An Iterative Method of Extracting Chinese ISA Relations for Ontology Learning

Lei Liu
College of Applied Sciences, Beijing University of Technology, Beijing, China
Email: liuliu_leilei@bjut.edu.cn

Sen Zhang[1], Luhong Diao[1], Cungen Cao[2].
[1]College of Applied Sciences, Beijing University of Technology, Beijing, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
Email: {zhangsen, diaoluhong}@bjut.edu.cn, cgcao@ict.ac.cn

***Abstract*— Automatic acquisition of ISA relations is a basic problem in knowledge acquisition from text. We present an iterative method extracting ISA relations from large Chinese free text for ontology learning. Firstly, it initially discovers a set of sentences using several special Chinese lexico-syntactic patterns from free text corpus. Secondly we combine outside layer removal and inside layer gathering for acquiring concepts of constituting ISA relation. Finally, ISA relations are verified with multiple features. Extracted ISA relations will be selected for new relation extracting cycle. Experimental results demonstrate good performance of the method for extracting ISA relation from large Chinese corpus.**

***Index Terms*—Relation Acquisition, Knowledge Acquisition, Information Extraction, Ontology Learning, ISA Relation.**

I. INTRODUCTION

Automatic acquisition of concepts and semantic relations from text has received much attention in the last ten years. Especially, ISA relation acquisition is a more interesting and fundamental because ISA relation play a crucial role in various NLP (Natural Language Processing) systems, such as systems for information extraction, information retrieval, and dialog systems. ISA relations are important in accuracy verification of ontologies, knowledge bases and lexicons [1] [2].

The types of input used for ISA relation acquisition are usually divided into three kinds: the structured data or text (e.g. database), the semi-structured data or text (e.g. dictionary), and free text (e.g. Web pages). Human knowledge is mainly presented in the format of free text at present, so processing free text have become a crucial yet challenging research problem.

There are two main approaches for automatic/ semi-automatic ISA relation acquisition. One is pattern-based (also called rule-based), and the other is statistics-based. The former uses the linguistics and natural language processing techniques (such as lexical and parsing analysis) to obtain hyponymic patterns, and then makes use of pattern matching to acquire ISA relation, and the latter is based on corpus and statistical language model, and uses clustering algorithm to acquire ISA relation.

At present the pattern-based approach is dominant. Among hyponymic patterns, "isa" patterns are more important. In this paper we present a iterative method of extracting ISA relation. Experimental results show that the method is adequate of extracting ISA relation from Chinese free text.

The rest of the paper is organized as follows. Section 2 describes related work, section 3 elaborates on Chinese-isa patterns for this work, section 4 presents an iterative method of extraction ISA relations, section 5 conducts a performance evaluation of the proposed method, and finally section 6 concludes the paper.

II. RELATED WORK

ISA relation is a semantic relation between concepts. Given two concepts x and y, there is the ISA relation between x and y if the sentence "x is a (kind of) y" is acceptable. ISA relation is also called as subordination, or hyponymy. We denote an ISA relation as ISA(x, y). For example ISA(apple, fruit).

At present the pattern-based approach is dominant, and its main idea is ISA relations can be extracted from text as they occur in detectable syntactic patterns. The so-called patterns include special idiomatic expressions, lexical features, phrasing features, and semantic features of sentences. Patterns are acquired by using the linguistics and natural language processing techniques. Using these acquired patterns, we can use pattern matching to acquire ISA relation.

There have been many attempts to develop automatic methods to acquire ISA relations from text corpora. One of the first studies was done by Hearst[3]. Hearst proposed a method for retrieving concept relations from unannotated text (Grolier's Encyclope-dia) by using predefined lexico-syntactic patterns, such as

…NP_1 is a NP_2… ---ISA (NP_1, NP_2)
…NP_1 such as NP_2… ---ISA(NP_2, NP_1)
…NP_1 {, NP_2}*{,} or other NP_3 …---ISA (NP_1, NP_3), ISA (NP_2, NP_3)

Manuscript received July 1, 2009; revised August 1, 2009; accepted September 21, 2009.

doi:10.4304/jcp.5.6.870-877

Other researchers also developed other ways to obtain ISA relations. Most of these techniques are based on particular linguistic patterns.

Caraballo used a hierarchical clustering technique to build a ISA relations hierarchy of nouns like the hypernym-labeled noun hierarchy of WordNet automatically from text [4]. Nouns are clustered into a hierarchy using data on conjunctions and appositives appearing in text corpus. The internal nodes of the resulting tree are labeled by the syntactic constructions from Hearst [3].

Morin and Jacquemin produced partial ISA relations hierarchies guided by transitivity in the relation, but the method works on a domain-specific corpus [5]..

Llorens and Astudillo presented a technique based on linguistic algorithms, to construct hierarchical taxonomies from free text. These hierarchies, as well as other relationships, are extracted from free text by identifying verbal structures with semantic meaning [6].

Sánchez presented a novel approach that adapted to the Web environment, for composing taxonomies in an automatic and unsupervised way. [7].

