The Research and Realization about Question Answer System based on Natural Language Processing

Qinglin Guo[1,2], Kehe Wu[1,2] and Wei Li[1,2]

1. School of Computer Science and Technology, North China Electric Power University, Beijing, 102206, China
2. Key Laboratory of Condition Monitoring and Control for Power Plant Equipment, Ministry of Education, Beijing, 102206
qlguo88@sohu.com

Abstract

Automatic Question Answer System(QAS)is a kind of high-powered software system based on Internet. Its key technology is the interrelated technology based on natural language understanding, including the construction of knowledge base and corpus, the Word Segmentation and POS Tagging of text, the Grammatical Analysis and Semantic Analysis of sentences etc. This thesis dissertated mainly the denotation of knowledge-information based on semantic network in QAS, the stochastic syntax-parse model named LSF of knowledge-information in QAS, the structure and constitution of QAS. And the LSF model parameters were exercised; it proved that they are feasible. At the same time, through "the limited-domain QAS" which was exploited for banks by us, these technologies are proved effective and propagable.

1. Introduction

More and more information are now available in machine-readable form. It makes the technology of Information Retrieval and Information Extraction more important for effectively looking up and making use of these information. But there exist some shortcomings with traditional search engines. The users' requirements are expressed with the keywords, which may result in the loss of semantic information. Search engine returns the relevant links or document lists, and users need more efforts to acquire the needed information. The research of Question Answering is to resolve these problems. It accepts the questions in natural language that denoting user requirements and returns the exact answers after analysing the document information. This is a challenging task to computers although it seems simple. We have a deep research on the technique of question answering based on this premise.

Automatic Question Answer System (QAS) is a kind of high-powered software system based on Internet [1]. After the special management aiming at one domain Knowledge Base, users can ask some questions on natural language through browser. The system can offer automatically answers of multimedia, make correlative statistics and give proper suggestions. It may be widely used in digital city construction, client consultation, long-distance on-line automatic answering, mobile wireless INTERNET business, and etc. Its key technology is the interrelated technology based on natural language understanding, including the construction of knowledge base and corpus [2], the Word Segmentation and POS Tagging of text, the Grammatical Analysis and Semantic Analysis of sentences etc. Question Answering System is a very hot spot and difficulty spot in the research community of natural language processing; it combines natural language processing techniquese and information retrieval techniques etc. A Question Answering System can return user a concise and accurate answer for question in natural language. But there is still no mature Question Answering System exploited by now, because we know that let a computer to understand human language is so difficult. This thesis dissertated mainly the denotation of knowledge-information based on semantic network in QAS, the stochastic syntax-parse model named LSF of knowledge-information in QAS, the structure and constitution of QAS. And the LSF model parameters were exercised.

2. The structure and constitution of QAS

QAS is of multiplayer B/S structure. Its logical structure includes Web Server, Application Server, and Database Server. Web Server supplies the service of input and output; Application Server supplies Web Server with the service of logical computing. And Database Server administers Application database, semantic Knowledge Base, and limited-domain Knowledge Base.

0-7695-2882-1/07 $25.00 ©2007 IEEE

QAS includes: (1) user interface: can communicate with uses in Chinese or English, receive and answer their natural language questions. (2) Questions analyzing subsystem: according to known question rules, frequent question corpus, sentence model and relevant environment, in addition to combine concrete terms and resources of grammar and meaning, so we may analyze users' questions, give asking points and information collection leading to corresponding keys after settlement of various meanings as well as indicative relations. (3) Dictionary and Knowledge base: store limited-domain or non limited-domain knowledge resource. (4) XML service subsystem: wanted keys are gained from corresponding electrical resources according to reference information collection of question analyzing subsystem. (5)keys processing model: semantic relations among various keys are analyzed as well as combine and cut properly. (6) Knowledge resources administrating subsystem [3]: it's working at words wareroom, rules wareroom, and domain Knowledge Base and corpus, which are used in QAS. It should provide video knowledge maintenance surface, can show structure and relations among different knowledge resources in order to fetch and maintenance Knowledge Base [4]. Every model in this system is involved in natural language technology such as knowledge information expressing and syntax analyzing which are sticking points of QAS.

3. The denotation of knowledge-information based on semantic network in QAS

QAS means searching for keys to questions in Knowledge Base, organizing keys and answering automatically for users' questions. Keys or key points in Knowledge Base are knowledge-information that is information including knowledge, and then how can knowledge information be put? It is known that knowledge-information may be expressed in several ways, for example, logic method, producing rule method, frame method and concept secondary method. They have different merits, but a common problem is that knowledge-information denotation is so fixing and isolating that knowledge points are torn to pieces and difficult to set up contact. Above cases are not adapt to QAS because QAS does not take out existing keys from Knowledge Base, but organize integrated answers after searching for knowledge points. Certainly there should be some contacts among knowledge points. So we use knowledge-information semantic network denotation in QAS.

