Spatial Structure Analysis and Identification of Candidate Hyponymy Relations

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

Lu Hong Diao
College of Applied Sciences, Beijing University of Technology, Beijing, China
diaoluhong@bjut.edu.cn

***Abstract*—Automatic acquisition of semantic relations is an important problem in knowledge acquisition. We present a method of spatial structure analysis of candidate Chinese hyponymy based on concept space and use them to identify candidate hyponymy. Firstly a group of candidate hyponymy relations is imported into concept space. Secondly we analyze the basic spatial structure of concept space including concept nodes and relation edges. More identified features based on the spatial structure analysis are given and used to identify hyponymy relations. Experimental results show that the spatial structure is very useful to the identification of hyponymy.**

***Index Terms*—hyponymy; relation acquisition; concept space; hyponymy identification**

I. INTRODUCTION

Automatic acquisition of knowledge from free text is becoming increasingly important recently. In knowledge acquisition research field, hyponymy as a basic semantic relations is a more interesting and fundamental problem. Hyponymy relations play a crucial role in various natural language processing systems, such as systems for machine translation, information retrieval, and intelligent query. Especially, hyponymy relations are important in the identification of ontologies, knowledge bases, and lexicons [1][2].

Hyponymy is a semantic relation between concepts. Given two concepts X and Y, there is the hyponymy relation between X and Y if the sentence "X is a (kind of) Y" is acceptable. Y is a hypernym of X, and X is a hyponym of Y. We denote a hyponymy relation as hr(X, Y). For example:

Mango is a kind of fruit---hr (mango, fruit)

Human knowledge is mainly presented in the format of free text at present, so processing free text has become a crucial yet challenging research problem. In free text, constituting hyponymy concept represented by the word, known as the concept word. The concept words in hyponymy relations may have multiple senses. Different word senses point different entities. The error hyponymy in acquired hyponymy will affect the building of hyponymy lexicon.

In our research, the problem of hyponymy identify-cation is described as follows:

Given a group of candidate hyponymy acquired based on rule-based or statistics-based, we denoted these relations as CHR= $\{(c_1, c_2), (c_3, c_4), (c_5, c_6), \ldots\}$, where c_i is the concept word of constituting candidate hyponymy. So the problem of hyponymy relation identification is how to identify correct hyponymy relations from CHR with well-formed validation algorithm.

Here we propose the concept of concept space, all the candidate hyponymy relations are imported to the concept space, and then identified through the analysis of spatial structure of concept space.

In this paper, we present a spatial structure analysis of candidate Chinese hyponymy based on concept space and use them to identify hyponymy. The rest of the paper is organized as follows. Section 2 describes related work in the area of hyponymy relations acquisition, section 3 gives the definition of concept space and analyzes the basic spatial structure of concept space including concept nodes and relation edges, section 4 gives a group of identified features and presents how to identify candidate hyponymy relations, section 5 conducts a performance evaluation, and finally section 6 concludes the paper.

II. RELATED WORK

There are two main approaches for hyponymy acquisition. One is rule-based (also called pattern-based), and the other is statistics-based. The rule-based approaches uses the linguistics and natural language processing techniques to obtain related patterns, and then makes use of pattern matching to acquire hyponymy relations, and the latter is based on statistical language model, and uses clustering algorithm to acquire hyponymy from corpus [3][4].

The main idea of rule-based approach is the hyponymy can be extracted from text as they occur in detectable syntactic patterns. The so-called patterns include special lexical features, idiomatic expressions, phrasing features, and semantic features of text.

doi:10.4304/jsw.8.6.1495-1502

There have been many researchers to research automatic methods to acquire hyponymy from free text corpora. One of the first studies was done by Hearst [5]. Hearst proposed a method for retrieving concept relations from Grolier's Encyclopedia by using lexico-syntactic patterns, such as

…NP_1 is a NP_2… ---hr(NP_1, NP_2)

…NP_1 such as NP_2… ---hr(NP_2, NP_1)

Other researchers also proposed other methods to obtain hyponymy. Most of these techniques are based on particular linguistic patterns [6].

Sánchez proposed a novel approach for composing taxonomies in an unsupervised way. It uses different types of linguistic patterns for hyponymy extraction and designed measures to infer information relevance [7].

Elghamry's method bootstraps the acquisition process by searching the Web for the lexico-syntactic patterns. A corpus-based hyponymy lexicon with partial hierarchical structure for Arabic was created from the Web [8].

Hattori proposed a method to acquire hyponymy relations from the Web based on property inheritance. Property inheritance from a concept to its hyponyms is assumed to be necessary and sufficient conditions of hyponymy relations to achieve high recall and not low precision[9].

