An Approach to Acquire Semantic Relationships Between Words from Web Document[1]

Xia Sun, Qinghua Zheng, Haifeng Dang, Yunhua Hu, and Huixian Bai

Shaanxi Provincial Key Laboratory of Satellite and Terrestrial Networks Tech,
School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China
sx@mailst.xjtu.edu.cn
{qhzheng, yunhuahu}@mail.xjtu.edu.cn
{xjtu_hfdang, baihuixian}@163.com

Abstract. In this paper, we focus on the semantic relationships acquisition from Chinese web documents motivated by the large requirement of web question answering system in e-Learning. With our scheme, we dwindle in numbers of text to be analyzed and obtain initial sentence-level text in pre-process phase. Then linguistic rules, which are broken down into unambiguous and ambiguous, designed for Chinese phrases are applied to these sentence-level text to extract the synonymy relationship, hyponymy relationship, hypernymy relationship and parataxis relationship. Lastly, candidates are refined using two heuristics. Compared to other previous works, we apply not only strict unambiguous linguistic rules but also loose ambiguous linguistic rules to extract relationships and proposed efficient approach to refine the outputs of these rules. Experiments show that this method can acquire semantic relationships efficiently and effectively.

1 Introduction

Web question answering is one of key issues in e-Learning due to the rapid growth of network education [1]. They provide direct answers to e-Learning users' questions during the leaning process of courseware, as well as associational learning through expanding keywords semantically. For instance, when users submit "network protocols" to a question answering system, they may expect to know some information about "FTP" or "HTTP" (hyponymy expansion); when users submit "operator system" to a question answering system, they may want to know some information about "computer software" (hypernymy expansion); When users submit "HTTP" to a question answering system, they may know some information about "FTP" (parataxis expansion). Through communicating with users, system can offer better services and induct users' learning. Practical experiences show that this ways of associational learning can advance the users' learning efficiency significantly. Thus, we proposed a method of acquiring semantic relationships between words from web documents to implement semantic expansion and provide associational learning.

[1] Funding for this work was provided by NSF grant 60373105 and 60473136.

R.W.H. Lau et al. (Eds.): ICWL 2005, LNCS 3583, pp. 236 – 243, 2005.

There have been several approaches to discover semantic relationships between words from text [2~6]. WordNet [7] describes semantic relationships between words, such as synonymy, hypernymy, hyponymy and antonymy. However, manually constructed repositories are time and manpower consuming, furthermore, have the limitation of broad-coverage lexicon. Researchers prefer to (semi-)automatic approaches of acquiring semantic relationships between words. Pantel and Lin proposed an algorithm, called CBC, for automatically extracting semantic classes by computing the similarity between words based on their distribution in a corpus. The output of this program is a ranked list of similar words to each word. A problem of such approach is that it only shows the degree of similarity between words, rather than differentiates the synonymy, hyponymy and hypernymy. Yuan obtained semantic relationships relying on HowNet, which is a large Chinese semantic lexicon. Because HowNet describes the semantics of common words, it misses many domain words, which results in the limitation of this method.

2 Architecture of Semantic Relationships Acquisition

We focus in this paper on acquiring semantic relationships between words from Web documents, including synonymy relationship, hyponymy relationship, hypernymy relationship and parataxis relationship. Figure 1 shows the architecture of semantic relationships acquisition.

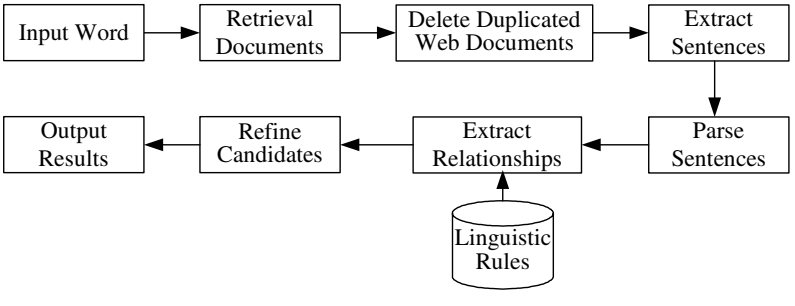

Fig. 1. the architecture of semantic relationships acquisition

As illustrated in Figure 1, our basic idea is as follows: collect the corpus that contain the potential semantic relationships using search engine; then delete duplicated Web documents and wipe off the sentence that have no relevancy with the word, which is submitted to search engine, in order to acquire fine initial data; next, match each sentence to be parsed against linguistic rules to extract relationships; finally, compute the candidates' scores and output the results. Following is description of every module.

Input Word (*IW*)**:** Considering abundant information on the Web, we decide to make use of search engine to gain the documents that contain semantic relationships.

Retrieval Documents: The search engine Google [8] is used to return and rank the top 500 documents.

