learning framework to harvest relations from Chinese UGCs. We identify is-a relations via word embedding based projection and inference, extract non-taxonomic relations and their category patterns by graph mining. We conduct experiments on Chinese Wikipedia and achieve high accuracy, outperforming state-of-the-art methods.

1 Introduction

UGCs are descriptive phrases related to entities, frequently appearing in online encyclopedias and vertical websites. These texts are concise and informative, reflecting the way people organize and characterize entities (Xu et al., 2016a).

UGCs (especially Wikipedia categories) are important sources for knowledge harvesting. Previous approaches (Flati et al., 2014; Ponzetto and Strube, 2007; Ponzetto and Navigli, 2009) focus on inferring is-a relations between entities and UGCs for taxonomy construction. A few others extract multiple types of relations from Wikipedia categories (Nastase and Strube, 2008; Suchanek et al., 2007). These methods are mostly designed for English language by employing language-specific patterns or linguistic rules.

For Chinese, harvesting semantic relations from texts poses different challenges. There is no distinction between singular and plural forms and no word spaces in Chinese. Word orders can be arranged in multiple ways with very flexible expressions. As illustrated in Qiu and Zhang (2014); Chen et al. (2014), the research of relation extraction from Chinese texts makes less significant process than the research for English. Although several approaches are proposed to construct Chinese taxonomies from Wikipedia categories (Li et al., 2015; Wang et al., 2014), extracting fine-grained and multi-typed relations from UGCs still needs further study. This is because there exist very few high-quality lexical patterns for relation identification in Chinese UGCs (in contrast to Nastase and Strube (2008); Suchanek et al. (2007)). Hence this problem is similar to "open relation extraction" (Etzioni et al., 2011) from Chinese short texts, without pre-defined relation types.

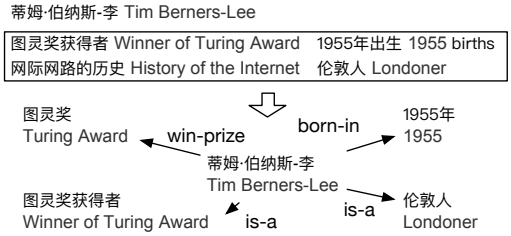


Figure 1: An illustrative example with respect to "Tim Berners-Lee" in Chinese Wikipedia.

In this paper, we propose a weakly supervised learning framework to mine fine-grained and multiple-typed relations from Chinese UGCs. A simple example is illustrated in Figure 1[1]. Inspired by Fu et al. (2014); Wang et al. (2017), is-a relations are extracted based on word embedding

*Corresponding author.

[1]The category "Winner of Turing Award" can serve as a class of "Tim Berners-Lee" (similar to Wu et al. (2012)) and be treated as a relational category (similar to Suchanek et al. (2007)). We regard both are valid and extract two relations.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2567–2577
Copenhagen, Denmark, September 7–11, 2017.

based projection models. We further refine prediction results by collective inference and hypernym expansion. For non-taxonomic relations, relation types and corresponding category patterns are identified jointly based on graph clique mining. Finally, these mined "raw" relations are mapped to canonicalized relation triples. In our work, except for a set of heuristic rules, the proposed approach is weakly supervised without manual labeling.

In the experiments, given only 0.6M entities and their respective 2.4M categories in Chinese Wikipedia, our method extracts 1.52M relations with an overall accuracy of 93.6%. The experiments also show that our approach outperforms previous methods for both is-a and non-taxonomic relation extraction from Chinese UGCs. The extracted relations and the labeled test set are publicly available[2].

The rest of this paper is as follows. Section 2 summarizes related work. Details of our approach are described in Section 3 to Section 5, with experiments in Section 6. Finally, we conclude our paper and discuss the future work in Section 7.

2 Related Work

In this section, we overview the related work on relation extraction from UGCs.

2.1 Is-a Relation Extraction

Is-a relations are backbones in taxonomies. In YAGO (Suchanek et al., 2007), a Wikipedia category is regarded as conceptual if it matches the pattern "pre-modifier + head word + post-modifier". WikiTaxonomy (Ponzetto and Strube, 2007) constructs a taxonomy from Wikipedia categories using multiple types of features. The taxonomy is reconstructed and improved in Ponzetto and Navigli (2009). Other similar projects use classifiers and rule based inference to predict is-a relations for taxonomy learning (Flati et al., 2014; Mahdisoltani et al., 2015; Nastase et al., 2010; Alfarone and Davis, 2015; Shwartz et al., 2016; Gupta et al., 2016). Since harvesting English is-a relations is not our focus, we do not elaborate here.

For Chinese, this task is more challenging because there are few category patterns that can be used to extract is-a relations from UGCs. Based on the word formation of Wikipedia categories, Li et al. (2015) propose a classification method to build a large Chinese taxonomy from Wikipedia. A similar approach is presented in Lu et al. (2015). Besides encyclopedias, Fu et al. (2013) generate candidate hypernyms and employ an SVM-based ranking model to detect the most likely hypernym of an entity. These methods have relatively high precision but require careful feature engineering and a large amount of human work.