Elghamry showed how a corpus-based hyponymy lexicon with partial hierarchical structure for Arabic can be created directly from the Web with minimal human supervision. His method bootstraps the acquisition process by searching the Web for the lexico-syntactic patterns [8].

III. Multiple Chinese ISA Relation Patterns

In Chinese, one may find several hundreds of different ISA relations patterns based on different quantifiers and synonymous words, which is equivalent to the single ISA pattern (i.e. (<?C1> is a <?C2>), (<?C3> such as <?C1>,<?C2>)) in English. Fig.1 depicts a few typical Chinese ISA relation patterns.

```
Pattern 1: 〈?C1><是一|为一><种|个|名|篇|片|块|部|颗|本
|…><?C2>
 (Pattern: 〈?C1〉 is a 〈?C2〉)

Pattern 2: <?C3><如|象|包括|包含|囊括|涵盖><?C1>{<或|
及|以及|和|与|、><?C2>}*<等>
 (Pattern: 〈?C3〉 such as 〈?C1〉,〈?C2〉…)
```

Figure 1. Defining Chinese ISA patterns

In Fig.1, Pattern 1 means "Pattern: <?C1> is a <?C2>". Pattern 2 means "Pattern: <?C3> such as <?C1>, <?C2>…". These items of the pattern are divided into constant item and variable item. Constant item is composed of one or more Chinese words or punctuations. Variable item is a non-null string variable. "<?C1>" is a variable item in the pattern. "|" expresses logical "or". In pattern1,"是一|为一"means "is a"; "种|个|名|篇|片|块|部|颗|本…" is a group of quantifiers. In pattern 2, "如|象" means "such as"; "包括|包含|囊括|涵盖" means "include"; "和|与" means "and"; "或" means "or"; "等" means "etc."; "及|以及" denotes "as well as"; Chinese dunhao "、" is a special kind of Chinese comma used to set off items in a series.

Chinese ISA relation patterns will be used to capture concrete sentences from Chinese free corpus. In this process, variables <?C> will be instantiated with words or phrases in a sentence, in which real concepts may be located. Let c and c' be the real concept in <?C>. If ISA(c, c′) is true, then we tag c by c_L, and c′ by c_H, as shown below.

{众所周知，{中国}c_L}$_{<?C1>}$/是一个/{社会主义{国家}c_H}$_{<?C2>}$

({It is well-known that {China}c_L}$_{<?C1>}$/is a/ { socialist nation}c_H }$_{<?C2>}$)

{{农作物}c_H 主要}$_{<?C5>}$/有/{{水稻}c_L}$_{<?C1>}$、{{玉米}c_L}$_{<?C2>}$、{{红薯}c_L}$_{<?C3>}$、{{烟叶}c_L}$_{<?C4>}$/ 等/

(The farm crop mainly includes paddy rice, corn, sweet potato, tobacco leaves etc.)

We can acquire ISA(中国,国家), ISA(水稻,农作物), ISA(玉米,农作物), ISA(红薯,农作物) and ISA(烟叶,农作物) from the above example.

IV. Our Method

Our method consists of four phases. In Phase I, we preprocess the raw corpus. In Phase II, we present a semi-automatic method for acquiring and analyzing non-concept composition, and converting them into removable lexicon and sentence patterns. We use an algorithm that combines outside layer removal and inside layer gathering to extract concepts c_L and c_H. Phase III, we combine multiple features of ISA relation together for ISA relation verification, and each relation that satisfies the threshold is stored in the ISA relation database. In the final phase, some seed relations that extracted from verified relations randomly are used to retrieve new corpus from the Web based on the Google API. The new corpus will be processed using our method again. The algorithmic framework is presented in Fig.2.

A. Building Corpus

Raw corpus is gathered from Chinese free text, and is preprocessed in a few steps, including word segmentation, part of speech tagging, and splitting sentences according to periods. Then we acquire the processed corpus by matching Chinese ISA relation patterns. Finally, processed corpus is divided into two groups: training corpus and testing corpus.

B. Concept Acquisition

As we know, Chinese is a language different from any western language. A Chinese sentence consists of a string of characters which do not have any space or delimiter in between [9]. And for ensuring the recall of ISA relation, we define Chinese-isa patterns without extra restriction rules. To handle these features, we have developed the following strategies.

We use an algorithm that combines outside layer removal and inside layer gathering to acquire concept c_L and c_H.

