Lexico-Syntactic Analysis of Query Types in Online Q&A Systems

Seung-Hye Shim, Sebastien Paumier, Jee-Sun Nam

Lexico-Syntactic Analysis of Query Types in Online Q&A Systems

Seung-Hye Shim*, Sebastien Paumier and Jee-Sun Nam***

*DICORA, Hankuk University of Foreign Studies, Korea

**IGM, University of Paris-Est, France

ssh4808@naver.com, paumier@univ-mlv.fr, namjs@hufs.ac.kr

Abstract

This paper suggests that a search engine should be able to recognize query sentences, rather than some keywords, in order to provide reliable result for users' needs. We aimed to analyze semantic relationships and syntactic patterns of query types in one specific domain. Based on this analysis, we show that how many diverse query patterns can be used to carry one idea. We observed different sentence types, especially questions both with and without the wh-phrase. The syntactic and semantic patterns we describe here can be used to improve existing search engines, as well as in analyzing other domains in future works.

1. Introduction

Current information retrieval (IR) search engines recognize keywords in the question, and then provide a ranked set of documents consisting those keywords. Since there is an enormous number of documents, users must also examine the data to find relevant answers to their queries. We need question answering (Q&A) systems, beyond the limited abilities of these IR systems, that are more reliable, understand sentence-level questions, and can analyze the users' intent. For example, consider "*ssangkhaphwul swuswul eti-ka calhayyo*?" ("Where is the best clinic for double eyelid surgery?") and *"ku pyengwenun etil-lul calhayyo?" ("*Which part of the body does the clinic do well at operating upon?"). The meaning of each sentence is rather different, despite utilizing the same word, "*eti"* ("where, which part of the body"). If we searched using this keyword, the search engine would not be able to distinguish between "*eti"* ("where") and "*eti"* ("which part of the body"). Therefore, we should examine sentence-level searches.

One of the important aspects of researching natural language Q&A systems is exploring the exact meanings of questions. However, this is not simple; for instance, we can generate various types of sentences using these concepts (SAMSUNG, 2009, MAKE, MOBILEPHONE), such as *"In 2009, what kinds of mobile phones were made by Samsung?"* or *"What types of cell phones did Samsung put out last year?"* To investigate the exact meanings of a wide range of questions, we should examine each domain and understand the syntactic properties and lexical characteristics of the sentences within each. This is why Q&A systems require more complicated natural language processing (NLP) techniques than do classical IR systems that use basic keywords.

In this paper, we describe various question types in a given domain by using a Local Grammar Graph (LGG) model. The LGG methodology, founded by Maurice Gross (Gross 1997, 1999), can effectively formalize linguistic expressions that violate general syntactic rules and are difficult to explain in the usual way. The LGG model allows us to describe more partial phenomena than whole sentences and to analyze and process compound

nouns, frozen idioms, collocations, and synonyms of certain phrases.

One common architecture for Q&A systems consists of three main modules: question processing, document processing, and answer processing. The important elements of question processing are question classification and reformulation. In this paper, we show how important question processing is in providing accurate answers to users' queries.

To accomplish this, we begin by collecting questions from one restricted domain, then study the syntactic patterns and lexical characteristics, and, finally, build the LGGs.

2. Related Works

Existing research on question classification is mostly based on machine learning approaches. Dell Zhang et al. (2003) has suggested the use of automatic question classification through a Support Vector Machine (SVM), taking advantage of two models, the bag of words model and the bag of n-grams model. They introduced the tree kernel that helps the SVM to utilize the syntactic structures of questions. Ulf Hermjakob (2001) has proposed another approach to the machine learning method. His paper suggested that parse trees need to be semantically richer and, structurally, more semantically-oriented than what most treebanks offer (Ulf, 2001). Another group of researchers have worked on generating clinical questions typically asked by patients (Ely et al., 2000). In order to offer doctors good examples about patient care, John W. Ely and associates categorized these question types. Kim Gi-Cheol (2006) has analyzed questions with wh-phrases, classifying the types of wh-phrases.

Since current research on natural language Q&A systems are mostly based on statistical models, which use limited syntactic rules and lexical information in open domains, we should underline further linguistic studies on sentence structures and semantic constraints.

