Semantic Computation in a Chinese Question-Answering System

LI Sujian (李素建)[1], ZHANG Jian (张 健)[2], HUANG Xiong (黄 雄)[2], BAI Shuo (白 硕)[2] and LIU Qun (刘 群)[2]

[1] *Institute of Computational Linguistics, Peking University, Beijing 100871, P.R. China*

[2] *Software Department, Institute of Computing Technology, The Chinese Academy of Sciences Beijing 100080, P.R. China*

E-mail: lisujian@pku.edu.cn

Received May 10, 2001; revised October 19, 2001.

Abstract This paper introduces a kind of semantic computation and presents how to combine it into our Chinese Question-Answering (QA) system. Based on two kinds of language resources, *Hownet* and *Cilin*, we present an approach to computing the similarity and relevancy between words. Using these results, we can calculate the relevancy between two sentences and then get the optimal answer for the query in the system. The calculation adopts quantitative methods and can be incorporated into QA systems easily, avoiding some difficulties in conventional NLP (Natural Language Processing) problems. The experiments show that the results are satisfactory.

Keywords similarity, relevancy, Hownet, question-answering, natural language processing

1 Introduction

With the explosion of information available on the Internet, Question-Answering systems can help us to find what closely matches users' needs. Since both questions and answers are mostly expressed in natural languages, Q/A methodologies have to incorporate NLP techniques, including syntactic and semantic computations. Due to the encouragement of the Text Retrieval Conference (TREC) and the Message Understanding Conferences (MUCs), some QA systems have achieved good performance[1]. However, these systems mainly aim at English. In this paper, based on these characteristics and some language resources, we build a Chinese Question-Answering system through the computation of semantic similarity and relevancy.

2 Overview of Language Resources

Hownet is a free Chinese-English bilingual resource which is released recently on the Internet[2~4]. It is a knowledge base describing the relations between concepts and the relations between the attributes of concepts. In our Chinese QA system we mainly use the knowledge base which includes 66,681 concepts. Every word sense is represented by the combination of several sememes. A sememe is a basic semantic unit that is indivisible in *Hownet*. According to the view of ontology, about 1500 sememes are extracted to compose an elementary set forming the basis of the Chinese glossary, as over 100 kinds of chemical elements constitute all the substances in nature. We describe several definitions in *Hownet* as follows:

$$SS = \{s_1, s_2, \ldots, s_n\}, \quad n = 1541$$
$$WS = \{c_1, c_2, \ldots, c_m\}, \quad m = 66,681$$
$$REL = \{*, @, ?, !, \sim, \#, \$, \%, \hat{}, \&, NULL\}$$
$$c_i \Rightarrow r_{i1}s_{i1}, r_{i2}s_{i2}, \ldots, r_{ik}s_{ik}, \quad r_{it} \in REL, \quad s_{it} \in SS \quad (1 < t < k)$$

This work is supported by the NKBRSF of China (Grant Nos.G1998030510, G1998030507-4).

where SS represents the set of the sememes which includes 1,541 elements; WS represents the set of the word senses in *Hownet* whose size is 66,681; REL is the set which describes the relations between a concept and a sememe or the relations between sememes. For every word sense c_i, a concept, its definition is composed by k items, each of which includes a relation symbol in REL and a sememe in SS.

In our system, another language resource available is *Cilin*[5], a Chinese Thesaurus, which conducts semantic classification for Chinese words. It comprises 12 major categories, 94 medium categories, and 1,428 minor categories. And the minor categories can be further divided into synsets according to their meanings. Every synset includes several words with the same or similar meanings. This hierarchical classification embodies synonymous relations and hyponym relations, and provides convenience for the expansion and semantic computation of word senses. We formalize several definitions as follows:

$$WS' = \{c_1, c_2, \ldots, c_{m'}\}, \quad m' = 61,125$$
$$SC = \{sc_1, sc_2, \ldots, sc_p\}, \quad p = 11,832$$

where WS' represents the set of word senses in *Cilin*, whose size is 61,125, and SC represents the set of synsets whose size is 11,832.