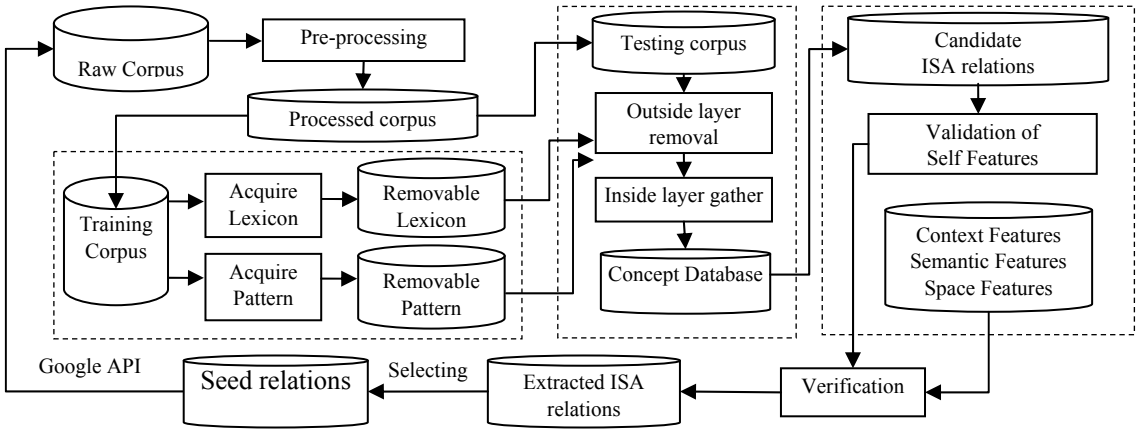

Figure 2. Framework of extract ISA relations

Outside layer removal: There are many non-concept components in <?C>. Most concepts are compound words without explicit boundaries, but the composition of the non-concept components is more fixed. So we make use of lexicon and sentence patterns on <?C> to carry on the outside layer processing. The processing is to remove non-concept components from the outside layer to inside layer just like peeling an onion.

Inside layer gathering: After the outside layer removal to <?C >, we continue to analyze the structure of the remainder by lexical analysis (such as word segmentation, and part of speech tagging), and make use of these information as the proof to judge whether the concepts are correct.

We use a semiautomatic method to acquire removable lexicon and sentence patterns. Let s be a sentence that matches Chinese ISA patterns, w be a word, p be a pattern. If w belongs to a part of non-concept components in s, then w is called removable word(w_r). If p belongs to a part of non-concept components in s, then p is called removable pattern (p_r).

The details of acquisition of removable lexicon and removable patterns are presented in algorithm 1.

Algorithm 1. Acquisition of removable lexicon and removable patterns

Let D be a list of removal words initially empty

Let P be a list of removal patterns initially empty

Input**:** Training corpus Cor_{train}, thresholds θ_1 and θ_2

Step1: Let $S = \{s_1, s_2, \ldots, s_n\}$ be a list of separate string in Cor_{train}.

Step2: For each pair of strings $s_1, s_2 \in S$, compute common substrings S_{sub}.

If there exists $s_{sub} \in S_{sub}$ is a common prefix or suffix of s_1 and s_2, add s_{sub} to D.

If there exists $S'_{sub} \subseteq S_{sub}$ satisfy conditions (i) the element amount of S'_{sub} >=2; (ii) exist a common substring is a prefix or suffix of s_1 or s_2, then add S'_{sub} to P.

Step3: Automatic filter D and P according to a set of rules (such as the length of common substring $<\theta_1$, the frequency of common substring $> \theta_2$)

Step4: Attached additional rules using the interactive manual,.

Output: removal lexicon and sentence patterns

An example of removable lexicon is as follows:

Removable **word**	**position**	**tense**
即将(be about to)	t	f
的确(certainly)	a	c
譬如(for example)	h	c

In above removable lexicon, we also use other tow tags, i.e. position and tense. The position tag indicates where the lexical term may possibly appear relatively to the c_L or c_H in a sentence. It takes h (before), t (after), and a (before or after). The tense tag indicates when the ISA(c_L, c_H) is true.

A removable sentence pattern is defined as follows.

removable pattern 005{

Pattern: <?w1><如|就如|正如|恰如|正像|像|就像|正向|正像><?w2><评价|所述|所说|一样|所言|所示|所讲|所见><，|。|. |？|！|,|.|?|!|；|;><?w3>

}

The algorithm combines outside layer removal and inside layer gathering to acquire concepts of constituting ISA relation. The details are described in algorithm 2.

Algorithm 2. Acquiring concepts

Let R be a list of candidate relations initially empty

Input: Testing corpus Cor_{test}, removal lexicon D, sentence patterns P

Step1: For each sentence $s \in Cor_{test}$, process <?C> according to Step2 – Step5. If all sentences have been processed, jump to Step6.

Step2: Make use of D and P to carry the removal transaction on <?C>.

(1) Discover patterns that match with <?C> using P. **If exist** a pattern, add elimination pattern tag to <?C>

(2) According to the principle of word length precedence using D, carry on the processing that remove non-concept components from <?C> outside layer to inside layer until it can't discover any removal word further.

Step3: Process tagged sentence. If <?C> is tagged completely, jump to Step2

Step4: Gather noun phases and remove adjective fractions.

Step5: Acquire candidate c_L and candidate c_H that is no-tagged components of <?C>, and add them to R

Step6: Return R

An example of acquisition concepts using removable lexicon and sentence patterns is shown below:

{正如页框一样，}$_{p\_e}$ 表格{也}$_{w\_e}$/是一种/{很好的}$_{adj}$ 容器对象

({Just as the page frame,}$_{p\_e}$ the table / is also a / {good}$_{adj}$ container object.)

C. Verification of ISA relation

There are still many error relations in the acquired ISA relations from free text. They will affect the building of ontologies. For the error relations in the phase of acquiring ISA relation, a verification method of ISA relations based on multiple features is given.

