A Chinese Intelligent Question Answering System Based on Domain Ontology and Sentence Templates

Xin-hua Zhu, Qing-hua Cao, Fang-fang Su

College of Computer Science and Information Technology
Guangxi Normal University, Guilin, Guangxi 541004, China
E-mail: zxh429@263.net

Abstract

With the development of network technology, distance education is becoming increasingly important. Intelligent question answering system is an important part of remote learning system. Currently, there have been a lot of mature English question answering systems. Owing to the Chinese natural language complexity and processing technical limitations, there is still no mature Chinese QA system exploited by now. This paper designs and implements a Chinese intelligent question answering system based on domain ontology and sentence templates according to Chinese characteristics. This system packages question template, semantic template and answer model into a sentence template, adopts automatic word segmentation and sentence templates matching method to understand the users’ questions, and uses domain ontology as knowledge base which provide domain vocabulary for question analysis and knowledge retrieval for answers’ generation. Our system achieves to question with limited natural language, and finally returns to the user a precise and concise answer.

Keywords: Domain ontology, Question answering system, Automatic word segmentation, Sentence template, Natural language understanding

1. Introduction

Along with the development of network technology, more and more people are using the information on the web with search engines, but the search engines exists many drawbacks [1]. Firstly, it returns a large amount of documents that include indexing terms in a user’s query. Hence, the user should look over carefully the whole text to find a short phrase that meets his/her needs. Secondly, the retrieval bases on a logical combination of several keywords, so the query itself can’t clearly express the users’ real intention; In addition, the matching algorithm of retrieval doesn’t involve semantic, which leads to the hard promotion of query efficiency. In these cases, the question answering (QA) system is put forward [2]. As the same time, distance education which changed people traditional ways of learning and made learners away form time, space and geographical constraints has been accepted by more people [3]. Intelligent question answering system is a key part of remote learning system. Intelligent QA system combines natural language processing technology to understand the meaning of question with natural language form, gives concise and accurate answers to the users and automatically replies learners’ questions [4].

Internationally, nowadays QA system is to be in the ascendant and many large research institutions and leading companies are involved in this field of research. Relatively, there exists some mature QA systems such as Massachusetts Artificial Intelligent Lab Start[5], university of Michigan’s AnswerBus [6], IBM based on statistical question-answering system[7] etc. Owing to the Chinese natural language complexity and processing technical limitations, there is still no mature Chinese QA system exploited by now [4]. Compared with English QA system, Chinese QA system has the following characteristics and difficulties [8]:

- Continuous: Unlike English and other European languages, Chinese sentences are composed of continuous characters strings without space between words. Word segmentation is the foundation of Chinese language processing, so it is as an initial work and one of the most important technologies in Chinese QA system.
- The form: Chinese lacks of narrow form changes that the tag facilitates computers processing, such as voice and tense of English, etc.
The grammar: The Chinese grammars are flexible and the relationship between sentences depends on word order, ‘meaning close’ and function word.

The semantics: There are many polysemies, brachylogies, context high dependence and meaning more than one word in Chinese. These are difficulties for computer to process.

Grammatical research: Chinese lack of computer-processing-oriented grammar, such as the research material on formal of Chinese sentence and conversion of different patterns.

Related information: Chinese is short of grammar, semantics dictionary of the linguistics material and related raw corpus and annotated corpus etc.

Through the above we can learn that some English QA system’s mature technologies can’t be directly applied in Chinese QA system.

Natural Language Understanding [9] is to study how people make computer understand and generate our daily use of natural language. Natural Language Processing (NLP) involves a wide range of techniques, and has a great practical value. It can be used as Knowledge Engineering, Expert System, and Computer Aided Instruction and other natural language man-machine interface. QA system needs NLP technology to deal with questions. At present, there are mainly four kinds of methods to analyze questions in Chinese QA system, which based on keywords, HPSG grammar, HNC and sentence template matching.

In computer science and information science, ontology is a formal explicit description of concepts and relations among them. Ontology definition contains four meanings: Conceptualization, Explicit, Formal and Share. Conceptualization refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit refers to the used concepts and its constraints are clearly defined. Formal refers to the fact that ontology is machine-readable, which means the ontology can be handled by computer. Share refers to the knowledge of common recognition that ontology reflected and reflects the concept sets of related field. Ontologies are the structural framework for organizing information and are used in artificial intelligent, the Semantic Web, System Engineering, Software engineering, Library science, and information artificial intelligence as a form of knowledge representation about the world or some part of it[10].

