Chinese Hypernym-Hyponym Extraction from User Generated Categories

Chengyu Wang, Xiaofeng He*
School of Computer Science and Software Engineering
East China Normal University, Shanghai, China
chywang2013@gmail.com, xfhe@sei.ecnu.edu.cn

Abstract

Hypernym-hyponym ("*is-a*") relations are key components in taxonomies, object hierarchies and knowledge graphs. While there is abundant research on *is-a* relation extraction in English, it still remains a challenge to identify such relations from Chinese knowledge sources accurately due to the flexibility of language expression. In this paper, we introduce a weakly supervised framework to extract Chinese *is-a* relations from user generated categories. It employs piecewise linear projection models trained on a Chinese taxonomy and an iterative learning algorithm to update models incrementally. A pattern-based relation selection method is proposed to prevent "semantic drift" in the learning process using bi-criteria optimization. Experimental results illustrate that the proposed approach outperforms state-of-the-art methods.

1 Introduction

A hypernym-hyponym ("*is-a*") relation is a word/phrase pair (x, y) such that x is a hyponym of y. These relations are extensively employed in machine reading (Etzioni et al., 2011), query understanding (Hua et al., 2015) and other NLP tasks. The extraction of *is-a* relations is necessary to construct taxonomies for Web-scale knowledge graphs (Suchanek et al., 2007; Wu et al., 2012; Wang et al., 2015).

In previous work, *is-a* relations were obtained by either using expert-compiled thesauri such as WordNet (Miller, 1995), or harvested automatically from the Web. Since knowledge in thesauri is usually limited in quantity and variety, it is more prevalent to harvest *is-a* relations from online encyclopedias (Ponzetto and Strube, 2007), Web corpora (Wu et al., 2012), etc. Currently, a majority of existing methods focus on syntactic, lexical and/or semantic analysis on English corpora, but most of these approaches are language dependent. It is not easy to apply methods for one language to knowledge sources in another language directly. For example, in Chinese, the word formation, grammar, semantics and tenses are flexible and more irregular. Thus patttern-based methods can only cover few linguistic circumstances. As pointed out by Li et al. (2013), the performance of syntactic analysis and named entity recognition on Chinese corpora still needs to be improved to support robust relation extraction. Furthermore, it is still difficult to use machine translation-based methods to extract such relations because there are great differences in word orders between English and Chinese (Cai et al., 2014).

More recently, word embedding (or distributed word representation) has been empirically proved effective in modeling some of the semantic relations between words by offsets of word vectors (Mikolov et al., 2013a; Mikolov et al., 2013b). The learning of word embeddings only requires shallow processing of a large text corpus. As Fu et al. (2014) suggest, the representation of *is-a* relations is more complicated than vector offsets. By studying the relations of word embeddings between hyponyms and their respective hypernyms, *is-a* relations can be identified by learning semantic prediction models.

In this paper, we consider the problem of harvesting Chinese *is-a* relations from user generated categories, which frequently appear in online encyclopedias and vertical websites. These category names

* Corresponding author.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1350–1361, Osaka, Japan, December 11-17 2016.

are classes, concepts or topics manually added by human contributors. For instance, in *Baidu Baike*[1], the page "奥巴马(Barack Obama)" has the following categories: "政治人物(Political Figure)", "外国(Foreign Country)", "元首(Leader)" and "人物(Person)". Given an entity and its category set, we aim to predict whether each category name is the hypernym of the entity. We observe that vector offsets of *is-a* relations are quite different in varied data sources and domains (discussed in Section 3). This implies that using a single model is difficult to preserve all the linguistic regularities of *is-a* relations. Furthermore, models learned from one knowledge source are not necessarily effective to extract *is-a* relations from another source, while it is a common practice to construct large-scale taxonomies from multiple Web sources (Fu et al., 2013; Dong et al., 2014; Wang et al., 2014).

To address this problem, we propose a weakly supervised framework to extract *is-a* relations automatically. In the initial stage, we build piecewise linear projection models trained on samples from an existing Chinese taxonomy (Li et al., 2015). In this stage, a K-means based incremental clustering technique is employed to group *is-a* relations with similar semantics together. In each cluster, a separate model maps entities to their respective hypernyms in the embedding space. After that, clustering results are updated incrementally with projection models retrained in an iterative manner. In each iteration, we extract previously unseen *is-a* relations from a collection of unlabeled <entity, category> pairs. To avoid "semantic drift" (Carlson et al., 2010b), a bi-criteria optimization method is proposed such that only those extracted *is-a* relations that are validated by three types of Chinese patterns in a corpus can be labeled as "positive" and added to the training set. In this way, projection models for the target knowledge source are trained without any labeling efforts.

The rest of this paper is organized as follows. Section 2 summarizes the related work. Details of our approach for addressing the *is-a* relation extraction problem are described in Section 3. Experimental results are presented in Section 4. We conclude our paper and discuss the future work in Section 5.

2 Related Work

The *is-a* relation extraction problem has been addressed by identifying hyponyms and their hypernyms from various data sources. Here, we present a summarization on methods on *is-a* relation extraction.

