Abstract: The Dynamic-Semantic-Framework (DSF) is a strategy used in the analysis of Chinese question sentences in a Question Answer System (QAS). With DSF, the system omits the traditional analysis on the syntax level, which is always quite complex in Chinese oral, and directly extracts keywords from the sentences dynamically. Because the type of the question is a essential element in the query, the system first decide the basic type of the question and fills it to the relevant slot of the main framework of the DSF. Then, the system would abstract the key word of the question by some special methods and these key words will be analyzed to form the concept fragments, which will be translated to the expression of semantic codes and filled into the slots of the DSF. The DSF is translated to query mode in order to get information from the database which contains the information of the Chinese authors and their works. The DSF can be readjusted if the query in the database fails, and with this strategy, we can avoid the query failure caused by the incorrectly analysis of the keywords. The answer abstracted from the database would be output by the answering process which is not discussed in this paper.

Key words: DSF, QAS, unknown word, key word, question type

1. Introduction *

The QAS (Question Answering System) is a hot subject in the NLP field, and the Chinese QAS is a new subject which has not been researched clearly. As oral Chinese is irregular in grammar and is unlikely to analyze the sentence in syntax, we choose framework to get keywords of the sentence with little syntax analysis and to obtain enough query objects as keywords. Meanwhile, in case that the literature database contains many special nouns which are not involved in the normal dictionary and as a result the normal word-segment is not always exact, we adopt the dynamic framework which will readjust its infill values and create new frameworks for query in order to get a reasonable result if the query fails. Finally, in order to understand the sentences in a sense level, we use word semantic information to fill the framework. Only if the word does not exist in the semantic dictionary will the word itself be filled. As a result, we call this strategy Dynamic-Semantic-Framework. If this strategy fails, the system returns an error report. And the difficulties in this strategy are how to decide the types of the framework and how to extract keywords to be filled into the right slots in the framework. The PRI of the query subsequence is also considered to avoid early query failure.

2. Realization of dynamic semantic framework

2.1 Preliminaries

Before the filling of framework, the system has to search for concept fragment which will be filled into the slots. In order to describe the question more accurate, we must treat the single-character which is always a feature of wrong segment which is caused by unknown words. Unknown word is the string which is not in the dictionary,
and can not be recognized. In the LES, they are mainly person name and appellations such as “落花生”, “鸿来”, “毛驴”, “小山”. The latter two will do noting to the segment but the former two will influence the result is not cut down in correct way.

As follow, we will introduce the method which is used as the process of the unknown word:

When the unknown word is cut wrongly, it may exist in three ways:

- Series of single-character or single-word: “李颉人”, “地山”, “朝花夕拾”
- A single-word and a double-word: “周树人”, “落花生”
- A single-character and a double-word: “王忘我”

The follow diagram will introduce the process of the LES to treat with unknown word:

When there is a series of single-word or single character, it is probably that there is an unknown word. But not all the series of single-word is an unknown word, it depends much on the tagging of the word and the context. The LES takes some special methods to deal with this problem:

1: Looking for the special word, and tag it.

2: If x₁,x₂,⋯,xₙ is a Chinese string between two special words, and the length is no less than 3, it is unlikely to be a unknown word, and it will not be processed, else, the prerequisite to decide whether the string(xᵢ⋯⋯xⱼ(1≤i<j≤n)) is a unknown word is that whether WORD(xᵢ⋯⋯xⱼ)=0, that is to say, it contains at least one single-character or single-word. If found, goes to step 2.

3: If i=j, that is to say, there is only one single-word or single-character, we treat the character as part of an unknown word, and the process goes to step 4, else, goes to step 5

4: The system will analyze the context. If length≤2, the two words will be combined to be an unknown word, else, if the length is 3, we combine the single one with its following word ,as an experiential rule, then the process would goes to step 6.

5: The process will combine the string from i to j as a whole, which will be regard as an unknown word.

6: If an unknown word is confirmed, the process treats it as a noun with semantic code “UK000000”, and the process is over.

2.2 Question types of Chinese dialogue

One of the differences between Chinese and other western language is that the interrogatives are not easy to sort. From table 1, we can conclude that many confirmations of the types should be decided from their context, and the relative information is always after the interrogative. As a result, when we found an interrogative, we will go afterwards until we found a type which matches the table 1.

