INTENTION EXTRACTION AND SEMANTIC MATCHING FOR INTERNET FAQ RETRIEVAL USING SPOKEN LANGUAGE QUERY

Yu-Sheng Lai, Kuen-Lin Lee and Chung-Hsien Wu

Department of Computer Science and Information Engineering, Cheng Kung University, Tainan, Taiwan, China.

ABSTRACT

An FAQ (frequently-asked question) pattern consists of a question and a text document that answers the question and contains some additional remarks. As a query is similar to the FAQ’s question, the FAQ’s answer gives a possible answer or parts of the answer of the query. On the other hand, an FAQ’s answer may also contain information not concerning with the corresponding FAQ’s question but embed the answer for other questions. For a given query, therefore, the answer can be obtained from both FAQ question and answer. In this paper, we propose a framework for Internet FAQ retrieval by using spoken language query. We aim at two points: (1) extraction of the main intention embedded in a query sentence and (2) semantic comparison between a query sentence and an FAQ pattern. To evaluate the system performance, a collection of 1022 FAQ patterns and a set of 185 query sentences are collected for experiment. In intention extraction, 91.9% of intention segments can be extracted correctly. Compared to the keyword-based approach, an improvement from 78.06% to 95.28% in recall rate for the top 10 candidates is obtained.

1. INTRODUCTION

In past years, the keyword-based query has been widely used in information retrieval. However, there are two main problems in the keyword-based query for information retrieval. Firstly, using only keywords cannot convey user’s intention completely and clearly compared to a whole sentence. For this sake, search engines of nowadays always output lots of unexpected results. Secondly, it is not easy for users to find appropriate keywords for querying. People are used to have a query by using spoken language directly rather than finding appropriate keywords out.

Compared to the keyword-based query, the spoken language-based query is a natural and human way for people. The spoken language-based approach is a still growing issue. Several webs, such as Ask Jeeves [1], FAQ Finder [4], Dr. E [3], etc., provide natural language query also. Even though the systems using spoken language for querying is not very practicable as a result of the difficulties in natural language understanding, the spoken language-based IR approach is still a trend in the future. By combining speech recognition with natural language understanding, the task for information retrieval will become more convenient and practical.

In this paper, we propose a framework for Internet FAQ retrieval by using Mandarin spoken language query. We aim at two points: (1) extraction of the main intention embedded in a query sentence and (2) semantic comparison between a query sentence and an FAQ pattern. To extract user’s intention, we define a semantic grammar to decompose a query sentence into two components, intention segment (IS) and keyword segment (KS). The intention segment conveys the main intention in a short clause or phrase. The keyword segment consists of the important words in the query sentence. To compare two question sentences, a parse tree-matching method that compares two parse trees recursively is proposed.

2. THE FRAMEWORK

Fig. 1 shows the framework of a spoken language query-based Internet FAQ retrieval system. By analyzing the grammatical forms of the query questions, a set of productions is generalized. Using the semantic grammar and a set of predefined stopping words, the semantic analyzer extracts two important components, the intention segment and the keyword segment, from the input query sentence. Then these two components are fed into two matching processes, a keyword matching and a semantic matching, respectively. In intention matching, a parse tree-matching method is proposed to measure the similarity between the IS parse tree of the query and those of the FAQ questions. In keyword matching, a five-level thesaurus constructed from an electronic Chinese dictionary and a knowledge database named HowNet are adopted to compare the KS in the query with those in the FAQ questions. Similarly, a matching between KS in the query and those in the FAQ answers is also performed. Finally, these matching scores are integrated together to give a list of ranked hyperlinks to the web pages containing relevant FAQ patterns.

3. SEMANTIC ANALYZER

Generally, in terms of the syntactic structure, questions can be divided into seven types. In terms of the function, questions can be identified by 14 types and can be grouped into the four categories: external, talk, relational, and expressive questions, where the external questions seek factual information external to the conversation and are more uncertain than the other kinds of questions. According to a study on discourse analysis of questions in Mandarin conversation [2], people tend to seek external information by using the following three types of questions: question-word questions, questions with sentence-final particle, and A-not-A questions. For Internet FAQ retrieval, most of the queries possess higher uncertainty, i.e. approaching to the external questions. Therefore we analyze the grammatical forms of the three types of questions and generalize a set of productions to extract the intention component from a query.