When a candidate ISA relation is correct or error, it often satisfies some features. We combine the semantic features, context features and space structure features of ISA relation together. If a candidate ISA relation satisfies a certain threshold with matching those features, we think that it is a real ISA relation. The features of ISA relation are defined as follows:

Definition 1: The feature of ISA relation is a 3-tuple ISAF= {SMF, CTF, STF}, where SMF is a group of semantic features, CTF is a group of context features and STF is a group of space structure features.

(1) SMF is constructed by the assumption that x and y are semantically similar in ISA(x, y). A candidate ISA relation can be verified by computing the semantic similar measure between x and y. SMF is subdivided into three features, i.e. SMF= {WF, SF, AF}, where WF represents word-formation feature, SF represents synonymous word feature and AF represents attribute feature.

(a) WF (word-formation feature)

As we know, Chinese word order is strict; Chinese lacks morphological change, and has no the explicit variety tag of plural, possessive and part of speech. A concept consists of one or several certain sequence Chinese characters. To some extent, Chinese characters can appear the semantic feature of concept.

So for each pair of candidate ISA relation (c_1, c_2), we assume the common substrings between c_1 and c_2 could imply the semantic similar measure between them. If there exists some substrings, the position (such as prefix and suffix), length and amount of substrings will provide the evidence for the existence of a ISA relation.

Given a candidate ISA relation (c_1, c_2), where $c_1=a_1a_2a_3\ldots a_n$, $c_2=b_1b_2b_3\ldots b_m$, the a_i and b_i both represent single character. We can give some features as follows:

CoChar(c_1, c_2) = $\{a_1, a_2, a_3, \ldots, a_n\}\cap\{b_1, b_2, b_3, \ldots, b_m\}$, |CoChar($c_1$, c_2)| denote the number of common character of c_1 and c_2.

Obviously, the above definition doesn't consider the sequence of character. So we add the following definitions.

CoPrefix(c_1, c_2) = $a_1a_2\ldots a_i$, where $a_1a_2\ldots a_i = b_1b_2\ldots b_i$, $a_{i+1}\neq b_{i+1}$, i<=n, i<=m. Especially, if CoPrefix(c_1, c_2)= c_2, then c_2 is the prefix c_1; if CoPrefix(c_1, c_2)= c_1, c_1 is the prefix c_2.

CoSuffix(c_1, c_2) = $a_j\ldots a_{n-1}a_n$, where $a_{n-j}\ldots a_{n-1}a_n = b_{m-j}\ldots b_{m-1}b_m$, $a_{n-j-1} \neq b_{m-j-1}$, i<n, i<m. Especially, if CoSuffix(c_1, c_2)= c_2, then c_2 is the suffix c_1; if CoSuffix (c_1, c_2)= c_1, c_1 is the suffix c_2.

For example:

CoSuffix (定时*炸弹*, *炸弹*) =*炸弹*

(CoSuffix (Time bomb, bomb) = bomb)

CoChar (话费优惠*业务*, *业务*套餐) ={*业*,*务*}

(CoChar (cheap-charge-of-calls service, service) = {service, plan})

CoPrefix(诗人屈原, 诗人) =诗人

(CoPrefix (poet QuYuan, poet) = poet)

(b) SF (synonymous word feature): We can use Cilin(a dictionary of Chinese synonymous words) to compute the semantic similarity of candidate (c_1, c_2). Cilin provides the mandarin synonym sets in a hierarchical structure. It contains approximately 70,000 Chinese words, and describes a five levels semantic hierarchy from common word to concrete word [10]. The fifth level is for the basis of synonym feature words. We denote Syn(c_1,c_2) as the common synonymous words between c_1 and c_2. "医生"(doctor) and "大夫"(doctor) are synonymous word in the below example.

Syn(主治*医生*, *大夫*) = {*医生*|*大夫*}

(Syn(doctor in charge of a case, doctor)= {doctor|doctor})

(c) AF (attribute feature): The attributes of concept can be used to discriminate different concept. If two concepts have the same attributes, they should be semantic similar. The attributes of concept can be acquired by an attribute acquiring system [11]. We denote common attributes as CoAttr(c_1,c_2). For example:

CoAttr(黄河, 河流) = {上游}

(CoAttr(yellow river, river) = { upriver })

CoAttr (比利时, 国家) = {面积, 首都, 人口}

(CoAttr (Belgium, country) = {area, capital, population})

(2) CTF is subdivided into two features, CTF= {FF, DF}.