Pattern matching based methods employ syntactic/lexical patterns to extract *is-a* relations. The early work introduced by Hearst (1992) utilizes manually designed patterns to obtain *is-a* relations from text corpora. For instance, based on "NP_1 such as NP_2", it can be inferred that NP_2 is the hypernym of NP_1, where NP_1 and NP_2 are noun phases. These patterns are effective for English and are used to build the largest taxonomy Probase (Wu et al., 2012). However, it is hard to handcraft all valid *is-a* patterns. Ortega-Mendoza et al. (2007) use "seed instances" (i.e., *is-a* word pairs) to discover lexical patterns from the Web using search engines and harvest new instances automatically. Snow et al. (2004) detect syntactic *is-a* patterns by analyzing the parse trees and train a hypernym classifier based on syntactic features. Similar approaches have been adopted in a variety of research (Caraballo, 1999; Etzioni et al., 2004; Sang, 2007; Ritter et al., 2009; Pantel and Pennacchiotti, 2006; Kozareva and Hovy, 2010). As Fu et al. (2014) suggest, many *is-a* relations are expressed in highly flexible manners in Chinese and these approaches have limited extraction accuracy.

Thesauri and encyclopedias can serve as knowledge sources to construct object hierarchies. Suchanek et al. (2007) link concepts in Wikipedia to WordNet synsets (Miller, 1995) by considering the textual patterns of Wikipedia categories. Ponzetto and Strube (2007) design lexical, syntactic and connectivity features to predict whether there is an *is-a* relation between a Wikipedia entity and its category. For Chinese language, Li et al. (2015) introduce a set of language-specific features to predict *is-a* relations using a SVM classifier and construct a large-scale Chinese taxonomy from Wikipedia. Fu et al. (2013) utilize multiple data sources such as encyclopedias and search engine results to design a ranking function in order to extract the most possible hypernym given an entity. Cross-lingual links in Wikipedia are leveraged in (Wang et al., 2014) to derive a bilingual taxonomy by a dynamic boosting model. These methods are more precise than free text extraction but have limited scope constrained by sources.

[1]Baidu Baike (http://baike.baidu.com/) is one of the largest encyclopedias in China, with over 13M entries up till July, 2016.

Text inference approaches make use of distributional similarity measures, which go beyond pattern matching and instead compare the contexts of word pairs in a corpus to infer their relations indirectly. Kotlerman et al. (2010) consider the asymmetric property of *is-a* relations and design directional similarity measures to make lexical inference. Other directional measures are proposed in (Bhagat et al., 2007; Szpektor et al., 2007; Clarke, 2012; Lenci and Benotto, 2012). These methods assume that a hyponym can only appear in some of the contexts of its hypernym and a hypernym can appear in all contexts of its hyponyms. One potential limitation is that the contexts in Chinese are usually flexible and sparse.

To tackle the data sparsity problem, word embedding based approaches have been proposed to solve a series of NLP tasks, such as sentiment classification (Zhou et al., 2015), machine translation (Zhang et al., 2014) and question answering (Yang et al., 2014). In these approaches, words are mapped to a low dimensional space by training neural network based language models, such as *CBOW* and *Skip-gram* models (Mikolov et al., 2013a). The dense word representations are more likely to deal with the context sparsity issue in Chinese stemmed from the flexible expressions. The state-of-the-art method in (Fu et al., 2014) is most related to ours, which takes a Chinese thesaurus as a-priori knowledge and train piecewise linear projection models based on word embeddings. In this paper, we further improve the performance of the word embedding based method by iterative learning of projection models and *is-a* relation selection based on Chinese textual patterns.

3 Weakly Supervised Is-a Relation Extraction

In this section, we describe the formal definition of our problem. The motivation of our method is discussed and the detailed steps introduced, namely, initial model training and iterative learning process.

3.1 Problem Statement

A taxonomy is a *direct acyclic graph* $G = (E, R)$ where nodes E represent entities/classes and edges R denote *is-a* relations. Following the work in Fu et al. (2014), *is-a* relations are regarded as *asymmetric* and *transitive* relations. Therefore, all correct *is-a* relations derived from G are in the transitive closure of R, denoted as R^* where $R^* = \bigcup_{i=0}^{\infty} R^{(i)}$ and $R^{(i+1)} = R \circ R^{(i)}$ with initial condition $R^{(0)} = R$ and $\circ$ being the composition operator of relations.

To extract *is-a* relations from user generated categories, we obtain the collection of entities E^* from the knowledge source (such as *Baidu Baike*). The set of user generated categories for each $e \in E^*$ is denoted as $Cat(e)$. Thus we need to design a learning algorithm F based on R^* to predict whether there is an *is-a* relation between e and c where $e \in E^*$ and $c \in Cat(e)$. In this way, we can utilize an existing taxonomy to harvest new *is-a* knowledge automatically.

3.2 Motivation of Our Method

The state-of-the-art method for Chinese *is-a* relation extraction is the word embedding based approach in (Fu et al., 2014). In their work, the projection parameters of a piecewise linear projection model learned from a Chinese thesaurus are used to identify *is-a* relations in encyclopedias. In this paper, we take a deeper look at the word vectors of hyponyms and hypernyms. As a preliminary experiment, we randomly sample *is-a* relations from a Wikipedia-based Chinese taxonomy (Li et al., 2015) and a Chinese thesaurus *CilinE*[2]. We compute the offsets of embedding vectors (i.e., $\vec{v}(x) - \vec{v}(y)$) where x is the hyponym of y. We have three observations, with examples shown in Table 1[3].

