Identifying Basic Patterns of Korean Natural Language Query

Jinseok Chae and Sukho Lee
Dept. of Computer Eng., Seoul National University
Seoul, 151-742, Korea
E-mail: {wahr, shlee}@ce2.snu.ac.kr

Abstract

Korean natural language queries are composed of a number of basic building blocks. This paper describes the process to identify the basic patterns considered to be basic building blocks constructing Korean queries. Two sets of Korean queries generated by two groups of senior-level students were experimented. Questions from the first set were produced by students who had no knowledge about databases and schema. Students from the second group attended a short lecture and understood that the Korean queries would be executed by a computer. By analyzing these experimental queries, seven basic patterns are identified. Korean queries combined by these basic patterns cover more than 80% of all questions.

1 Introduction

Recently a number of Korean natural language query interfaces which transform Korean queries into formal database queries such as QUEL or SQL have been developed.

NHI (Kim and Lee, 1985) accepts Korean natural language queries and generates QUEL. K-NLQ (Chae *et al.*, 1993) transforms Korean queries into SQL. Kim *et al.* (1994) proposes Korean Natural Language Query System which also transforms Korean queries into SQL. KID (Chae and Lee, 1995) transforms Korean queries into query graphs used in object-oriented databases. However, these interfaces do not fortify their opinions with any empirical data including a large number of sample queries.

Korean natural language queries are supposed to be composed of a number of basic building blocks. The term, basic building blocks, means that there are essential minimal pieces to construct Korean queries. In other words, Korean queries can be decomposed into a number of basic building blocks. However, there has been no research on identifying these basic building blocks of Korean queries. Identifying the basic patterns is important because it provides the basis to interpret Korean queries efficiently. In order to identify basic patterns considered to be basic building blocks, two sets of Korean queries generated by two groups of senior-level students were collected and these all queries are analyzed. The reason collecting different types of queries is for observing whether the preliminary training has an effect on the system's performance. The method used to collect sample queries in this paper is based on Dekleva (1994). Sample queries were generated by graduate and undergraduate students majoring in computer engineering. This task was assigned as a homework exercise which was part of the students' course grade.

Students from the first group had no training about databases and schema information. This assured that they wrote questions very naturally. But the purpose of this exercise, some cautions to form a query and a set of sample queries were given. The first group of students provided 142 questions.

The second group of students was introduced to the topics of databases and schema information through an hour lecture. An object-oriented model of a sample database consisting of 15 classes and a set of sample queries are also presented. The second group was also provided a list of successful queries and cautions to form a query. At the end of the hour, the same homework as assigned

to the first group was given to the second group. One hundred and fifty-one queries were generated from this second group.

Chae and Lee (1995) identified three basic patterns but they were insufficient to process various Korean queries because they were recognized through small set of sample questions. By experiments, only 60% of Korean queries were transformed correctly. In this paper, by analyzing the collected all sample queries, seven basic patterns of Korean queries are identified and more than 80% of all sample queries belong to these basic patterns. The system's processing capability is better than the previous one based on fewer sample queries.

The remainder of this paper is organized as follows. The overview of the KID and the schema of a sample database are explained in Section 2. In Section 3, the process of identifying basic patterns of Korean queries are described. In Section 4, the experimental results are presented. Finally, the conclusion is given in Section 5.

2 Overview of KID

2.1 System Architecture

The KID consists of three modules: *natural language analyzer*, *semantic interpreter* and *OQL generator*. The natural language analyzer accepts Korean queries and generates appropriate parsing trees. The semantic interpreter decomposes the parsing results into query phrases by referring to the database dictionaries and builds query frames for each query phrase. The OQL generator integrates these query frames and produces OQL proposed in ODMG-93 (Cattell, 1993).

The block structure of the KID is shown in Figure 1. In the figure, rectangles indicate modules and arrows the flow of processing.