(a) FF (frequency features): If candidate (c_1, c_2) appears frequently in a kind of ISA relation pattern or in various ISA relation patterns, the probability of ISA(c_1, c_2) is higher. The type number of pattern that can acquire (c_1, c_2) is denoted by lpf(c_1,c_2). The total of number of pattern that can acquire (c_1, c_2) is denoted by lef(c_1, c_2). For example:

lpf (玉米, 农产品) = 4

(lpf(corn, farm produce)= 4)

lef (新加坡, 国家) = 109

(lef(Singapore, country) = 109)

(b) DF (domain features): Our corpus comes from Web and includes some error knowledge. We can acquire many error ISA relations, such as (美丽, 罪) ((beauty,

evil)). If candidate (c_1, c_2) appears in a certain scientific domain-specific context, (c_1, c_2) may be a true piece of scientific knowledge; otherwise it may be a pair of general concepts and may not have any value. The domain-specific context is discriminated with a domain dictionary [2]. Given a group of context $CT(c_1,c_2)$ = {ct_1,ct_2, …, ct_n}, where ct_i is the i context of (c_1,c_2), $fw(ct_i)$ is the number of domain word in ct_i, and $length(ct_i)$ is the byte length of ct_i. The classify formula is as follows.

$$\text{Classify}(c_1,c_2) = \frac{\sum_{i=1}^{n} fw(ct_i)}{\sum_{i=1}^{n} length(ct_i)} \times 1000 \quad (1)$$

For example:

Classify (人参, 药材) = 25.5

(Classify (panax, medicinal materials) = 25.5)

Classify (钾, 元素) = 65.4

(Classify (kalium, element) = 65.4)

(3) STF is a group of space structure features. When a group of candidate ISA relations are correct or error, they often satisfy some space structure feature. In space structure analysis, we use the coordinate relation between concepts. The coordinate relations are acquired using a set of coordinate relation patterns including "、". Chinese dunhao "、" is a special kind of Chinese comma used to set off items in a series. For example:

In a sentence of matching a coordinate pattern, if there exists concept c_1 and concept c_2 divided by "、", then c_1 and c_2 are coordinate, denoted as cr(c_1, c_2). An example is as shown below.

农作物主要有{水稻}c_1、{玉米}c_2、{红薯}c_3、{烟叶} c_4等

(The farm crop mainly includes paddy rice, corn, sweet potato, tobacco leaves etc..)

cr(水稻, 玉米, 红薯, 烟叶) (cr(paddy rice, corn, sweet potato, tobacco leaves)) is acquired from the above example.

We can add some space structure features on the basis of above coordinate relations. A few important features are as follows:

Structure 1: (c_1, c_2), (c_2, c_3), (c_1, c_3). For example:

(番茄, 蔬菜), (番茄, 食品), (蔬菜, 食品)

((tomato, vegetable), (tomato, food), (vegetable, food))

Structure 2: (c_1, c_2), (c_2, c_3), (c_3, c_1). For example:

(游戏, 生活), (生活, 童话), (童话, 游戏)

((game, life), (life, fairy tale), (fairy tale, game))

Structure 3: (c_1, c), (c_2, c),…, (c_m, c), cr(c_1, c_2, …, c_m), $\exists c_i \in \{c_1, c_2, \ldots, c_m\}$, CoSuffix ($c_i$, c). For example:

c=*马*, cr(千里*马*, 骏*马*, 老骥, 白驹), CoSuffix(千里*马*, *马*), CoSuffix(骏*马*, *马*)

(c=horse, cr(swift horse, courser, nag, white horse), CoSuffix(swift horse, horse), CoSuffix(courser, horse))

Structure 4: (c_1, c), (c_2, c), …, (c_m, c), cr(c_1, c_2, …, c_m), $\exists c_i \in \{c_1, c_2, \ldots, c_m\}$, lpf ($c_i$, c) ≥1. For example:

c=职业, cr(警察, 医生, *歌手*, 护士), lpf(*歌手*, 职业)=3

(c=profession, cr(policeman, doctor, singer, nurse), lpf(singer, profession)=3)

Structure 5: (c_1, c), (c_2, c), …, (c_m, c), cr(c_1, c_2, …, c_m), (c'_1, c), (c'_2, c), …, (c'_n, c), cr(c'_1, c'_2, …, c'_n), $\{c_1, c_2, \ldots, c_m\} \cap \{c'_1, c'_2, \ldots, c'_n\} \neq \varnothing$.

For example:

c=职业, cr(时装, 休闲装, 礼服), cr(裤装, 时装, 休闲服) $\{c_1,\ldots, c_m\} \cap \{c'_1,\ldots, c'_n\}$ = {时装, 休闲装}

(c= clothing, cr(fashionable dress, sportswear, full dress),(trousers, fashionable dress, sportswear), $\{c_1,\ldots, c_m\} \cap \{c'_1, \ldots, c'_n\}$ = {fashionable dress, sportswear})

Because the above features are all uncertainty knowledge, they must be converted into a set of creation type heuristic rules used in uncertainty reasoning. Here we use CF (certainty factors) that is the most common approach in rule-based expert system. The CF formula is as follows:

$$CF(CR, f) = \begin{cases} \dfrac{P(CR|f)-P(CR)}{1-P(CR)}, & P(CR|f) \geq P(CR) \\ \dfrac{P(CR|f)-P(CR)}{P(CR)}, & P(CR|f) < P(CR) \end{cases} \quad (2)$$

Where CHR is a set of candidate ISA relations, which has a precision P(CHR). P(CHR|f) is the precision of a subset of CHR satisfying feature f. CF is a number in the range from -1 to 1. If there exists CF(CHR, f)≥0, then we denote f as positive feature and CF(CHR, f) denotes the support degree of feature f; if there exists CF(CHR, f)<0, then we denote f as negative feature and CF(CHR, f) denotes the no support degree of feature f. We take word-formation feature as example. In the word-formation feature test, P(CHR)=0.74.