The two language resources introduced above are of great help to our computation in semantic similarity and relevancy of any two Chinese words.

3 System Description

At present, the processing mechanism of most QA systems are based on sentences[6], and at the same time, it absorbs the techniques of information retrieval, information extraction and natural language processing[7]. As shown in Fig.1, for the large quantity of information from the Internet, keywords and mood words such as those extracted from queries are inputted to the process of Information Retrieval to reduce the scope of searching, and at the same time the sentences whose mode or negative/positive mood is not consistent with the query sentence are also filtered out. Then the results obtained and the question needed to query are submitted simultaneously to the modules involved in natural language processing. These modules include the segmentation module, the entity recognition module, and the semantic annotation module. After the processing of these modules, we can get sentences with semantic annotation which can enter the module of semantic computation (SC). The SC module gets the relevancies between sentence pairs. Then we select the sentence pairs with the largest value of relevancy.

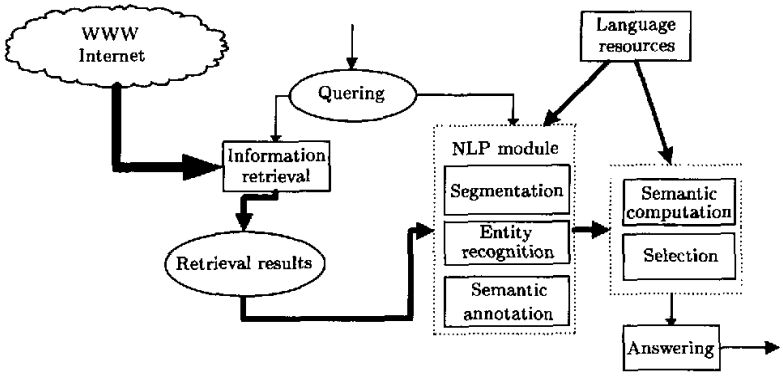

Fig.1. System structure.

In Fig.1, the thicker the line, the more information it represents. The language resources include *Hownet* and *Cilin*. According to the characteristics of the Chinese language, we must conduct segmentation for sentences. At the same time or after segmentation, the named entity should also be

picked out and semantic annotation is conducted for the segmented words and the named entity. The three natural language modules do not have explicit boundary. Based on the semantic information collected in the three NLP modules, we conduct semantic computation between query and relevant sentences. The main function of the semantic computation module is to get the relevancy value between sentence pairs and to sort them. This paper mainly discusses the techniques concerning how to conduct semantic computation.

4 Semantic Computation

Semantic computation is the kernel of our system, which is conducted by three steps. The first step is to conduct the computation of the similarity and association between sememes. Second, similarity and relevancy between words are computed. In the last step, based on the results of the two steps above, we can calculate the relevancy between sentences and get the sentence pairs with the maximal value of relevancy.

4.1 Similarity and Associativity Between Sememes

In *Hownet*, the relations among sememes are built through several feature files. The sememes in one feature file construct a tree structure. As shown in Fig.2, this is a sample structure of nodes that belong to the feature files. Relations between sememes can be obtained from these hierarchical trees and based on these relations we can compute the similarity and association between sememes within this mechanism. Every node is called a *main sememe*. Every main sememe is followed by some sememes included in the square brackets, which we can see as its explanation, called *explanatory sememes*. Every explanatory sememe is usually preceded by a symbol which describes its relation with the main sememe. Both main sememes and their explanatory sememes have hyponyms and hypernyms, thus we can get association between sememes in different feature files. It is followed that all the sememes in *Hownet* construct a network structure.