Acosta presented a method for extracting hyponymy relations from definitions situated on specialized texts in Spanish. The set of extraction hypernyms from analytical definitions is employed as a seed to extract an additional set of relations from a domain-specific corpus [10].

Costa focused on the extraction of hyponymy relations from individual user sessions by examining, search behavior. Those extracted relations reflect the knowledge that the user is employing while searching the web [11].

In previous studies, we have done some research work about hyponymy acquisition. We presented an iterative method extracting hyponymy from large Chinese free text. We combined outside layer removal and inside layer gathering for acquiring concepts. Hyponymy relations were verified with multiple features [12].

III. The Analysis of Concept Space

A. Building Concept Space

In Chinese, one may find several hundreds of hyponymy patterns based on different quantifiers and synonymous words, which is equivalent to the single hyponymy pattern (i.e. (<?C1> is a <?C2>), (<?C3> such as <?C1>,<?C2>)) in English. **Fig.1** depicts a few typical Chinese hyponymy relation patterns [12].

Pattern 1: <?C1><shiyi|weiyi><zhong|ge|ming|pian|kuai|bu|ke|ben|…><?C2>
(Pattern 1: <?C1> is a <?C2>)
Pattern 2:<?C3><ru|xiang|baokuo|baohan|hangai ><?C1>{<huo|ji|yiji|he|yu|dunhao><?C2>}*<deng>
(Pattern 2: <?C3> such as <?C1>, <?C2>…)

Figure 1. Defining Chinese hyponymy 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. "|" expresses logical "or". "<?C1>" is a variable item in the pattern.

In pattern1,"shiyi|weiyi"means "is a"; "zhong|ge|ming|pian|kuai|bu|ke|ben…" is a group of quantifiers. In pattern 2, "ru|xiang" means "such as"; "baokuo|baohan|hangai" means "include"; "he|yu" means "and"; "huo" means "or"; "deng" means "etc."; "ji|yiji" denotes "as well as"; Chinese "dunhao" is a special kind of Chinese comma used to set off items in a series.

Chinese hyponymy patterns will be used to capture concrete sentences from 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 hr(c, c′) is true, then we tag c by c_L, and c′ by c_H, as shown below.

{As everyone knows, {China}c_L }$_{<?C1>}$/is a/ {{ country }c_H with a long history }$_{<?C2>}$

{The {farm crop}c_H mainly}$_{<?C5>}$ includes {{paddy rice}c_L}$_{<?C1>}$, {{corn}c_L}$_{<?C2>}$, {{sweet potato}c_L}$_{<?C3>}$, {{tobacco leaves} c_L}$_{<?C4>}$ etc.

We can acquire hr(China, country), hr(paddy rice, crop), hr(corn, crop), hr(sweet potato, crop) and hr(tobacco leaves, crop) from the above example.

There are still many error relations in the acquired hyponymy relations from text. They must be identified for the building of lexicon.

It is well known that Chinese is a language different from any western language. A Chinese sentence is made up of a string of characters which do not have any space or delimiter in between. Firstly we initially acquire a set of candidate hyponymy relation from large Chinese free text based on Chinese lexico-syntactic patterns. Then we build concept space using those candidate hyponymy relations [13][14].

Definition 1: The concept space is a directed graph G = (V, E, W) where nodes in V represent concepts of the hyponymy and edges in E represent relationships between concepts. A directed edge (c_1, c_2) from c_1 to c_2 corresponds to a hyponymy from concept c_1 to concept c_2. Weights in W are used to represent varying degrees of certainty.

Definition 2: For each node c in a graph G, (c, c') ∈E, c' is a direct hypernym concept of c, and c is a direct hyponym concept of c', the set of direct hypernym concept of c is denoted by $\mu^h(c)$, the set of direct hyponym concept of c is denoted by $\mu_h(c)$. The number of direct hypernym concept of c is denoted by $|\mu^h(c)|$, and the number of direct hyponym concept of c is denoted by $|\mu_h(c)|$.

Definition 3: For each edge (c_1, c_2) in a graph G= (V, E, W), (c_1, c_2) ∈E, if $|\mu^h(c_1)|$ =1, $|\mu_h(c_1)|$ =0, $|\mu^h(c_2)|$ =0, $|\mu_h(c_2)|$ =1, then (c_1, c_2) is an isolated edge. If (c_1, c_2) ∈E

is not an isolated edge, (c_1, c_2) is denoted by adjacent edge.

