Semantic-based Query Generation For Information Retrieval

Seung-Eun Shin*, Young-Hoon Seo
School of Electrical & Computer Engineering
Chungbuk National University, Cheongju, Korea

ABSTRACT

In this paper, we describe a generation mechanism of semantic-based queries for high accuracy information retrieval and question answering. It is difficult to offer the correct retrieval result because general information retrieval systems do not analyze the semantic of user's natural language question. We analyze user's question semantically and extract semantic features, and we generate semantic-based queries using them. These queries are generated using the se-mantic-based question analysis grammar and the query generation rule. They are represented as semantic features and grammatical morphemes that consider semantic and syntactic structure of user's questions. We evaluated our mechanism using 100 questions whose answer type is a person in the TREC-9 corpus and Web. There was a 0.28 improvement in the precision at 10 documents when semantic-based queries were used for information retrieval.

***Keywords**: Query Generation, Semantic Feature, Information Retrieval and Natural Language Question.*

1. INTRODUCTION

IR(Information Retrieval) technique has been rapidly developed with the growth and commercial application of Internet. Especially in the field such as Web whose information is linked mutually, IR technique has been intensively developed to find information fast and exactly. But, we can often find that high ranked documents may be far from the user's intention on the traditional IR system, therefore effective retrieval rank technique to consider the user's intention has been demanded. In the case of Google(www.google.com) that is offering Web search service, the rank system as Page Rank to reflect characteristics of Web is introduced and used to determine a web page's relevance or importance[1].

IR techniques are researched vigorously about various issues such as meta-search, distributed information retrieval, expert optimization search etc. And many researches for QA(Question Answering) system have been done recently. After analyzing user's natural language question, QA system finds the answer from IR results and offers it to user. Practical QA systems are being tested in a limited domain.

There are a variety of techniques for allowing natural language questions in IR systems[2, 3, 4]. The simplest approach is to remove the "function words" from the question and uses the remaining words in a standard keyword search (Alta Vista). In more complex approaches, pattern matching (the Electric Knowledge search engine), parsing (Ask Jeeves), and machine learning[2] techniques can support the association of more appropriate keywords with a query. But present natural language search engines perform poorly because most do not reflect the semantic of user's question on the query.

Generally when humans induce a relationship between query and document, they decide the relevance of document not based on existence of query terms, but based on semantics of query terms in documents. Thus, the retrieval method which intends to consider semantics of query terms has been studied such as query expansion, latent semantic indexing (LSI) and mutual information[5, 6, 7, 8, 9].

There is a research about the generation of lexical paraphrases of queries posed to an Internet resource[10]. These paraphrases are generated using WordNet and part-of-speech information to propose synonyms for the content words in the queries. Statistical information, obtained from a corpus, is then used to rank the paraphrases. But, this approach generates only lexical paraphrases that don't consider the semantic of queries. And this is similar to the same approach that uses only synonyms for query expansion.

Many research efforts have been made on how to analyze user's question as the query for the effective IR and QA. But most IR systems do not reflect the semantic of questions because it index user's question and documents by keywords extracted from morphological analysis or n-gram method. Therefore, it is difficult to offer the correct IR result because most researches do not analyze the semantic of user's natural language question.

In this paper, we describe semantic-based query generation from user's question for high accuracy IR and QA. We analyze user's question semantically and extract semantic features, and we generate some queries using them. We use the semantic-

**Corresponding author. E-mail: seshin@dcenlp.chungbuk.ac.kr*

Manuscript received Oct 4, 2005 ; accepted Oct 25, 2005

based question analysis grammar and the query generation rule for semantic-based query generation. They are represented as semantic features and grammatical morphemes that consider semantic and syntactic structure of user's questions. We evaluated our mechanism using 100 questions whose answer type is a person in the TREC-9 corpus and Web. There was a 0.28 improvement in the precision at 10 documents when semantic-based queries were used for information retrieval.

2. SEMANTIC-BASED QUESTION ANALYSIS

We recommend a font of 9 points. The main text of this document is set in 9 point Times New Roman. If absolutely necessary, we suggest the use of condensed line spacing rather than smaller point sizes. Some technical formatting software prints mathematical formulas in italic type, with subscripts and superscripts in a slightly smaller font size. This is acceptable.

