Extracting Hyponymic Relations from Chinese Free Corpus

LEI LIU[1,2], CUNGEN CAO, HAITAO WANG[1,2]
Institute of Computing Technology, Chinese Academy of Sciences[1]
Graduate School of the Chinese Academy of Sciences[2],
Beijing 2704# 420, Post Code 100080, China
CHINA

Abstract: - Research on hyponymy acquisition is a basic and crucial problem in knowledge acquisition from text. In this paper we present a method of hyponymic relation acquisition and verification based on Chinese lexico-syntactic patterns. Firstly, we make use of removable lexicons and sentence patterns that have been semi-automatically obtained to analyze Chinese-isa patterns. Then we use an algorithm that combines outside layer removal and inside layer gathering to acquire hyponymic concept. In the final phase, we combine self features and context features together for the verification of hyponymy. Experimental results show that the method is adequate of extracting hyponymy from Chinese free text.

Key-Words: - Hyponymic relation; Relation acquisition; Knowledge acquisition; Information extraction

1 Introduction

Automatic acquisition of concepts and semantic relations from text has received much attention in the last ten years. Especially, hyponymy acquisition is more interesting and fundamental because hyponymic relations are important in accuracy verification of ontologies, knowledge bases and lexicons [1] [2] [3].

The types of input used for hyponymy 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) [4][5][6][7]. 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 hyponymy acquisition. One is pattern-based (also called rule-based), and the other is statistics-based.

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

The rest of the paper is organized as follows. Section 2 describes related work in the area of automatic hyponymy acquisition, section 3 elaborates on Chinese-isa patterns for this work, section 4 presents an algorithm to acquire and verify hyponymy based on Chinese-isa patterns, section 5 conducts a performance evaluation of the proposed method, and finally section 6 concludes the paper.

2 Related Work

Hyponymy is a semantic relation between word meanings. Given two concepts X and Y, there is the hyponymy between X and Y if the sentence "X is a (kind of) Y" is acceptable. X is a hyponym of Y, and Y is a hypernym of X. Hyponymy is also called as subordination, or the "isa" relation[1]. We denote a hyponymic relation by HR(X, Y), as in the following example:

中国是一个发展中国家 ---HR(**中国**,**发展中国家**)
(China is a developing country ---HR(China, developing country))

Hyponymy 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.

There have been many attempts to develop automatic methods to acquire hyponymy from text corpora. One of the first studies was done by Hearst[8][9]. Hearst proposed a method for retrieving concept relations from unannotated text (Grolier's Encyclopedia) by using predefined 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 developed other ways to obtain hyponymy. Most of these techniques are based on particular linguistic patterns [10][11].

Caraballo used a hierarchical clustering technique to build a hyponymy hierarchy of nouns [10]. The internal nodes are labeled by the syntactic constructions from Hearst [8].

Morin produced partial hyponymy hierarchies guided by transitivity in the relation, but the method works on a domain-specific corpus [11].

3 Chinese-isa Patterns

Though “ isa” patterns are simple ones, the amount of sentences matching “ isa” patterns is more dominant than matching other patterns. In Chinese, one may find several hundreds of different “isa” patterns based on different quantifiers (**量词**), which is equivalent to the single “isa” pattern (i.e. <?C1> is a <?C2>) in English. Fig.1 depicts a few typical Chinese-isa patterns.

```
1. defpattern Chinese-isa
2. {
3.  <量词>=种|个|名|篇|片|块|堆|群|批|章|节|环|部|次|步|颗|套|本|条|张|幅|款|代|缕|位|卷|册|只|双|件|台|门|棵|株|朵|根|头|尾|…
4.  模式: <?C1><是一|为一><量词><?C2> (Pattern: <?C1> is a <?C2>)
5. }
```

Fig.1: Defining Chinese-isa patterns

In Fig.1, line 3 defines a group of quantifiers denoted as <!quantifiers>, which can be referenced in line 4. Line 4 defines a group of “isa” patterns, and it means “Pattern: <?C1> is a <?C2>”. “|” expresses logical “or”, “?C1” and “?C2” are two variables in the pattern.