- **Observation 1.** For a fixed x, if y_1 and y_2 are hypernyms of different levels, it is likely that $\vec{v}(x) - \vec{v}(y_1) \not\approx \vec{v}(x) - \vec{v}(y_2)$. For example, "Country" is a high-level hypernym of "Japan" while "Asian Country" covers a narrow spectrum of entities.

- **Observation 2.** If $(x_1, instanceOf, y_1)$ and $(x_2, subClassOf, y_2)$ hold, it is likely that $\vec{v}(x_1) - \vec{v}(y_1) \not\approx \vec{v}(x_2) - \vec{v}(y_2)$. Although both *instanceOf* and *subClassOf* are *is-a* relations in a broad

[2]http://www.ltp-cloud.com/download/

[3]We use the l_2 norm of vector offsets to quantify the difference.

	Example with English Translation	Vector Offsets
True Positive	$\vec{v}$(日本) − $\vec{v}$(国家) ≈ $\vec{v}$(澳大利亚) − $\vec{v}$(国家) $\vec{v}$(Japan) − $\vec{v}$(Country) ≈ $\vec{v}$(Australia) − $\vec{v}$(Country)	$1.03 \approx 0.99$
Observation 1	$\vec{v}$(日本) − $\vec{v}$(国家) $\not\approx$ $\vec{v}$(日本) − $\vec{v}$(亚洲国家) $\vec{v}$(Japan) − $\vec{v}$(Country) $\not\approx$ $\vec{v}$(Japan) − $\vec{v}$(Asian Country)	$1.03 \not\approx 0.71$
Observation 2	$\vec{v}$(日本) − $\vec{v}$(国家) $\not\approx$ $\vec{v}$(主权国) − $\vec{v}$(国家) $\vec{v}$(Japan) − $\vec{v}$(Country) $\not\approx$ $\vec{v}$(Sovereign State) − $\vec{v}$(Country)	$1.03 \not\approx 1.32$
Observation 3	$\vec{v}$(日本) − $\vec{v}$(国家) $\not\approx$ $\vec{v}$(西瓜) − $\vec{v}$(水果) $\vec{v}$(Japan) − $\vec{v}$(Country) $\not\approx$ $\vec{v}$(Watermelon) − $\vec{v}$(Fruit)	$1.03 \not\approx 0.39$

Table 1: Examples of three observations.

sense, the *is-a* relations between (i) entities and classes, and (ii) classes and classes are different in semantics.

- **Observation 3.** For *is-a* pairs in two different domains (x_1, y_1) and (x_2, y_2), it is likely that $\vec{v}(x_1) - \vec{v}(y_1) \not\approx \vec{v}(x_2) - \vec{v}(y_2)$. This implies that *is-a* relations can be divided into more fine-grained relations based on their topics, such as politics, grocery, etc. A similar finding is also presented in (Fu et al., 2014).

These situations bring the challenges in modeling *is-a* relations correctly. Furthermore, *is-a* relations across different knowledge sources vary in characteristics. For example, *is-a* relations in a taxonomy are mostly *subClassOf* relations between concepts, while a large number of *is-a* relations derived from online encyclopedias are *instanceOf* relations, especially in the emerging domains, such as the Internet, new technologies, etc. The differences of *is-a* representations between knowledge sources suggest that a simple model trained on the taxonomy is not effective for *is-a* extraction from encyclopedias. The observations prompt us to take a two-stage process to deal with this problem. In the initial stage, we train piecewise linear projection models based on the taxonomy, aiming to learn prior representations of *is-a* relations in the embedding space. Next, we iteratively extract new *is-a* relations from user generated categories using models in the previous round and adjust our models accordingly. The characteristics of *is-a* relations of the target source are learned in a step-by-step manner.

3.3 Initial Model Training

We first train a *Skip-gram* model over a Chinese text corpus with over 1 billion words to obtain word embeddings. We randomly sample *is-a* relations from R^* as training data, denoted as $R^{'} \subset R^*$. In previous work, Mikolov et al. (2013b) and Fu et al. (2014) employ vector offsets and projection matrices to map words to their hypernyms, respectively. In this paper, we further combine the two relation representations together in the embedding space. For a pair (x_i, y_i), we assume a projection matrix $\mathbf{M}$ and an offset vector $\vec{b}$ map x_i to y_i in the form: $\mathbf{M} \cdot \vec{v}(x_i) + \vec{b} = \vec{v}(y_i)$.