Another thread of related work is cross-lingual approaches, which use larger English knowledge sources to supervise Chinese is-a relations extraction. For example, Wang et al. (2014) propose a dynamic adaptive boosting model to learn taxonomic prediction functions for English and Chinese. Xu et al. (2016b) link Chinese entities with DBpedia types based on cross-lingual links between Chinese and English entities. Other approaches can be found in Wu et al. (2016); Mahdisoltani et al. (2015). These methods take advantages of languages with richer resources but are constrained by cross-lingual links.

To capture linguistic regularities of is-a relations, deep learning approaches map the vectors of entities to the vectors of their hypernyms. Fu et al. (2014) design piecewise linear projection models to learn Chinese semantic hierarchies based on word embeddings (Mikolov et al., 2013). Wang and He (2016) improve this approach by adding an iterative update strategy and a pattern-based validation mechanism. Wang et al. (2017) design a transductive learning approach by considering the semantics of both is-a and not-is-a relations, linguistic rules and the unlabeled data jointly. In this work, we further propose a word embedding based model that consider the word formation of UGCs to improve the prediction results.

2.2 Non-taxonomic Relation Extraction

Unlike the case of is-a relations, the task of extracting non-taxonomic relations from UGCs has rarely been addressed. A possible cause is that harvesting relations from short texts is more challenging. The pioneer work Nastase and Strube (2008) extracts relations by lexical pattern matching and inference. Pasca (2017) studies how to decompose Wikipedia categories into attribute-value pairs. YAGO (Suchanek et al., 2007) uses regular expression based matching to harvest relations. While patterns in English are more regular, enumerating patterns for Chinese requires a large amount of human labor. In our work, we solve this problem by graph mining, which has high pre-

[2]https://chywang.github.io/data/emnlp17.zip

cision and requires minimal human intervention. Note that our work is also similar to open relation extraction (Etzioni et al., 2011) due to the unknown number of relation types. The difference is that our work focuses on UGCs which are very short phrases rather than sentences.

3 General Framework

In Wikipedia, each entity e is associated with a collection of UGCs $Cat(e)$. We first learn a prediction model $f(e, c)$ to distinguish is-a relations from not-is-a relations where $c \in Cat(e)$, and extract all is-a relations (Section 4). For example, we can obtain is-a relations "(Tim Berners-Lee, is-a, Londoner)" and "(Tim Berners-Lee, is-a, Winner of Turing Award)", as shown in Figure 1.

After that, we mine non-taxonomic relations from Wikipedia UGCs (Section 5). Our algorithm first makes a single pass over all categories to mine significant category patterns (Section 5.1). For example, the pattern "[E]获得者(Winner of [E])" is extracted, which frequently appears in UGCs and may refers to a type of relation where "[E]" is a placeholder for entities. Candidate relation instances for such patterns are obtained by a graph clique mining algorithm (Section 5.2). The instances extracted based on the previous pattern are "(Tim Berners-Lee, Turing Award)", "(Albert Einstein, Nobel Prize for Physics)", etc. Finally, the extracted "raw" instances are mapped to canonicalized triples (Section 5.3). In this step, a relation predicate "win-prize" is defined for the pattern and these pairs are mapped to "win-prize" relations.

4 Mining Is-a Relations

In this section, we introduce how to learn $f(e, c)$ and extract is-a relations from UGCs.

4.1 Training Data Generation

The training of $f(e, c)$ requires positive and negative entity-category pairs. To avoid the time-consuming labeling process, we generate the training set automatically. The first part is borrowed from Fu et al. (2014), containing 1,391 positive pairs and 4,294 negative pairs. However, the number of positive pairs is not sufficient for our propose. We design a heuristic rule to generate more positive pairs from Wikipedia categories. We treat a pair (e, c) as positive if the following two conditions hold:

- The category c matches the pattern "pre-modifier + 的+ head word" or the head words of e and c are the same[3].
- The head word of a category name is a noun and is *not* in a Chinese thematic lexicon extended from the dictionary used in Li et al. (2015), containing 184 thematic words (e.g., "军事(Military)", "娱乐(Entertainment)".

In total, we sample 5,000 pairs to add to our training set. The TP rate is 98.7%, estimated over 300 pairs, indicating the effectiveness of rules.

4.2 Projection-based Model Prediction

Except for the previous pattern, other Chinese is-a relations can not be directly extracted by lexical matching. Inspired by Wang et al. (2017), we employ projection models to learn the semantics of is-a and not-is-a relations.

A projection model is a linear model that maps the embedding vector of a word to the vector of another where the two words satisfy a particular relation (Fu et al., 2014). In Wikipedia, most category names are relatively long and fine-grained, making it difficult to learn the embeddings precisely. We find that given a pair (e, c), if the head word of category c is a valid hypernym of e, so it is for c itself, e.g., "英格兰计算机科学家(CS scientist in England)" for "Tim Berners-Lee". Denote $\vec{v}(e)$ as the embedding vector of entity e, with the dimensionality as n. Let c_h be the head word of c. For each pair in the positive training set $(e, c) \in D^+$, assume there is a positive projection model such that $\mathbf{M}^+\vec{v}(e) + \mathbf{B}^+ \approx \vec{v}(c_h)$ where $\mathbf{M}^+$ is an $n \times n$ projection matrix and $\mathbf{B}^+$ is an $n \times 1$ bias vector. Similarly, for pairs in negative training set $(e', c') \in D^-$, we learn a negative model $\mathbf{M}^-\vec{v}(e') + \mathbf{B}^- \approx \vec{v}(c'_h)$. Note that we do not impose explicit connections between two models because the semantics of Chinese is-a and not-is-a relations are very complicated and difficult to model (Fu et al., 2014; Wang and He, 2016). In our work, we let the algorithms to learn representations of is-a/not-is-a relations.