Table 1. Sample semantic-based queries

<table>
<tr><th>Question 1</th><th>Who wrote Hamlet?</th></tr>
<tr><td>Semantic-based queries</td><td>write|wrote|to write|... Hamlet
compose|composed|... Hamlet
writer of Hamlet
author of Hamlet</td></tr>
<tr><td>Question 2</td><td>Author of Hamlet?</td></tr>
<tr><td>Semantic-based queries</td><td>write|wrote|to write|... Hamlet
compose|composed|... Hamlet
writer of Hamlet</td></tr>
</table>

Question 2 is a question whose the predicate is omitted. Users of QA systems or IR systems generally use natural language questions as the queries. User's question is a form of Question 1 or Question2. Other approaches that don't analyze the semantic of user's question cannot generate queries that are different in the syntactic structure. Our approach can generate queries whose the syntactic structure is different from user's question as Table 1.

2.1 Semantic Feature

We defined categories and semantic features for questions whose answer type is a person in TREC-8/TREC-9 corpus and Web, and we constructed the semantic feature dictionary of noun and verb from those questions.

Table 2 shows the sample of semantic features. In Table 2, the category means the subcategory of questions whose answer type is a person. The category is classified by thirty three categories such as author, family, prizewinner, politician, developer, inventor, scholar, entertainer, player etc. "Common" includes semantic features which are used commonly in all categories. And total semantic features are 125.

We expanded semantic feature nouns and verbs using synonym dictionary. The se-mantic feature dictionary consists of 1761 nouns and 278 verbs.

Table 2. Sample categories and semantic features

Category	Semantic Features
Common	Nationality, Time, Sex, Person, ...
Author	Title, Pen name, Author noun, Author verb, ...
Family	Relationship, Standard person, Person Info, Relationship Info, ...
Prizewinner	Prize, Prize noun, Prize verb, Place, ...
Politician	Position, Event, Organization, Elect noun, ...

2.2 Semantic-based Question Analysis Grammar

We designed the semantic-based question analysis grammar to analyze user's question semantically. That grammar is represented as semantic features and grammatical morphemes in order to consider semantic and syntactic structure of user's questions.

Table 3 shows the example of the semantic-based question analysis grammar, and it considered characteristic of Korean. In Table 3, it is a simple grammar for 'author' that is represented as extended BNF. We can analyze user's question semantically using it.

Table 4 shows the example of a question analysis result. As Table 4, we can analyze user's question semantically using that grammar and extract semantic features from user's question. Extracted semantics features are used to generate semantic-based queries for IR and QA.

Table 3. The example of the semantic-based question analysis grammar

<table>
<tr><th>Category</th><th>Semantic-based Question Analysis Grammar</th></tr>
<tr><td>Author</td><td><Author> ::= <Author_Info> [<Who>][?]
<Author_Info> ::= <Author_VP><etm> <Author_N><sj>
| <Author_VP><etm> <Person_N><sj>
| "Title"<pj> <Author_N><sj>
<Author_VP> ::= "Title"<oj> <Author_V>
|"Book_Info"<xj> <About> <Genre><oj> <Author_V></td></tr>
<tr><td></td><td>Author_N : *nouns whose semantic feature is 'author'*
Author_V : *verbs whose semantic feature is 'author'*
Genre : *nouns whose semantic feature is 'genre'*
etm : *adnominal suffix*
sj, : *subjective postposition*
pj : *possessive postposition*
oj : *objective postposition*
xj : *postposition('e')*</td></tr>
</table>

" " means the semantic feature that we must extract from user's question.
"Title" : Title of a book
"Book_Info" : additional information about the book in question

Table 4. The example of a question analysis result

Question 1	Hamlet-ul jeo-sul-han sa-ram-eun nu-gu-in-ga? (Who wrote Hamlet?)	
Grammar	<Author> ::= <Author_Info> [<Who>][?] <Author_Info>::=<Author_VP><etm> <Person_N><sj> <Author_VP> ::= "Title"<oj> <Author_V>	
Question Analysis Result	Category	Author
	Title	Hamlet
Question 2	Hamlet-ui jeo-ja-neun? (Author of Hamlet?)	
Grammar	<Author> ::= <Author_Info> [<Who>][?] <Author_Info> ::= "Title"<pj> <Author_N><sj>	
Question Analysis Result	Category	Author
	Title	Hamlet

Our approach can analyze equally user's questions whose the syntactic structure is different as Table 4. As Table 4, we can decide answer type(Category: "Author") and extract semantic feature(Title: "Hamlet") of Question 1 and Question 2.