In Fig.2, the relation between a main sememe and its hypernym or hyponym is called *Vertical Relation*, we measure the sememes with Vertical Relations with similarity; other relations which span different feature structures are called *Horizontal Relation* which can be measured by association between sememes.

```
- entity | 实体
  ⊢ thing | 万物 [#time | 时间, #space | 空间]
  ···⊢ physical | 物质 [!appearance | 外观]
    ···⊢ animate | 生物 [*alive | 活着, !age | 年龄, *die | 死, *metabolize | 代谢]
      ···⊢ AnimalHuman | 动物 [!sex | 性别, *AlterLocation | 变空间位置, *StateMental | 精神状态]
        ···⊢ human | 人 [!name | 姓名, !wisdom | 智慧, !ability | 能力, ! occupation | 职位, *act | 行动]
          |⊢ humanized | 拟人 [fake | 伪]
          ⊢ animal | 兽 [^* GetKnowledge | 认知]
            ⊢ beast | 走兽 [^*GetKnowledge | 认知]
        ···
- event | 事件
    ⊢ static | 静态
    |⊢ relation | 关系
  ···
```

Fig.2. A sample tree structure of feature sememes.

For the two sememes in the tree structure of Fig.2, there exist three possible relations:

1) When the two sememes are in different trees, the similarity will be 0;

2) The two sememes at least have one common ancestral node, but they are in different branches of the ancestral node;

3) One sememe is the ancestral node of the other one.

Then, we compute the similarity between sememes as equations in (1):

$$\mathrm{sim}(s_1, s_2) = \begin{cases} \alpha/\mathrm{dist}(s_1, s_2), & t(s_1) = t(s_2) \\ 0, & t(s_1) \neq t(s_2) \end{cases} \quad s_1, s_2 \in SS' \tag{1}$$

where, for any two sememes s_1, s_2 in the sememe set SS, $\mathrm{sim}(s_1, s_2)$ represents the similarity between s_1 and s_2, $t(s_1) = t(s_2)$ represents that the two sememes are in one tree structure and their similarity is inversely proportional to their distance.

Like the structure of Fig.2, the explanatory sememes build a bridge for two sememes in different trees. For example, there should exist some relation between the sememes 'animate | 生物' and 'alive | 活着' which do not have any similarity at all. Here we introduce a new measure called association to represent those relations spanning different trees. In doing so, the tree structure becomes a net structure. In order to compute associations, we need expanded the current sememes in two directions. One is expanded to the hypernyms of explanatory sememes, which is called Horizontal Associative Expansion (HAE), the other expansion is to the explanatory sememes of the hypernyms, which is called Vertical Associative Expansion (VAE). We compute associations according to the equations in (2):

$$\begin{cases} \mathrm{ext}(s_j) = \{s_i | \mathrm{REL}(s_j, s_i)\} \\ \mathrm{Asso}(s_1, s_2) = \sum\limits_{s_1 \in \mathrm{ext}(s_1)} w_i \mathrm{sim}(s_i, s_2) + \sum\limits_{s_j \in \mathrm{ext}(s_2)} w_j \mathrm{sim}(s_1, s_j) \end{cases} \tag{2}$$

where $\mathrm{ext}(s_j)$ is an extension set of the sememe s_i which includes HAE and VAE. We endow a weight to every relation in REL which describes how this kind of relation has an influence on association. In computing the association between s_1 and s_2, the first part represents the association between s_2 and the extensive set of s_1; and the second part is for s_1 and the extensive set of s_2.

4.2 Similarity and Relevancy Between Words

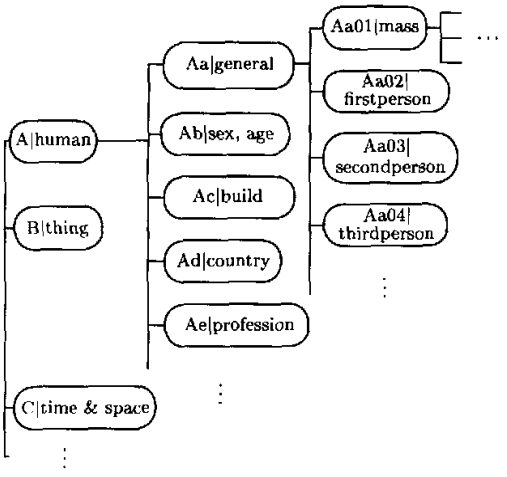

Fig.3. A sample structure in Cilin.