This approach learns is-a and not-is-a relation representations implicitly and does not require deep NLP analysis on UGCs, which is suitable to deal with the flexible expressions in Chinese. In

[3]The head word of a category name is the root word in the dependency parsing tree. "的" is an auxiliary word in Chinese, usually appearing between adjectives and nouns.

the training phase, we aim to minimize the objective function for positive projection learning:

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

where $\lambda > 0$ gives an additional Tikhonov smoothness effect on the projection matrices (Golub et al., 1999). For negative model, we have

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

After model training, for an unlabeled pair (e, c), if the category c is the correct hypernym of the entity e, the vector $\vec{v}(c_h)$ will be close to $\mathbf{M}^+\vec{v}(e)+\mathbf{B}^+$ and far away from $\mathbf{M}^-\vec{v}(e)+\mathbf{B}^-$. Denote $d^+(e, c)$ and $d^-(e, c)$ as:

$$d^+(e,c) = \|\mathbf{M}^+\vec{v}(e) + \mathbf{B}^+ - \vec{v}(c_h)\|_2$$

$$d^-(e,c) = \|\mathbf{M}^-\vec{v}(e) + \mathbf{B}^- - \vec{v}(c_h)\|_2$$

The prediction score is calculated as follows:

$$s(e,c) = \tanh(d^-(e,c) - d^+(e,c))$$

where $s(e, c) \in (-1, 1)$. High prediction score means a large probability of the existence of an is-a relation between e and c.

4.3 Collective Prediction Refinement

As indicated in Fu et al. (2013); Levy et al. (2015), some categories naturally serve as "prototypical hypernyms", regardless of the entities. To encode this assumption into our method, we refine the previous prediction results by collective inference.

Consider the category "伦敦人(Londoner)" in Figure 1, which can be literally translated as "伦敦(London)人(person)". "人(person)" is the "prototypical hypernym" here. Other categories whose head words are "人(person)" such as "哥本哈根人(Copenhagen person[4], people from Copenhagen)" are likely to be conceptual categories, too.

Denote H as the head word set of all Wikipedia UGCs. For each $h \in H$, let $D_h = \{(e, c)\}$ be the collection of unlabeled pairs (i.e., pairs not in the training set) where the head word of category c is h. D_h^+ is the collection of positive pairs with h as the head word of c in the training set (generated based on Section 4.1). We define the unnormalized global prediction score $\tilde{g}(h)$ for each $h \in H$:

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

In this formula, each unlabeled data instance $(e, c) \in D_h$ has the weight of $s(e, c)$ and each training data instance $(e, c) \in D_h^+$ has the weight of 1. $\frac{|D_h^+|+\sum_{(e,c)\in D_h} s(e,c)}{|D_h|+|D_h^+|}$ is the average prediction score for categories with the head word h. $\ln(1 + |D_h| + |D_h^*|)$ gives a larger impact to $\tilde{g}(h)$ when the head word h appears more frequently in Wikipedia categories. This heuristic setting is inspired by transductive learning which takes both training and unlabeled data into consideration (Chapelle et al., 2006). It is also similar to the prior probability feature (Fu et al., 2013).

We normalize the global prediction score $g(h)$ as follows:

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

The prediction function $f(e, c)$ for the entity e and the category c with the head word h is defined in a combination of $s(e, c)$ and $g(h)$:

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

where $\beta \in (0, 1)$ is a tuning parameter that controls the relative importance of the two scores.

We predict there is an is-a relation between entity e and category $c \in Cat(e)$ if at least one of the two conditions holds:

- (e, c) meets the two conditions in Section 4.1.
- $f(e, c) > \theta$ where θ is a threshold.

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. This step (called hypernym expansion) increases the number of hypernyms and hence the number of is-a relations.[5]

5 Mining Non-taxonomic Relations

In this section, we present our approach to extract non-taxonomic relations from Wikipedia UGCs.

[4] Literal translation.

[5] We do not extract all the entity-head word pairs (e, c_h) as is-a relations because word segmentation, tagging and parsing errors may occur when we extract head words by NLP tools. We observe that if c_h is also a Wikipedia concept, the head word extraction process is most probably correct.

5.1 Single-pass Category Pattern Miner

This module automatically learns important category patterns that appear frequently in Wikipedia and have a probability to represent certain semantic relations. Formally, a category pattern p is an ordered sequence of common words and entity tags. For example, the pattern of the category "图灵奖获得者(Winner of Turing Award)" is "[E]获得者(Winner of [E])". Define $R_p = \{(e_p, c_p)\}$ as the collection of entity pairs such that in Wikipedia page e_p, a category containing c_p matches the pattern p[6]. c_p is in the place of "[E]". Consider the previous example. In Wikipedia page "Tim Berners-Lee", there is a category "Winner of Turing Award" that matches the pattern "Winner of [E]". "Turing Award" is the "[E]" here. Thus we have e_p ="Tim Berners-Lee" and c_p ="Turing Award" as an entity pair in R_p. We can see that R_p is the collection of all candidate relation instances that may have the relation that p represents.