To understand users’ intention in depth, two important components, a keyword segment and an intention segment, are extracted from a query sentence. The KS contains several keywords in the query sentence and some of the keywords in the KS are automatically extracted by an unknown word
3.1. Semantic Grammar

We define a 6-tuple semantic grammar as $G=\langle N,\Sigma,P,\delta,S\rangle$, where $N$ is a set of non-terminal symbols, $\Sigma$ is a set of terminal symbols disjoint from $N$, $P$ is a set of productions, each of the form $A \rightarrow \alpha$, where $A$ is a non-terminal and $\alpha$ is a string of symbols from the infinite set of strings $(N \cup \Sigma)^*$. $S$ is a designated start symbol and is interpreted as the “query or question” in this paper, and $P_P$ and $P_K$ are a set of intention segment productions and a set of keyword segment productions respectively, each of the form IS (or KS): $\cdot \cdot \cdot$, where $\cdot$ is a string of symbols from $(N \cup \Sigma)^*$ and the string should be derived from the designated symbol $S$.

Fig. 2 lists a sample lexicon for the FAQ retrieval system. Fig. 3 lists a context-free grammar with example phrases for each production, where the $|$ symbol indicates a non-terminal has alternative possible expansions and a symbol in parentheses indicates the symbol can be omitted.

<table>
<thead>
<tr>
<th>Productions</th>
<th>Example Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow (NP) Why \mid VP$</td>
<td>+ + (What reason are people with normal liver function possible to have the cirrhosis?)</td>
</tr>
<tr>
<td>$NP \rightarrow ProperNoun \mid (Det) Nominal$</td>
<td>(cirrhosis of the liver)</td>
</tr>
<tr>
<td>$VP \rightarrow (Verb) NP$</td>
<td>+ (liver function)</td>
</tr>
</tbody>
</table>

Figure 2. A sample lexicon for the FAQ retrieval system

| 3.2. Intention and Keyword Extraction |

Two sets of important productions, IS productions and KS productions, are combined into a context-free grammar shown in Fig. 3 to form the semantic grammar for extraction of the main intention and the keywords embedded in a sentence. Figs. 4 and 5 show examples of IS productions and KS productions respectively.

By the IS production, the intention segment can be easily extracted from each grammatical query sentence. Similarly, the keyword segment can be extracted from a grammatical query sentence also. The keywords for keyword matching can be extracted in two steps. Firstly, the keyword segment is segmented into few words including unknown words extracted by an unknown word extraction algorithm [5]. Secondly, the words defined in the stopping word set are discarded and the remainings are the extracted keywords for keyword matching.

<table>
<thead>
<tr>
<th>IS Production</th>
<th>IS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(NP) Why \mid VP</td>
<td>(reason for the cirrhosis of the liver)</td>
</tr>
<tr>
<td>\rightarrow VP de5 ywan2 yin1</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4. A sample of IS production and the extracted IS

Figure 5. A sample of KS production, the extracted KS, and the finally extracted keywords for keyword matching

4. WORD SIMILARITY

In HowNet, a concept is represented as a combination of features and symbols, where the features are divided into two types, the primary feature and the secondary feature. The primary feature indicates the concept’s category that can be represented as a tree. Based on the category, the secondary features give more concrete description. The symbols describe the relationship between the concept and its own secondary features. In our approach, the word similarity in semantics is estimated from the comparisons of both the primary features and the secondary features.

4.1. Primary Feature Similarity

In the hierarchical structure, to measure the similarity between two primary features can be considered from following two view aspects. From the node-based view aspect, each node represents a unique concept and contains a certain quantity of information. The similarity between two concepts can be considered as the quantity of the common information between these two concepts. For each two nodes in the hierarchical structure, the common information is defined as the information content (IC) [7] of the nearest parent node. In this paper, the information content of a feature $x$ is defined as follows:

$$IC(x) = −\log p(x)$$  

where $p(x)$ is the probability of encountering a word with the feature $x$ or with a feature subsuming the feature $x$.

From the edge-based view aspect, the similarity between two concepts can be certainly measured by the distance between them, in which the distance is often the summation of the edge costs. In the hierarchical structure, the depth also affects the distance. For the classification of concepts, the lower concepts are finer than the upper concepts. Therefore the distance should be decreased as the depth is increased.

Combining the above two view aspects, the distance between two primary features $pf_1$ and $pf_2$ is defined as follows:

$$Dist(pf_1,pf_2) = \sum_{i=1}^{\text{Super}(pf_1,pf_2)} \text{Cost}(\text{Super}(pf_1,pf_2))$$  

where $\text{Super}(pf_1,pf_2)$ denotes the least common super node of $pf_1$ and $pf_2$, $p(c)$ denotes the parent node of $c$, and $\text{Cost}(c,p(c))$ denotes the edge cost between the node $c$ and its parent, defined as follows:

$$\text{Cost}(c,p(c)) = \left[ IC_{pf}(c) − IC_{pf}(p(c)) \right]$$  

where $\gamma$ is an adjusting factor which controls the depth effect upon the edge cost, $\gamma=1$ in this paper, and $IC_{pf}$ denotes the information of a primary feature.