The precision of candidate ISA relations satisfying the feature |CoSuffix(c_1, c_2)|= c_2 is 98%, namely P(CHR|f)=0.98, the result of CF is (0.98-0.74)/(1-0.74)=0.92. The f is a positive feature.

The precision of candidate ISA relations satisfying the feature | CoPrefix (c_1, c_2)|= c_1 is 1%, namely P(CHR|f)=0.01, the result of CF is (0.01-0.74)/ 0.74= -0.99. The f is a negative feature.

After those features are converted into a set of production rules, we can carry uncertainty reasoning in candidate ISA relations.

The iterative verification of ISA relations can be realized by a production system. The rule database of production system is composed of the above production rules. Here we mainly focus on the control mechanism of the whole iterative verification, that is, how to select the rules which can be activated and use these rules to update the CF value of candidate ISA relations. The basic control process is shown in algorithm 3.

Algorithm 3. The basic process of iterative verification of ISA relations

Input*:* the set of candidate ISA relations CHR, the set of production rules Rule, the initial judgment threshold α, the incremental threshold β, the terminal threshold γ;

Output*:* the set of correct ISA relations HR, the set of error ISA relations FR.

Step1*:* For each ISA relation $r \in$CHR, set its certainty factor CF(*r*) to be 0;

Step2*:* For each ISA relation $r \in$CHR, continue Step3 - Step4;

Step3*:* Find the production rules *rulelist*$\subseteq$Rule which r can satisfy.

Step4*:* Execute all the rules in *rulelist* and modify the CF of *r*. The execution orders of rules may lead to the different CF. So these rules can be executed in order according to the descending sort of certainty factors.

Step5*:* If there not exists $r \in$CHR, it satisfy CF(r)<α, goto Step7.

Step6*:* Move each r that satisfies CF(r)<α from CHR to FR. Then set the certainty factor of r in CHR to be 0 and goto Step2.

Step7*:* If α<γ, then α=min{α+β,γ}, for each r∈CHR, set CF(r) =0 and goto Step2. If α≥γ, then move each r∈CHR to HR.

Step8*:* return HR and FR.

In algorithm 3, if CF(r)<α, r is an error ISA relation. When the error ISA relation r is deleted from CHR, the space structure of CHR has changed. So it need to rerun the rules until the number of ISA relations which satisfy CF(r)<α to be 0. Then the first inside cycle is terminated.

Next, α is updated using β. If α<γ, continue the next inside cycle. The outside cycle is end until α≥γ, and the algorithm terminates. We adopt two-layer cycle instead of directly using γ as the judgment threshold. Because the certainty factors of some correct ISA relations may decrease by error ISA relations which activates the space feature rules. So it is necessary to remove the ISA relations with the lowest CF in every cycle. It can avoid the removal of correct ISA relations.

D. Iterative extracting ISA relations

The ISA relations in HR are considered as the final correct ISA relations. Some seed relations that extracted from HR randomly will be changed into a group Chinese query strings with ISA patterns. We used those query strings to retrieve more corpus from the Web based on the Google API. The new corpus will be processed using our method again.

The iterative process will be terminated until satisfied some threshold, such as the cycle times, the scale of ISA relations.

V. EVALUATION

We adopt three kinds of measures: R (Recall), P (Precision), and F (F-measure). They are typically used in information retrieval. Let H be the total number of correct ISA relations in the test corpus. Let H_1 be the total number of ISA relations acquired. Let H_2 be the total number of correct ISA relations acquired. We can give the measure of evaluation metrics as follows:

(1) Recall is the ratio of H_2 to H, i.e. $R = H_2/H$

(2) Precision is the ratio of H_2 to H_1, i.e. $P = H_2/H_1$

(3) F-measure is the harmonic mean of precision and recall, i.e. $F = 2RP/(R+P)$

TABLE I
THE PROCESSED RESULT OF TESTING CORPUS

	Number(ratio)	P	R	F
Testing corpus	38,458(100%)	58.6 %	100 %	69.8%
Out layer removal	30,455(79.2%)	69.5%	93.9 %	79.9%
Inside layer gather	28,672 (74.7%)	72.1 %	91.7 %	80.7%

A. Concept Acquisition Analysis

Firstly, we used the People's Daily Corpus (1998-1999, 91503 papers) as raw corpus. Processed corpus contains about 190,000 sentences acquired by matching Chinese ISA patterns. Then we divided it into two groups: training corpus (80%) and testing corpus (20%). After the training corpus was processed, we acquired 1892 removable words and 73 removable patterns. In concept processing phase, we manually evaluated the processed result of testing corpus. The detailed result is shown in Table I.

In the following, we analyze data from three aspects.

(1) Out layer removal

We use removable words and sentence patterns to acquire concepts of constituting ISA relation.