Let L_p be the number of common words in pattern p. We define the support of the pattern $supp(p)$ as follows:

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

where $\ln(1 + L_p)$ gives larger support values to longer patterns because longer patterns tend to be more specific and may contain richer semantics.

In the implementation, we employ a CRF-based Chinese NER tagger (Qiu et al., 2013) and a dictionary consisting of all Wikipedia entities to recognize the entities and obtain these patterns. This step processes all the categories within a single pass and calculates their support values. It keeps top-k highest support patterns as the input of the next step, together with the matched entity pairs.

5.2 Graph-based Raw Relation Extractor

In this part, for each top-k highest support pattern p, we select a subset of pairs R_p^* from R_p as seed relation instances for an underlying relation that the pattern p may represent. After that, we filter out low quality patterns and extract relation instances $R_p^{'}$ from R_p as the final result.

5.2.1 Seed Relation Instance Extraction

To select seed relation instances R_p^*, we propose an unsupervised graph mining approach. Let $G_p = (C_p, L_p, W_p)$ be a weighted, undirected graph where C_p, L_p and W_p denote vertices, edges and edge weights, respectively. The vertices correspond to the matched entities in categories for pattern p, i.e., $C_p = \{c_p | (e_p, c_p) \in R_p\}$. The edge weights reflect the semantic similarities among entities in C_p. Because the link structure in Chinese Wikipedia is relatively sparse (Wang et al., 2016), we estimate the similarity between entities c_p and $c_p^{'}$ semantically as follows:

$$sim(c_p, c_p^{'}) = \frac{\sum_{c \in Cat(c_p)} \sum_{c' \in Cat(c_p^{'})} \cos(\vec{v}(c_h), \vec{v}(c_h^{'}))}{|Cat(c_p)| \cdot |Cat(c_p^{'})|}$$

where $\cos(\cdot)$ is a cosine function to compute the similarity of two words in the embedding space.

Given a similarity threshold τ, iff $sim(c_p, c_p^{'}) > \tau$, we have $(c_p, c_p^{'}) \in L_p$ and $w(c_p, c_p^{'}) = sim(c_p, c_p^{'})$. In this way, entities in C_p are interconnected if they are similar in semantics.

In this paper, we model the problem of mining R_p^* from R_p as a Maximum Edge Weight Clique Problem (MEWCP) (Alidaee et al., 2007), which detects a maximum edge weight clique C_p^* from C_p in R_p to form R_p^*. Recall that in an undirected graph with edge weights, a maximum edge weight clique is a clique in which the sum of edge weights in the clique is the largest among all the cliques.

To produce a solution for MEWCP, several algorithms have been proposed in the optimization research community, e.g., unconstrained quadratic programming (Alidaee et al., 2007) and the branch-and-cut algorithm (Sørensen, 2004). However, they suffer from high computational complexity due to the NP-Hardness of the problem (Alidaee et al., 2007). In this paper, we introduce an approximate algorithm based on Monte Carlo methods. The general procedure is shown in Algorithm 1. It starts with an empty graph G_p^* to store the clique. In each iteration, it selects an edge $(c_p, c_p^{'})$ from G_p with the probability proportional to its weight $w(c_p, c_p^{'})$. After a particular edge $(c_p, c_p^{'})$ is chosen, the algorithm adds the edge to G_p^*, and removes the edge and other edges that do not connect with any nodes in C_p^* from G_p. This process iterates until no more edges in G_p can be added to G_p^*. Thus, the vertices in G_p^* form the desired clique C_p^*.

Because it is a random, approximate algorithm, the average runtime complexity depends on the input graph structure. We can see that the worst-case runtime complexity is $O(|L_p|^2)$. We run it k

[6]Without ambiguity, we use e_p to represent both the Wikipedia page with the title as e_p and the entity e_p itself.

Algorithm 1 Algorithm for MEWCP

Input: Graph $G_p = (C_p, L_p, W_p)$.
Output: Maximum edge weight clique C_p^*.

```
Initialize temp graph G_p^* = (C_p^*, L_p^*) with C_p^* = ∅ and L_p^* = ∅;
while L_p ≠ ∅ do
    Sample (c_p, c_p') from L_p with prob ∝ w(c_p, c_p');
    C_p = C_p \ {c_p, c_p'}, C_p^* = C_p^* ∪ {c_p, c_p'};
    L_p = L_p \ {(c_p, c_p')}, L_p^* = L_p^* ∪ {(c_p, c_p')};
    for each (c̃_p, c̃_p') ∈ L_p do
        if c̃_p ∉ C_p^* and c̃_p' ∉ C_p^* then
            C_p = C_p \ {c̃_p, c̃_p'};
            L_p = L_p \ {(c̃_p, c̃_p')};
        end if
    end for
end while
return Maximum edge weight clique C_p^*;
```

times and produce multiple results. We select the clique with largest edge weights as the maximum edge weight clique for G_p. The seed relation instance collection is defined as $R_p^* = \{(e_p, c_p) | c_p \in C_p^*, (e_p, c_p) \in R_p\}$. Thus the total runtime complexity is $O(k|L_p|^2)$. In this way, the NP-hard problem is effectively solved in quadratic time.