3.1 The structure of knowledge-information semantic network

Semantic network consist of nodes and arcs among nodes [5]. Commonly, nodes are used to show physical entity, concept or state, while arcs showing relations among them. For instance, Fig.1 is a semantic network for describes "MY-CHAIR".

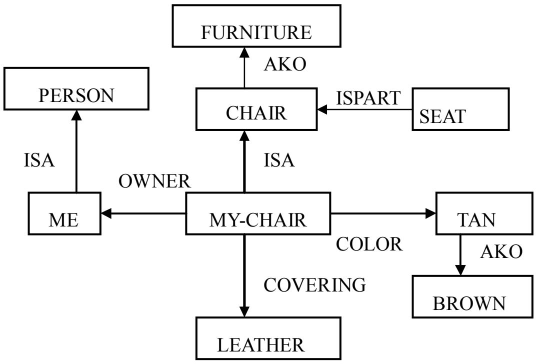

Fig.1 The semantic network of "chair"

In Fig.1, the parts above the node "MY-CHAIR" mean "MY CHAIR IS A CHAIR","A CHAIR IS A KIND OF FURNITURE", " SEAT IS A PART OF A CHAIR"; The left parts of " MY-CHAIR" mean "MY CHAIR BELONG TO ME", "I AM A PERSON"; The right parts of "MY CHAIR" mean "MY CHAIR IS PALM BROWN", "PALM BROWN IS A SORT OF BROWN"; the parts below "MY-CHAIR" mean "MY CHAIR IS COVERD WITH LEATHER". "ISA" and "AKO" are normal relations in semantic network. "ISA" which is read "is…an example" means that a certain individual is one element in some set, such as "MY CHAIR IS A CHAIR"; "AKO" is abbreviation of A-KIND-OF, it denoting a set is a subset of another set, an good example is "A CHAIR IS A KIND OF FURNITURE". The relations in Fig.2 namely "ISPART", "OWNER", "COLOR" and "COVERING" denote attributes of node object. We can see from Fig.2 that semantic network can describe and express knowledge-information distinctly

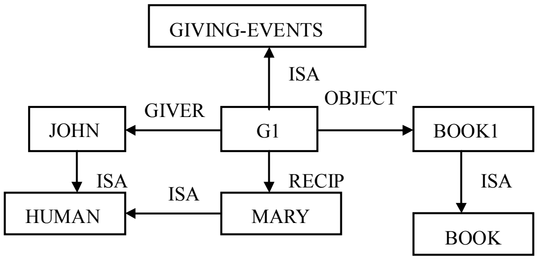

Fig.2 The semantic network of the sentence

3.2 The denotation of semantic network's predicate logic

Semantic network is good for AKO [6], but both FURNITURE and multi ones can be denoted by AKO. For instance, FURNITURE (CHAIR) may be denoted by AKO (CHAIR, FURNITURE). Multi ones may be changed into FURNITURE and AKO. For example, the predication of “John gives a book to Mary” is denoted logically:

(∃x)[GIVE(JOHN, x, MARY)□BOOK(x)]

May be denoted by AKO:

ISA (G1, GIVING-EVENTS)
GIVER (G1, JOHN) RECIP (G1, MARY)
OBJECT (G1, BOOK1) ISA (BOOK1, BOOK)

Of course, more information may be added. For instance:

HUMAN (JOHN) HUMAN (MARY)

Fig.2 denotes corresponding semantic network.

4. The stochastic syntax-parse model named LSF of knowledge-information in QAS

Local environment information is regarded as an important means to WSD in sentence structure all along [7]. But in some lingual models, which are assigned by probability on the basis of rules traditionally, the probability of grammar-producing model is only decided by non-terminal, while is independent of glossarial example in analyzing tree. This quality of non-vocabulary makes lingual phenomena description inadequate for probability model. Therefore, QAS adopts the stochastic syntax-parse model named LSF.