3. Query Analysis

As discussed above, it is vital for Q&A systems to examine a question's intent and find answer documents related to that question. Despite the voluminous existing research, Web search engines still provide answers by matching the keywords included in the questions. If one submits the question, "*ssankhaphwul cal hanun pyengwen etiinkayo?"* ("Where is the best clinic for double eyelid surgery?") on Google, the results show that the search engine recognized certain keywords, such as *ssankhaphwul, cal, ha (nun), pyengwen,* and *eti (inkayo)*, and then offered documents including those keywords. The unstable Q&A system phenomenon is related to the difficulty of classifying of a variety of query types. Therefore, a Q&A system needs to recognize and categorize diverse query types representing one idea. For this paper, we built up a corpus of questions about plastic ("cosmetic") surgery, collected from online Websites such as *NAVER-In (http://kin.naver.com)* and *DAUM.cafe (http://cafe.daum.net)*. The corpus used in this study consisted of about 400 sentences. The question types can be divided into 2 groups:

[1] Questions with "WH" words, "*munges"* ("what"), "*eti"* ("where"), "*ngencey* ("when"), "*nwuku"* ("who")

[2] Questions without "WH" words.

It is obvious that we cannot expect only wh-questions, because plenty of question types can be used for one idea. For instance, a variety of sentence types can express the idea, "questions about clinic names."

[1] Questions with "WH" words:

(1) *"kasumswuswul yumyenghan pyengwen etiinkayo."*
("Where is the most famous bad-breath surgery clinic?")

(2) "*kho swuswul etika cal hanunci allyecwuseyyo"*
("Where is the best clinic for a nose job?")

[2] Questions without "WH" words:

(1) *"Kho swuswul cal hanun pyengwen epsulkkayo."*
("Is there any good clinic for a nose job?")

(2) *"ssangkhaphwul cenmwun pyengwen asinun pwun chwuchenhay cwuseyyo."*
("If you know a good clinic for double eyelid surgery, please recommend it to me").

Below are some examples of question types with "*eti" ("*where") and without "*eti*." In the questions with "*eti*," the sentences are not only interrogatives but are also declarative sentences, and the position of "*eti"* varies, anywhere from the end of the sentence to in the middle. In the questions without "WH" words, we observe declarative sentences. Thus, we confirmed that these diverse question types represent one idea, and, therefore, we need to examine the corpus of a specific domain.

For this task, we first built the corpus and classified about 400 sentences according to what each question was about, because the answer type relates to the question's focus. We looked at each question and categorized its focus. Based upon such observations, we extracted the focus of each question, which we divided into four sub-domains, such as the following:

Ⅰ. Questions about Clinic Names
Ⅱ. Questions about Results After Surgery
Ⅲ. Questions about Costs
Ⅳ. Questions about Surgical Methods

In this paper, we aimed to describe the query patterns regarding "Clinic Names." To accomplish this purpose, we analyze lexical characteristics and syntactic patterns in the next section, based on 220 queries about clinic names.

4. Lexical and Syntactic Descriptions of Questions about Clinic Names

4.1 Lexical Characteristics

In this section, we analyze the lexical characteristics observed in the "plastic surgery" corpus, since it is important to know the keywords included in questions. By using a Korean morphological analyzer, *Geuljabi*, we extracted the high-frequency nouns and separated them into three classes, as follows.

Class 1. Nouns regarding the part of body

After analyzing the questions, we found one of the most frequent noun types concerned the part of the body that the users wanted to change. The nouns specifying the body part should be specified when the users make a query. Questions about clinic names usually included nouns of this class, especially those pertaining to face parts: "*kho"* ("nose"), "*nwun"* ("eyes"), "*anmyen"* ("face"), etc. These words mostly appeared with verbs such as "*hata"* ("to do") and "*patta"* ("to receive").

Class 2. Nouns related to facility type

Words about places frequently appeared in the questions. The location classifiers that appeared in the questions included "*kos"* ("space") and "*tey"* ("place"). Furthermore, people usually expressed some generic terms, like "*pyengwen"* ("hospital") or "*senghyengoykwa"* ("surgery clinic"), in order to inquire regarding the clinics' names.

Class 3. Geographic Names

When users asked about clinic names, they usually included the local area like, "*Pwusan*" or "*Kangnam*." The nouns for geographic names act as options, which the users provide in their questions in order to learn the clinics' names.