Table 1: Types of Chinese questions which would be asked in the Literature QAS

<table>
<thead>
<tr>
<th>TYPE</th>
<th>Interrogative set</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAME (include old name, penname, ZI, HAO)</td>
<td>“谁”, “什么-人”, “哪个-人”</td>
</tr>
<tr>
<td>TIME(include birth and death time, publishing time of an article)</td>
<td>“何时”, “什么-时候”, “哪一年”</td>
</tr>
<tr>
<td>PLACE(include birth and death place, native place)</td>
<td>“何地”, “什么-地方”, “哪里”</td>
</tr>
<tr>
<td>OPUS</td>
<td>“什么-作品”</td>
</tr>
<tr>
<td>ATTRIBUTE(background)</td>
<td>“怎么样” “如何”</td>
</tr>
<tr>
<td>JUDGEMENT</td>
<td>NULL</td>
</tr>
</tbody>
</table>

After the type can be confirmed, the system put it into the relevant slot in the main framework of the DSF and the research of the database would based on it. The words used in the type deciding process would be tagged so that they would not be searched as key word afterwards.If the type can not be confirmed, the system returns an error report.
2.3 The search of keywords and concept fragment

The purpose of searching the keywords is to match the semantic query system with the basic query of the database. As the system scans the input sentence, it will extract nouns, adjectives and verbs as the keywords. With a semantic dictionary, the system can get the semantic information of the words. And then, the system would analyze these words, combining them into a concept fragment. The strategy is based on several rules. For example, the word group “最早” would be combined as one fragment as the rule that an adverb followed by an adjective would be combined. The system would put these concept fragments into the child frameworks. Moreover, when there is a word whose sense is “V*******” (******* is the code of the attribution of the value in the semantic dictionary), it means it is the specific value of the database and the system treat it as a field value, and the infill the attribution slot, else, the attribution slot will be null.

The query of the database adopts the VSM and D1 is

2.4 The structure of DSF and the query mode

Conferring the intention of the asker, the framework includes main framework and several child frameworks. The framework is build up by slots and the slots are of two kinds: the attribution-value slot and the link slot. The attribution-value slots give an attribution and its value as well as the link slot points to another slot, which is its child framework or its brother framework.

The main framework includes 4 slots:
1. TYPE slot: infill the type of query
2. Query Attribution Slot: infill the type of question.
3. Link Slot: point to a child framework.
4. Number Slot: describe whether the result is single or not.

The child framework also contains three slots:
1. Restriction attribution slot: infill the restriction of the query attribution.
2. The query-type slot: including “=”, “LIKE”, which decide the query method in the query mode.
3. Link slot: point to the next child framework, if there is no framework left, the value is NULL.

The vector of a line of information in the database, D2 is of the input question:

\[
Sim(D_1, D_2) = \cos \theta = \frac{\sum_{k=1}^{n} W_{1k} \times W_{2k}}{\sqrt{\left(\sum_{k=1}^{n} W_{1k}^2\right) \left(\sum_{k=1}^{n} W_{2k}^2\right)}}
\]

2.5 Readjustment of DSF

As the first framework may be filled incorrectly and the query may fail, we decide to readjust the framework if the query fails. This is called readjustment.

When the query fails, it means that the framework is filled wrongly, and the system will refill the framework and create a new query. The readjustment depends mainly on the concept segment, which is at first formed to be the longest part considering that the longer the concept fragment is, the more accurate it expresses the question. But sometimes the concept would be too long and the word included may be not existed in the database. As we choose the longest fragment as the query keywords at first, if the query fails, the system will reduce the length of the fragment by the segment process and fill them again into the slots of the frameworks to form a new query. For example, in a question: “鲁迅是绍兴人么？”, the “绍兴人” is treated as a concept fragment at first. But in case that there is no information such as “绍兴人” in the database, and the database just records a birth place as “绍兴”. In this case, the keyword is reduced to “绍兴” and “人”, and “人” would in no use because it is not a keyword in the system.

If the concept is cut down to single word and can not be separated at all, the readjustment is over and the system would output an error report.
3. Experimental results

The LES system is programmed with visual C++ 6.0 and works in Windows 2000. As the system mainly faces to the people who have certain knowledge of Chinese culture, we use a experiment of 37 oral Chinese questions gathered from several students in Chinese Culture Major to test the system’s capability. In order to show the analysis process clearly, we reserve not only the final output answer line, but also the content of the frameworks in a visible text. If the framework has been readjusted, we would also record the framework before readjustment. After the test, we get a conclusion that the system could analyze 87.1% sentences correctly without readjustment, 93.5% with readjustment. Analyzing the failed answers, we find that the main reason for the failure is that the keywords are extracted wrongly. The two sentences, which are correct with readjustment, are at first extracted with too large fragment and the match in the queries fail. One of the two which still fail with readjustment is because the word segment is completely wrong and the other is because the question type is judged wrongly.

Table 2: Results of a test with 31 questions that would be asked in a Literature QAS

<table>
<thead>
<tr>
<th>total number</th>
<th>Query without readjustment</th>
<th>Query with readjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Success</td>
<td>fail</td>
</tr>
<tr>
<td>31</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>100%</td>
<td>87.1%</td>
<td>12.9%</td>
</tr>
</tbody>
</table>

4. Conclusions

The DSF is a very suitable method of semantic analysis in QAS system, and in order to extract keywords filled into the framework correctly, the pre-process is pivotal, and the method used to extract keyword still need ameliorate. Also, how to decide the type of the sentences is not correct every time. Finally, the scale of the semantic dictionary is also important and how to readjust the framework still needs improvement.

5. References