As we can see from Table I, it has a higher precision (69.5%) than original testing corpus (58.6%) after out layer removal. Some sentences tagged completely are filtered and are no longer processed further. Though the recall is decreased to 93.9%, filtered sentences are important to improve precision. An example is as follows:

{当然}$_{w\_r}${现在}$_{w\_r}${只}$_{w\_r}$/是一种/意向.

({Certainly}$_{w\_r}$ {it}$_{w\_r}$ {now}$_{w\_r}$ { just}$_{w\_r}$ /is a kind of/ intention).

Partially tagged sentences are removed tag directly. The effect of removal is satisfactory mostly. An example is shown as follows:

{例如}$_{w\_r}${清末的}$_{p\_r}$ 任伯年{便}$_{w\_r}$/是一位/{雅俗共赏的}$_{p_r}$大画家.

({For example,}$_{w\_r}$ Ren Bo Nian {in the end of the Qing Dynasty}$_{p\_r}$ {surely}$_{w\_r}$ /is a / painter {of suit both refined and popular tastes.}$_{p\_r}$)

(2) Inside layer gather

The removed tag and no tagged sentences continued to be processed with inside layer gather. Gathered sentences are the sentences under the influence of gathering noun phase (n) and removing adjective fraction (adj) according to the result of part of speech. To some extent, the gather makes up the shortage of the removal. For example:

{美丽富饶的}$_{adj}$ {海南}$_{n}$ /是一座/{历史悠久的}$_{p\_e}${岛屿}$_{n}$.

({Beautiful and resourceful}$_{adj}$ {Hainan}$_{n}$ /is an/ {island}$_{n}$ {with long history}$_{p\_r}$)

A group of candidate ISA relations were acquired after out layer removal and inside layer gather, and had a precision of 72.1% and a recall of 91.7%.

TABLE II
THE RESULT OF ITERATIVE VERIFICATION

Input: CHR 28,672 P(CHR)=72.1% α=-0.8 β=0.3 γ=0			
Threshold		Result	
		HR	The add of FR
α= -0.8			
	The first cycle	28,672	386
	The second cycle	28,286	12
α= -0.5			
	The first cycle	28,274	644
	The second cycle	27,630	28
	The third cycle	27,602	4
α= -0.2			
	The first cycle	27,598	612
	The second cycle	26,986	102
	The third cycle	26,884	12
α= 0			
	The first cycle	26,872	522
	The second cycle	26,350	12
HR: 26,338 (91.9%) P(HR)=77%, R(HR)=98.1%, F(HR)=86.3% FR:2,334 (8.1%) P(FR)=16%, R(FR)=1.8%, F(FR)=3.2%			

B. Relation Verification Analysis

We acquired candidate ISA relations CHR after concept acquisition. We manually evaluated the initial set CHR and the classified sets. The detailed result is shown in Table II.

As we can see from Table II, there are 2,334 relations in FR finally after 4 outside cycle and 10 inside cycle. Because FR saves error relations, its recall and precision must be very low. FR has the precision of 16% and recall of 1.8%. That is to say, our methods can throw away many error ISA relations under the condition of skipping a few correct relations. HR saves correct relations and has the precision of 77% and recall of 98.1%. If we want to increase the precision of HR, we can augment γ value.

For analyzing the influence of threshold γ, we choose several different values. The detailed result is shown in Table III.

As we can see from Table III, there are 28,672 ISA relations initially. With the increase of threshold γ, the precision is also increase. If we want to increase the precision, we can augment γ value. For example, when γ=0.8, the precision is up to 95%, and but its recall decreases to 26%. That is to say, when threshold γ is a small value, our methods can throw away many error ISA relations under the condition of skipping a few correct relations. But when threshold γ is a large value, our methods can throw away many error ISA relations and also skip many correct relations at same time.

C. Iterative Extracting Relations Analysis

TABLE III
THE INFLUENCE OF THRESHOLD γ

Input: CHR 28,672 P(CHR)=72.1% α=-0.8 β=0.3				
γ	The result of verified ISA relations			
	The number	P	R	F
γ=0	26,338(91.9%)	77%	98%	0.86
γ=0.2	19,268 (67.2%)	81%	76%	0.78
γ=0.4	15,568(54.3%)	88%	66%	0.75
γ=0.6	8,658(30.2%)	92%	39%	0.54
γ=0.8	5,562(19.4%)	95%	26%	0.41

TABLE IV
THE RESULT OF ITERATIVE CYCLE

Input: test corpus 38,458 α=-0.8 β=0.3 , θ =6				
cycle	The result of extracted ISA relations			
	γ=0.2, P		γ=0.8, P	
1	19,268	81%	5,562	95%
2	72,466	80%	10,246	94%
3	120,564	79%	20,554	94%
4	162,344	79%	34,855	92%
5	193,223	78%	43,555	92%
6	215,588	77%	54,745	90%

We selected ISA relations from CHR(satisfy condition α=-0.8, β=0.3, γ=0.8) as seed relations, and changed them into a group Chinese query strings with ISA patterns. We used those query strings to retrieve more corpus from the Web based on the Google API. The iterative process will be terminated until the cycle times θ =6. The detailed result is shown in Table IV.