In Section 2 we have introduced two kinds of language resources. For *Hownet* it is easier to construct a net structure for sememes and then to get their similarities and associations. Because every word sense is composed of sememes, it is difficult for *Hownet* to expand the similar or same word senses. Now we utilize the second language resource — *Cilin* — to make expansion of conceptions. As in Fig.3, it is a sample structure of conceptions in *Cilin*. Every node is a semantic class. The nearer to the root node, the more abstract the conception that the node represents. Unlike *Hownet*, not every node in the structure represents a concrete word sense, and only the leaf node is a collection of Chinese words with the same or similar sense.

Similar to the computation of sememes, we have the following equation:

$$\mathrm{sim}(c_1, c_2) = \begin{cases} \alpha/\mathrm{dist}(c_1, c_2), & t'(c_1) = t'(c_2) \\ 0, & t'(c_1) \neq t'(c_2) \end{cases} \quad c_1, c_2 \in WS' \tag{3}$$

where c_1 and c_2 are any two word senses in *Cilin*, $t'(c_1) = t'(c_2)$ represents that the two conceptions belong to the same semantic class and their similarity is inversely proportional to their distance.

Here we adopt a measure called relevancy to represent the associative relation between word senses. The goal of computing the similarity and association between sememes is to get the relevancy of word senses according to the equations in (4):

$$\begin{cases} \mathrm{Rele}(c_1, c_2) = \mathrm{Rele}(\mathrm{def}(c_1), \mathrm{def}(c_2)) \\ \mathrm{Rele}(\mathrm{def}(c_1), \mathrm{def}(c_2)) \approx \sum\limits_{s_i \in \mathrm{def}(c_1)} \max\limits_{s_j \in \mathrm{def}(c_2)} \mathrm{Rele}(s_i, s_j) \\ \mathrm{def}(c) = \{s_i | \mathrm{REL}(c, s_i)\} \\ \mathrm{Rele}(s_i, s_j) = w_s \mathrm{sim}(s_i, s_j) + w_a \mathrm{asso}(s_i, s_j) \end{cases} \tag{4}$$

where $\mathrm{Rele}(c_1, c_2)$ is the relevancy between two word senses c_1 and c_2, and $\mathrm{def}(c)$ is a set of explanatory sememes for the word sense c, w_s and w_a are the weights of the similarity and association between sememes respectively, and we can get a relevancy between sememes $\mathrm{Rele}(s_i, s_j)$. To get the relevancy of two sets of sememes, we pick out the possible sememe pairs with the maximal value and sum them up.

4.3 Relevancy Between Sentences

We assume that the filtered sentences s_1 and s_2 have been segmented, resolved anaphorically and annotated semantically. Then s_1 and s_2 can be regarded as two sequences of m and n keywords: $w_{11}w_{12}\ldots w_{1m}$ and $w_{21}w_{22}\ldots w_{2n}$.

To compute the relevancy of a sentence pair, we use the similarity and relevancy of word pairs. We select the word pairs that contribute most to the relevancy of sentence pairs. The word pairs are connected with lines as in Fig.4.

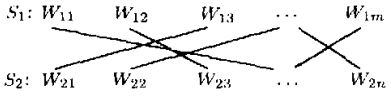

Fig.4. Word pairs in two sentences.