Natural language analyzer: This module performs morphological analysis (Kang and Kim, 1994) and parsing to create the internal representations such as parsing trees from Korean queries. The parsing mechanism uses a variation of the CYK-algorithm (Yang and Kim, 1993). The KID system employs a general natural language analyzer used in Korean-English machine translation (Lee and Kim, 1993). The natural language analyzer generates two structures: *tree structure* and *predicate argument structure* (Allen, 1988). Among these, the semantic interpreter accepts the predicate argument structure.

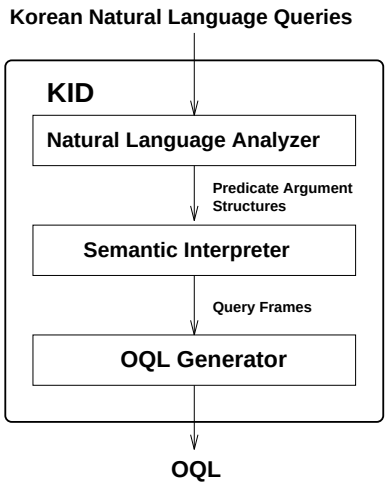

Figure 1: System architecture

Frame:	Frame Name	
Parents:	Parents Frame	
Properties:	H-CLASS	NIL
	R-CLASS	NIL
	H-ATTR	NIL
	R-ATTR	NIL
	VALUE	NIL
	OPERATOR	NIL
	CLASS-OF	NIL
	FUNCTION	NIL

Figure 2: Format of a query frame

Semantic interpreter: This module decomposes the predicate argument structures into *query phrases* (QPs) and builds *query frames* (Q-Frames). It utilizes two database dictionaries: *schema dictionary* and *domain dictionary* (Chae and Lee, 1995). The schema dictionary is used to specify the schema related information and the domain dictionary is used to determine the domain of unknown terms having the semantic ambiguities.

OQL generator: This module integrates all query frames and generates OQL. A query frame is designed to have the information about the class-attribute hierarchy such as classes, attributes, relationships, values, and operators. The format of a query frame is shown in Figure 2.

2.2 Class-Attribute Hierarchy

Figure 3 shows a sample class-attribute hierarchy used in this paper. It consists of fifteen classes

and '*' indicates multi-valued attributes. In this class-attribute hierarchy, each class has a number of attributes including the reference attribute representing the attribute-domain relationship.

3 Identifying Basic Patterns

3.1 Predicate Argument Structures of Basic Patterns

A Korean queries can be decomposed into a number of QPs and each QP is one of the identified basic patterns. By analyzing sample queries, we identify seven basic patterns: *head phrase I* (HP1), *head phrase II* (HP2), *noun modifier phrase* (NMP), *verb modifier phrase* (VMP), *adverb modifier phrase* (AMP), *verb phrase* (VP), and *comparative phrase* (CP).

Among these, NMP, VP and CP were identified in Chae and Lee (1995) and other phrases are identified additionally in this experiment. The predicate argument structures of basic patterns are as follows.

HP1: HEAD → [HT1 | HT2 | HT4]
  ⊢ [SUB | OBJ] → NOUN Noun
HP2: HEAD → HT3
  ⊢ (MOD → ADV 모두)
  ⊢ MOD → NOUN 몇
NMP: QP-HEAD → Qp-head
  ⊢ MOD → NOUN Noun
VMP: QP-HEAD → Qp-head
  ⊢ MOD → VERB Verb
AMP: QP-HEAD → Qp-head
  ⊢ MOD → ADV Adverb
VP: QP-HEAD → Qp-head
  ⊢ [MOD | VCON] → VERB Verb
    ⊢ [MOD | OBJ |
       SUB | NCON] → NOUN Noun
CP: QP-HEAD → Qp-head
  ⊢ [MOD | VCON] → VERB Verb
    ⊢ SUB → NOUN Noun1
    ⊢ MOD → NOUN Noun2

HEAD represents a head word of a sentence. It is classified into four types: HT1, HT2, HT3 and HT4. QP-HEAD indicates a head word of a QP. MOD denotes modifiers, SUB subject, OBJ object, VCON verb conjunction, NCON noun conjunction and ADV adverb. The sign '|' denotes 'OR'.