The QA system this paper approached has following characters:

1. The QA system uses natural language interface which meets the needs of the user knowledge expression and gives users accurate answers according to their questions input.
2. This paper revolves around design and implementation of Intelligent QA System in a restricted domain and does a series of researches aiming at the construction of domain ontology knowledge.
3. The AQ System has a strong versatility and portability, and can be transplanted into different fields according to vary the fields of knowledge base.
4. The system doesn’t support continuous question.

2. The Intelligent QA System framework model

The QA System, which based on search engine, obtains a great deal of question-related or unrelated documents from the queried information. However, if we get user-satisfied answers, a further information extraction is required. So, this kind of QA system consists of four steps: question analysis, information retrieval, information extraction and answer generation [2]. Compared with this traditional system, due to its structural organization and explicitly semantics, this new QA system which based on domain ontology, can obtain correct answers. It doesn’t require another processing. Consequently, it only has these three modules: question analysis, information retrieval and answer generation. We design a framework model of QA System based on ontology knowledge base according to the three components, as shown in Figure 1.
We can learn the whole process of the Intelligent QA System from the figure 1: First, the system gets users’ questions from the interface, and makes automatic words segmentation of the questions and removes the stop words and holiday words. Then, match the processed question with sentence templates predefined in the sentence template library to acquire query semantic information table and answer template of the question. We use the semantic information table to retrieve and reason the ontology knowledge base which has been built. The last is to combine retrieval answer with answer template to generate answer and return it to the user’s interface.

3. Collect sentence templates

Question analysis is an important part of QA system, and its target is to understand the users’ query semantics. As we know, question analysis is composed of lexical analysis, syntax analysis and semantic analysis, and the semantic analysis is the hardest part in Natural Language Processing so far. Question analysis method based on sentence template matches the questions with sentence templates already existed in the sentence template library. After successful matches, the system will extract the semantic information according to sentence template. So, this method avoids the complicated lexical analysis, syntax analysis, and semantic analysis. Apparently, it is suitable for QA system which has specific domain [11][12]. In this method, the quality and quantity of sentence template library has great thing to do with questions understanding. As a result, collecting sentence templates plays a key role in this QA system.

In our paper, we formulate that the sentence template is made up of question templates, semantic templates and answer models which are correspond with each other. The BNF definition of sentence template is as follow:

\[
< \text{sentence template}>::= (<\text{question template}>, <\text{semantic template}>, <\text{answer model}>)
\]

The question template represents surface form of sentence template, the semantic template stands for semantic information, and answer model is used to reorganize answers which should be returned to users. We make a set of rules to establish semantic templates by referencing the logical structure of ontology. Such as the concept is expressed with C (if there are so many, we use C1, C2...to represent), and A represents properties of the concept (if there are so many properties using A1, A2...). The content of property A is expressed by A.V where the point symbol says the subordinate relationship. For example, the property of concept C is C.A and the content of C.A is C.A.V [11].

We have collected a large number of sentence templates which cover almost the users’ questions and basically meet the users’ demand. Part of the sentence templates are shown in Figure 2:
In the sentence templates, the @ says the parameters, language variable U%V represents U or V, the angle bracket <> represents the same semantics in different expression, the parentheses () indicates the optional part of the questions, the number 1 and 2 represents the location of parameters and the following English letters are the semantic logo of the parameters. The semantic information of <question template 1> is to find the property A of the concept 1, and <question template 2> is to find the relationship between the concept 2 and concept 3 under the condition of the concept 1.

4. The Construction of Domain Ontology

The domain ontology is a complete description of domain concept and its relation. From the semantic, the concept represents the collection of object, and its meaning can be described by the properties. The relation represents related knowledge, which is used to show the mapping relationship of domain concept. The concepts make up concept tree of subject areas by being related with subordinate relationship, and the son concepts can inherit all attributes of their father concept.

Ontology knowledge base is benefit to organize, manage, maintain, reason and query the knowledge. The knowledge base of Intelligent QA System is a limited ontology knowledge base, which belongs to the domain ontology category [13]. Combining with the characteristics of intelligent teaching system’s domain knowledge base, we divide the domain ontology into two levels: the subject upper ontology and the subject application ontology. The subject upper ontology is located to the top of the structure of teaching domain ontology, which is used to describe the meta-level subject, abstraction, common concept and its relationships. The different subjects have the different concepts of subject upper ontology. We take computer science as an example, the subject upper ontology has “program”, “algorithm”, “protocol”, “command”, “data” etc. As well as, the subject application ontology describes the specific concepts and relationships of different courses, and it is located under the level of the subject upper ontology, such as the concepts of “arithmetic unit”, “controller", "memory" etc in the Principles of Computer Composition Course, the concepts of “ TCP/IP protocol", ”Bridge", “LAN” etc in the Computer Network Course.