These classes can be summarized as follows:

Class 1	Face: "*kho"* ("nose"), "*nwun"* ("eyes")	Body: "*kasum"* ("breath"), "*hepekci"* ("thigh")	
Class 2	*"Pyengwen"* ("hospital")	*"kos"* ("space")	*"tey"* ("place")
Class 3	Local area: "*Pwusan*," "*Sewul,"* "*Kangnam," etc.*		

Table 1. Lexical Characteristics

In the next section, we describe the syntactic patterns for each question type.

4.2 Syntactic Patterns

The syntactic patterns of the questions about clinic names can be classified into two groups: interrogative and declarative sentences:

[1] Interrogative sentences, with suffixes such as "*-eyyo*?"
[2] Declarative sentences, with suffixes such as "*-e cwuseyyo*"

There are several question types, but users want to know where the best clinic is. In this section, we analyze the syntactic patterns of each query type and present some examples, using LGGs. Classifying the varieties of query types will help enrich precision and recall ratios in answer extractions. First, we considered questions containing "*eti″* ("where"), "*etten″* ("which"), and "*enu″* ("which").

4.2.1 Interrogative Sentences

First, the interrogative pronoun "*eti″* ("where") can appear in many syntactic positions in interrogative sentences. The most frequent position is in the subject, followed by a nominative postposition such as "*ka*" or "*nun,*" as in "*eti-ka cal-hayyo?″* ("Where is doing well?"). However, this pronoun appears in other syntactic positions. For example, "*eti*" can appear with the copular term "*ita″* ("to be"), as in "*N eti-i-nkayo*?" ("N where-be-Sfx" ["Where is (N)?"]). Moreover, it can be followed by a locative postposition, such as "*e″* ("at") or "*eyes″* ("in"). Compare (1) to (2) and (3):

(1) *"eti-ka kho-swuswul cal-ha-nayo?"*
("Where [is the clinic that] does nose jobs well?")

(2) *"kho-swuswul hako siph-untey eti-lo ka-yahanayo?"*
("Where do I have to go for a nose job?")

(3) *"kasum swuswul eti-eyse ha-nunkey coh-ulkkayo?"*
("My bad-breath surgery should be done where?" ["Where should my bad-breath surgery be done?"])

In the above examples, all questions are about the clinic names, and "*eti″* ("where") presents the focus of the question. However, this pronoun is followed not only by a nominative case marker but also by a locative postposition, such as "*lo″* ("to") or "*eyse″* ("in"). When "*eti″* ("where") appears with the locative case, such as "*lo″* ("to"), the predicate is generally carried by a verb like "*kata″* ("to go"), a verb of movement. Therefore, we suggest representing their syntactic patterns as follows:

(1) $[eti]_1$-*ka*[1] $[N]_2$-Post V? (" $[Where]_1$ V Prep $[N]_2$?)

[1] The notation we use here should be read as follows: *ka* = nominative proposition, *lo* = locative proposition (e.g. "*lo*/*e*/*eyes*"), *W* = any sequences.

(2) ([N]$_1$) W [*eti*]$_2$-lo V?	(" ([N]$_1$) V to-[Where]$_2$?)
(3) ([N]$_1$) W [*eti*]$_2$-eyse V?	(" ([N]$_1$) V in-[Where]$_2$?)

Table 2. The syntactic patterns of Query Types with Pronoun "*eti*"

Second, there are question types that begin with interrogative adnominals, such as "*etten"* ("which") or "*enu"* ("which"). Here are some examples:

(4) *"etten pyengwen-i ssangkhaphwul cal ha-nayo?"*
("Which clinic does the double eyelid operation well?")

(5) *"kho swuswul-un enu pyengwen-i yumyeng-hanayo?"*
("Which clinic is famous for [its] nose jobs?")

(6) *"Anmyen-swuswul-un enu pyengwen-e ka-yahayo?"*
("In order to have facial surgery, which clinic do I have to go to?")

In this situation, the wh-words, such as "*etten"* ("which") and "*enu"* ("which"), appear with a noun; the sequence "*etten/enu"* ("which") + *"N"* corresponds to "*eti"* ("where"), as we mentioned above. Below, we present suggested syntactic patterns for questions with "*etten"* ("which") and "*enu"* ("which"):

(4) [*etten/enu* N]$_1$ -*ka*[2] [N]$_2$-Post V?	(" [Which N]$_1$ V Prep [N]$_2$?)
(5) ([N]$_1$) W [*etten/enu* N]$_2$-e V?	(" ([N]$_1$) V at-[Where]$_2$?)