3. SEMANTIC-BASED QUESTION GENERATION

Most IR systems do not reflect the semantic of a sentence because they index user's question and documents by keywords extracted from morphological analysis or n-gram method. Therefore, it is difficult to offer the correct IR result because most do not analyze the semantic of user's natural language question. To improve the accuracy of IR and QA using user's natural language question, we generate semantic-based queries from the user's question and expand them.

3.1 Query Generation using Rule

We designed the query generation rule, and it is represented as semantic features and grammatical morphemes that consider semantic and syntactic structure of question category. And it considers characteristic of Korean.

Table 5. The example of query generation rule

Category	Query Type	Rule of query generation
Author	A	<Title><oj> <Author_V> <Title><pj> <Author_N> <Book_Info><xj> <About> <Genre><oj> <Author_V>
	B	[Title] <Author_N> <Book_Info> <Genre>

[Word] means that the relevant document must include "Word" at least one.

Table 6. The example of query generation result

Question 1	Hamlet-ul jeo-sul-han sa-ram-eun nu-gu-in-ga? (Who wrote Hamlet?)	
Question 2	Hamlet-ui jeo-ja-neun? (Author of Hamlet?)	
Question Analysis Result	Category	Author
	Title	Hamlet
Semantic-based Query	A	Hamlet-eul jeo-sul-ha (write Hamlet) Hamlet-ui jeo-ja (author of Hamlet)
	B	[Hamlet] Jeo-ja (author)

The relevant document must include "Hamlet" at least one.

Table 5 shows the example of the query generation rule. In Table 5, the query generation rule is a simple rule for 'author' category. We can generate some queries using the rule and semantic features extracted from user's question. Query type A means a phrase that must be matched exactly for document retrieval, and query type B includes queries except for query type A.

Table 6 shows the example of semantic-based query generation result using the rule. Our approach can generate same semantic-based queries from questions (Question1 and Question 2) whose syntactic structure is different.

3.2 Query Expansion using Dictionary

After query generation using the rule, we expand queries using semantic feature dictionary. We except the query that the relevant document must include(that is marked by []) from query expansion.

Table 7. The example of the expanded query

Question 1	Hamlet-ui jeo-ja-neun nu-gu-in-ga? (Who wrote Hamlet?)	
Question 2	Hamlet-ui jeo-ja-neun? (Author of Hamlet?)	
Semantic-based Query	A	Hamlet-eul jeo-sul-ha (write Hamlet) Hamlet-ui jeo-ja (author of Hamlet)
	B	[Hamlet] Jeo-ja (author)
Expanded Semantic-based Query	A	Hamlet-eul sseu (write|wrote|to write|... Hamlet) Hamlet-eul jeo-jag-ha (compose|composed|to compos|... Hamlet) Hamlet-ui jag-ga (writer of Hamlet) Hamlet-ui geul-sseun-i (writer of Hamlet)
	B	[Hamlet] Jeo-ja, jag-ga, geul-sseun-i (author, writer) ......

Table 7 shows the example of query expansion. As Table 7, we can expand queries using semantic feature dictionary. We can obtain expanded queries such as 'write Hamlet', 'wrote Hamlet', 'to write Hamlet', 'author of Hamlet', 'writer of Hamlet' etc. from 'author' and 'Hamlet' in query expansion.

We use expanded semantic-based queries to improve the accuracy of IR and QA using user's natural language question.

4. EXPERIMENTS

We select 100 questions as test set from natural language questions whose answer type is a person. We experiment on precision at N documents.

In experiments, we used Google as IR system. Table 8 shows the precision at N documents.