5.2.2 Relation Extraction and Filtering

After the seed relation instances R_p^* are detected, we employ a confidence score to quantify the quality of pattern p. Intuitively, if pattern p represents entity pairs with the same clear semantic relation, the size of R_p^* and the sum of edge weights in C_p^* will be sufficiently large. Here, we define the confidence score of pattern p as follows:

$$conf(p) = \frac{\ln(1 + |R_p^*|)}{|R_p^*| \cdot (|R_p^*| - 1)} \sum_{c_p, c_p^{'} \in C_p^*, c_p \neq c_p^{'}} sim(c_p, c_p^{'})$$

Based on the formula, patterns with low confidence scores can be filtered. For the remaining patterns, given each $(e_p, c_p) \in R_p$, we add it to the final extracted relation instance collection $R_p^{'}$ if $(e_p, c_p) \in R_p^*$ or it is similar enough to entity pairs in R_p^*. Denote γ as a parameter that controls the precision-recall trade-off. The criteria is:

$$\frac{\sum_{c_p^{'} \in C_p^*} sim(c_p, c_p^{'})}{|C_p^*|} > \frac{\gamma \sum_{c_p^{'}, c_p^{''} \in C_p^*, c_p^{'} \neq c_p^{''}} sim(c_p^{'}, c_p^{''})}{|R_p^*| \cdot (|R_p^*| - 1)}$$

In general, our method detects most probably correct pairs as "seeds" and extract other pairs that are similar enough to seeds. Because it is difficult to ensure high precision for short text relation extraction, we do not use iterative extraction method to avoid "semantic drift" (Carlson et al., 2010).

5.3 Relation Mapping

The final step is to map $R_p^{'}$ to relation triples with a proper relation predicate. Based on category patterns, we have three types of mappings:

Direct Verbal Mapping If the head word of the pattern is a verb, we can use it as the relation predicate. For example, in "[E]出生([E] births)", "出生(born in)" is expressed as a verb in Chinese and is taken as a predicate.

Direct Non-verbal Mapping If the category pattern does not contain a verb but expresses a semantic relation by one/many non-verbs, we define the relation predicate and map the entity pairs to relation triples by logical rules. For example, in the pattern "[E]获得者(Winner of [E])", "获得者(winner)" is a noun that indicates the "得奖(win-prize)" relation.

Indirect Mapping Similar to Suchanek et al. (2007), a few patterns do not describe relations between entity pairs, but should be mapped to other relations indirectly[7]. In "[E]军事([E] military)", it indicates that the entity is related to the topic "军事(military)". Thus, we define a new relation predicate "话题(topic-of)" and establish the relations between entities and "军事(military)".

As seen, the only manual work in our approach is to define relation predicates for direct non-verbal mappings and indirect mappings. In our work, such logical mapping rules are required for only a couple of relation types. Therefore, the proposed approach needs very minimal human work.

6 Experiments

In this section, we conduct experiments to evaluate our method and compare it with state-of-the-art approaches. We also present the overall extraction performance to make the convincing conclusion.

[7]There are also a few is-a relations that are generated by indirect mapping. For example, the pattern "[E]单曲([E] digital single)" infers that the entity that is associated with the category is a song. However, most of the cases are related to non-taxonomic relations.

6.1 Data Source and Experimental Settings

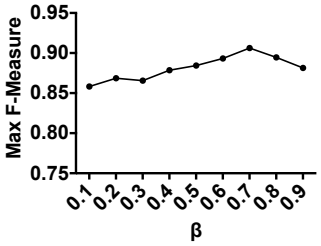
The data source is downloaded from the Chinese Wikipedia dump of the version January 20th, 2017[8]. Because some Wikipedia pages are not related to entities, we use heuristic rules to filter out disambiguation, redirect, template and list pages. Finally, we obtain 0.6M entities and 2.4M entity-category pairs. The open-source toolkit FudanNLP (Qiu et al., 2013) is employed for Chinese NLP analysis. The word embeddings are trained via a Skip-gram model using a large corpus from Wang and He (2016) and set to 100 dimensions.

6.2 Is-a Relation Extraction

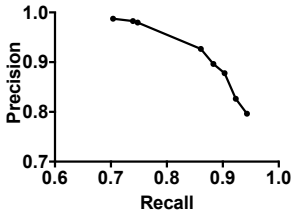
Test Set Generation We randomly select 2,000 entity-category pairs and ask multiple human annotators to label the relations (i.e., is-a and not-is-a). We discard all the pairs that have inconsistent labels across different annotators and obtain a dataset of 1,788 pairs. 30% of the data are used for parameter tuning and the rest for testing. The dataset is publicly available for research.[9]

Parameter Analysis Two parameters are required to be tuned in our method, i.e., β and θ. We vary the value of β from 0.1 to 0.9. With a fixed value of β, we change the value of θ to achieve the best performance over the development set. Figure 2(a) illustrates the maximum F-measure. Experimental results show our method is generally not very sensitive to the selection of β. When $\beta = 0.7$, it has the highest performance, indicating a good balance between the local and global prediction scores. Additionally, Figure 2(b) illustrates the precision-recall curve with respect to the change of θ when $\beta = 0.7$. The highest F-measure is achieved when we set $\theta = 0.05$.