There are many methods of ontology development which are put forward based on different angles and application, such as IDEFS [14] method, Uschold method [15], Methontology method [16], TOVE method [17], etc. In our intelligent QA system, we select the knowledge engineering method that is proposed by Noy and McGuinness [18]. Constructing the teaching domain ontology based on this method can play greater function of ontology.

Taking Computer Science and its “Principles of Computer Composition” courses as an example, we adopt knowledge engineering method to establish teaching domain ontology, and this model part of concepts and their relations is demonstrated in Figure 3.
5. Question analysis

In the system we use sentence templates matching algorithms to process sentences. It needs to preprocess the question which includes word segmentation, remove stop words and holiday words before using the algorithms. The whole question analysis consists of three steps which are the word segmentation, remove stop words and holiday words, and sentence templates matching algorithms [19]. It will get the semantic information table of the questions after question analysis. In order to speed up finding the matched sentence template, the system uses a fuzzy matching method.

5.1 Automatic word segmentation

Word segmentation of Chinese is an initial work, and it is one of the most important technologies in computer processing of Chinese language. At present, there are several types of methods in word segmentation of Chinese text field. They include mainly three kinds of methods, such as character string matching method, artificial intelligence method and statistical linguistic model method. In our paper, combining the characteristics of restricted domain knowledge, we adopt a maximal matching words segmentation method based on dictionary. At the same time, the dictionary is divided into general dictionary and computer professional dictionary. The general dictionary uses common dictionary CoreDict.dct of Chinese Academy of Sciences’ research achievements which contains about ten millions entries and punctuation with part of speech, and the computer professional dictionary uses Google’s Computer professional dictionary which contains more than 9,000 computer professional vocabularies.

Long word priority is the most characteristic of forward matching segmentation algorithm. It reduces segmentation ambiguities to some extent. The restricted domain contains a large of professional vocabularies which are not easy to generate segmentation ambiguities and follow the principles of long-term priority, so we select the maximal forward matching segmentation algorithm.

The maximal matching segmentation algorithm is slightly different from the previous because there are two dictionaries. The improved algorithm is as follow:

First step: Set the length of the longest word in the dictionary is 8 and denote MaxLength=8. Take the character string whose length is equal to 8 from the input text InputString.

Second step: Match the character string with computer professional dictionary.

Third step: If matched successful, the character string is a word, then remove the character string from InputString and return the first step.

Fourth step: If matched unsuccessful, we will match the character string with general dictionary. If they matched successful, the character string is a word. Remove the character string from InputString and return the first step.
Fifth step: If matched unsuccessful with general dictionary, it shows that the character string doesn’t exist in the word library. Then remove the last character of the character string and assign it to the character string. Return the second step.

Sixth step: The character string is a single word when its length is 1. Output the single word and remove it from InputString, then return the first step.

5.2 Removing stop words and holiday words

In the users’ questions, it often contains some not clear meaning function words and polite language, such as“呢，啊，吗”,“请等等一下，我想请教”, etc, and the function words and polite language do not help understanding the questions and increase complexity of the template library module. So we need to create stop words and holiday words table to filter these words.

5.3 Sentence templates fuzzy matching

In our system, we select fuzzy sets method to match the user questions with the sentence templates in the library [2]. The class attributes of different sentence templates constitute a fuzzy set before to be identified, and the different types of label make up the support set of optional templates fuzzy set. Before the templates identified, we can use membership functions of fuzzy set methods to determine it possibility of being classified as a certain type. As we match the sentence templates, set the threshold of a successful templates matching is Tq, use the indicators Mq to indicate the current degree of questions matched. It will determine that the current question is consistent with the matching template if the matching degree exceeds the predetermined threshold, as Mq>Tq.

6. Ontology knowledge base retrieval and answers generation

The ontology is a shared conceptualization of knowledge, which has certain knowledge structure and is convenient to retrieve. At present, it has special ontology query languages and application programming interface. So we can realize the retrieval of ontology knowledge base even reasoning [20]. In the ontology knowledge base, search the related information, and then use answer templates from sentences analysis to generate answer by restructuring the information, finally, the answers with natural language form will be returned to the user.