The basic process about building concept space is shown in Algorithm 1.

Algorithm 1. The process of building concept space

Input: the set of candidate hyponymy relations CHR from Chinese free text based on lexico-syntactic patterns;

Output: the concept space G.

Step1*:* Initialize G = (V, E, W), let V=∅, E=∅, W=∅;

Step2*:* For each (c_1, c_2)∈CHR, repeat Step3-Step4;

Step3*:* If $c_1 \notin V$, $c_2 \in V$, then $V=V\cup\{c_1\}$; $E=E\cup\{(c_1,c_2)\}$; If $c_1 \notin V, c_2 \notin V$, then $V=V\cup\{c_1,c_2\}$; $E=E\cup\{(c_1,c_2)\}$; If $c_1 \in V$, $c_2 \notin V$, then $V=V\cup\{c_2\}$; $E=E\cup\{(c_1,c_2)\}$;

Step4*:* CHR= CHR-$\{(c_1,c_2)\}$;

Step5*:* For each $r \in E$, set its $w(r) \in W$ to be 0.

Step6: return G;

With the concept space scale increases, its structure becomes more complex, but this complex spatial structure can be split into a number of simple structures, we used two ways to analyze the structure of the concept space: node and edge.

B. The Spatial Structure of Edge

Let (c_1, c_2) is an adjacent edge in the concept space G.

According to the definition of the adjacent edge, (c_1, c_2) is an adjacent edge , at least one of the following conditions are satisfied: (i) $|\mu^h(c_1)|>1$; (ii) $|\mu_h(c_1)| >0$; (iii) $|\mu^h(c_2)| >0$; (iv) $|\mu_h(c_2)| >1$. The adjacent structure of (c_1, c_2) is divided into: 2-adjacency, 3-adjacency, and 4-adjacency.

(1) **2-adjacency:** $|\mu_h(c_1)| =0$, and $|\mu^h(c_2)| =0$. There are three basic 2-adjacency structures, as shown in **Fig. 2**. The dashed arrow is used to represent (c_1, c_2).

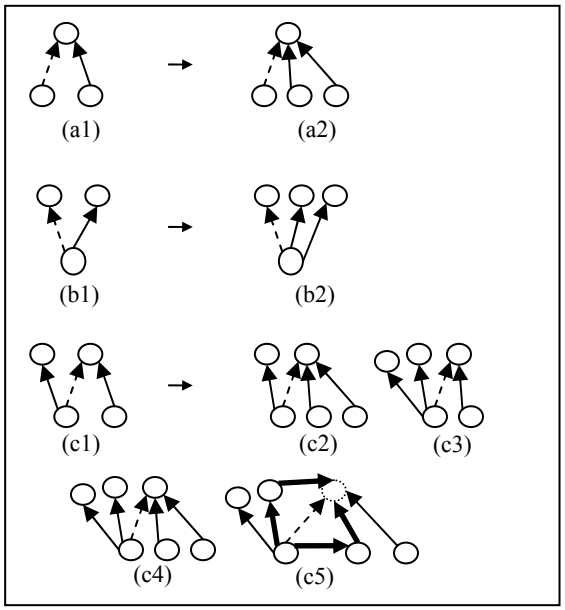

Figure 2. The analysis of 2-adjacency

Fig. 2(a): $|\mu^h(c_1)| =1$, $|\mu_h(c_2)| =2$, in **Fig. 2(a1)**. This structure can be extended to $|\mu^h(c_1)| =1$, $|\mu_h(c_2)| >2$, as shown in **Fig. 2(a2)**;

Fig. 2(b): $|\mu^h(c_1)| =2$, $|\mu_h(c_2)| =1$, in **Fig. 2(b1)**. This structure can be extended to $|\mu^h(c_1)| >2$, $|\mu_h(c_2)| =1$, as shown in **Fig. 2(b2)**;

Fig. 2(c): $|\mu^h(c_1)| =2$, $|\mu_h(c_2)| =2$, in **Fig. 2(c1)**. This structure can be extended to: $|\mu^h(c_1)| >2$, $|\mu_h(c_2)| =2$, as shown in **Fig. 2(c2)**;

$|\mu^h(c_1)| =2$, $|\mu_h(c_2)| >2$, as shown in **Fig. 2(c3)**;

$|\mu^h(c_1)| >2$, $|\mu_h(c_2)| >2$, as shown in **Fig. 2(c4)**;

In particular, if $|\mu^h(c_1) \cap \mu_h(c_2)| >1$, then there is redundant edges, as shown in **Fig. 2(c5)**, where the thick line arrows indicate the edge;

(2) **3-adjacency:** There are two basic 3-adjacency structures, as shown in **Fig. 3**. The dashed arrow is used to represent (c_1, c_2).