Comparative Study We set up the following strong baselines to compare our method with state-of-the-art approaches. The experimental results are shown in Table 1. To represent entity-category pairs with word embedding based features, we implement several state-of-the-art methods: the *concat* model $\vec{v}(e) \oplus \vec{v}(c_h)$, the *sum* model $\vec{v}(e) + \vec{v}(c_h)$ and the *diff* model $\vec{v}(e) - \vec{v}(c_h)$ (Baroni et al., 2012; Roller et al., 2014; Mirza and Tonelli, 2016). l_2-regularized logistic regression is trained to make the prediction due to the high performance in previous research. This approach achieves the highest F-measure of 72.6%. We also test the *piecewise projection* model proposed in Wang and He (2016) over Chinese Wikipedia, which is state-of-the-art for predicting is-a relations between Chinese words. It has a slight improvement in performance. As seen, our method without the hypernym expansion step (i.e., "Our Method (w/o Exp)" in Table 1) increases the F-measure by 13.2% (with $p < 0.01$) compared to Wang and He (2016). The full implementation of our method has the F-measure of 89.0%, which shows the effectiveness of our approach.



(a) Maximum F-measure when β varies from 0.1 to 0.9



(b) Precision-recall curve when θ varies from -1 to 1

Figure 2: Parameter analysis.

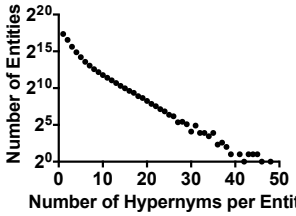
Method	Precision	Recall	F1
Concat Model	79.5%	64.2%	67.2%
Sum Model	80.9%	70.1%	72.6%
Diff Model	78.3%	69.0%	71.5%
Piecewise Projection	78.9%	72.3%	75.5%
Our Method (w/o Exp)	89.2%	88.1%	88.7%
Our Method	**89.8%**	**88.3%**	**89.0%**

Table 1: Performance comparison over test set.

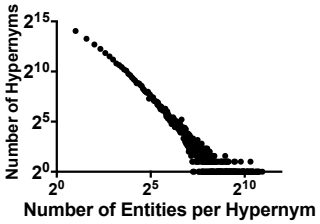
Overall Results In total, we extract 1.17M is-a relations from Chinese Wikipedia categories, consisting 412K entities and 113K distinct categories. In Figure 3(a), we present how many entities have a particular number of hypernyms. In average, each entity has 2.84 hypernyms. We can see that this distribution fits in a semi-log line, defined by

[8] http://download.wikipedia.com/zhwiki/20170120/

[9] There exist a few public datasets for Chinese is-a relations (Fu et al., 2013, 2014). But they aim to learn is-a relations between short concepts/terms and are not suitable for evaluating our work. We focus on understanding (relatively long) categories for entities.



(a) Distribution of number of hypernyms per entity



(b) Distribution of number of entities per hypernym

Figure 3: Distributional analysis on is-a relations.

Category Pattern	Relation Predicate
[E]校友(Alumni)	毕业(graduated-from)
[E]队教练(Coach)	执教(coach-team)
[E]省市镇 (City/Town in Province)	位于(located-in)
[E]获得者(Winner)	获奖(win-prize)

Table 2: Manually defined relation mappings.

a log scale on the y-axis and a linear scale on the x-axis. Similarly, each hypernym has 10.35 entities in average, with the distribution illustrated in Figure 3(b). The number of entities per hypernym follows the power-law distribution with a long tail.

6.3 Non-taxonomic Relation Extraction

Detailed Steps We first run the single-pass pattern miner and extract the category patterns with top-500 highest support values. This is because only fewer than 20 entities are matched for the rest of the patterns. For each of these patterns, we fix $\tau = 0.7$ and run the MEWCP algorithm three times to ensure the high reliability of the seed relation instances, and select top-250 most confident category patterns. To determine the value of γ, we carry out a preliminary experiment, which samples 200 entity pairs to estimate the accuracy. It shows that even we set γ to a relatively low value (i.e., 0.2), the accuracy is over 90%. Finally, 26 relation types are created automatically based on direct verb mapping. We design the mapping rules and relation predicates for the remaining 16 relation types manually, with examples in Table 2.

Evaluation To evaluate the correctness of extracted relations, we carry out two experimental tests: accuracy test and coverage test. Following Suchanek et al. (2007), in the accuracy test, we randomly sample 200 relation instances for each relation type and ask human annotators to label. We discard the results if human annotators disagree. The coverage test is to determine whether the extracted relations already exist in Chinese knowledge bases. Low coverage score means these relations are not present in existing Chinese knowledge bases. In the experiments, we take CN-DBpedia V2.0 (Xu et al., 2017) as the ground truth knowledge base. Up till February 2017, it contains 41M explicit semantic relations of 9M entities, excluding entity summaries, synonyms, etc. We use the CN-DBpedia API[10] to obtain relations for each entity and report the coverage of relation r as:

$$cov(r) = \frac{\#\text{Matched extractions in CN-DBpedia}}{\#\text{Correct extractions generated by our approach}}$$

For fair comparison, because relations in different knowledge base systems may express differently, we ask human annotators to determine whether the relations extracted by our approach and CN-DBpedia match or not. In Table 3, we present the size, accuracy and coverage values of eight non-taxonomic relations, each with over three thousand relation instances.