4.1 Model describing

Here, we describe a sort of basic probability depending model. It is named lexical semantic frame (LSF for short [8,9]) in order to be put easily. LSF is supposed as a result of character string $s=w_i\ldots w_j$, SR(R, h, w_i)denotes that wi among LSF relies on the word h through semantic relation, thus we can write down the function SR(i)=SR(R, h, w_i). Analyzing semantic probability p(SR(i)|h, w_i)among words is on the basis of this model. The model supposes that there exists high conjunction between depending relation R and Hyponym node, the contradiction of data sparsely is less. So we can give LSF the analyzing probability from $w_i\ldots w_j$:

$$\mathrm{P}(\mathrm{LSF}|w_i\ldots w_j)=\prod_{k=i,w_k\neq h}^{j}\mathrm{P}(\mathrm{SR}(k)|h,\ w_k) \qquad (1)$$

If we input $S=w_1\ldots w_n$, the task of stochastic analyzer lies in finding the best analysis T*:

$$\mathrm{T}^*=\arg\max_{T\in Parse(w_1^n)} p(\mathrm{T}|\,w_1^n) \qquad (2)$$

Among those, Parse (w_1^n) should be all possible structure-parses for the input sentence, p (T, w_1^n) is defined as the product of all used LSF probability in analysis. We make out that the probability-parse model is based on Bi-vocabulary depending relation.

4.2 Exercise the model parameters

Unlike rules probability model, the probability model parameter based on vocabulary association is usually gained from supervised training as well as using tagged corpus. In fact, The reasons that we use both the words in corpus and their Hyponym POS information to estimate P(LSF |$w_i\ldots w_j$) are:

vocabulary information plays a vital WSD role on matching of depending frame.

considering the limit to corpus scale, words repetition has little probability in sentence analysis, we must deal with statistic result smoothly [10]. Vocabulary information is needed to “magnify” to reduce the degree of data sparseness with the help of Hyponym part of speech. But the close word class such as preposition or adverb uses statistic information of words.

Ordering W to be the set of whole vocabulary in training corpus, SUBCAT to be the set of all Hyponym part of speech, CORPUS to be training set, RS to be set of semantic depending relation, LSF to be the semantic frame of vocabulary, we can define the following functions:

$$\mathrm{CSR}(\mathrm{SR},\ \mathrm{LSF})=\begin{cases}0, & \mathrm{SR}(R,\ h,\ w_i)\notin \mathrm{LSF}\\ 1, & \mathrm{SR}(R,\ h,\ w_i)\in \mathrm{LSF}\end{cases} \qquad (3)$$

Above CSR can denote the counting condition of depending relation in corpus training [11].

Counting function C(<w_h, sc_h>, <w_c, sc_c>)denotes the occurring times of upright depending relation (<w_h, sc_h>, <w_c, sc_c>) in training corpus, and w_h, $w_c\in W$, sc_h, $sc_c\in$ SUBCAT in that function.

$$\mathrm{C}(<w_h,\ sc_h>,\ <w_c,\ sc_c>)=\sum_{\substack{LSF\in CORPUS\\ R\in RS}}\mathrm{CSR}(\mathrm{SR}(R,\ w_h,\ w_c),\mathrm{LSF}) \qquad (4)$$

Counting function C(R, <w_h, sc_h>, <w_c, sc_c>) tells the common occurring times of depending relation R, headword w_h and adjacent sub-node w_h in corpus.

$$\mathrm{C}(R,\ <w_h,\ sc_h>,\ <w_c,\ sc_c>)=\sum_{LSF\in CORPUS}\mathrm{CSR}(\mathrm{SR}(\mathrm{R},\ w_h,\ w_c),\mathrm{LSF}) \qquad (5)$$

Function F(R|<w_h, sc_h>, <w_c, sc_c>) denotes the common probability of depending relation R, headword w_h and adjacent sub-node w_c, $\hat{F}$ being its most likely estimate, then

$$\hat{F}\,(R|<w_h,\ sc_h>,\ <w_c,\ sc_c>)=\frac{\mathrm{C}(R,\ <w_h,\ sc_h>,\ <w_c,\ sc_c>)}{\mathrm{C}(<w_h,\ sc_h>,\ <w_c,\ sc_c>)} \qquad (6)$$

Finally, we can conclude:

$$P(\mathrm{LSF}|w_i\ldots w_j)=\prod_{k=i,w_k\neq h}^{j} P(\mathrm{SR}(k)|h,\ w_k)=$$
$$\prod_{k=i,w_k\neq h}^{j} \hat{F}\ (R|<w_h,\ sc_h>,\ <w_c,\ sc_c>) \qquad (7)$$

In above formula, we may use parameter smoothing technology that is the Back-off –based method [12]. In analysis course, dynamic scheming pruning process and probability computing process are similar to rules probability model. If the analysis of the two parts in one cell case having the same attribute structure, then the analysis result of the part which has lower probability will be cast aside and will not participate in the following analyzing-combining process.