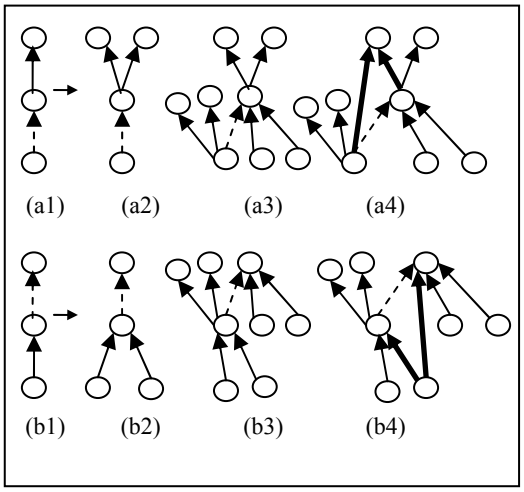

Figure 3. The analysis of 3-adjacency

Fig. 3(a): $|\mu_h(c_1)|=0$, $|\mu^h(c_1)| =1$,$|\mu_h(c_2)|=1$,$|\mu^h(c_2)|=1$, in **Fig. 3(a1)**, This structure can be extended to $|\mu_h(c_1)|=0$, $|\mu^h(c_1)| =1$, $|\mu_h(c_2)|=1$, $|\mu^h(c_2)|>1$, as shown in **Fig. 3(a2)**; If it exists a 2-adjacency structure, this structure can be extended to **Fig. 3(a3)**. In particular, if $|\mu^h(c_1) \cap \mu_h(c_2)| >1$, then there is redundant edges, as shown in **Fig. 3(a4)**, where the thick line arrows indicate the edge;

Fig. 3(b): $|\mu^h(c_2)|=0$, $|\mu^h(c_1)| =1$, $|\mu_h(c_2)|=1$, $|\mu_h(c_1)|=1$, in **Fig. 3(b1)**, This structure can be extended to $|\mu_h(c_1)|=0$, $|\mu^h(c_1)| =1$, $|\mu_h(c_2)|=1$, $|\mu_h(c_1)|>1$, as shown in **Fig. 3(b2)**; If it exists a 2-adjacency structure, this structure can be extended to **Fig. 3(b3)**. In particular, if $|\mu_h(c_1) \cap \mu_h(c_2)| >1$, then there is redundant edges, as shown in **Fig. 3(b4)**, where the thick line arrows indicate the edge;

(3) **4-adjacency:** As shown in **Fig. 4**. The dashed arrow is used to represent (c_1, c_2).

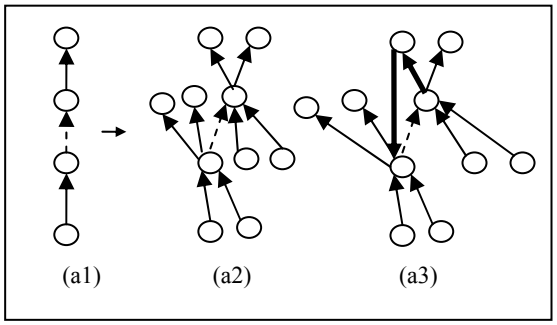

Figure 4. The analysis of 4-adjacency

Fig. 4(a): $|\mu_h(c_1)| = 1$, $|\mu^h(c_1)| = 1$, $|\mu_h(c_2)| = 1$, $|\mu^h(c_2)| = 1$, in **Fig. 4(a1)**, If it exists 2-adjacency structure and 3-adjacency structure , this structure can be extended to **Fig. 4(a2)**. In particular, if $|\mu_h(c_1) \cap \mu^h(c_2)| > 1$, then there is redundant edges, as shown in **Fig. 4(a3)**, where the thick line arrows indicate the edge.

C. The Spatial Structure of Node

Let c is a node in the concept space G. There are at least adjacent to node c-connected nodes. Here the number of $\mu^h(c)$ and $\mu_h(c)$ are considered, as shown in **Fig. 5**. The black solid point is used to represent c.

