AUTOMATIC EXTRACTION OF LEXICAL RELATIONS FROM CHINESE MACHINE READABLE DICTIONARY

SU-JIAN LI and Qun LIU and SHUO BAI and XUE-QI CHENG

Software Department, P.O.Box 2704, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, China 100080

Abstract

Lexical relations are very important for NLP. Most previous work to get them is done by hand. In this paper, we describe an automated strategy which exploits a machine readable dictionary (MRD) to construct a richly-structured network of lexical relations. In our system lexical relations include five basic semantic relations, two phonetic relations and one orthographic relation. These relations constitute the basic framework of our lexical network. Then we present an approach to use heuristic functions to extract semantic relations while we conduct syntactic parsing. Experimental results demonstrate that our method is effective.

Keywords

Natural Language Processing (NLP), machine readable dictionary (MRD), semantic relation, lexical relation, syntactic parsing

1 Introduction

At present there are many resources such as machine readable dictionaries (MRDs) available, which we can use to extract lexical relations automatically. WordNet (Miller G., 1990) organizes English nouns, verbs, and adjectives into synonym sets linked with different relations. On the base of Wordnet (Miller 1990), FreeNet (Beeferman 1998) enhanced the discovery of lexical connections between words and concepts. Microsoft Research has developed MindNet (Richardson 1998) through Longman's Dictionary of Contemporary English (LDOCE) and automatically constructed a highly interconnected network of words linked by semantic relations. These resources and systems all provide rich information for use in natural language (NL) analysis.

As for Chinese, Mei (1983), Dong (1999) have conducted research on lexical relations and built the semantic mechanism of Chinese words respectively, but most of those work are finished by hand and can only get a limited number of lexical relations. Few researches have been reported to automatically extract lexical relations for Chinese words. In this paper, we describe an automated system, which is designed to enhance the discovery of lexical relations in the Chinese language and construct a richly structured network labelled with lexical relations.

2 Basic Framework of the System

The centerpiece of our system is to get a large-scale knowledge base that we construct automatically from the electronic edition of Modern Chinese Dictionary (MCD). In fact the knowledge base is a network through which all possible relations between Chinese words are represented. The lexical network is composed of nodes and arcs. A node is simply a legal Chinese word. A legal Chinese word might be a word in our lexicon, or a proper noun not in the lexicon such as a Chinese name, a name of an institute et al. Here arcs mean the relations. A relation is a finite set of ordered pairs of nodes, or links. Each relation has a name, expressing a certain connection between words.

First of all, we construct a mechanism of lexical relations. These relations include semantic, phonetic and orthographical relations. At the same time, we adopt MCD as our MRD to extract lexical relations. MCD is a dictionary which includes about 60,000 entries of words, idioms or phrases. Every entry in it is a definition of one word comprising its phonetic notation and meaning. Section 3 and 4 describe the details of these lexical relations, and section 5 illustrates the methods of extracting lexical relations.

3 Basic Semantic Relations

Semantic relations are the main lexical relations. We divide the Chinese words into different categories. A category is one basic unit of classification and one net structure of categories contains several basic semantic relations of Chinese words. We conclude these relations as Synonymous Relation (SR), Antonymous Relation (AR), Constitutive Relation (CR), Logical multi-Hierarchical Relation (LHR), and Non-Monotonous Reasoning Relation (NMRR).

0-7803-7087-2/01/$10.00 © 2001 IEEE

3.1 Several Definitions

In order to describe the semantic relations, first we introduce several definitions:

1) Set of Chinese words $W = \{w1, w2, \ldots, wm\}$, $m > 0$. W is composed of all the Chinese words. $wi(0<i\leqslant m)$ represents the Chinese name of an entity, or an activity, or a feature, in the real world.
2) Set of categories $C = \{C1, C2, \ldots, Ct\}$, $t > 0$. $Ci(0<i\leqslant t)$ represents one particular category which can be looked as one class with some attributes. We suppose every Chinese word can be mapped into one or more than one category. There exist various relations among these categories.
3) Set of attributes of words $A = \{a1, a2, \ldots, an\}$, $n > 0$. A represents a set of the possible attributes of Chinese words. $ai(0<i\leqslant n)$ is one kind of attribute, which might be syntactic, semantic or phonetic. Att_of(a,X), $a \in A, X \in C$, means a is one attribute of the category C.