Supposing that we inputting a sentence in QAS: "She eats pizza without anchovies", now we have:

$$P(T_1)=\prod P(LSF_i)=\mathrm{P(AGT|eat,\ she)P(OBJ|eat,\ pizza)P(MOD\ |\ pizza,\ anchovies)} \qquad (8)$$

$$P(T_2)=\prod P(LSF_i)=\mathrm{P(AGT|eat,\ she)P(OBJ|eat,\ pizza)P(MOD|eat,\ anchovies)} \qquad (9)$$

Supposing that we can gain the model parameter through corpus statistics such as Table 1, then:

$$P(T_1)=0.0025\times0.002\times0.003=1.5\times10^{-6}$$
$$P(T_2)=0.0025\times0.002\times0.0001=5\times10^{-8}$$

T_1 may be chosen to be the right result according to this. If we convert "anchovies" to "hesitation", then $P(T_1)=5\times10^{-8}$, $P(T_2)=4\times10^{-7}$. We find that language model may also help us to choose sound analysis result with the change of words in sentence. This is just about its merit.

Table 1 Interrelated model parameters

P(AGT|eat, she)	0.0025
P(OBJ|eat, pizza)	0.002
P(MOD|pizza, anchovies)	0.003
P(MOD|eat, anchovies)	0.0001
P(MOD|pizza, hesitation)	0.0001
P(MOD|eat, hesitation)	0.0008

5. Conclusion

In a word, the key technology of QAS is the interrelated technology based on natural language understanding, including the construction of knowledge base and corpus, the Word Segmentation and POS Tagging of text, the Grammatical Analysis and Semantic Analysis of sentences etc. This thesis dissertated mainly the denotation of knowledge-information based on semantic network in QAS, the stochastic syntax-parse model named LSF of knowledge-information in QAS, the structure and constitution of QAS. And the LSF model parameters were exercised; it proved that they are feasible. At the same time, through "the limited-domain QAS" which was exploited for banks by us, these technologies are proved effective and propagable.

Acknowledgment

We would like to acknowledge the support from the National High Technology Research and Development Program (863 Program) (2004AA1Z2450) , the National Natural Science Foundation of China (70572090, 60305009) and the Ph.D Degree Teacher Foundation of North China Electric Power University(H0585)in China.

References

[1] K. Lai and M. Baker, "Statistical Parsing a Context-free Grammar and Word Statistics", *Proceedings of the 21th National Conference on Artificial Intelligence*, Pittsburgh in USA, vol. 20, no. 1, January 2005, pp. 78~79.

[2] A.B. Downey, "Modeling user's Interests in Information Filtering", *Proceeding 5th Text Retrieval conference*, New York, May 2001, pp. 101~103.

[3] R. Bellman and L. A. Kalaba, "Abstraction and pattern classification", *JMAA*, Elsevier Science, Amsterdam in Netherlands, vol. 52, no. 1, January 2005, pp. 1-7.

[4] J. Philip, *A tutorial on Techniques and Application for Natural Language Processing*, 2^{rd} ed., Pennsylvania: Carnegie-Mellon University Press，June 2004, pp.36-78.

[5] S.F. Tian, *Artificial Intelligence And Knowledge Engineering*, Beijing: Peking University Press, March 2001, pp.5-33.

[6] T.S. Yao, *Natural Language Understanding*, Beijing: Tsinghua University Press, July 2002, pp.65-91.

[7] M.A. Collins, "new Statistical Parser Based on Bigram Lexical Dependencies", *Proceedings of the 44th Annual Meeting of the ACL*. Washington, August 2006, pp.184~191.

[8] E. Charniak, "Statistical Techniques for Natural Language Parsing", *AI Magazine*, vol. 26, no. 4, pp. 33-43, April 2005.

[9] B. Terje and A. Jon, "Natural language analysis for semantic document modeling", *Data and Knowledge Engineering*, Elsevier Science, Amsterdam in Netherlands, vol. 42, no. 1, January 2005, pp. 45-62.

[10] Z. Shichao and J. Mnhammed, "Mining Multiple Data Sources: Local Pattern Analysis", *Data Mining and Knowledge Discovery*, Berlin Heidelberg New York: Springer-Verlag, vol. 12, no. 8, August 2006, pp. 121-125.

[11] A. Bouchachia, "Learning with hybrid data", *Proceedings of the 5th International IEEE Conference on Intelligent Hybrid Systems. Magn. Japan*, vol. 20, August 2005, pp. 193-198.

[12] A. Laura and A.J. Weljters, "A Rule-Based Approach for Process Discovery: Dealing with Noise and Imbalance in Process Logs", *Data Mining and Knowledge Discovery*, Berlin Heidelberg New York: Springer-Verlag, vol. 12, no. 8, August 2006, pp. 67-87.