Delete Duplicated Web Documents: Reprinting of information between websites produces a great deal redundant web documents. If we extract relationships from duplicated Web documents, the statistic value of association between words is invalid. Therefore, string of feature code based algorithm [9] is adopted to remove the duplicated web documents.

Extract Sentences: To decrease the amount of text to be processed and make the procedures of relationships extraction easy, the documents are broken into sentences that contain *IW*.

Parse Sentences: For each sentence to be extracted, following processes are performed. Word segmentation is the first step, since there are no blanks to mark word boundaries in Chinese text. After segmentation, *POS* (Part Of Speech) tagging and shallow parser should be run. Interested reader can refer to our previous works [10~12].

Linguistic Rules Database: It stores many linguistic rules, which are deduced from expressions linked by semantic relationships by manually.

Extract Relationships: We match each sentence to be parsed against linguistic rules to extract words (Output Word, *OW*) that associate with *IW*. Different semantic relationship corresponds a set of linguistic rules.

Refine Candidates: Two heuristics are applied to refine the candidates (see section 3.2).

Output Results: The format of ultimate results are defined as:

Results:={*Syn* (*IW*, *OW*) | *Hypon* (*IW*, *OW*) | *Hyper* (*IW*, *OW*) | *Par* (*IW*, *OW*)}

Where *Syn*, *Hypon*, *Hyper* and *Par* denote the synonymy relationship, hyponymy relationship, hypernymy relationship and parataxis relationship respectively.

3 Semantic Relationships Acquisition

3.1 Description of Linguistic Rules

Expressions that reflect specific semantic relationships are either explicit or implicit, depending on whether exists the distinct characteristic. For examples, “door knob” is an implicit expression, while “water consists of hydrogen and oxygen” is an explicit expression with a characteristic term “consist of”. The explicit ones are further broken down into unambiguous and ambiguous. Correspondingly, linguistic rules are divided into unambiguous rules and ambiguous rules.

Unambiguous rules always convey a specific semantic relationship. The illustrations of unambiguous rules are list below.

IF *Unambig* < *Syn* (*IW*+ ([punctuation] CT_{Usyn} [punctuation])+ *OW*) > THEN *Syn* (*IW*, *OW*)

CT_{Usyn}:={ “又称为(in other words)” | “或称 (alias)” | “简称(abbreviate)” …}

Where, *Unambig* denotes the type of rules, i.e., unambiguous rule. Identifying a specific semantic relationship according to a set of terms (Characteristic Term, *CT*) is an efficient way. CT_U denotes a set of characteristic terms in unambiguous rules. And CT_{Usyn} indicates a set of characteristic terms to recognize synonymy relationship. A pair of bracket denote optional item. The meaning of this rule is: if a term that belongs to CT_{Usyn} follows the *IW*, then subsequent noun to CT_{Usyn} is the target. An example is: "Hyper Text Transmission Protocol is commonly *abbreviated* to HTTP". A simple matching against the rules leads to the discovery of synonymy relationship, *Syn* (Hyper Text Transmission Protocol, HTTP).

IF *Unambig* < *Hypon* (*IW*+([punctuation | quantifier] CT_{Uhypon} [punctuation | quantifier])+*OW*) > THEN {*Hypon* (*IW*, *OW*), *Hyper* (*OW*, *IW*), [*Par* (OW_i, OW_j)]}

CT_{Uhypon}:={ "包括(such as)" | "分为(divide into)" | "由…组成" (consist of)…}

We take an example to demonstrate the rule 2. The example sentence is: "*such* animals *as* cats and dogs". The rule 2 is applied to extract the hyponymy relationship from this example sentence. Two relationships are acquired: *Hypon* (animal, cat), *Hypon* (animal, dog). It is noteworthy that hyponymy relationship and hypernymy relation is relative. If word A is hyponymy relationship of word B, then word B is hypernym relationship of word A. Thus, two hypernym relationships are produced: *Hyper* (cat, animal), *Hyper* (dog, animal). Considering "cat" and "dog" have the hypernymy relation with the same word "animal", we gain a parataxis relationship additionally, *Par* (cat, dog).

IF *Unambig* <*Par* (*JOP*(*IW*, *OW*) + CT_{Upar} > THEN *Par* (*IW*, *OW*)

CT_{Upar}:= {"联系(relate to)" | "区别(discriminate from)" | "相比较" (compare with)…}

The expression of *JOP*(*IW*, *OW*) means a juxtaposition of phrases containing *IW* and other noun words, which are targets. For instance, "*discriminate* HTTP *from* FTP", a parataxis relationship additionally is derived: *Par* (HTTP, FTP).