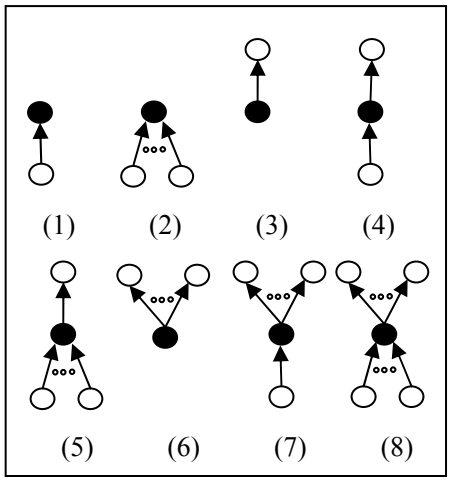

Figure 5. The analysis of node

(1) $|\mu^h(c)| = 0, |\mu_h(c)| = 1$
(2) $|\mu^h(c)| = 0, |\mu_h(c)| > 1$
(3) $|\mu^h(c)| = 1, |\mu_h(c)| = 0$
(4) $|\mu^h(c)| = 1, |\mu_h(c)| = 1$
(5) $|\mu^h(c)| = 1, |\mu_h(c)| > 1$
(6) $|\mu^h(c)| > 1, |\mu_h(c)| = 0$
(7) $|\mu^h(c)| > 1, |\mu_h(c)| = 1$
(8) $|\mu^h(c)| > 1, |\mu_h(c)| > 1$

D. Spatial Structure Analysis

We used about 8GB of raw corpus from the Chinese Web pages. Raw corpus is preprocessed in a few steps, including word segmentation, part of speech tagging, and splitting sentences according to periods. Then we acquired more than 60,000 candidate hyponymy relations (Precision 73.3%) from processed corpus by matching Chinese hyponymy patterns.

(1) Edge Analysis

The detailed result is shown in Table I. The percentage of edge is the ratio of the number special edges and the number of all edges in concept space. The percentage of node is the ratio of the number of node meeting special edges and the number of all nodes in concept space.

It can be seen from Table 1, the number of isolated edges is very low (4.3%) and the number of adjacent edges is very high (95.7%). It indicates that the correlation of knowledge.

Isolated edge

The precision of isolated edges is 65%.The correct rate is lower than the average rate (8.3%). Analysis from the isolated side (c_1, c_2), c_1 is generally the instance concept, and c_2 is generally the class concept. Both the level of the difference is small. For example:

(Shanghai modern theaters, private professional troupes)

(Big Spring Bay coal mine, village-run joint-stock coal mine)

TABLE I
THE RESULT OF EDGE ANALYSIS

Edge structure category	The number of edges	The percentage of edge	Precision	The number of node	The percentage of node
all the edges	62,201	100%	73.3%	40,390	100%
isolated edge	2,677	4.3%	65%	5,331	13.2%
adjacent edge	59,524	95.7%	73.5%	35,059	86.8%
2-adjacency	16,064	25.8%	76%	17,045	42.2%
(a)	7,066	11.4%	75%	10,178	25.2%
(b)	1,733	2.8%	71%	2,867	7.1%
(c)	7,265	11.7%	78%	6,381	15.8%
3-adjacency	32,440	52.1%	77%	23,506	58.2%
(a)	22,165	35.6%	79%	18,135	44.9%
(b)	10,273	16.5%	73%	6,825	16.9%
4-adjacency	11,020	17.7%	57%	2,221	5.5%

2-adjacency

(a) The percentage of edges satisfying this structure is 11.4%, Analysis from the (c_1, c_2), c_1 is generally the instance concept, and c_2 is generally the class concept. Both the level of the difference is small. For example:

(c_2: Shrub species c_1: thyme cuckoo, David Rose, Alpine Green Line Ju, gray plum)

(b) The percentage of edges satisfying this structure is only 2.8%, Analysis from the (c_1, c_2), this structure does not meet the usual structure of semantic relations.

((Oracle, The archaeological data) (Oracle, the software manufacturer))

Here "Oracle" is a concept of polysemous words.

(c) The percentage of edge satisfying this structure is only 11.7%. Analysis from the (c_1, c_2), this structure satisfies the structure (a) and (b) at the same time.

(c_2: Animal ingredients c_1: leopard bone, antler, bear bile, musk)

(c_2: Rare medicinal herbs, the state banned the export of goods, aromatic medicine c1: musk)

3-adjacency

The percentage of edges satisfying this structure is 52.1%(3-adjacency(a) 35.6%, 3-adjacency(b) 16.5%). Analysis from the 3-adjacency(a), c_1 is generally the instance concept, and c_2 is generally the class concept. Analysis from the 3-adjacency(b), c_1 and c_2 are both generally the class concept. It indicates that more hyponymy (c_1, c_2), c_1 is generally the instance concept, and c_2 is generally the class concept.