Because the similarity is opposite to the distance, the similarity between two primary features can be derived from the distance as follows:
4.2. Secondary Feature Similarity

Except for the primary feature, the secondary features can be used to measure the similarity between two words. However, the secondary features are not hierarchical and the definition usually contains one or more secondary features. The similarity between two secondary features cannot be estimated like that between primary features. The secondary features of a concept are viewed as a binary vector. Thus, the similarity between two sets of secondary features can be estimated by comparing two binary vectors.

There are many methods for estimating the similarity between two binary vectors, such as Dice coefficient, Jaccard coefficient, Overlap coefficient, Cosine function, etc. [6]

Combining the information content with the Dice coefficient, the similarity between two sets of secondary features is defined as follows:

\[ \text{Sim}_{sf}(s_f, s'_f) = 2 \times \frac{\sum_{f \in s_f} \text{IC}_{sf}(f)}{\sum_{f \in s_f} \text{IC}_{sf}(f) + \sum_{f' \in s'_f} \text{IC}_{sf}(f')} \]

where \( s_f \) and \( s'_f \) denote two sets of secondary features, \( f \) and \( f' \) indicate secondary features in the sets \( s_f, s'_f, s_f, \) and \( s'_f \) respectively, and \( \text{IC}_{sf} \) denotes the information content of a secondary feature.

4.3. A Broad-Sense Thesaurus

By analyzing the similarity between two words in morphology, in syntax, and in semantics, we divide the similarity between two words into five levels and each is assigned a measurement by considering the difference between adjacent levels. The word similarities can be leveled as follows:

L1. If two words \( w_1 \) and \( w_2 \) are identical in morphology, in syntax, and in semantics, 1 is certainly assigned as the word similarity.

L2. If two words \( w_1 \) and \( w_2 \) are different in morphology, but identical both in syntax and in semantics, \( \leq 1 \) is assigned as the word similarity.

L3. If two words \( w_1 \) and \( w_2 \) are identical in syntax and in semantics, the word similarity is defined as follows:

\[ \text{Sim}_{sid}(w_1, w_2) = \mu \cdot \max_{d \in \text{del}(w_1), d' \in \text{del}(w_2)} \text{Sim}_{sid}(d, d') \]

where \( \text{Sim}_{sid}(d, d') \) denotes the similarity between two word definitions \( d \) and \( d' \) as defined as follows:

\[ \text{Sim}_{sid}(d, d') = \text{Sim}_{pf}(p_f, p'_f) \cdot \text{Sim}_{sf}(s_f, s'_f) \]

where \( p_f \) and \( p'_f \) denote the primary features of two word definitions \( d \) and \( d' \) respectively, \( s_f \) and \( s'_f \) denote the secondary features of two word definitions \( d \) and \( d' \) respectively, and \( \mu \) is a combination parameter.

L4. If two words \( w_1 \) and \( w_2 \) are similar in semantics but different in syntax, the word similarity is defined as follows:

\[ \text{Sim}_{wss} = \nu \cdot \text{Sim}_{wss}^{(1)} \]

where \( \nu \) is a constant reflecting the syntactic effect upon the word similarity.

L5. If two words \( w_1 \) and \( w_2 \) are different in morphology, in syntax, and in semantics, the word similarity should approach to 0.

5. MATCHING PROCESS

After extracting the IS and the KS, the retrieval task can be divided into two matching processes, an intention matching and a keyword matching.

5.1. Intention Matching

Two intention segments, \( IS_1 \) and \( IS_2 \), extracted from a query sentence and the definition of an FAQ pattern respectively, are parsed into two parse trees, \( T_1 \) and \( T_2 \). The similarity between these two intention segments is estimated by a parse tree-matching formula defined as follows:

\[ \text{Sim}_{is}(IS_1, IS_2) = \text{Sim}_{is}(T_1, T_2) \]

where \( \text{Sim}_{is}(T_1, T_2) \) denotes the similarity between both single-vertex trees, \( T_1 \) and \( T_2 \) represent subtrees of \( T_1 \) and \( T_2 \) respectively, \( |T_1| \) and \( |T_2| \) denote the numbers of the subtrees of \( T_1 \) and \( T_2 \) respectively, and \( \text{Sim}_{is}(T_1, T_2) \) denotes the similarity between two non-single-vertex trees defined as follows:

\[ \text{Sim}^{\text{tree}}_{is}(T_1, T_2) = 1 - \frac{\sum_{i=1}^{n} \text{Dist}_{pf}(p_f, p'_f)}{|T_1| \cdot |T_2|} \]

where \( \text{Dist}_{pf}(p_f, p'_f) \) denotes the distance between two primary features, \( p_f \) and \( p'_f \) respectively, \( \mu \) is a combination parameter.