As we can see from Table IV, with the increase of cycle time, the number of extracted ISA relations is also increase. There are 215,588 extracted ISA relations under γ=0.2 cycle=6. But its precision decreases to 77%, and the precision of seed relations also decreases to 90%. The reason is that some error ISA relations were extracted from new corpus based on Google API, and then they were selected as seed relations. Furthermore, more error ISA relation may be extracted by next cycle.

VI. CONCLUSION

Automatic acquisition of ISA relations is regarded as a basic problem in knowledge acquisition from text. In this paper we present an iterative method extracting ISA relations from large Chinese free text for ontology learning. We combine outside layer removal and inside layer gathering for acquiring concepts of constituting ISA relation. ISA relations are verified with multiple features. Extracted ISA relations were selected for new relation extracting cycle. Experimental results demonstrate good performance of the method for extracting ISA relation from large Chinese corpus.

Because the language and corpus is different from other work, it is difficult in the comparison of the proposed method with previous approaches.

There are still some inaccurate relations in the result. There are mainly two reasons to cause those errors. First, the structure of a sentence is so complicated that removable patterns can't handle, and more syntactical information will be helpful for resolving this problem. Second, relations may represent a kind of metaphor, and more sophisticated verification methods are needed.

So it is necessary for us to solve some problems, such as the ambiguity of tag, the reasonable usage of semantic information, and the polysemy of concept word. In future, we will combine some methods (such as the morpheme analysis, web page tag, concept space etc.) to the further verification of ISA relation.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China under Grant No.60573064, and 60773059; the National 863 Program under Grant No. 2007AA01Z325, and the Beijing University of Technology Science Foundation (grant nos. X0006014200803, 97006017200701).

REFERENCES

[1] Beeferman D, Lexical discovery with an enriched semantic network, In Proceedings of the Workshop on Applications of WordNet in Natural Language Processing Systems, ACL/COLING, 1998, pp.358--364.

[2] Cungen Cao and Qiuyan Shi, Acquiring Chinese Historical Knowledge from Encyclopedic Texts, *In Proceedings of the International Conference for Young Computer Scientists*, 2001, pp.1194-1198.

[3] Marti A. Hearst, Automatic acquisition of hyponyms from large text corpora, *In Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, 1992, pp.539-545.

[4] Sharon A. Caraballo, Automatic construction of a hypernym-labeled noun hierarchy from text, *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999, pp.120-126.

[5] Emmanuel Morin, Christian Jacquemin, Projecting corpus-based semantic links on a thesaurus, *In Proceedings of the 37th* Annual *Meeting of the Association for Computational Linguistics*, 1999, pp.389-396.

[6] Juan Lloréns and Hernán Astudillo, Automatic generation of hierarchical taxonomies from free text using linguistic algorithms. *Advances in Object-Oriented Information Systems*, OOIS 2002 Workshops, Montpellier, France, 2002, pp.74-83.

[7] David Sánchez, Antonio Moreno. Patterned automatic taxonomy learning from the Web. *AI Communications*, 21(3), 2008. pp27-48

[8] Khaled Elghamry. Using the Web in Building a Corpus-Based Hypernymy-Hyponymy Lexicon with Hierarchical Structure for Arabic. *Faculty of Computers and Information*, 2008. pp157-165.

[9] ZHANG Chun-xia, HAO Tian-yong, The State of the Art and Difficulties in Automatic Chinese Word Segmentation, *JOURNAL OF SYSTEM SIMULATION*, Vol.17, No.1, 2005, pp.138-143.

[10] Mei JJ, Zhu YM, Gao YQ, Yin HX, *Tongyici Cilin (Dictionary of Synonymous Words)*, Shanghai Cishu Publisher China, 1983.

[11] Guogang Tian, Cungen Cao, Lei Liu, Haitao Wang, MFC: A Method of Co-referent Relation Acquisition from Large-scale Chinese Corpora, *ICNC'06-FSKD'06*, Xi'an, China 2006.

Lei Liu was born in 1979, and received his M.S. degree in computer science from Shandong normal university in 2003. He received his Ph.D degree in computer software from institute of Computing Technology, the Chinese Academy of Sciences in 2007. Now he is a lecturer in college of Applied Sciences, Beijing University of Technology. His current research interests include: knowledge acquisition and ontology learning.

Sen Zhang was born in 1963, and received his Ph.D degree in computer software from institute of Computing Technology, the Chinese Academy of Sciences in 1998. Now he is a associate professor in college of Applied Sciences, Beijing University of Technology. His research interests include multimedia signal processing and computational linguistics.

Luhong Diao was born on June 18, 1978. He received his B.Sc. degree and M.S. degree in the Dept. Computer Science of Shandong University and Ph.D. degree in Institute of Computing Technology, Chinese Academy of Sciences. His main research interests are computer graphics, pattern recognition and imaging processing.

Cungen Cao was born in 1964, and received his M.S. degree in 1989 and Ph.D. degree in 1993 both in mathematics from the Institute of Mathematics, the Chinese Academy of Sciences. Now he is a professor of the Institute of Computing Technology, the Chinese Academy of Sciences. His research area is large scale knowledge processing.