To capture the multiple implicit language regularities in the training data, we follow the piecewise model training technique in (Fu et al., 2014). We first partition $R^{'}$ into K groups by K-means, denoted as $R^{'} = \bigcup_{k=1}^{K} C_k$ where C_k is the collection of *is-a* pairs in the kth cluster. Each pair $(x_i, y_i) \in R^{'}$ is represented as the vector offset $\vec{v}(x_i) - \vec{v}(y_i)$ for clustering. In each cluster, we assume *is-a* relations share the same projection matrix and vector offset. Therefore, we aim to learn K projection matrices and offset vectors as representations of *is-a* relations. For each cluster C_k $(k = 1, 2, \cdots, K)$, we aim to minimize the following objective function:

$$J(\mathbf{M}_k, \vec{b}_k; C_k) = \frac{1}{|C_k|} \sum_{(x_i, y_i) \in C_k} \|\mathbf{M}_k \cdot \vec{v}(x_i) + \vec{b}_k - \vec{v}(y_i)\|^2$$

where $\mathbf{M}_k$ and $\vec{b}_k$ are the projection matrix and the offset vector for C_k, learned via Stochastic Gradient Descent.

3.4 Iterative Learning Process

In the iterative learning process, we train *is-a* relation projection models on a series of dynamically enlarged training set $R^{(t)}$ ($t = 1, 2, \cdots, T$). The main idea is to update clustering results and prediction models iteratively in order to achieve a better generalization ability on the target knowledge source.

Initialization. We have two datasets: (i) the positive dataset $R^{(1)} = R^{'}$ and (ii) the unlabeled dataset $U = \{(x_i, y_i)\}$, which is created from user generated categories. Usually, we have $|U| \gg |R^{(1)}|$. Denote $C_k^{(t)}$ as the collection of *is-a* pairs, $\vec{c}_k^{(t)}$ as the cluster centroid, and $\mathbf{M}_k^{(t)}$ and $\vec{b}_k^{(t)}$ as model parameters in the kth cluster of the tth iteration. We set $C_k^{(1)} = C_k$, $\vec{c}_k^{(1)} = \frac{1}{|C_k|} \sum_{(x_i,y_i) \in C_k} \vec{v}(x_i) - \vec{v}(y_i)$, $\mathbf{M}_k^{(1)} = \mathbf{M}_k$ and $\vec{b}_k^{(1)} = \vec{b}_k$ as the initial values.

Iterative Process. For each iteration $t = 1, \cdots, T$, the models are updated as follows:

- **Step 1.** Randomly sample $\delta \cdot |U|$ instances from U and denote it as $U^{(t)}$ where δ is a sampling factor. For each $(x_i, y_i) \in U^{(t)}$, compute the cluster ID as $p_i = \arg\min_{k=1,\cdots,K} \|\vec{v}(x_i) - \vec{v}(y_i) - \vec{c}_k^{(t)}\|$. We first compute the difference $d^{(t)}(x_i, y_i)$ as $d^{(t)}(x_i, y_i) = \|\mathbf{M}_{p_i} \cdot \vec{v}(x_i) + \vec{b}_{p_i} - \vec{v}(y_i)\|$. The prediction result is $f_M^{(t)}(x_i, y_i) = I(d^{(t)}(x_i, y_i) < \epsilon)$ where $I(\cdot)$ is an indicator function and ϵ is a pre-defined threshold. We use $U_-^{(t)}$ to represent word pairs in $U^{(t)}$ predicted as "positive" in this step.

- **Step 2.** For each $(x_i, y_i) \in U_-^{(t)}$, predict the label (*is-a* or *not-is-a*) by pattern-based relation selection method (introduced in Section 3.5), denoted as $f_P^{(t)}(x_i, y_i)$. Define $U_+^{(t)} = \{(x_i, y_i) \in U_-^{(t)} | f_P^{(t)}(x_i, y_i) = 1\}$. Update the two datasets as follows: (i) $U = U \setminus U_+^{(t)}$ and (ii) $R^{(t+1)} = R^{(t)} \cup U_+^{(t)}$.

- **Step 3.** Denote the collection of *is-a* pairs in $U_+^{(t)}$ that belongs to the kth cluster as $U_k^{(t)}$. Update the cluster centroid $\vec{c}_k^{(t)}$ as follows:

$$\vec{c}_k^{(t+1)} = \vec{c}_k^{(t)} + \lambda \cdot \frac{1}{|U_k^{(t)}|} \sum_{(x_i,y_i) \in U_k^{(t)}} (\vec{v}(x_i) - \vec{v}(y_i) - \vec{c}_k^{(t)})$$

  where λ is a learning rate in $(0, 1)$ that controls the speed of cluster centroid "drift" over time. Re-assign the membership of clusters $C_k^{(t+1)}$ for each $(x_i, y_i) \in R^{(t+1)}$ based on new centroids.

- **Step 4.** For each cluster $C_k^{(t+1)}$, update model parameters by minimizing the objective function:

$$J(\mathbf{M}_k^{(t+1)}, \vec{b}_k^{(t+1)}; C_k^{(t+1)}) = \frac{1}{|C_k^{(t+1)}|} \sum_{(x_i,y_i) \in C_k^{(t+1)}} \|\mathbf{M}_k^{(t+1)} \cdot \vec{v}(x_i) + \vec{b}_k^{(t+1)} - \vec{v}(y_i)\|^2$$

  with the initial parameter values $\mathbf{M}_k^{(t+1)} = \mathbf{M}_k^{(t)}$ and $\vec{b}_k^{(t+1)} = \vec{b}_k^{(t)}$.