We use a dynamic programming algorithm to get the relevancy of a sentence pair by the following equations:

$$\begin{cases} \mathrm{Rele}(S_1, S_2) = M_{m,n} \\ 1/d_{ij} = \alpha \mathrm{Rele}(w_{1i}, w_{2j}) + \beta \mathrm{sim}(w_{1i}, w_{2j}) \\ M_{0,j} = M_{i,0} = 0 \\ M_{1,1} = 1/d_{1,1} \\ M_{i,j} = \max\limits_{1 \le k \le n} \{1/d_{ik} + M_{i-1,j}\} \end{cases} \tag{5}$$

where α and β are weights representing the degree in which the similarity and relevancy of words contribute to the relevancy of sentence pairs, d_{ij} is the semantic distance between the i-th word in the first sentence and the j-th word in the second sentence. According to the recursive equation, we can finally get the value of $M_{m,n}$ which represents the relevancy between the two sentences s_1 and s_2.

After we get the relevancy of all sentence pairs, we compare their values. The larger the value of relevancy, the more relevant the two sentences. We get the sentence as the answer of the query that has the largest value of relevancy.

5 Experiments and Discussion

The semantic computation contains three steps and every step makes use of the computation of the last step. The three steps conform to the characteristics of the Chinese language: from morphemes to words to phrases.

We did experiments on every step above, and the results are satisfactory, reflecting the correlation between elements in every step. Here are some examples: Table 1 illustrates the similarity and association of some example sememe pairs, and the examples in Table 2 demonstrate the similarity and relevancy of some word pairs.

In Table 1 and Table 2, due to the difference of weights, the quantitative levels of different measures are different and we should compare them vertically.

Table 1. Example of Sememe Pairs and Their Similarity and Associativity

Sememe 1	Sememe 2	Sim	Sememe 1	Sememe 2	Asso
Discuss | 商讨	Debate| 辩论	0.80	Material | 材料	Consume | 摄取	0.35
TalkNonsense | 瞎说	Debate | 辩论	0.32	Human | 人	Act | 行动	0.80
Spread | 撒	Throw | 扔	0.40	Produce | 制造	Software | 软件	0.40
Cook | 吐出	Throw | 扔	0.533	Compile | 编辑	Software | 软件	0.80
Dream | 做梦	Cool | 制冷	0.114	Planting | 栽植	FlowerGrass | 花草	0.80
Mental | 精神	Machine | 机器	0.267	CauseToLive | 使活	FlowerGrass | 花草	0.267

Table 2. Example of Word Pairs and Their Similarity and Relevancy

Word 1	Word 2	Sim	Word 1	Word 2	Rele
摇动 (shake)	晃动 (rock)	0.90	致意 (give one's regards)	恰巧 (by chance)	0.0
摇动 (shake)	移动 (move)	0.64	实行 (implement)	笑 (smile)	0.267
病人 (patient)	医院 (hospital)	0.00	病人 (patient)	医院 (hospital)	51.995
医生 (doctor)	病人 (patient)	0.410	医生 (physician)	生病 (be ill)	50.107
医生 (doctor)	护士 (nurse)	0.64	勤劳 (diligent)	富裕 (wealthy)	51.307
揣测 (guess)	了解 (know)	0.512	贫穷 (poor)	懒惰 (lazy)	51.657
揣测 (guess)	推想 (suppose)	0.90	勤劳 (diligence)	贫穷 (poor)	27.457
反常 (abnormal)	奇怪 (strange)	0.64	写 (write)	作者 (author)	33.000

We use the IR module to retrieve 20 relevant documents and extract 50 sentences averagely. So for every query sentence there are about 1,000 sentences. We calculate and sort these 1,000 relevancy values between the retrieved sentences and the query sentence, and finally get one or more sentences with the largest value as answers. We illustrate 5 queries to show the effect of our Q-A system. 93 people are selected to evaluate whether these answers are reasonable. This evaluation is simplified with the following standard: if one person thinks the answer reasonable, the score is incremented by 1; otherwise, the score remains unchanged. Then the maximal score by which one answer can get is 93. In Table 3, the first column represents the No. of one query sentence; the second is the sum of the retrieved sentences; the third column represents the largest relevancy which we get by semantic computation; and the last column records the score of one answer.