The Korean words corresponding each head type are as follows.

- HT1 : 보여라(show), 검색하라(retrieve), 출력하라(output), 나열하라(list), ···
- HT2 : 누구인가(who), 무엇인가(what), 얼마인가(how), 어디인가(where), 언제인가(when), ···
- HT3 : 명(분)인가, 개인가, 번인가(what number of), ···
- HT4 : 있는가(is there), 없는가(is not there), ···

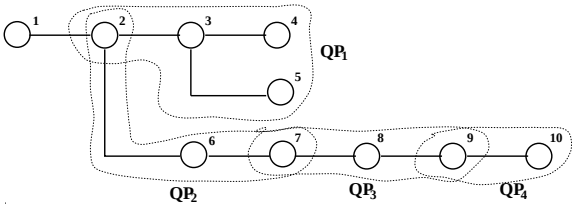

Figure 4: Pattern identification process

Some examples of QPs are:

- 과목을 등록한 학생 (students who enroll in course)

  VP: QP-HEAD → students
    ⊢ MOD → VERB enroll in
      ⊢ OBJ → NOUN course

- 키가 165보다 큰 교수 (professors who are taller than 165)

  CP: QP-HEAD → professors
    ⊢ MOD → VERB be taller
      ⊢ SUB → NOUN height
      ⊢ MOD → NOUN than 165

The pattern identification process employs the DFS (Depth First Search) algorithm. Figure 4 explains the process as an example when QP_1 is CP, QP_2 and QP_3 are VP, and QP_4 is NMP. The nodes of the tree structure in Figure 4 denote words in the questions and the numbers above the nodes the visiting sequences by DFS.

3.2 Definition of Korean Queries

A Korean query (KQ) consists of a head phrase (HP) and a main query (MQ). The MQ is classified into two kinds: *simple query* (SQ) and *composite query* (CQ). The SQ is a query which is a simple concatenation of several QPs without any conjunction (e.g., 'and', 'or' or 'among'), but the CQ has such conjunctions. If a query has the word '중' or '가운데' which means 'among' in English, then this query has an 'AMONG' indicator. The definition of Korean queries is as follows:

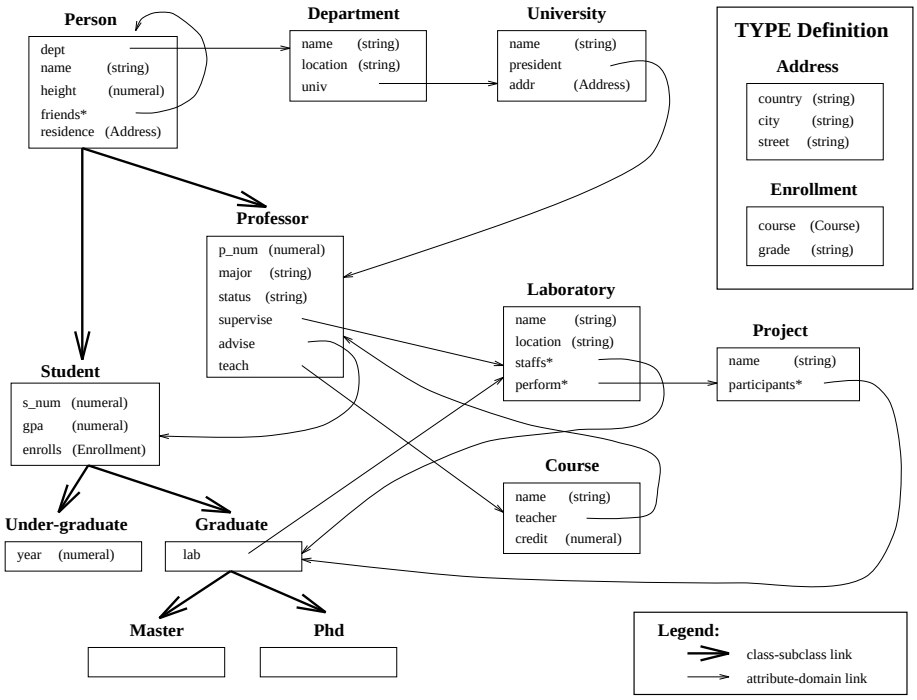

Figure 3: Class-attribute hierarchy

$$
\begin{aligned}
KQ &::= HP\ MQ \\
MQ &::= SQ \mid CQ \\
SQ &::= QP_1\, QP_2\, \cdots QP_n\ (n \geq 1) \\
CQ &::= SQ_1\, \Xi\, SQ_2\, \Xi\, \cdots\, \Xi\, SQ_m\ (m \geq 2) \\
HP &::= HP1 \mid HP2 \\
QP &::= NMP \mid VMP \mid AMP \mid VP \mid CP \\
\Xi &::= AND \mid OR \mid AMONG
\end{aligned}
$$

3.3 Statistics

Table 1 shows the distribution ratio of head types. About 45% of all questions were HT2. This shows that the head phrases of HT2 are the most frequently used representations in Korean queries. More than a third (34.8%) of the questions were HT3. The head phrases of HT3 are corresponding to the HP2 pattern. Almost all questions of HP2 are transformed into the aggregation function such as 'COUNT(*)'. The head phrases of HT1 and HT4 occupy relatively small portion.

Table 2 shows the results of performing the pattern identification process from sample queries generated by the first group. More than a half (54.9%) of the questions from the first set were composite queries including conjunctions. The distribution ratio of the composite queries is 70% of 'AMONG', 40% of 'AND' and 10% of 'OR'. Total percentage is more than 100%, because the conjunctions can appear in the queries more than once. More than 80% of the questions were recognized as appropriate for the interpretation. In other words, these questions can be made by combining seven basic patterns. About 18% of the questions have the unidentified patterns. Examples of such questions are:

	Number	%
HT1	52	17.7
HT2	133	45.4
HT3	102	34.8
HT4	6	2.1
Total	293	100.0

Table 1: Distribution ratio of head types

- 교수대 학생 비율이 가장 낮은 학과를 보여라. (Show the departments which have the lowest ratio of professors vs. students.)
- 여학생의 수가 남학생의 수보다 많은 학과를 보여라. (Show the departments having more female students than male.)

	Number	%
Simple queries	38	26.7
Composite queries	78	54.9
Unidentified queries	26	18.4
Total	142	100.0

Table 2: Identification results of the first group

	Number	%
Simple queries	65	43.0
Composite queries	61	40.4
Unidentified queries	25	16.5
Total	151	100.0

Table 3: Identification results of the second group

- 가장 최근에 생긴 학과를 보여라. (Show the department which is most recently established.)

Above questions have the unidentified patterns such as querying the ratio of two classes, comparing the number of two subsets, and having complex or unclear modifiers. Since they are considered to be exceptional cases and hard to interpret, we exclude them from the basic patterns.

Table 3 shows the results of the second group. About 40% of the questions from the second set were composite queries and so were simple queries. More than 80% of the questions were recognized as appropriate for the interpretation like the case of the first set. Only 16.5% of the questions have the unidentified patterns.

4 Experiments

The prototype system of KID is implemented on Sun Sparcstation using C language. Table 4 shows the results of the execution of the sample queries generated by the first group by KID. Only about 22% of the questions were interpreted correctly. More than a half (55.6%) of the questions referred to information that was simply not stored in the database. Examples of such questions are:

- 정보광장에 ID가 있는 학부생은 누구인가? (Which undergraduate students have the IDs in the Information Square?)
- '홍길동' 학생의 혈액형은 무엇인가? (What the blood type does the student 'G. D. Hong' have?)