Model Prediction. After the training phase, given a pair (x_i, y_i) in the test set, our method predicts that x_i is the hyponym of y_i if at least one of the following conditions holds:

1. (x_i, y_i) is in the transitive closure of $R^{(T+1)}$ (based on *transitivity* property of *is-a* relations).

2. $f_M^{(T+1)}(x_i, y_i) = 1$ (based on final model prediction).

Discussion. The key techniques of the algorithm lie in two aspects: (i) combination of *semantic* and *syntactic-lexico is-a* extraction and (ii) incremental learning. The positive relation selection method in Step 2 can also be regarded as a variant of coupled learning (Carlson et al., 2010a). We ensure only when the results of semantic projection and pattern-based approach are consistent, these relations are added to our training set. Also, at each iteration, the model parameters are updated incrementally. By solving the recurrent formula, the update rule of centroids in Step 3 is equivalent to:

$$\vec{c}_k^{(T+1)} = (1-\lambda)^T \cdot \vec{c}_k^{(1)} + \lambda \cdot \sum_{t=1}^{T} \left(\frac{(1-\lambda)^{T-t}}{|U_k^{(t)}|} \cdot \sum_{(x_i,y_i) \in U_k^{(t)}} (\vec{v}(x_i) - \vec{v}(y_i) - \vec{c}_k^{(t)}) \right)$$

We can see that $\vec{c}_k^{(T+1)}$ is a weighted average of vector offsets of *is-a* relations added into the cluster, where the weight increases exponentially over time. With cluster assignments and prediction models updated, our models gradually fit the semantics of new *is-a* relations extracted from the unlabeled dataset.

3.5 Pattern-Based Relation Selection

We now introduce the pattern-based approach used in Step 2 of the iterative learning process. Although Chinese patterns for relation extraction can not guarantee high precision and coverage, we employ them as a "validation" source for model-based extraction results. The goal of this method is to select only a small portion of relations as $U_+^{(t)}$ from $U_-^{(t)}$ with high confidence to add to the training set $R^{(t)}$.

Previously, Fu et al. (2013) design several Chinese Hearst-style patterns manually for *is-a* extraction. In this paper, we collect a broader spectrum of patterns related to *is-a* relations, and categorize them into three types: "Is-A", "Such-As" and "Co-Hyponym". The examples are shown in Table 2[4]. We have the following two observations:

- **Observation 4.** If x_i and y match an "Is-A" or "Such-As" pattern, there is a large probability that x_i is the hyponym of y. Let $n_1(x_i, y)$ be the number of matches for x_i and y in a text corpus.
- **Observation 5.** If x_i and x_j match a "Such-As" or "Co-Hyponym" pattern, there is a large probability that no *is-a* relation exists between x_i and x_j. Let $n_2(x_i, x_j)$ be the number of matches for x_i and x_j, and $n_2(x_i)$ be the number of matches for x_i and x^* where x^* is an arbitrary hyponym other than x_i.

Category	Examples	Corresponding English Translation
Is-A	x_i是一个y	x_i is a kind of y
	x_i是y之一	x_i is one of y
Such-As	y，例如x_i、x_j	y, such as x_i and x_j
	y，包括x_i、x_j	y, including x_i and x_j
Co-Hyponym	x_i、x_j等	x_i, x_j and others
	x_i和x_j	x_i and x_j

Table 2: Examples of Chinese hypernym/hyponym patterns.

In this algorithm, we utilize the prediction of projection models and Chinese hypernym/hyponym patterns jointly to decide which relations in $U_-^{(t)}$ should be added into $U_+^{(t)}$. For each $(x_i, y_i) \in U_-^{(t)}$, denote $PS^{(t)}(x_i, y_i)$ and $NS^{(t)}(x_i, y_i)$ as the positive and negative scores that indicate the level of confidence. We define the positive score based on model prediction and Observation 4:

$$PS^{(t)}(x_i, y_i) = \alpha \cdot \left(1 - \frac{d^{(t)}(x_i, y_i)}{\max_{(x,y) \in U_-^{(t)}} d^{(t)}(x, y)}\right) + (1-\alpha) \cdot \frac{n_1(x_i, y_i) + \gamma}{\max_{(x,y) \in U_-^{(t)}} n_1(x, y) + \gamma}$$

[4]In practice, there can be over two candidate hyponyms in "Such-As" and "Co-Hyponym" patterns. For simplicity, we only list two here, denoted as x_i and x_j.

where $\alpha \in (0,1)$ is a tuning weight to balance the two factors and γ is a smoothing parameter. For simplicity, we empirically set $\alpha = 0.5$ and $\gamma = 1$ in this paper. We define the negative score based on Observation 5 as follows:

$$NS^{(t)}(x_i, y_i) = \log \frac{n_2(x_i, y_i) + \gamma}{(n_2(x_i) + \gamma) \cdot (n_2(y_i) + \gamma)}$$

A high negative score between x_i and y_i means the strong evidence of the frequent co-occurrence of x_i and y_i in "Such-As" or "Co-Hyponym" patterns, where x_i and y_i are likely to be co-hyponyms. This indicates that there is a low probability of the existence of an *is-a* relation between them.