3.2 Semantic Relations

1) Synonymous Relation (SR): Synonymous words are structured in synsets, underlying a linguistic concept. Every synset is connected with a category, representing a textual definition that can be described in a logical form which is the building block of our knowledge base. This formulation that one category includes at least one word, can provide an elegant manner of localizing ambiguities. In fact Synonymous Relation is a relation of words and it can be seen as the basis of all the semantic relations.
2) Antonymous Relation (AR): if $X,Y \in C$, AR(X,Y) represents X and Y are two categories with antonymous conception. This kind of relation has the characteristic of symmetry. That is, if X is the antonymous category of Y, Y must be the antonymous category of X. When two categories have the antonymous relation, the words they contain also have the corresponding antonymous relation.
3) Logical multi-Hierarchical Relation (LHR): if $X,Y \in C$, IS_A(X,Y) represents that X is one offspring of Y and the attributes of X can inherit from Y. Because a category can inherit from more than one category, we can also have the conclusion: $X,Y1,Y2 \in C$, IS_A(X,Y1) and IS_A(X,Y2) and $Y1 \neq Y2$. Thus we can get a network of categories. This kind of relation is the basis to assign values to attributes in some categories.
4) Constitutive Relation(CR): We assume two categories X and Y satisfy this relation. They have three possible cases: X is the constitutive part of Y; X is a member of Y; X is the constitutive material of Y. No matter in which case, if $X,Y \in C$, CR(X,Y) is used to represent this relation. In our system this relation occupies a large number, because it is widely used to explain the meaning of an entry. CR also demonstrates the idea of inheritance and the ability of inductive learning.
5) Nonmonotonous Reasoning Relation (NMRR):

   $\exists X,Y \in C, \exists a \in A,$
   IS_A(X,Y) and (a in Y) and (a in X)
   if $X:<a> \neq Y:<a>$
   then NMRR(X,Y)

   we define this kind of relation as follows. Here (a in Y) represents that a is one attribute of Y, and X: <a> represents the value of the attribute a in the category of X. We assume that X is an offspring of Y and that both X and Y have the attribute a. By default, X:<a> inherits from Y. However, if X: <a> isn't equal to Y: <a>, but assigned a new value. Then we call the relation between X and Y as nonmonotonous reasoning relation. This relation isn't independent and above all it should satisfy IS_A() relation, but it represents a general phenomenon in nature.

A network of categories can represent those relations discussed above and finally extend to the relations between the glossary of Chinese words. On the other hand, those relations demonstrate inductive and reasoning ability, and bring convenience to semantic computation and analysis.

4 Other Lexical Relations

There are about 10,000 Chinese characters, each of which has its own phonetic notation. All Chinese words are assembled by these characters with variant length and we can get the phonetic notations of every word directly from those characters. Due to the revelation of FreeNet (Beeferman 1998), and from the relation above between Chinese characters and words, we add two phonetic relations: Homonymous relation (HR) and Rhythmic relation (RR), and one orthographic relation (OR).

- Homonymous Relation (HR): Homonymous phenomena are very popular in Chinese. This relation is computed by listing the pinyin (Chinese phonetic notation) of every word and conducting comparison. The Chinese words with the same pinyin notation have the relation of HR, e.g. "北京 (Peking)" and "背景 (background)" have the same phonetic notation "beijing".
- Rhythmic Relation (RR): The phonetic notation of every Chinese word is composed of two parts: an initial consonant, and a simple or compound vowel. This relation is computed by comparing the vowel part of the last character in one word. We can say that the words with the same last vowel have the relation of RR. For example, for Chinese words "语言(language)", "眷恋(be sentimentally attached to)", "收敛 (constringency)", "哀怨 (plaintive)", their phonetic notation are: "yuyan", "juanlian", "shoulian", "aiyuan" respectively. The italic font represents that all those words have the same rhyme "an".
- Orthographic Relation (OR): Different combination of several Chinese characters perhaps constitutes different words. The words with the same Chinese characters have the relation of OR. Take an example, the two Chinese characters "本" and "文" can constitute two words "本文(This paper)" and "文本 (text)", which have the relation of OR.

The phonetic relations allow us to master the rhyme of the Chinese language. They can contribute to finding the rhyming words with certain aims, useful for lyric poetry. And the orthographic relation can be helpful for mastering the Chinese words.

5 Extracting Semantic Relations from MCD

We have defined the basic lexical relations in our system, and the lexical network will be established based on them. Next we train and fill in the network using MCD. We can obtain phonetic relations and orthographic relations according to the notations of the Chinese words. The most important and difficult is to acquire semantic relations from MCD.

In our system, semantic information has been extracted from MCD in a two-step procedure, first selecting the appropriate text in the dictionary, and then parsing the text to identify the possible semantic relations.