Table 3. The syntactic patterns of Query Types with "*etten*/*enu"* ("which")

However, we find some questions where "*eti*" appears with a noun like "*etten/enu*." Let us consider these:

(7) *"eti coh-un pyengwen eps-ulkkayo?"*
("Is there a good clinic in that location?")

(8) *"eti cal ha-nun pyengwen asinayo?"*
("Do you know a good clinic in that location?")

We see "*eti*" appears as an interrogative adnominal in the above examples. However, it differs from the true interrogative pronoun "*eti,*" which cannot be omitted from interrogative sentences, because we can delete it in the above examples. Therefore, for (7) and (8), we can also have the following sentences:

[2] The notation we use here should be read as follows: *ka* = nominative proposition, *lo* = locative proposition (e.g. "*lo*/*e*/*eyes*"), *W* = any sequences.

(9) *"coh-un pyengwen eps-ulkkayo?"*
("Is there a good clinic?")

(10) *"cal ha-nun pyengwen asinayo?"*
("Do you know a good clinic?")

Thus, "*eti*" in (7) and (8) is a pseudo-pronoun, an appositive pronoun to the noun. We shall treat this phenomenon in detail in our future works.

4.2.2 Declarative Sentences

Some queries appear as declarative sentences. They contain certain location classifiers, such as "*kos*" ("place") and "*tey*" ("location"), instead of the interrogative wh-words. Consider the following:

(11) *sewul-ey kho swuswul cal hanun kos chwuchen com ha-ycwuseyyo.*
("Please recommend to me a good place to get a well-done nose job operation in Seoul.")

(12) *inchen-ey nun-senghyeng cal ha-nun tey allkosipheyo.*
("I want to know a good location for eye surgery in Incheon")

In the above examples, we see questions without wh-words but with location classifiers. The classifiers require sentential modifiers to become noun phrases. However, they express the information that the users want to know. We summarize their syntactic patterns as follows:

(6)	*S-Modifier [CLAS]$_2$-Acc V!*	(" V [S-Modifier Location]!)
(7)	([N]$_1$) *S-Modifier [CLAS]$_2$-Acc V.*	("[N]$_1$V [S-Modifier Location])

Table 4. The syntactic patterns of Query with Classifiers of Location

5. Conclusion

In this paper, we described and classified question types regarding "*Clinic Names*" in the domain of "plastic surgery." We divided the question types into 2 classes: those with and without wh-words. The former consisted of interrogative sentences, sub-divided into two groups. The first consisted of those containing interrogative pronouns, such as "*eti*" ("where") or interrogative adnominals, such as "*etten*" ("which") and "*enu*" ("which"). The second consisted of declarative sentences that appear with location classifiers, such as "*kos*" ("place") and "*tey*" ("location").

As we mentioned above, for each query type there exists several linguistic expressions corresponding to one idea, and we need a formal method for presenting them in an effective way. LGG (Local-Grammar Graph)

formalism should be a perfect mechanism for handling the variations in the sentences that nevertheless carry the same idea. The LGG model, created by Maurice Gross (Gross 1997, 1999), can effectively formalize linguistic expressions that violate general syntactic rules and are difficult to explain in the regular way. For example, the following LGG presents 285 questions about one idea: "What is the best clinic for surgery on BodyPart?" which can be written as (CLINIC, BODYPART, SURGERY).