A bi-criteria optimization problem can be formed where positive and negative scores should be maximized and minimized simultaneously, which is hard to optimize. We further covert it into a positive score maximization problem with negative score constraints:

$$\begin{aligned} \max \quad & \sum_{(x_i,y_i)\in U_+^{(t)}} PS^{(t)}(x_i, y_i) \\ \text{s. t.} \quad & \sum_{(x_i,y_i)\in U_+^{(t)}} NS^{(t)}(x_i, y_i) < \theta, U_+^{(t)} \subset U_-^{(t)}, |U_+^{(t)}| = m \end{aligned}$$

where m is the size of $U_+^{(t)}$ and θ is used to constrain negative score limits. This problem is a special case of the *budgeted maximum coverage problem* (Khuller et al., 1999), which is NP-hard. Based on the proof in (Khuller et al., 1999), the objective function is *monotone* and *submodular*. Therefore, we design a greedy relation selection algorithm to solve this problem with the accuracy of $1 - \frac{1}{e}$, shown in Algorithm 1. Finally, for each $(x_i, y_i) \in U_+^{(t)}$, we make the prediction as: $f_P^{(t)}(x_i, y_i) = I((x_i, y_i) \in U_+^{(t)})$.

Algorithm 1 Greedy Relation Selection Algorithm

```
Initialize U_+^(t) = ∅;
while |U_+^(t)| < m do
    Select candidate is-a pair with largest PS: (x_i, y_i) = arg max_{(x_i,y_i)∈U_+^(t)} PS^(t)(x_i, y_i);
    Remove the pair from U_-^(t): U_-^(t) = U_-^(t) \ {(x_i, y_i)};
    if NS^(t)(x_i, y_i) + Σ_{(x,y)∈U_+^(t)} NS^(t)(x, y) < θ then
        Add the pair to U_+^(t): U_+^(t) = U_+^(t) ∪ {(x_i, y_i)};
    end if
end while
return Collection of is-a relations U_+^(t);
```

4 Experiments

In this section, we conduct comprehensive experiments to evaluate our method on publicly available datasets. We also compare it with state-of-the-art approaches to make the convincing conclusion.

4.1 Experimental Data

In the experiments, we use four datasets consisting of word pairs, and a large Chinese text corpus. The statistics of our datasets are summarized in Table 3.

Dataset	Positive	Negative	Unknown
Wiki Taxonomy	7,312	-	-
Unlabeled Set	-	-	78,080
Validation Set	349	1,071	-
Test Set	1,042	3,223	-

Table 3: Datasets summarization.

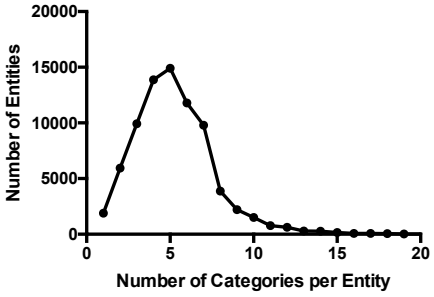


Figure 1: Unlabeled data statistics.

To learn word embeddings, we crawl Web pages from *Baidu Baike* and extract the contents to form a Chinese text corpus, consisting of 1.088B words. We use the open source toolkit *Ansj*[5] for word segmentation. Finally, we train a *Skip-gram* model to obtain 100-dimensional embedding vectors of 5.8M words. We calculate statistics for the pattern-based method in Section 3.5 using the same corpus.

The taxonomy data is obtained from the authors (Li et al., 2015), which consists of 1.3M *is-a* relations derived from Chinese Wikipedia. We use 7,312 *is-a* relations sampled from the taxonomy to train the initial projection models. To construct the unlabeled set, we randomly sample 0.1M entities from our *Baidu Baike* data, filter out entities without user generated categories and extract 78K <entity, category> pairs. The distribution of the number of categories per entity is illustrated in Figure 1.

To our knowledge, the only publicly available dataset for evaluating Chinese *is-a* relation extraction is published in (Fu et al., 2014), containing 1,391 *is-a* relations and 4,294 unrelated entity pairs. We use it to evaluate our method by splitting the dataset into 1/4 for validation and 3/4 for testing randomly.

4.2 Evaluation of Our Method

To tune the parameters of our method, we first run the K-means algorithm several times and train projection models. When we set the cluster number $K = 10$, our initial model achieves the best performance with a 73.9% F-measure. We also vary the value of parameter ϵ from 0.5 to 2 and find that the highest F-measure is achieved when $\epsilon = 1.05$.