5.1 Sentence Selection

In the MCD, most entries are defined regularly, and there exist a few sentences which will cause difficulties in extracting lexical relations. We pick out these sentences and ignore them. The selection of proper sentences is based on two aspects: the first is to consider the length of one sentence; and the second is the syntactic structure of the sentence.

For example, if there exists one sentence S, |S| denotes the length of S. For the convenience of parsing, we prescribe that |S| should be less than or equal to 10, that is, every sentence includes no more than 10 words. As for syntactic structure, we refer to the formula 1:

$$P(H=h,S=s)=P(H=h)*P(S=s) \qquad (1)$$

where H means a head word whose value h is its POS (part of speech), and S represents a sentence in the definition that corresponds to the head word H. The value of S adopts a kind of sentence pattern, which means the POS sequence of the words in one sentence. We suppose that the occurrence of the two events H and S is independent. P(H=h, S=s) is the joint probability that the POS of one head word H is h and one sentence S takes a certain sentence pattern s. According to formula 1, we get a series of possibilities for the combination of the type of the head word and the sentence structure. We define a threshold t, filter out the sentences with the joint possibility less than t, and the remaining sentences are the ones that we want.

$$\begin{cases} |S|\le 10 \\ P(H=h,S=s)\ge t \end{cases} \qquad (2)$$

Thus, according to formula 2 we get a collection of sentences which satisfy the conditions.

5.2 Identifying Semantic Relations

In MCD most words are defined by virtue of their hypernyms, synsets, antonyms and/or constituent parts. If some attribute of a word is different from its ancestor, the dictionary would often make an explanation in the entry. In one entry we call the defined word as a *head word*, and the key words that define the head word as *content words*. A *content word* is called so relatively, and when the content word is defined, then it will also be called a *head word*. In MCD we mainly extract the five semantic relations between head words and content words from the definition of head words. Here we illustrate the following example definitions and see how the sentences in them express the semantic relations.

1. "鸵鸟 (ostrich): 现代鸟类中最大(the largest)的鸟 (bird), 高可达 3 米 (3 metres high), 颈长(a long

neck), 头小(a small head), 嘴扁平(a flat rostra), 翼短小(short wings), 不能飞(can't fly)…"
This sentence is typical, which captures inheritance, meronym, and exception. From this sentence, we can derive the semantic relations as in figure 1.

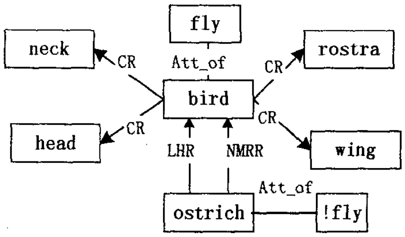

Figure 1:Semantic Relations in Example 1

Here, the words "neck", "head", "rostra" and "wing" are the constituent parts of the word "bird", and therefore they have the relation of CR with "bird" respectively. There exists the relation of LHR between "ostrich" and "bird", and thus "ostrich" has the constituent parts of "neck", "head", "rostra" and "wing". The word of "fly" is one attribute of "bird", but "ostrich" doesn't have the attribute "fly", and so there is the relation of NMRR between "ostrich" and "bird".

2. "好(good): 优点多的(having many virtues), 使人满意的(satisfying), 跟'坏'相对(compared with 'bad')"

From the definition of "good", we can extract one SR and one AR as in figure 2.

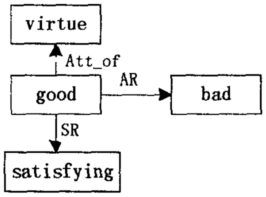

Figure 2:Semantic Relations in Example 2

5.3 Acquiring Semantic Relations

We acquire semantic relations from two levels. In the first level we conclude about 20 simple rules to get semantic relations. These simple rules are related to some fixed phrases, for example, from the Chinese phrases "与…相对 (compared with)", "与…相反 (opposed to)", we often get antonymous relations; on the other hand, the phrases such as "也就是说 (in other words)", "即 (that is)" and so on are connected with synonymous relations.

The second level is that we need a robust parser. After the process of selection, the text we get is composed of short regular phrases, which are easily parsed. And there are fewer ambiguities in the selected text than the general text either from the word level or from sentence level. Different from English, in order to acquire semantic relations, the system involves segmenting every sentence into words, tagging part of speech (POS) for every word, and then parsing them according to our Chinese phrase grammar. During the parsing process, some heuristic functions are conducted to identify the occurrence of possible semantic relations, which are consistently associated with fixed syntactic and lexical patterns.