From the experimental results, we can see that the accuracy is over 90% for all the eight relations. Especially the accuracy values of some relations are over 98% or even equal to 100%. This means it is reliable to extract relations from Chinese UGCs based category pattern mining. The results of the coverage tests present a large variance among different relations. While some relations such as "born-in" have a relatively high coverage in CN-DBpedia, other relation instances that we extract are rarely present in the knowledge base. Overall, the average coverage is approximately 21.1%. This means although the Chinese knowledge base is relatively large in size, it is far from complete. Furthermore, most relations in Chinese knowledge bases are extracted from infoboxes, in the form of attribute-value pairs (Fang et al., 2016; Niu et al., 2011; Wang et al., 2013). Thus, the knowledge harvested from UGCs can be an important supplementary for these systems.

[10] http://knowledgeworks.cn:20313/cndbpedia/api/entityAVP

Relation	Size	Accu.	Cov.	Relation	Size	Accu.	Cov.
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%
逝世(died-in)	5,597	100.0%	18.4%	得奖(win-prize)	3,262	90.0%	27.3%

Table 3: Size, accuracy and coverage values of eight extracted relation types.

Type	Category Pattern	Example
Member pattern	[E]成员/总统 Member/President of [E]	中国科学院成员 Member of Chinese Academy of Sciences
Verb-NP pattern	[E]+Verb+(的)+Noun Phrase	1990年建立的组织 Organization founded in 1990
Verb pattern	[E]+Verb	1980年出生1980 births

Table 4: Category patterns that we design for CN-WikiRe.

Currently, we only focus on Chinese Wikipedia categories. We will study how to extend our approach to UGCs for other knowledge sources, especially domain-specific sources in the future.

Overall Results In summary, our approach extracts 1.52M relations, including 1.17M is-a relations and 0.36M others. The estimated accuracy values of is-a, other and all relations are 92.2%, 97.4% and 93.6% respectively. The accuracy values are estimated over random samples of 500 relations.

Comparison Harvesting non-taxonomic relations from UGCs is non-trivial with no standard evaluation frameworks available. Furthermore, the significant difference between English and Chinese makes it difficult to compare our method with similar research. Pasca (2017) focuses on modifier in categories and is not directly comparable to our work. In YAGO (Suchanek et al., 2007), relations in categories are extracted by handcrafting regular expressions. They extract nine non-taxonomic relations, with accuracy values of around 90%-98%. Our approach avoids the manual work to a large extent and harvests more types of relations with a comparable accuracy.

Next we compare our work with Nastase and Strube (2008), which heavily relies on prepositions in patterns such as "Verb in/of" and "Member/CEO/President of" to discover relations. In Chinese, prepositions are usually expressed implicitly and hence these patterns are not directly applicable. We implement a variant for Chinese (denoted as CN-WikiRe). The patterns that we used in CN-WikiRe are shown in Table 4. In the experiments, we extract 165,048 non-taxonomic relation instances using CN-WikiRe, containing 631 relation types. Although the number of relation types may seem large at the first glance, only 14% of them are actual relation predicates, with the rest being either incorrect or uninformative. The reasons are twofold: i) word segmentation and POS tagging for Chinese short texts still suffer from low accuracy and ii) not all verbs extracted by CN-WikiRe can serve as relation predicates (e.g., "传导(transmit)", "缩小(shrink)"). We sample 500 relations from the collection where the extracted verbs are labeled as real relation predicates. The accuracy is 58.6%, much lower than our method. Furthermore, the partially explicit and implicit patterns (see (Nastase and Strube, 2008)) do not have their counterparts in Chinese. Therefore, our method is superior to existing systems.

7 Conclusion and Future Work

We propose a weakly supervised framework to extract relations from Chinese UGCs. For is-a relations, we introduce a word embedding based method and refine prediction results using collective inference. To extract non-taxonomic relations, we design a graph mining technique to harvest relation types and category patterns with minimal human supervision. Future work includes: i) improving our work for short text knowledge extraction and ii) designing a general framework for cross-lingual UGC relation extraction.

Acknowledgements

This work is supported by the National Key Research and Development Program of China under Grant No. 2016YFB1000904. Chengyu Wang would also like to thank the ECNU Outstanding Doctoral Dissertation Cultivation Plan of Action (No. YB2016040) for the support of his research.

References

Daniele Alfarone and Jesse Davis. 2015. Unsupervised learning of an IS-A taxonomy from a limited domain-specific corpus. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*. pages 1434–1441.

Bahram Alidaee, Fred Glover, Gary A. Kochenberger, and Haibo Wang. 2007. Solving the maximum edge weight clique problem via unconstrained quadratic programming. *European Journal of Operational Research* 181(2):592–597.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of 13th Conference of the European Chapter of the Association for Computational Linguistics*. pages 23–32.

Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Coupled semi-supervised learning for information extraction. In *Proceedings of the Third International Conference on Web Search and Web Data Mining*. pages 101–110.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. 2006. *Transductive Inference and Semi-Supervised Learning*. MIT Press.

Yanping Chen, Qinghua Zheng, and Wei Zhang. 2014. Omni-word feature and soft constraint for chinese relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. pages 572–581.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. 2011. Open information extraction: The second generation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. pages 3–10.

Zhijia Fang, Haofen Wang, Jorge Gracia, Julia Bosque-Gil, and Tong Ruan. 2016. Zhishi.lemon: On publishing zhishi.me as linguistic linked open data. In *Proceedings of the 15th International Semantic Web Conference*. pages 47–55.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2014. Two is bigger (and better) than one: the wikipedia bitaxonomy project. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. pages 945–955.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. pages 1199–1209.