Chinese-isa patterns will be used to capture concrete sentences from Chinese free corpus. In this process, variables <?C1> and <?C2> will be instantiated with words or phrases in a sentence, in which real concepts may be located. Let c be the real concept in <?C1>, and c′ in <?C2>. If HR (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 }$_{<?C2>}$)

The problem now is that after a sentence matches an “isa” pattern, how can we identify the real concepts from the sentence and how can we verify that they satisfy the hyponymic relation? It is difficult to resolve those problems for several reasons [12]:

(1) 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.

(2) The structural degree of free text is very weak, and the expression is vivid and diverse.

(3) For ensuring the recall of hyponymy, we define Chinese-isa patterns without extra restriction rules, therefore the degree of generalization of “ isa” patterns is higher.

To handle these features, we have developed the following strategies.

(1) For concept acquisition, we use an algorithm that combines outside layer removal and inside layer gathering to acquire concept c_L and c_H because “isa” patterns have already specified the context that concepts appear in.

Outside layer removal: There are many non-concept components in <?C1> and <?C2>. Most concepts are compound words without explicit boundaries, but the composition of the non-concept components is more fixed. So we present a new semi-automatic algorithm for acquiring and analyzing non-concept components, and then convert them into lexicon and sentence patterns. Finally, we make use of lexicon and sentence patterns on <?C1> and <?C2> 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. At the same time, lexicon and sentence patterns also can provide the semantic information for hyponymy verification

Inside layer gathering: After the outside layer removal to <?C1> and <?C2>, 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.

(2) For hyponymy verification, we combine the self features and context features of hyponymy together.

4 Method and Algorithms

Our method consists of four phases. In Phase I, we pre-process the raw corpus from the Web. In Phase II, we present a semi-automatic method for acquiring and analyzing non-concept components, and converting them into removable lexicon and sentence patterns. In Phase III, we use an algorithm that combines outside layer removal and inside layer gathering to extract concepts c_L and c_H. In the final phase, we combine self features and context features of hyponymy together for hyponymy verification, and each relation that satisfies the threshold is stored in the hyponymy database. The algorithmic framework is presented in Fig.2.

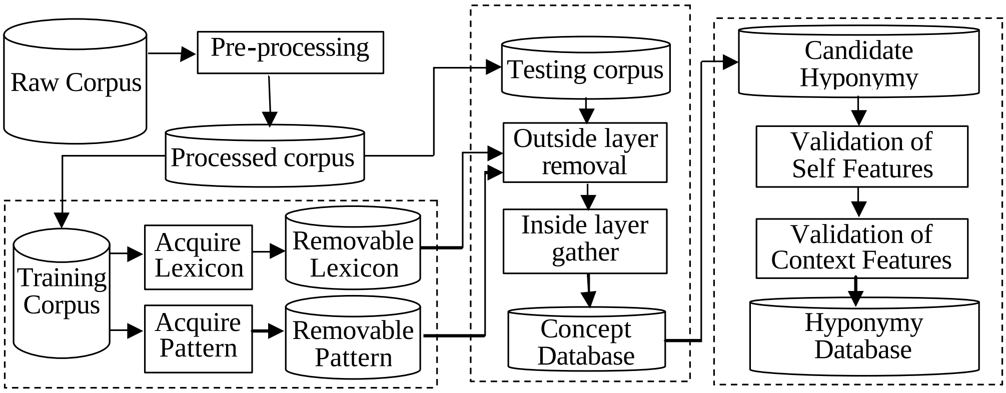

Fig.2: Framework of hyponymy acquisition

4.1 Phase I: Building Corpus

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

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: Separate each of sentences in Cor_{train} according to punctuation marker. Let $S = \{s_1, s_2, \dots, s_n\}$ be a list of separate string.

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

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

If exist $S'_{sub} \subseteq S_{sub}$ satisfy conditions (i) the element amount of S'_{sub} >=2; (ii) the sequence that common substrings appear is consistent, and can't cross; (iii) exist a common substring is a prefix or a 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: Using the Interactive Manual, D is divided into filter word, elimination word, and ambiguous word, and attached some necessary tag. P is divided into filter pattern and elimination pattern, and attached additional rules. Otherwise, we carry on the merge and generalizations to the similar patterns in form.

Output: removal lexicon and sentence patterns.

Fig.3: Details of acquisition of removable lexicon and patterns

4.2 Phase II: Acquisition of Removable Lexicon and Sentence Patterns

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.