4-adjacency

The percentage of edges satisfying this structure is 17.7% and the percentage of nodes meeting this structure is 5.5%. This is caused by two reasons. Firstly 4-adjacency structure is the most complex in all structures, and therefore consistent with a relatively small number of edges of this structure. Secondly nodes are reused, so it makes the number of nodes have a low proportion.

Analysis from the 4-adjacency, c_1 and c_2 are both generally the class concept. It can be reasonable explained, indicating that more hyponymy (c_1, c_2), c_1 is generally the instance concept, and c_2 is generally the class concept.

The precision is relatively low(57%) in 4-adjacency. It is easy to encounter the loop structure.

For example:

((disaster, history) (history, memories) (memories, beauty) (beauty, disaster))

(2) Node Analysis

The detailed result is shown in Table II.

It can be seen from Table 2, the different structures of the node are a great difference in the proportion, and the precision of concepts also have some difference.

a. structure (1) and structure (2)

The percentage of structure (1) nodes is 15.2%, and its precision (86.5%) is lower than the precision (92%) of all nodes. Its precision is the lowest one. For the reason, it is no more adjacent nodes in addition to a hyponym adjacent node.

If a node has more adjacent nodes, it instructions this node can be obtained from more text. It identifies the correctness of this node. In structure 2, it is due to the increase in the number of adjacent nodes, making the precision of nodes reached 95%.

b. structure (3)

Nodes in structure (3) have obvious characteristics of the instance concept, and it has the highest proportion(54.1%). It indicates that the relations between instances and classes have a higher proportion.

c. structure (4) and structure (5)

Structure (4) and structure (5) node's percentage is very low, 0.8% and 1.4% respectively, but it exists hypernym and hyponym adjacent node, so the precision of node is 99% and 98% respectively. In addition, the hypernym adjacent node and the hyponym adjacent node may constitute hyponymy relationship.

d. structure (6)

Structure (6) node's percentage is 11.3%, but it exists

TABLE II
THE RESULT OF NODE ANALYSIS

Node structure category	The number of node	The percentage of nodes	The Precision of concept	The mean of $\|\mu h(c)\|$	The mean of $\|\mu h(c)\|$
all the nodes	40,390	100%	92%	1.5	1.5
(1) $\|\mu^h(c)\| = 0, \|\mu_h(c)\| = 1$	6,149	15.2%	86.5%	0	1
(2) $\|\mu^h(c)\| = 0, \|\mu_h(c)\| > 1$	5,422	13.4%	95%	0	4.2
(3) $\|\mu^h(c)\| = 1, \|\mu_h(c)\| = 0$	21,860	54.1%	91%	1	0
(4) $\|\mu^h(c)\| = 1, \|\mu_h(c)\| = 1$	323	0.8%	99%	1	1
(5) $\|\mu^h(c)\| = 1, \|\mu_h(c)\| > 1$	565	1.4%	98%	1	6.7
(6) $\|\mu^h(c)\| > 1, \|\mu_h(c)\| = 0$	4,574	11.3%	97%	4.2	0
(7) $\|\mu^h(c)\| > 1, \|\mu_h(c)\| = 1$	326	0.8%	99%	7.0	1
(8) $\|\mu^h(c)\| > 1, \|\mu_h(c)\| > 1$	1,171	2.9%	99%	15.4	24.4

hypernym and hyponym adjacent node, so the precision of node is 99% and 98% respectively. In addition, the hypernym adjacent node and the hyponym adjacent node may constitute hyponymy relationship.

Structure (6) node has no hyponym adjacent node, often as an instance concept. The instance concept often has different hypernym nodes based on the different classification criteria. For example:

((Hemingway, artist) (Hemingway, writer)
(Hemingway, an adventurer) (Hemingway, the world celebrity))

e. structure (7) and structure (8)

Structure (7) and structure (8) node's percentage is very low, 0.8% and 2.9% respectively. Similar to the structure of (4) (5), it exists hypernym and hyponym adjacent node, so the precision of node is 99%.

However, such the precision of the edge of these nodes is very low. In structure (7), the precision of (c', c)(the node c and its hyponym c') is only 40%.

In addition, Table 2 also gives the mean of $|\mu^h(c)|$ and the mean of $|\mu^h(c)|$. Especially the structure (8), the mean of $|\mu^h(c)|$ and the mean of $|\mu^h(c)|$ is 15.4 and 24.4, respectively. It shows that the node having more hyponym may have more hypernym nodes, and the node having more hypernym may have more hyponym nodes.