Now we take some examples to see how to acquire various semantic relations. In the example sentence (2), there is a fixed phrase "跟…相对 (compared with…)", and thus we can get the antonymous semantic relation between "好(good)" and "坏(bad)".

For the example definition 1, the head word is "ostrich", we parse one sentence of its definition according to the following fragment of grammar rules as in figure 3.

```
1)  S   →NP
2)  S   →Noun  VP              //  Att_of(2,1) CR(1,*)
3)  NP  →AP    '的    Noun  //  LHR(3,*)
4)  NP  →NP    '的    Noun  //  LHR(3,*)
5)  NP  →Noun                  //  SR(1,*)
6)  VP  →Verb
7)  AP  →DP    Adj
8)  DP  →Noun  Noun  '中
```

Figure 3. Fragment of Grammar Rules

Here 'Noun', 'Verb' and 'Adj' in the grammar rules are POS which we tag for nouns, verbs, and adjectives respectively. And NP, VP, AP, DP mean some kind of phrase respectively. The functions after symbol "//" are heuristic functions which can be used to acquire semantic relations. In the parsing process after every reduction completes, the result such as one syntactic structure will be submitted to some heuristic functions to obtain possible semantic relations. Every heuristic function contains two parameters which are the pairs of one relation, and an Arabic numeral marks the relative position of some content word, and the symbol '*' denotes the current head word. In rules, the Chinese words, which are quoted in the rules, are usually function words commonly used. Now we parse the first sentence in the example definition 1 and see the results in Table 1.

	Example Sentence and parsing results	Rule No.
Original Sentence	鸵鸟: 现代鸟类中最大的鸟	
After segmentation	鸵鸟/noun: 现代/noun 鸟类/noun 中/最大/adj 的/鸟/noun	
Parsing process	鸵鸟/noun: (现代/noun 鸟类/noun 中/) DP 最大/adj 的/鸟/noun	8
	鸵鸟/noun: ((现代/noun 鸟类/noun 中/) DP 最大 /adj) AP 的/鸟/noun	7
	鸵鸟/noun: (((现代/noun 鸟类/noun 中/) DP 最大 /adj) AP 的/鸟/noun)NP	3

Table 1. parsing process of an example sentence

In Table 1, we can see that the final result obtains one NP, using Rule 3 and then calling the function LHR(3,*) to arrive at one relation of LHR between the two words "鸟 (bird)" and "鸵鸟 (ostrich)". The other short sentences in the definition can also be parsed as the example sentence in table 1, and then we can get other semantic relations between the content words and the head word "ostrich".

In order to acquire semantic relations, we need to scan every sentence two times: one to find the fixed pattern and one to parse the sentence. At present, our system includes more than 120 grammar rules and according to these rules get about 300,000 semantic relations. In addition, we have obtained about 3,960 HR relations, 968,750 RR relations and 656 pairs of words with OR relation.

6 Conclusion

This paper discussed a strategy to automatically construct a large knowledge base with lexical relations. Especially we present the approach of combining grammar rules with heuristic functions to extract semantic relations from selected sentences in a machine readable dictionary. In view of the regularity of the texts in a dictionary, this approach is very effective, sparing much human labour at the same time. Although our system is designed specially for extracting Chinese lexical relations, the method can be applied to other languages. The knowledge base we get is useful for many applications, such as machine translation system, information retrieval system etc. The limitation of our system is that the classification of semantic relations between words is too coarse, and our next work is to consider expanding the types of semantic relations.

References

1. Baker C.F., Fillmore C.J., and Lower J.B., The Berkeley FrameNet Project, In Proc. of COLING-ACL'98, pp. 86—90, 1998.

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.

3. Dolan W., L. Vanderwende, and S. Richardson, Automatically Deriving Structured Knowledge Bases from On-line Dictionaries, In Proceedings of First Conference of the Pacific Association for Computational Linguistics (Vancouver, Canada), pp. 5—14, 1993.

4. Dong Zhendong and Dong Qiang, The Introduction to Hownet, http://www.keenage.com, 1999.

5. Lv Shuxiang and Ding Shengshu et al, Modern Chinese Dictionary (revised edition), The Commercial Press, 1999.

6. Mei Jiaju, Chinese thesaurus «Tongyici Cilin», Shanghai thesaurus Press, 1983.

7. Miller G., Wordnet: An On-line Lexical Database. International Journal of Lexicography, 3(4), 1990.

8. Richardson S. D., Dolan W.B., and Vandervende L., Mindnet: acquiring and structuring semantic information from text, In Proc. Of COLING-ACL'98, pp. 1098—1102, 1998.