- **Filter word (w_f):** if the nonexistent hyponymy in s attributes to the occurrence of w in s, then w is called filter word.
- **Elimination word (w_e):** if w belongs to a part of non-concept components in s, then w is called elimination word.
- **Ambiguous word (w_a):** if w sometimes belongs to a part of the concept in s, sometimes belongs to non-concept components, then w is called ambiguous word.
- **Filter pattern (p_f):** if the nonexistent hyponymy in s attributes to the occurrence of p in s, then p is called filter pattern.
- **Elimination pattern (p_e):** if p belongs to a part of non-concept components in s, then p is called elimination pattern.

The details of acquisition of removable lexicon and patterns are presented in Fig.3.

An example of removable lexicon and sentence patterns is shown in Fig.4 and Fig.5. In Fig.4, we also use other three tags, i.e. position, pos/neg, 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 pos/neg tag shows that when a term is removed from <?C1> or <?C2>, and how the remainder will be logically handled. For example, when we remove the term "by no means" (where the tag is -1) from <?C1>, we obtain a negative $HR(c_L, c_H)$. It takes 1 (completely positive), + (possibly positive), 0 (neutral), - (partially negative), and -1 (completely negative). The tense tag indicates when the $HR(c_L, c_H)$ is true.

In Fig.5, we define a elimination pattern. We prescribe additional rules for improving the result of

word	**type**	**position**	**pos/neg**	**tense**
传说中(in legend)	w_f	a		
并不(by no means)	w_e	t	-1	c
即将(be about to)	w_e	t	0	f
真(really)	w_a	t		

Fig.4: Items of removable lexicon

```
  defpattern elimination pattern 016
  {
     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>
     additional rules: notcontain(<?w2>,。|.|？|！|，|,|.|?|!|；|;)
  }
```

Fig.5 An example of removable sentence patterns

pattern matching. In additional rules, notcontain(a,b) expresses that string b is not a substring of string a, and <?w1> and <?w2> are pattern variables. An example sentence matching elimination pattern 016 is shown below:

{**正如页框一样**, }$_{p\_e}$**表格也**/**是一种**/**容器对象**

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

4.3 Phase III: Acquisition of Hyponymic Concept

We use an algorithm that combine outside layer removal and inside layer gathering to acquire concept c_L and c_H. The details are described in Fig.6.

4.4 Phase IV: Verification of Hyponymy

We analyze the features of hyponymy when we establish and set up the heuristic rules. If a candidate hyponymy satisfies a certain threshold with matching those features, we say that it is a real hyponymic relation. The feature of hyponymy may be defined as follows:

Definition 1: The feature of hyponymy is a 2-tuple HRF= {SF, CF}, where

(1) **SF** (self features): It is constructed by the assumption that c_L and c_H are semantically similar in HR(c_L, c_H), and is subdivided into two features, i.e. SF={WF, SEF}, where

WF (word-formation features): If common substrings of c_L and c_H exist, the position (such as prefix and suffix) will provide the evidence for the existence of a hyponymic relation.

SEF (semantic features): Making use of dictionary of synonymous words [13] to compute the semantic similarity of candidate HR (c_L, c_H).

```
Let R be a list of candidate hyponymys initially empty
Input: Testing corpus $Cor_{test}$, removal lexicon D, sentence patterns P
Step1: For each sentence $s \in Cor_{test}$, process <?C1> and <?C2> respectively 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>. Here let <?C1> and <?C2> be <?C>
    (1) Discover patterns that match with <?C> using P. Here matching pattern may have several.
        If exist filter pattern, add filter pattern tag to <?C> and turning Step3
        If exist elimination pattern, add elimination pattern tag to <?C>
    (2) According to the principle of word length precedence using D. When satisfy as follows a arbitrary condition, the processing stop: (a) discover filter word; (b) discover two ambiguous word continuously; (c) can't discover any removal word further.
Step3: Process tagged sentence according to the following principle.
    If exist filter word or filter pattern tag in <?C>, jump to Step2
    If <?C> is tagged completely, jump to Step2
    If satisfy conditions (i) exist ambiguous word tag in most inside layer tag of <?C>; (ii) ambiguous word is separated solely in the result of word segmentation, then get rid of ambiguous word tag.
Step4: According to the result of part of speech tagging, we 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
```

Fig.6: Algorithm of acquiring concepts c_L and c_H

(2) **CF** (context features): Here we aim to use priori contextual knowledge to perform hyponymy verification. CF is also subdivided into two features, CF={FF, DF}, where

FF (frequency features): If candidate HR (c_L, c_H) appears frequently in a kind of hyponymic patterns or in various hyponymic patterns, the probability of HR(c_L, c_H) is higher.