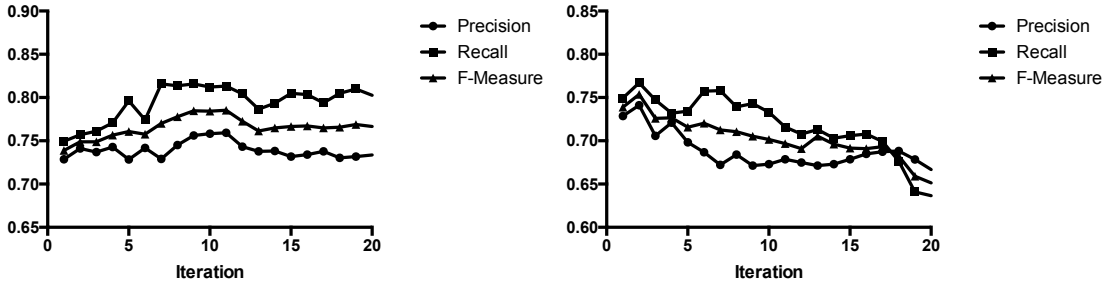
We report the performance of our method in 20 iterations to illustrate the effectiveness of the iterative process. We tune the parameters on the validation set and finally set $\delta = 0.2$, $\lambda = 0.5$ and add 500 new *is-a* relations into the training set in each iteration. In Figure 2(a), these new *is-a* relations are selected based on Algorithm 1. The F-measure increases from 74.9% to 78.5% in the first 10 iterations, which shows that newly extracted *is-a* relations can be of help to boost the performance of our models. The F-measure slightly drops and finally keeps stable after 15 iterations with F-measure around 76.7%. The possible cause of the drop is that a few false positive pairs are still inevitably selected by Algorithm 1 and added to the training set. After manual checking of these pairs, the average accuracy is 98.8%. Some of the erroneous cases include <脂肪(Fat), 健康(Health)>, <萧亚轩(Elva Hsiao), 时尚(Fashion)>, <信息(Information), 科学(Science)>, etc. They express *topic-of* relations rather than is-a relations. The performance becomes stable because the newly selected *is-a* relations tend to be similar to ones already in the training set after a sufficient number of iterations. In Figure 2(b), we directly sample 500 word pairs that are predicted as "positive" into our training set. Despite the slight improvement in the first iteration, the performance drops significantly because a large number of false positive instances are added to the training set for projection learning.

4.3 Comparison with Previous Methods

We evaluate our method and previous methods on the test set. The results are shown in Table 4.

We first re-implement three corpus-based *is-a* relation extraction methods on the *Baidu Baike* corpus. The pattern matching method for English *is-a* relations is originally proposed in (Hearst, 1992). For a Chinese corpus, we implement this method by employing Chinese Hearst-style patterns translated by Fu et al. (2013). The result shows that hand-craft patterns have low coverage for Chinese relation extraction

[5]http://nlpchina.github.io/ansj_seg/



(a) With the pattern-based relation selection method

(b) With no relation selection method

Figure 2: Performance of our method in each iteration.

because the language expressions are flexible. The automatic pattern detection approach in (Snow et al., 2004) improves the recall from 19.8% to 28.1%. However, the precision drops 28.9% because the syntactic parser for Chinese is still not sufficiently accurate, causing errors in feature extraction. The distributional similarity measure introduced in (Lenci and Benotto, 2012) has a 58.1% F-measure and is not effective for our task because the contexts of entities appeared in the free text are noisy. We also directly take the Chinese Wikipedia-based taxonomy (Li et al., 2015) to match *is-a* relations in the testing set. The result has a 98.5% precision but low recall due to the limited coverage of *is-a* relations in Chinese Wikipedia. The word embedding based approach in (Fu et al., 2014) achieves the highest F-measure 73.3% compared to all the previous methods. It shows the projection of word embeddings can model the semantics of Chinese *is-a* relations.

We now discuss our weakly supervised relation extraction method (WSRE) and its variants. In Table 4, *WSRE (Initial)* refers to the *is-a* extraction models trained in the initial stage. Although it is similar to (Fu et al., 2014), F-measure is improved by 2% because we consider both vector offsets and matrix projection in *is-a* representation learning, which is more precise. *WSRE (Random)*, *WSRE (Positive)* and *WSRE* employ the iterative learning process for *is-a* extraction. In *WSRE (Random)*, new *is-a* relations added to the training set are selected randomly from word pairs predicted as "positive" by our model. *WSRE (Positive)* considers only maximizing positive scores in relation selection, ignoring the effects of negative scores. *WSRE* is the full implementation of our method. Based on the results, the performance of *WSRE (Random)* decreases because of false positives in the training set. The F-measure of the latter two methods is increased by 2.3% and 3.3%, respectively, compared to *WSRE (Initial)*, which indicates that the proposed approach can improve prediction performance and generalization ability. *WSRE* outperforms *WSRE (Positive)* by 1% in F-measure, which shows the negative score constraint reduces the error rate in the relation selection process. Overall, our approach outperforms the state-of-the-art method (Fu et al., 2014) by 5.3% in F-measure. We further combine our method with the taxonomy (*WSRE+Taxonomy*) and achieve an 81.6% F-measure, which also has a better performance than Fu's method combined with the extension of a manually-built hierarchy, as shown in (Fu et al., 2014).