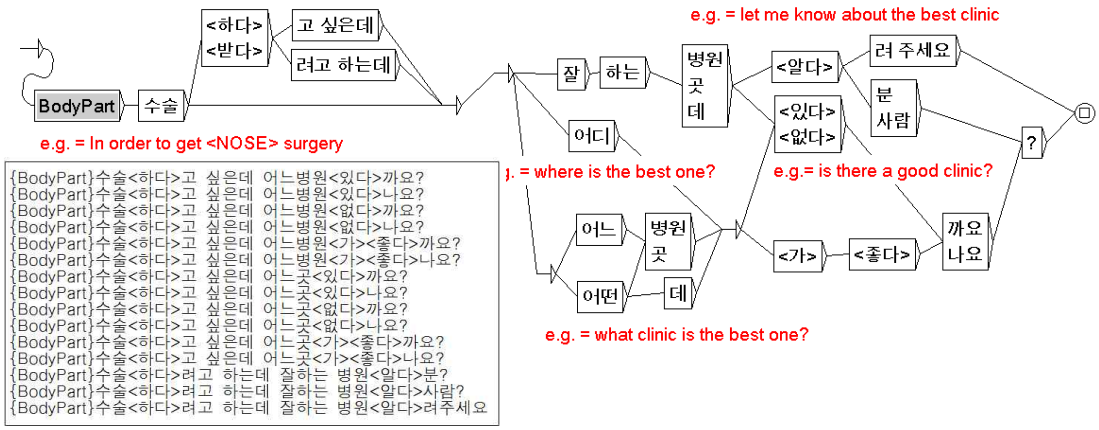

Figure 1. An LGG representing question types about CLINICS

In this way, we built main LGGs for the sub-domains about *Clinic Names*, which are called finite sub-graphs, to recognize all the different types of questions that have the same meaning. These LGGs were constructed by the Unitex Graph Editor and compiled into finite-state transducers (FSTs), for direct use in text processing (Paumier 2003; http://www-igm.univ-mlv.fr/~unitex/). To verify the accuracy of these LGGs, we built another set of corpora of on-line queries and examined the given LGGs' question-meaning recognition ratios.

The syntactic and semantic patterns we described here will be useful for improving Q&A systems' ability to recognize on-line queries. They apply to the study of other medical domains' query types, as well as to the plastic surgery domain we presented in this paper.

References

D. Zhang and W.-S. Lee, 2003. Question classification using support vector machines. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. P 26~32.

K.-C. Choung and Y.-H. Seo. 2003. Question analysis using information and noun semantic information. *The Korea Contents Association*. Vol.1 No.2. P 185~189.

U.-S. Lee. 2007. *Hyentay kwuke nhuymwunsauy mwunpepkwa uymi. ("Grammar and Semantics of Korean Interrogative pronouns)*. Korean Linguistics Association.

M. Gross, 1997. The construction of Local Grammars. *In Finite-State Language Processing*, E.Roche&Y.Schabes("eds). Language. Speech. and Communication. Cambridge. Mess. : MIT Press. P 329~354.

M. Gross, 1999. A bootstrap method for constructiong local grammars. In Contenproary Mathematics. *Proceedings of the Symposium*. 18~20 December 1998. Blegrade. Serbia. N.Bokan("ed). University of Belgrade. P 229~250.

J.-S. Nam. 2009. Study on Comparative Sentences based on Adjectival Predicates for Automatic Extraction of On-line Comparative opinions. *Journal of Language Science* 16-2, 63~95.

P. Jacquemart and P. Zweigenbaum. 2003. Towards a medical questions-answering system: a feasibility study. *Studies In Health Technology And Information*. Vol.95. P 463~468.

G.-C. Kim. 2008. Korean language quesry analysis with an interrogative pronoun fir infroemation retrieval. *The journal of the korean institue of communication science.* Vol.33 No.2D. p 48~54.

W.- K Kim. 2008. Discourse analysis of Q&A sentences for Question-Answering system. *Korean Language and Text Research*. Vol. 30. P 235~263

Y. Niu and G. Hirst. 2004. Analysis of semantic classes in medical text for question answering. *Proceedings of ACL workshop*.

S. Paumier, 2003. *De la reconnaissance de formes linguistiques à l'analyse syntaxique.* PhD Dissertation. University of Paris-Est Marne-la-Vallée.

S.-E. Shin, H.-G. Park and Y.-H. Seo. 2007. Question analysis and expansion based on semantics. *The Korean Contents Society*. Vol.7 No.7. P 50~59.

K. Woo and C. Park. 2004. A study on modifycation structure and query sentence for question-answering system. *Korean Society For Internet Information*. Vol.5.No.2. P 473~476.

J.- W. Ely et al. 2000. A taxonomy of generic clinical questions: classifiaction study. *BMJ*. Volume 321.

U. Hermjakob. 2001. Parsing and question classification for question answering. Proceeding of the workshop on open-domain question answering at ACL-2001.