Table 8. Precision at N documents

Document Level Average		
	Precision	
	Google	Our approach
At 5 docs	0.61	0.92
At 10 docs	**0.59**	**0.87**
At 15 docs	0.55	0.85
At 20 docs	0.52	0.82
At 30 docs	0.51	0.76

Precision at N documents: The percentage of documents retrieved in the top N that are relevant. If fewer than N documents are retrieved, then all missing documents are assumed to be non-relevant. Precision considers each retrieved relevant document to be equally important, no matter if is retrieved for a query with 500 relevant documents or a query with two relevant documents.

In experiment, we obtained the precision at 10 Docs of 0.87 when semantic-based queries were used for information retrieval. Our experiments showed that our approach results in a notable improvement in the precision at 10 documents (+0.28). It means that our approach can be used for intelligent IR or high accuracy QA to allow natural language questions.

5. CONCLUSION AND FUTURE WORKS

In this paper, we have described semantic-based query generation for intelligent IR or high accuracy QA. It is difficult to offer the correct IR result because general IR systems do not analyze the semantic of user's natural language question. We analyzed user's question semantically and extracted semantic features, and we generated semantic-based queries using them. Our experiments showed that our approach results in a notable improvement in the precision at 10 documents (+0.28). It means that our approach can be used for intelligent IR or high accuracy QA.

In the future work, we will expand the semantic-based question analysis grammar and the rule of query generation for various user's questions. Also, we are planning to apply our approach to other questions besides questions that answer type is a person.

REFERENCES

[1] Brin, S., and Page, L., "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Computer Networks and ISDN Systems, 30, 1998, 107–117.

[2] John M. Zelle and Raymond J. Mooney, "Learning semantic grammars with constructive inductive logic programming", In Proceedings of the 11th National Conference on Artificial Intelligence, 1993, 817–823.

[3] Tomek, S., Jose, P., Jussi, K., Anette, H., and Pasi, T. & Timo, L., "Natural Language Information Retrieval: TREC-8 Report", The Eighth Text REtrieval Conference (TREC-8), 2000, 35–56.

[4] Gann Bierner, "Alternative Phrases and Natural Language Information Retrieval", In Proc. Of the 39th ACL, 2001, 58–65.

[5] Kyung-Soon Lee, Young-Chan Park and Key-Sun Choi, "Re-ranking model based on document clusters", Information Processing and Management, 37, 2001, 1–14.

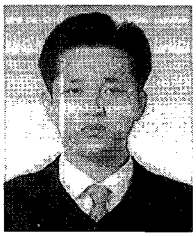
[6] Larry Fitzpatrick and Mei Dent, "Automatic Feedback Using Past Queries: Social Searching?", In Proc. 20'th ACM SIGIR International Conference on Research and Development in In-formation Retrieval, 1997, 306–313.

[7] C. Buckley and G. Salton and J. Allan, "The effect of adding relevance information in a relevance feedback environment", In Proc. 17'th ACM SIGIR International Conference on Research and Development in Information Retrieval, 1994, 292–298.

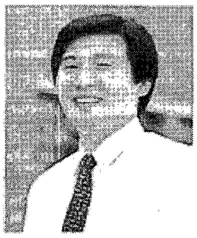
[8] Scott Deerwester and Susan T. Dumais and Richard Harshman, "Indexing by Latent Semantic Analysis", Journal of the American Society for Information Science, 41(6), 1990, 391–407.

[9] Michael L. Mauldin and Jaime G. Carbonell, Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing, Kluwer Academic Publishers, 1991.

[10] Ingrid Zukerman and Bhavani Raskutti, "Lexical Query Paraphrasing for Document Retrieval", The 17th International Conference on Computational Linguistics, COLING 2002.



Seung-Eun Shin
He received the B.E. and M.E. degrees in computer engineering from Chungbuk National University in 1999 and 2001, respectively. Currently, he is a Ph.D. candidate in computer engineering at Chungbuk National University. His research interests include information retrieval and natural language processing.



Young-Hoon Seo
He received the B.E, M.E. and Ph.D degrees in computer engineering from Seoul National University in 1983, 1985 and 1991, respectively. He had been the Visiting Scholar in Center for Machine Translation, Carnegie-Mellon University, U.S.A in 1994~1995. He is currently a professor in school of electrical and computer engineering at Chungbuk National University. His research interests include natural language processing, machine translation, spoken language processing and information retrieval.