IV. Spatial Structure Features

According to the above spatial structure analysis, we analyzed the degree of influence of the spatial structure to the hyponymy relation. When a group of candidate hyponymy relations are correct or error, they often satisfy some spatial structure feature. If a candidate hyponymy satisfies a certain threshold with matching those features, we think that it is a correct hyponymy.

In spatial structure analysis, we used the coordinate relation between concepts. The coordinate relations are acquired using a set of coordinate relation patterns including "dunhao". Chinese "dunhao" is a special kind of Chinese comma used to set off items in a series.

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

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

cr(corn, sweet potato, tobacco leaves) is acquired from the above example.

Figure 6 depicts a few typical structure features of hyponymy.

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

((apple, fruit), (fruit, food))

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

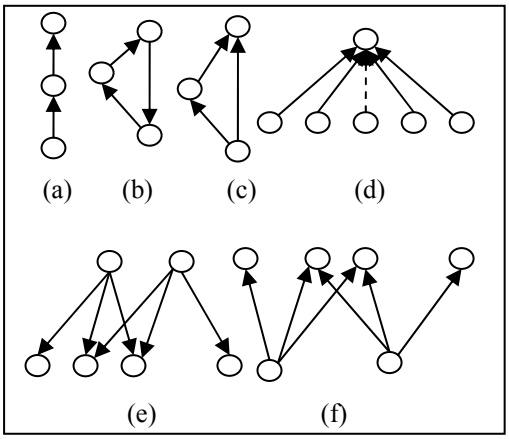

Figure 6. structure features of hyponymy partly

((game, life), (life, mythology), (mythology, game))

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

((potato, vegetable), (potato, foodstuff), (vegetable, food))

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

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

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

(c= food, $\{c_1,\ldots, c_m\}$={cake, bread, butter}, c'=product, $\{c'_1,\ldots, c'_n\}$={cake, brewis}, $\{c_1,\ldots, c_m\} \cap \{c'_1, \ldots, c'_n\}$ = {cake})

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

(c=tomato, $\{c_1,\ldots, c_m\}$={plant, vegetable, fruit}, c'= aubergine, $\{c'_1,\ldots, c'_n\}$={vegetable, food}, $\{c_1,\ldots, c_m\} \cap \{c'_1, \ldots, c'_n\}$ = {vegetable })

We use CF (certainty factors) that is the most common approach in rule-based expert system. The identification features of hyponymy are converted into a group of production rules used in uncertainty reasoning. The formula is as follows:

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

Where CHR is a set of candidate hyponymy, 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 exists CF(CHR, f)<0, then we denote f as negative feature and CF(CHR, f) denotes the no support degree of feature f.

If exists CF(CHR, f)≥0, then we denote f as positive feature and CF(CHR, f) denotes the support degree of feature f.

For example, P(CHR)=0.69, the precision of candidate hyponymy relations satisfying the certain feature is 95%, namely $P(CHR|f_b)=0.95$, the result of CF is (0.95-0.69)/(1- 0.69)=0.839. The f is a positive feature.

The precision of candidate hyponymy relations satisfying certain feature is 21%, namely P(CHR|f)=0.21, the result of CF is (0.21-0.69)/ 0.69= -0.696. The f is a negative feature.

After those features are converted into a group of production rules, we can carry uncertainty reasoning in concept space and calculate the weight of hyponymy.

V. Evaluation

We used three kinds of measures: R (Recall), P (Precision), and F (F-measure). They are typically used in information retrieval and information extraction.

Let h_1 be the total number of hyponymy relations in the classified set.

Let h_2 be the total number of correct hyponymy relations in the classified set.

Let h be the total number of correct hyponymy relations in the CHR.

Precision is the ratio of h_2 to h_1, i.e. $P = h_2/h_1$

Recall is the ratio of h_2 to h, i.e. $R = h_2/h$

F-measure is the harmonic mean of recall and precision. It is high if both precision and recall are high. $F = 2RP/(R+P)$

We still use 62,201 candidate hyponymy relations from 8G free text. For analyzing the influence of threshold α, we choose several different values. We manually evaluated 15% initial set CHR and 15% final result set. The detailed result is shown in Table III.

As we can see from Table III, there are 62,201 candidate relations in concept space. Threshold α is the judgment threshold α. With improve of threshold α, the precision is also improved. If we want to improve the precision, we can improve the value of α. For example, when α=0.9, the precision is up to 98%, and but its recall decreases to 19%. When threshold α is a relatively small value, the method can remove many error hyponymy relations under the condition of skipping a small amount of correct relations. But when threshold α is a relatively large value, the method can remove many error hyponymy relations and also lose many correct relations correspondingly.