5.2. Keyword Matching

Given two sets of keywords, \( K_A \) and \( K_B \), extracted from a query and a question respectively, the similarity between them are defined as follows:

\[ \text{Sim}_{kw}(K_A, K_B) = \frac{1}{|K_A|} \max_{a \in K_A} \sum_{f(a)} \text{Sim}_{kw}(a, f(a)) \]

where \( f \) is a one-to-one function from \( A \) to \( B \), in which \( A \) represents \( K_A \) and \( B \) represents \( K_B \), \( f \) is a function, \( |K_A| \) and \( |K_B| \) denote the numbers of the keywords of \( A \) and \( B \) respectively, \( \text{Sim}_{kw}(a, f(a)) \) denotes the word similarity between the word \( a \) and the corresponding word \( f(a) \).

On the other hand, the similarity between two sets of keywords extracted from a query and an answer is estimated by a vector space model, in which the TF-IDF is adopted to measure the term weighting.

6. EXPERIMENTAL RESULTS

A collection of 1022 FAQ patterns and a set of 185 query sentences are collected for experiment. The FAQ patterns are collected from several medical web sites and grouped into 10 categories. 20 persons are asked for giving some queries on two assigned categories each. To evaluate the system performance, each FAQ is tagged as relevance/irrelevance corresponding to the queries.

6.1. Correctness of Intention Segment

The intention segments are very important for understanding users’ underlying intention further. By analyzing the commonly used question types, a semantic grammar
is defined as the query and the FAQ pattern. In case the similarity is determined by the keyword similarity between the query and the FAQ pattern, the three similarity measures, keyword segments only. In case the similarity is determined by comparing both intention segments only. In case the similarity is determined by the similarities of the three parameters, therefore, we define a temporal score to estimate the similarity between two questions as follows:

\[
S' = \gamma \text{Sim}_{q} + (1-\gamma) \text{Sim}_{e_k}
\]

where is a weight.

In this approach, an average of 1.36 answers corresponding to each query sentence are correct. Instead of the precision, therefore, an average rank is proposed to measure the system performance. The average rank is defined as follows:

\[
\bar{R}(\Omega) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \text{rank}(p_j)
\]

where denotes the testing set, \(N\) is the total number of the query sentences in \(\text{Sim}_{q}\), \(n_i\) is the number of the answers corresponding to the \(i\)th query, \(p_j\) denotes the \(j\)th answer of the \(i\)th query, and \(\text{rank}(\cdot)\) denotes the ranking of the answer.

By varying the weight from 0 to 1 stepped by 0.1, the solutions of the three parameters for the top 3 average ranks are obtained. As the system achieves the best performance, the parameters are distributed over the following intervals: [0.5, 0.7], [0.7, 1.0], and [0.4, 0.7]. The averages, \(\alpha=0.63\), \(\beta=0.84\), and \(\gamma=0.6\) are computed respectively and adopted in the following experiments.

### 6.3. Overall Performance

To measure the final similarity between a query and a FAQ pattern, the three similarity measures, \(\text{Sim}_{w}, \text{Sim}_{e_k},\) and \(\text{Sim}_{e_k}\), are integrated into a score \(S\) that is defined as follows:

\[
S = \alpha \text{Sim}_{w} + \beta \text{Sim}_{e_k} + \gamma \text{Sim}_{e_k}
\]

where , and indicate the weights for each individual similarity measure. In case \(\alpha=1\), the final similarity is determined by comparing both intention segments only. In case \(\alpha=1\), the final similarity is determined by comparing both keyword segments only. In case \(\alpha=1\), the final similarity is determined by the keyword similarity between the query and the FAQ answer only, i.e., a keyword-based IR approach. In case \(\alpha=0\), the final similarity is determined by the keywords in the query and the FAQ pattern. In case \(\alpha=0\), the final similarity is determined by the similarity between the query sentence and the FAQ question, not concerning with the FAQ answer.

By varying the three weights from 0 to 1, stepped by 0.1, the best case was obtained. In the best case, the system achieves an average rank of 3.027 and a recall rate of 95.28% for the top 10 results. Compared to the keyword-based IR approach, improvements of around 9 in average rank and 22.06% in recall rate were obtained.

### 7. CONCLUSIONS

In this paper, we proposed a framework to retrieve Internet FAQ patterns by semantic matching and obtained a significant improvement compared to the keyword-based approaches. The contributions mainly contain: (1) intention extraction of a query sentence, (2) a parse tree-matching method and a one-to-one mapping function for semantic matching. In intention extraction, 92% of intention in the collected corpus can be extracted correctly. Since several types of questions are not considered, it results that some out-of-grammar questions cannot be parsed successfully. The problem can be solved by analyzing other types of questions to generalize more appropriate IS productions. In semantic matching, the matching process is reliable in word level. In the future, we will devote to the development of a semantic matching method in phrase level and in sentence level.

### REFERENCES