DF (domain features): If candidate HR (c_L, c_H) appears in a certain scientific domain-specific context, HR (c_L, c_H) may be a true piece of scientific knowledge; otherwise it may be a pair of general concepts and may not have any value.

5 Evaluation

We conduct a performance evaluation of the proposed method.

5.1 Evaluation Method

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 hyponymic relations in the corpus.

Let H_1 be the total number of relations acquired.

Let H_2 be the total number of correct relations acquired.

(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 H1, i.e. $P = H_2/H_1$

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

5.2 Experimental Results

We used about 15GB of raw corpus from the Web. Processed corpus contains about 1,180,000 sentences acquired by matching Chinese-isa patterns. Then we divided the processed corpus into two groups: training corpus (80%) and testing corpus (20%).

After the training corpus was processed by PhaseII, we acquired removable lexicon and removable patterns (their amount in brackets), shown as follows Table 1:

input	Training corpus (943,580)

output	removable lexicon(5828):
	filter word(1995);
	elimination word(3586);
	ambiguous word(247)

	removable pattern(124): filter pattern(35); elimination pattern(89)

Table1: The number of removable lexicon and pattern

The testing corpus was processed by Phase III and Phase IV. We manually evaluated a 1% random sample of each classified corpus. The detailed result is shown in Table 2.

Note that in Table 2, hyponymys had not yet been acquired in results with an asterisk sign "*", so the criterion was as follows:

(1) H is the total number of sentences implying correct relations in testing corpus.

(2) H_1 is the total number of sentences acquired.

(3) H_2 is the total number of sentences implying correct relations acquired.

While hyponymy had been acquired in results with a plus sign "+", so the criterion H_2 was changed：H_2 is the total number of correct relations acquired.

In the following, we analyze each classified corpus in different phases.

5.2.1 Filtered sentences (Phase III)

Filtered sentences are no longer processed further including sentences with filter tags and sentences tagged completely. From the result of Table 2, filtered sentences have a very low recall (3.5%) and precision (5.6%). Filtered sentences are important for other sentences to improve quality. An example is as follows:

电梯{左边}$_{w\_f}$ /是一个/大工作室.

({On the left side of}$_{w\_f}$ the elevator / is a / big utility room.)

5.2.2 No tagged sentences (Phase III)

As we can see from Table 2, the quality of no tagged sentences is the best in all classified corpus of removal phase. Its recall and precision is 36.2 % and 80.1% respectively. Furthermore, most of those sentences' hyponymy can be acquired directly. For example:

氢氟酸/是一种/腐蚀剂

(Hydrofluoric acid /is a kind of/ erosion)

5.2.3 Partially tagged sentences (Phase III)

Partially tagged sentences are tagged by removable lexicon and pattern. Partially tagged sentences are also a very important part and have a higher

Processing Phase		Classified Corpus	Number(ratio)	Precision	Recall	F-measure
Phase III	Input	Testing corpus*	235,895(100%)	50.6 %	100 %	67.2%
	Removal	Filtered sentences*	75,068(31.8%)	5.6 %	3.5 %	4.3%
		No tagged sentences*	53,985(22.9%)	80.1 %	36.2 %	49.9%
		Partially tagged sentences*	106,842(45.3%)	67.6 %	60.5 %	63.9%
	Middle result1	No tagged+Partially tagged*	160,827(68.2%)	71.8 %	96.7 %	82.4%
	Gather	Gathered sentences*	36,563 (15.5%)	53.3 %	16.3 %	25.0%
		No gathered sentences*	124,264 (52.7%)	76.3 %	79.4 %	77.8%
	Middle result2	Candidate hyponymy$^+$	162,235	68.5 %	93.1 %	78.9%
Phase IV	Features	Filtered hyponymy$^+$	37,584	9.2 %	2.9 %	4.4%
		Acquired hyponymy$^+$	124,651	86.5 %	90.3 %	88.4%

Table 2: The processed result of testing corpus

F-measure of 63.9%. From the result of removal, the effect of removal is satisfactory mostly. An example is shown as follows:

{例如}$_{w\_e}${清末的}$_{p_e}$任伯年{便}$_{w\_a}$/是一位/{雅俗共赏的}$_{p_e}$大画家.