4.4 Error Analysis

We analyze errors occurred in our algorithm. The majority of the errors (approximately 72%) stems from the difficulty in distinguishing *related-to* and is-a relations. Some word pairs in the test set have very close semantic relations but are not strictly *is-a* relations. Such cases include <中药(Traditional Chinese medicine), 药草(Herb)>, <元帅(Marshal), 军事家(Strategist)>, etc. For example, most major components in traditional Chinese medicine are herbs, however, "药草(Herb)" is not a hypernym of "中药(Traditional Chinese medicine)" from a medical point of view. These cases are difficult to handle without additional knowledge. The errors in the iterative learning process (discussed in Section 4.2) also contribute to inaccurate prediction of this type.

The rest of the errors are caused by the inaccurate representation learning for fine-grained hypernyms. Take an example of the hyponym "兰科(Orchids)" in the test set, our algorithm recognizes that "植物(Plant)" is a correct hypernym, but it fails for "单子叶植物纲(Monocotyledon)". The possible causes

Method	Precision (%)	Recall (%)	F-Measure (%)
Previous Methods			
Hearst (Hearst, 1992)	96.2	19.8	32.8
Snow (Snow et al., 2004)	67.3	28.1	39.6
Taxonomy (Li et al., 2015)	**98.5**	25.4	40.4
DSM (Lenci and Benotto, 2012)	48.5	58.1	52.9
Embedding (Fu et al., 2014)	71.7	74.9	73.3
Our Method and Its Variants			
WSRE (Initial)	74.1	76.7	75.3
WSRE (Random)	69.0	75.7	72.2
WSRE (Positive)	75.4	80.1	77.6
WSRE	75.8	81.4	78.6
WSRE+Taxonomy	78.8	**84.7**	**81.6**

Table 4: Performance comparison between different methods.

is that "单子叶植物纲(Monocotyledon)" rarely appears in the corpus and is not well represented in the embedding space. We will improve learning of word and relation embeddings in the future.

5 Conclusion and Future Work

In this paper, we propose to extract Chinese *is-a* relations from user generated categories. Specifically, the task can be divided into two steps: initial model training and iterative learning. In the initial stage, word embedding based piecewise linear projection models are trained on a Chinese taxonomy to map entities to hypernyms. Next, an iterative learning process combined with a pattern-based relation selection algorithm is introduced to update models without human supervision. Experimental results show that this approach outperforms state-of-the-art methods. However, our experiments illustrate that free-text Chinese relation extraction still suffers from low coverage. We aim to address this issue by learning generalized pattern representations under the guidance of existing relations in the future.

Acknowledgements

This work is supported by the National Key Research and Development Program of China under Grant No. 2016YFB1000904.

References

Rahul Bhagat, Patrick Pantel, and Eduard H. Hovy. 2007. LEDIR: an unsupervised algorithm for learning directionality of inference rules. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 161–170.

Jingsheng Cai, Masao Utiyama, Eiichiro Sumita, and Yujie Zhang. 2014. Dependency-based pre-ordering for chinese-english machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 155–160.

Sharon A. Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *27th Annual Meeting of the Association for Computational Linguistics*.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010a. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*.

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

Daoud Clarke. 2012. A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, 38(1):41–71.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 601–610.

Oren Etzioni, Michael J. Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2004. Web-scale information extraction in knowitall: (preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, pages 100–110.

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.

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.

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.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *14th International Conference on Computational Linguistics*, pages 539–545.

Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. 2015. Short text understanding through lexical-semantic analysis. In *31st IEEE International Conference on Data Engineering*, pages 495–506.

Samir Khuller, Anna Moss, and Joseph Naor. 1999. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Zornitsa Kozareva and Eduard H. Hovy. 2010. Learning arguments and supertypes of semantic relations using recursive patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1482–1491.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 543–546.

Hai-Guang Li, Xindong Wu, Zhao Li, and Gong-Qing Wu. 2013. A relation extraction method of chinese named entities based on location and semantic features. *Appl. Intell.*, 38(1):1–15.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 746–751.

George A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the Acm*, 38(11):39–41.

Rosa M. Ortega-Mendoza, Luis Villaseñor Pineda, and Manuel Montes-y-Gómez. 2007. Using lexical patterns for extracting hyponyms from the web. In *Proceedings of Mexican International Conference on Advances in Artificial Intelligence*, pages 904–911.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

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.

Alan Ritter, Stephen Soderland, and Oren Etzioni. 2009. What is this, anyway: Automatic hypernym discovery. In *Learning by Reading and Learning to Read, the 2009 AAAI Spring Symposium*, pages 88–93.

Erik F. Tjong Kim Sang. 2007. Extracting hypernym pairs from the web. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems 17, NIPS 2004*, pages 1297–1304.

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.

Idan Szpektor, Eyal Shnarch, and Ido Dagan. 2007. Instance-based evaluation of entailment rule acquisition. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, page 456–463.

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.

Chengyu Wang, Ming Gao, Xiaofeng He, and Rong Zhang. 2015. Challenges in chinese knowledge graph construction. In *31st IEEE International Conference on Data Engineering Workshops*, pages 59–61.

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.

Min-Chul Yang, Nan Duan, Ming Zhou, and Hae-Chang Rim. 2014. Joint relational embeddings for knowledge-based question answering. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 645–650.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained phrase embeddings for machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 111–121.

Huiwei Zhou, Long Chen, Fulin Shi, and Degen Huang. 2015. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 430–440.