Ambiguous rules often express several distinct semantic relationships by equal opportunity. That is to say, the same ambiguous rule reflects different relationships in different contexts. Except for different *CT*, the formats of ambiguous rules and unambiguous rules are quite similar. CT_A denotes a set of characteristic terms in ambiguous rules. Different from the elements of CT_U, the elements of CT_A are identifications of a certain semantic relationship, as well as another semantic relationship in some contexts. For instance, "connects two cities, *namely*, New York and Chicago" is hyponymy, whereas "Hyper Text Transmission Protocol, *namely*, HTTP" is synonymy. Thus, it is inconsistent because a pair of words has several kinds of semantic relationships. The following heuristics is applied to eliminate the inconsistency.

3.2 Heuristics

Heuristics 1 is designed to filter out the noisy candidates acquired using the unambiguous rules. This noise is primarily due to overgeneralization of the unambiguous patterns. Some sentences match a certain unambiguous rule, but extracted pairs of terms take none of pre-definitional semantic relationships (synonymy, hyponymy, hypernymy and parataxis). For the purposes of removing these false pairs occurring in this circumstance, we compute the semantic association

degree of each pair of candidate, *SUP*, and retain only the pairs that the number of times occurred in collects is bigger than the value of *SUP*. Finally, calculate the value of *CON* of the each rest candidate, and get rid of potential false pairs.

Definition 1
SUP is the minimal value that determines whether a pair of terms takes on a certain semantic relationship. If $f(R(w_1, w_2)) \geq SUP$ then delete $R(w_1, w_2)$. Where, $f(*)$is the number of times a pair of terms occurred in text that are consistent with a rule, and we set *SUP* to 2.

Heuristics 1
If a pair of terms, such as $R(W_p, W_q)$, was extracted using more than one specific rule, then the bigger the number of rules and the value of $f(R(W_p, W_q))$are, the more possible $R(W_p, W_q)$ is a correct relationships pair.

Definition 2
According to the heuristics 1, *CON* is defined as:

$$CON(R\ (W_p, W_q)) = SUM + \frac{M \times SUM}{M + SUM} \tag{1}$$

$$SUM = \sum_{i}^{M} f(R\ (W_p, W_q)) \tag{2}$$

M is the number of rules. We set *CON* to 15. If the value of *CON* of each candidate is less than 15, then this pair is regard as false one and is deleted.

Heuristics 2 is designed to filter out the noisy candidates extracted using the ambiguous rules. There have mainly two kinds of noisy candidates: i) since the same ambiguous rule reflects different relationships in different contexts, some extracted pairs are marked wrong relationships. That is, we acquire a pair of terms using a specific ambiguous rule, such as hyponymy, but this pair of terms actually reflects other pre-definitional relationship; ii) some extracted pairs of terms take on none of pre-definitional semantic relationships, as a result of overgeneralization of the patterns. In order to remove these false candidates, the following operation will be applied.

Heuristics 2

IF $Set^A \bigcap \tilde{Set}^U = C \neq \varnothing$ THEN $\tilde{Set}^A = Set^A - C$

Where, Set^A and Set^U are the set of pairs that are extracted using ambiguous rules and unambiguous rules respectively. $\tilde{Set}$ is the set of pairs that have been refined. The heuristics 2 means: If there exists some extracted pairs that match against not only unambiguous rules but also ambiguous rules, then these extracted pairs are removed from Set^A. Because refined pairs that were extracted by unambiguous rules are more desirable. After that, we adopt the heuristics 1 to remove the rest false candidates.

4 Experimental Results and Analysis

To verify the validity of the proposed approach, we design three experiments. 50 nouns were selected from a lexicon at random, and then were submitted to Google. Top 500 documents were returned for each selected noun. After deleting duplicated Web documents and wiped off the sentences that have no relevancy with the selected word, the following matching strategies were performed:

Unambiguous rules and ambiguous rules are applied to extract the relationships, as well as the candidates were refined with heuristics. The experimental results are in Table 1.

Only unambiguous rules are applied to extract the relationships. We recorded the two sets of experimental results respectively according as candidates were refined or unrefined with heuristics, shown as Table 2.

Only ambiguous rules are applied to extract the relationship. We recorded the two sets of experimental results respectively according as candidates were refined or unrefined with heuristics, shown as Table 3.

The *precision* performance metrics is defined as:

$$precision = \frac{M}{N} \tag{3}$$

Where, M is the number of correct retrieved relations (human annotator judge the correctness of results provided by system). N is the number of retrieved relations.

The *recall* performance metrics is not computed. Because the goal of our approach is to extract as many valid pairs as possible from the text collection. We do not attempt to capture every instance of such pairs. Instead, we exploit the fact that these pairs will tend to appear multiple times in the collections. As long as we capture one instance of such a pair, we will consider the system to be successful.