This system selects SPARQL as ontology query language. SPARQL is a query language for RDF that W3C recommendation, which provides a powerful query functions based on graph pattern matching. It includes solution sequences and modifiers (ORDER BY, PROJECTION, DISTINCT, REDUC-D, OFFSET, LIMIT), optional pattern matching, matching alternatives, restricting the value and so on. A SPAQLR Query is a quadruple (GP, DS, SM, R) where: GP is a graph pattern which expresses the query semantics. DS is a set of RDF graphs, which provides information of RDF data sources. SM is a set of solution modifier to indicate the constraints of query results. R is a result format to specify the output form of the query results. The following shows the four parts of SPARQL query [21].

<table>
<thead>
<tr>
<th>Result format</th>
<th>R</th>
<th>SELECT ?sage ?sno</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set</td>
<td>DS</td>
<td>FROM <a href="http://www.example.com/Mary/foaf.rdf">http://www.example.com/Mary/foaf.rdf</a></td>
</tr>
<tr>
<td>Graph pattern</td>
<td>GP</td>
<td>WHERE {?id1 foaf:sage “23”; foaf:knows ?id2, ?id2 foaf:sage ?sage; foaf:mbox ?sno.}</td>
</tr>
<tr>
<td>Solution modifier</td>
<td>SM</td>
<td>ORDER BY ?sage</td>
</tr>
</tbody>
</table>

The SPARQL query language bases on graph pattern matching. The basic graph pattern is the triples pattern mode. While more complex graph patterns can be composed of basic patterns in various ways. SPARQL graph pattern matching is defined in terms of combining the results from matching basic graph patterns. Any graph pattern terminates a basic graph pattern. The triples patterns are composed of subject, predicate and object, which like a RDF triple. But it may have variables in any subject, predicate or object. For example, query a simple triple {计算机硬件 组成 部件} {computer hardware, compose, ?component}, the SPARQL query language are as follow:
In order to get clear and accurate answers, the query results with SPARQL language cannot be directly returned to the user, which needs to combine with the answer model. Through the front description we know that the System gets semantic template of the question when analyzing the question. These semantic templates not only provide basis for information retrieval, but also play an important part in answers generating. For example, the above mentioned <question template 1> corresponds to the answer mode is “C的 A是”。The retrieved knowledge will be reorganized with the answer model. Then the answer with a brief natural language statement is backed to the users. The <question template 1> returns to the users’ answer is “C的 A是<SPARQL查询结果>”.

7. The implementation of Intelligent QA System

In this system, the main development tools used NetBena6.0, Jena, protégé and Racer. The structure of the system is made up of foreground natural-language interface and background ontology knowledge base. The foreground natural-language interface is mainly responsible for the interaction between the users and system, which accepts the users’ natural language questions and returns answers to the users. The background ontology knowledge base is in charge of storing domain knowledge that is the basis of answering users’ questions in the QA system.

According to the above described methods, collecting sentence templates, constructing knowledge base based on domain ontology, adopting question analysis method based on sentence template, and using SPARQL query language, we develop a simple Chinese QA System whose results are shown in figure 4.

![Figure 4. Experiment results of the QA system](image)

8. Conclusions

This Paper analyzes the Chinese QA system from these three modules: questions analysis, information retrieval and answer generated. The first and most important thing of QA System is question analysis. At present, it is hard to do semantic analysis for Natural Language Processing technology. In our paper we adopt question analysis method based on sentence templates. This method does not require the complicated process such as syntactic analysis and semantic analysis, but can understand question’s semantic by its surface form, and it’s especially suit to specific field of QA System. Our QA system can answer most of the questions, but cannot answer various questions like human. First, Unlike English and other European languages, Chinese sentences are composed of continuous characters strings without space between words. Word segmentation is the foundation of Chinese language processing. Although our QA System is based on the limited field, and the dictionary is divided into general dictionary and computer professional dictionary, there are also unknown words and the inevitable segmentation ambiguity. So the accuracy of the word segmentation directly impacts
the success of the matching sentence template. Second, the way of the users' expression is complicated, so the templates library cannot exhaustively list all the sentence templates. Thus the number of sentence templates restricts the efficiency of sentences understanding. These need to be further studied in our future work.

9. References