From Table 3, we can see that the answers are reasonable for most people. The largest values of relevancy for every query are very different, which is because our computation is dependent on the length and words of one sentence.

Table 3. Results of Several Queries in Q-A

Query No.	Relevant sentences	Largest relevancy	Score
1st	1,029	205.127	89
2nd	986	232.411	93
3rd	997	334.826	92
4th	1,003	602.133	93
5th	1,002	603.329	91

6 Conclusions

This paper mainly introduces the application of semantic computation in our Question-Answering system. We can compute the similarity and relevancy between words, and get the optimal result by calculating the relevancy between sentences. Our method conforms to the characteristics of the Chinese language, combining semantic information with the computation at three levels and avoiding a lot of complexities in language processing. At the same time, the results of the intermediate process, such as the similarity and association between sememes, and the similarity and relevancy between word senses, are also very helpful in other research fields, e.g., polysemous disambiguation clustering, and bilingual alignment, to name a few.

Acknowledgements The authors would like to thank Dr. LU Song, Ms. LIANG Yan for their help on this work, and thanks also go to anonymous reviewers for their valuable comments on this paper.

References

[1] Voorhees E. The TREC-8 question answering track report. National Institute of Standards and Technology, 1999, p.77.
[2] Dong Zhendong. Hownet. http://www.keenage.com, 1999.
[3] Zhou Qiang, Feng Songyan. Building a relation network representation for How-net. In *Proc. 2000 International Conference on Multilingual Information Processing*, Urumqi, China, 2000, pp.139-145.
[4] Gan K W, Wong P W. Annotating information structures in Chinese texts using HowNet. In *Second Chinese Language Processing Workshop*, Hong Kong, China, 2000, pp.85-92.
[5] Mei Jiaju. Tongyici Cilin. Shanghai Thesaurus Press, 1983.
[6] Katz B. From sentence processing to information access on the World Wide Web. AAAI Spring Symposium on Natural Language Processing for the World Wide Web, Stanford University, Stanford CA, 1997.
[7] Rohini Srihari, Wei Li. Information extraction supported question answering. (Cymfony Inc.) In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, National Institute of Standards and Technology, Gaithersburg MD, 1999.

LI Sujian received her B.S. and M.S. degrees in computer science from Shandong University of Technology in 1996 and in 1999 respectively. And she received her Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences in 2002. She is now a postdoctor in the Institute of Computational Linguistics, Peking University. Her current research interests include machine translation, information extraction, and machine learning.

ZHANG Jian received his B.S. degree in physical oceanography from the Ocean University of Qingdao, China in 1998, and his M.S. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences in 2001. Now he is a Ph.D. candidate at the School of Computer Science, Carnegie Mellon University. His research interests include machine learning, information retrieval and data mining.

HUANG Xiong received his B.S. and M.S. degrees from Peking University in 1992 and 1995 respectively. He received his Ph.D. degree from Beijing University of Aeronautics and Astronautics in 1999. From May, 1999 to May, 2001 he conducted research in the Institute of Computing Technology as a postdoctoral fellow. His major interests lie in analysis and design of combinatorial algorithms, computational complexity, Web information retrieval and Web application development.

BAI Shuo received his M.S. and Ph.D. degrees in computer science from Peking University in 1987 and 1990 respectively. Then he conducted research as a postdoctoral fellow in the Mathematics Department of Peking University. He has published more than 60 papers in refereed journals and conferences. His research interests are computational linguistics, natural language processing and network security.

LIU Qun is an associate professor in the Institute of Computing Technology, Chinese Academy of Sciences. He received his B.S. degree in computer science from the University of Science and Technology of China in 1989 and his M.S. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences in 1992. He is pursuing his Ph.D. degree in computer science in Peking University from 1999 till now. His research interests include machine translation, natural language processing and Chinese information processing.