Ruiji Fu, Bing Qin, and Ting Liu. 2013. Exploiting multiple sources for open-domain hypernym discovery. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pages 1224–1234.

Gene H. Golub, Per Christian Hansen, and Dianne P. O'Leary. 1999. Tikhonov regularization and total least squares. *SIAM J. Matrix Analysis Applications* 21(1):185–194.

Amit Gupta, Francesco Piccinno, Mikhail Kozhevnikov, Marius Pasca, and Daniele Pighin. 2016. Revisiting taxonomy induction over wikipedia. In *Proceedings of the 26th International Conference on Computational Linguistics*. pages 2300–2309.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 970–976.

Jinyang Li, Chengyu Wang, Xiaofeng He, Rong Zhang, and Ming Gao. 2015. User generated content oriented chinese taxonomy construction. In *Proceedings of the 17th Asia-Pacific Web Conference*. pages 623–634.

Weiming Lu, Renjie Lou, Hao Dai, Zhenyu Zhang, Shansong Yang, and Baogang Wei. 2015. Taxonomy induction from chinese encyclopedias by combinatorial optimization. In *Proceedings of the 4th CCF Conference on Natural Language Processing and Chinese Computing*. pages 299–312.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. 2015. YAGO3: A knowledge base from multilingual wikipedias. In *Proceedings of the Seventh Biennial Conference on Innovative Data Systems Research*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*. pages 3111–3119.

Paramita Mirza and Sara Tonelli. 2016. On the contribution of word embeddings to temporal relation classification. In *Proceedings of the 26th International Conference on Computational Linguistics*. pages 2818–2828.

Vivi Nastase and Michael Strube. 2008. Decoding wikipedia categories for knowledge acquisition. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*. pages 1219–1224.

Vivi Nastase, Michael Strube, Benjamin Boerschinger, Cäcilia Zirn, and Anas Elghafari. 2010. Wikinet: A very large scale multi-lingual concept network. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Xing Niu, Xinruo Sun, Haofen Wang, Shu Rong, Guilin Qi, and Yong Yu. 2011. Zhishi.me - weaving chinese linking open data. In *Proceedings of the*

10th International Semantic Web Conference. pages 205–220.

Marius Pasca. 2017. German typographers vs. german grammar: Decomposition of wikipedia category labels into attribute-value pairs. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. pages 315–324.

Simone Paolo Ponzetto and Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*. pages 2083–2088.

Simone Paolo Ponzetto and Michael Strube. 2007. Deriving a large-scale taxonomy from wikipedia. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*. pages 1440–1445.

Likun Qiu and Yue Zhang. 2014. ZORE: A syntax-based system for chinese open relation extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. pages 1870–1880.

Xipeng Qiu, Qi Zhang, and Xuanjing Huang. 2013. Fudannlp: A toolkit for chinese natural language processing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. pages 49–54.

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of the 25th International Conference on Computational Linguistics*. pages 1025–1036.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. page 2389–2398.

Michael M. Sørensen. 2004. New facets and a branch-and-cut algorithm for the weighted clique problem. *European Journal of Operational Research* 154(1):57–70.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*. pages 697–706.

Chengyu Wang and Xiaofeng He. 2016. Chinese hypernym-hyponym extraction from user generated categories. In *Proceedings of the 26th International Conference on Computational Linguistics*. pages 1350–1361.

Chengyu Wang, Junchi Yan, Aoying Zhou, and Xiaofeng He. 2017. Transductive non-linear learning for chinese hypernym prediction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Chengyu Wang, Rong Zhang, Xiaofeng He, and Aoying Zhou. 2016. Error link detection and correction in wikipedia. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. pages 307–316.

Zhigang Wang, Juanzi Li, Shuangjie Li, Mingyang Li, Jie Tang, Kuo Zhang, and Kun Zhang. 2014. Cross-lingual knowledge validation based taxonomy derivation from heterogeneous online wikis. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. pages 180–186.

Zhigang Wang, Juanzi Li, Zhichun Wang, Shuangjie Li, Mingyang Li, Dongsheng Zhang, Yao Shi, Yongbin Liu, Peng Zhang, and Jie Tang. 2013. Xlore: A large-scale english-chinese bilingual knowledge graph. In *Proceedings of the ISWC 2013 Posters & Demonstrations Track*. pages 121–124.

Tianxing Wu, Guilin Qi, Haofen Wang, Kang Xu, and Xuan Cui. 2016. Cross-lingual taxonomy alignment with bilingual biterm topic model. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. pages 287–293.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. pages 481–492.

Bo Xu, Chenhao Xie, Yi Zhang, Yanghua Xiao, Haixun Wang, and Wei Wang. 2016a. Learning defining features for categories. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. pages 3924–3930.

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. Cndbpedia: A never-ending chinese knowledge extraction system. In *Advances in Artificial Intelligence: From Theory to Practice - 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*. pages 428–438.

Bo Xu, Yi Zhang, Jiaqing Liang, Yanghua Xiao, Seung-won Hwang, and Wei Wang. 2016b. Cross-lingual type inference. In *Proceedings of 21st International Conference on Database Systems for Advanced Applications*. pages 447–462.