Table 1. Results of the First Matching Strategy

Type	Number of retrieved relations N_1	Number of correct retrieved relations M_1	Precision P_1
Synonymy	28	24	85.7%
Hyponymy	73	65	89.0%
Parataxis	32	26	81.3%

The results in Table 1 illustrate that our approach can acquire the most semantic relationships correctly. Hyponymy and hypernymy is a relative relationship. The precisions of these two relationships are same. So we only list the experimental results of hyponymy.

Where, P_{r2} is the precision when candidates were refined with heuristics in experiment 2. P_{u2} is the precision when candidates were unrefined in experiment 2.

As shown in Table 2, the average precision of (95.2%+92.3%+94.4%)/3 is exciting, which is increase of 7.9% than the P_{u2}. This indicates the heuristics to refine the candidates is effective. Nevertheless, comparing M_{r2} with M_1, we conclude that some correct relationships were failed to extract if only unambiguous rules were applied.

Table 2. Results of the Second Matching Strategy

Type	Number of retrieved relations		Number of correctly retrieved relations		Precision	
	N_{r2}	N_{u2}	M_{r2}	M_{u2}	P_{r2}	P_{u2}
Synonymy	21	24	20	20	95.2%	83.3%
Hyponymy	52	58	48	49	92.3%	84.5%
Parataxis	18	19	17	17	94.4%	89.5%

Table 3. Results of the Third Matching Strategy

Type	Number of retrieved relations		Number of correctly retrieved relations		Precision	
	N_{r3}	N_{u3}	M_{r3}	M_{u3}	P_{r3}	P_{u3}
Synonymy	10	13	7	8	70.0%	61.5%
Hyponymy	23	26	17	17	73.9%	65.4%
Parataxis	15	19	9	11	60.0%	57.9%

Where, P_{r3} is the precision when candidates were refined with heuristics in experiment 2. P_{u3} is the precision when candidates were unrefined in experiment 3.

Similarly, from the results of P_{r3} and P_{u3}, we can see that the heuristics can delete the most incorrect relationships, while legitimate ones maybe be eliminated sometimes, e.g. M_{r3} and M_{u3} in the first row and the third row of table 3.

5 Conclusions

We present an approach of semantic relationships acquisition from Chinese Web documents. We don't analyze every sentence in documents, but wipe off the sentences that have no relevancy with the *IW* to dwindle in numbers of text to be analyzed and obtain initial sentence-level text. For a distinct semantic relationship, corresponding unambiguous rules and ambiguous rules designed for Chinese phrases are applied to the initial sentence-level text. Lastly, candidates are refined with two heuristics according to different traits of rules. Compared to other previous works, we apply not only strict unambiguous rules but also loose unambiguous rules to extract relationships and proposed efficient approach to refine the outputs of these rules.

Although linguistic rules we used almost cannot cover all instances, we believe that there are many expressions that convey the same relationships on the Web. So the most semantic relationships can be recognizes through these rules.

The main shortcoming of our approach is that acquisition of rules. At present, these rules are derived from large number of corpus by human, which is inefficient. So we intend to the detection of extraction rules and to discover constrains for all the rules.

References

1. Xia Sun, Qinghua Zheng.: Semantics-based Answers Selection in Question Answering System. In: Proceedings of the 3rd International Conference on Web-Based Learning (ICWL2004), Tsinghua University, Beijing, China (2004)
2. Girju. R., Badulescu. A., Moldovan. D.: Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In: Proceedings of HLT-NAACL (2003)
3. Gildea. D, Jurafsky. D.: Automatically Labeling Semantic classes. In: Proceedings of Annual Conference of the Association for Computational Linguistics, ACL (2004)
4. Pantel. P., Lin, D.: Discovering Word Senses from Text. In: Proceedings of ACM Conference on Knowledge Discovery and Data Mining, SIGKDD (2002)
5. Matthew. B., Eugene. C.: Finding Parts in Very Large Corpora. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Maryland (1999)
6. Li. Y., He. Q., Zhongzhi Shi.: Association Retrieve Based On Concept Semantic Space. Journal of University of Science and Technology, Beijing (2001)
7. WordNet. http://www.cogsci.princeton.edu/~wn/index.shtml
8. Google. http://www.google.com/
9. Wu. P, Chen.Q., Ma. L.: The Study on Large Scale Duplicated Web Pages of Chinese Fast Deletion Algorithm Based on String of Feature Code. Journal of Chinese Information Processing, Beijing, (2003)
10. Qinghua Zheng, Sunjuan Zhang.: A Novel Algorithm of Eliminating the Chinese Word Segmentation Ambiguities for Web Answer. Computer Engineering and Applications (2004)
11. Zhaojing Wang, Qinghua Zheng.: An Approach of POS Tagging for Web Answer. Computer Engineering and Applications (2004)
12. Xia Sun, Qinghua Zheng.: A Method of Special Domain Lexicon Construction Based on Raw Materials. Mini-Micro Systems (2005)