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

5.2.4 Gathered sentences (Phase IV)

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}$_{p\_e}$)

Middle result2 is a group of candidate c_L and c_H that are acquired by removing no-tagged components of sentences. Middle result2 is an output of Phase III and also an input of Phase IV.

5.2.5 Filtered hyponymy (Phase IV)

We filtered some no-hyponymy by the self features and the context features of hyponymy. Filtered hyponymys have a recall of 2.9 % and a precision of 9.2%. The result implies the effect of features verification. For example:

{伦敦}$_{n}$ 的{牛津街}$_{n}$/是一条/{非常繁华的}$_{p\_e}${商业街道}$_{n}$

({The Oxford street}$_{n}$ of {London}$_{n}$ / is a / {very prosperous}$_{p\_e}$ {business street}$_{n}$.)

In Phase III, we acquired candidate HR (伦敦, 商业街道), HR (牛津街, 商业街道). In Phase IV, we may verify HR (牛津街, 商业街道) and filter HR (伦敦, 商业街道) by the word-formation feature of hyponymy.

5.2.6 Acquired hyponymy (Phase IV)

Acquired hyponymys are the final result of our algorithms. From Table 2, acquired hyponymys have a recall of 90.3%, a precision of 86.5%, and a F-measure of 88.4%.

As we can see from Fig.7, 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 lexicon and patterns can't handle, as in 8), and more syntactical information will be helpful for resolving this problem. Second, relations may represent a kind of metaphor, as in 7), more sophisticated verification methods are needed.

1) HR (北京,城市) HR (Beijing, city)	*correct*
2) HR (贝克汉姆,球星) HR (Beckham, Ball star)	*correct*
3) HR (宾馆服务业,行业) HR (hotel servicing business, business)	*correct*
4) HR (彩票, 无记名中奖凭证) HR (lottery, credentials of secret win a prize)	*correct*
5) HR (操作系统,系统软件) HR (operating system, system software)	*correct*
6) HR ((内存控制器,芯片) HR (memory control unit, chip)	*correct*
7) HR (法律, 象征) HR (law, symbol)	*incorrect*
8) HR (信息愈来愈,经济资源) HR (information more and more, economic resource)	*incorrect*

Fig.7: Examples of acquired hyponymy

6 Conclusion

In this paper we presented a method that acquires and verifies hyponymic relations from Chinese free text. It initially discovers a set of sentences using Chinese lexico-syntactic patterns. Removable lexicons and sentence patterns are semi-automatically obtained from these sentences. We combined outside layer removal and inside layer gathering for acquiring hyponymic concepts. In the final phase, hyponymic relations is verified with self features and context features. Experimental results demonstrate good performance of the method for extracting hyponymy from Chinese free text.

7 Acknowledgements

This work is supported by the Natural Science Foundation (grant nos. 60273019, 60496326, 60573063, and 60573064), and the National 973 Programme (grants no. 2003CB317008 and G1999032701).

References

[1] Miller G, WordNet: An On-line Lexical Database, *International Journal of Lexicography*, Vol.3, No.4, 1990, pp.235-244.

[2] 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.

[3] Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende, Mindnet: acquiring and structuring semantic information from text, *In proceedings of COLING-ACL'98*, 1998, pp.1098-1102.

[4] 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.

[5] Dolan William, Vanderwende Lucy, Richardson Stephen D, Automatically Deriving Structured Knowledge Bases From On-Line Dictionaries, *In Proceedings of the Pacific Association for Computational Linguistics, Vancouver, British Columbia* , 1993, pp.5-14.

[6] Keiji Shinzato and Kentaro Torisawa, Acquiring hyponymy relations from web documents. *In Proceedings of HLT-NAACL*, 2004, pp.73–80.

[7] Song Rou, Xu Yong, An Experiment on Knowledge Extraction from an Encyclopedia Based on Lexicon Semantics, *Computational Linguistics and Chinese Language Processing*, Vol.7, No.2, 1992, pp.101-112.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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