TABLE III
THE RESULT OF IDENTIFICATION

Concept space：62,201 hyponymy relations P(CHR)=73.3%				
	The result of identified hyponymy			
α	Number	P	R	F
α=0.0	57,112(91.8%)	77%	96%	0.86
α=0.1	52,021(83.6%)	79%	90%	0.84
α=0.2	47,232(75.9%)	82%	85%	0.83
α=0.3	40,782(65.6%)	85%	76%	0.80
α=0.4	33,632(54.1%)	88%	65%	0.75
α=0.5	26,766(43.0%)	90%	53%	0.67
α=0.6	19,154(30.8%)	93%	39%	0.55
α=0.7	15,758(25.3%)	94%	32%	0.48
α=0.8	13,370(21.5%)	96%	28%	0.44
α=0.9	9,072(16.2%)	98%	19%	0.33

VI Conclusion

On the basis of previous work, we present a method of spatial structure analysis of candidate Chinese hyponymy based on concept space. Firstly, we give a group of candidate hyponymy relations that is imported into concept space. Secondly we analyze the basic structure of concept space including concept nodes and relation edges. More identified features based on the spatial structure analysis are given and used to identify hyponymy relations.

For verify hyponymy relations, the hyponymy features are converted into a group of production rules. Experimental results show that the spatial structure is very useful to the identification of hyponymy. It will raise the precision of hyponymy relations and benefit to the building of ontologies and knowledge bases.

There are still some error relations in the final result set. In future, we will combine some other methods (such as knowledge database, context etc.) to the further identification of hyponymy.

Acknowledgments.

This work is supported by the National Natural Science Foundation of China (Grant No. 61105040, 61203284), the Beijing Natural Science Foundation (Grant No 4133085) and the Beijing University of Technology Science Foundation (Grant No. 00600051 4311002).

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] Xin Wang, Ying Wang. Research on Web Query Translation based on Ontology. Journal of Software, 2012. Vol 7(12). pp 2849-2856

[3] 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.
[4] Verginica, Barbu, Mititelu. Hyponymy Patterns Semi-automatic Extraction, Evaluation and Inter-lingual Comparison. TSD2008, 2008. pp37-44.
[5] Marti A. Hearst, Automated Discovery of WordNet Relations, To Appear in WordNet: An Electronic Lexical Database and Some of its Applications, Christiane Fellbaum (Ed.), MIT Press, 1998, pp.131-153.
[6] Yamada, I., Torisawa, K., Kazama, J., Kuroda, K., Murata, M., De Saeger, S., Bond, F. and Sumida, A. Hypernym discovery based on distributional similarity and hierarchical structures. the Conference on Empirical Methods in Natural Language Processing, 2009. pp.929–937,.
[7] David Sánchez, Antonio Moreno. Pattern-ed 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] Shun Hattori, Hiroaki Ohshima, Satoshi Oyama, and Katsumi Tanaka. Mining the Web for Hyponymy Relations Based on Property Inheritance. the 10th Asia Pacific Web Conference, 2008. pp99-110.
[10] Acosta, O., C. Aguilar, y G. Sierra. A Method for Extracting Hyponymy-Hypernymy Relations from Specialized Corpora Using Genus Terms. Proceedings of the Workshop in Natural Language Processing and Web-based Technologies 2010, pp1-10.
[11] Rui P. Costa, Nuno Seco. Hyponymy Extraction and Web Search Behavior Analysis Based on Query Reformulation. Proceedings of the 11th Ibero-American conference on AI: Advances in Artificial Intelligence, 2008. pp1-10.
[12] Lei Liu, Sen Zhang, LuHong Diao, CunGen Cao. An Iterative Method of Extracting Chinese ISA Relations for ontology learning. Journal of Computers, 5(6), 2010. pp870-877.
[13] Lei Liu, Sen Zhang, LuHong Diao, ShuYing Yan, CunGen Cao. A Verification Method of Hyponymy between Chinese Terms Based on Concept Space. Active Media Technology 2009, pp160-170.
[14] Lei Liu, Sen Zhang, LuHong Diao, ShuYing Yan , CunGen Cao, Using Concept Space to Verify Hyponymy in Building a Hyponymy Lexicon. Artificial Intelligence and Computational Intelligence 2009, pp479-486.

Dr. Lei Liu is currently an associate professor in College of Applied Sciences, Beijing University of Technology. His major research interests include knowledge acquisition and ontology learning.

Dr. Luhong Diao is currently an associate professor in College of Applied Sciences, Beijing University of Technology.. His major research interests include computer vision, image processing.