	Number	%
Correct interpretation	32	22.5
Inappropriate queries	79	55.6
Unknown words	7	5.0
Structural problems	24	16.9
Total	142	100.0

Table 4: Execution results of the first group

	Number	%
Correct interpretation	105	69.5
Inappropriate queries	18	11.9
Unknown words	3	2.0
Structural problems	25	16.6
Total	151	100.0

Table 5: Execution results of the second group

- 취미가 농구인 사람은 누구인가? (Who plays basketball as a hobby?)

Some other questions were simply unclear, confused or not specific enough in relation to the database. About 17% of the questions have the structural problems. These problems originate from the parsing of Korean queries. It is known as a hard task to analyze overly complex Korean sentences correctly. The remainder of interpretation problems was caused by the underdeveloped state of the lexicon. Many of the synonyms are unknown to the system will have to be added to the lexicon.

Table 5 shows the results of the execution of the second group by KID. About 70% (compared to 22.5 % of the first set) of the questions were interpreted correctly. It shows a strong indication: users, not just the system, have to be prepared to use the system effectively. The effects of an hour lecture is significant. This is why we use natural language query interfaces in spite of their limitations. Only 11.9% (compared to 55.6% in the first set) of the questions were still inappropriate.

5 Conclusion

In this paper, seven basic patterns of Korean queries are identified to interpret Korean natural language queries efficiently. The contributions of this paper are summarized as follows.

- Basic patterns of Korean queries are identified by analyzing a large number of various queries.
- KID system is implemented and the system's performance is better than the previous one based on fewer sample queries.
- Results of experiments show that the preliminary training is necessary.

In order to identify the basic patterns, two sets of Korean queries were collected. Questions from the first set were produced by students who had no knowledge of databases and schema. Students from the second group attended a short lecture and learned about databases and schema information. More than 80% of the questions are composed of identified basic patterns by analyzing these sample queries. The results of the execution by KID show that the preliminary training is essential for users who are willing to ask databases in natural language.

In the future, we will concentrate on upgrading the processing capability of the system. We have plan to devise methods to manipulate the exceptional cases and to add unknown words and synonyms to the lexicon.

References

Kim, S., and Lee, S. (1985) "The Design and Implementation of Interface for Processing Natural Hangul Query," (in Korean) Journal of the Korean Information Science Society, Vol. 12, No. 1, pp. 31-44.

Chae, J., Kim, S., and Lee, S. (1993) "Design and Implementation of a Natural Language DB Query System," (in Korean) Journal of the Korean Information Science Society, Vol. 20, No. 6, pp. 810-820.

Kim, J. M., Hyun, M. Y., and Lee, S. J. (1994) "Koran Natural Language Query System for Searching Database,"(in Korean) Proceedings of the 21st KISS Fall Conference, pp. 637-640.

Chae, J., and Lee, S. (1995) "Natural Language Query Processing in Korean Interface for Object-Oriented Databases," Proceedings of the First International Workshop on Applications of Natural Language to Data Bases, Versailles, France, pp. 81-94.

Dekleva, S. M. (1994) "Is Natural Language Querying Practical?," DATA BASE (The Journal of the ACM SIGBIT), Vol. 25, No. 2, pp. 24-36.

Cattell, R.G.G. (1993) The Object Database Standard: ODMG-93, Morgan Kaufmann Publishers.

Kang, S. S., and Kim, Y. T. (1994) "Syllable-based Model for the Korean Morphology," Proceedings of the COLING 94, pp. 221-226.

Yang, J., and Kim, Y. T. (1994) "Korean Analysis using Multiple Knowledge Sources," (in Korean) Journal of the Korean Information Science Society, Vol. 21, No. 7, pp. 1324-1332.

Lee, H. G., and Kim, Y. T. (1993) "Korean-English Machine Translation based on Idiom Recognition," Proceedings of IEEE Region 10 Conference (TENCON '93).

Allen, J. (1988) Natural Language Understanding, Benjamin/Cummings Co. Ltd.