하이브리드 방법의 사용자 질의 의도 분류

(A Hybrid Method for classifying User' s Asking Points)

김학수 [†] 안영훈 [†] 서정연 [††]
(Harksoo Kim) (An, Young Hun) (Jungyun Seo)

요 약 질의응답 시스템이 올바른 답변을 제시하기 위해서는 사용자의 의도를 정확하고 강건하게 파악하는 것이 매우 중요하다. 이러한 요구 사항을 만족시키기 위해서 본 논문에서는 실용적 질의응답 시스템을 위한 질의 유형 분류기를 제안한다. 제안된 질의 유형 분류기는 규칙 기반의 방법과 통계 기반의 방법을 접목시킨 하이브리드 방법을 사용한다. 제안된 방법을 사용함으로써 수동으로 규칙을 작성하는 시간을 줄일 수 있었고 정확률을 향상시킬 수 있었으며 안정성을 보장받을 수 있었다. 제안된 방법에 대한 실험에서 질의 유형을 분류하는데 80%의 정확률을 얻었다.

키워드 : 질의 의도 분류, 질의 응답 시스템, 하이브리드 방법

Abstract For QA systems to return correct answer phrases, it is very important that they correctly and stably analyze users' intention. To satisfy this need, we propose a question type classifier (i.e. asking point identifier) for practical QA systems. The classifier uses a hybrid method that combines a statistical method with a rule-based method according to some heuristic rules. Owing to the hybrid method, the classifier can reduce the time to manually construct rules, yield high precision rate and guarantee robustness. In the experiment, we accomplished 80% accuracy of the question type classification.

Key words : Classifying user's asking points, Question-answering system, Hybrid method

1. Introduction

One of the differences between an IR system and a QA system is whether the system has a classification module for identifying users' asking points or not. As shown in TREC QA systems[1, 2, 3, 4, 5, 6, 7, 8], most of current QA systems have the special modules to classify users' question types. As in these systems, the classification of question types is necessary for Question Answering (QA) systems because the systems should filter out inadequate answer candidates. For example, if a user asks a question like "*Who is the president of Yahoo?*", the QA system should return the names of person. If the QA system returns answer candidates that are included in semantic categories such as *country*, *date*, and *time*, he/she will keep a suspicious eye on usefulness of the system.

In real fields like World Wide Web (WWW), the application domains of QA systems should be easily shifted to other domains because the types of users' questions are different according to the domains. For example, if a QA system is installed at a university, users will often ask which score they obtained at the last exam. On the other hand, if the QA system is installed at a public office, users will often ask which types of documents they should submit to the office. This means that we should be able to easily add new question types to the QA system. In addition, it says that the QA system should be able to robustly analyze users'

[†] 비 회 원 : 다이퀘스트 연구소 선임연구원
hskim@diquest.com
cyllian@diquest.com
[††] 종신회원 : 서강대학교 컴퓨터학과 교수
seojy@ccs.sogang.ac.kr

논문접수 : 2002년 7월 19일
심사완료 : 2002년 10월 25일

queries even though the domain is changed. At the same time, we still expect high accuracy and easy-to-tune ability to the classifier. To satisfy these needs, we propose a hybrid system for classifying question types. The hybrid system combines two different method a rule-based method and a statistical method. The rule-based module uses regular expression rules that are manually constructed. The module provides fast classification and high precision rate. The statistical module uses induction rules that are learned by C4.5[9]. The module guarantees robust classification because it is learned from large amount of data.

This paper is organized as follows. In Section 2, we review the previous works of question classification. In Section 3, we describe how to construct a named entity dictionary that is an essential resource of our system. In Section 4, we propose a hybrid system for classifying question types. In Section 5, we analyze the result of our experiments. Finally, we draw some conclusions.

2. Previous Works

The current approaches for identifying users' asking points can be classified into two groups; rule-based approaches and statistical approaches.

The rule-based approaches[1, 3, 4, 5, 7, 8] generally use finite state recognizers for matching lexico-syntactic patterns that are manually generated. MURAX[8] is a representative QA system that uses a rule-based question classifier. Most of rule-based systems are similar to that of MURAX and have some advantages as follows:

- They can promptly classify users' questions into predefined semantic categories (i.e. question classes or question types).
- They can be easily tuned to good performance for specific domain.

However, the time required for classification will linearly grow up, and the maintenance of the rules will become more and more difficult as the number of the rules increases. Furthermore, the handcrafted rules may be fragile and have to be manually rebuilt or optimized for domain changes.

The statistical approaches[2, 10, 11, 12] use a large amount of training data for question classifi- cation. The training data is manually annotated with semantic tags. Generally, statistical systems can robustly classify users' queries because the systems are based on reliable information that is obtained from a large amount of training data. However, the statistical system has a weak point that sometimes it gives unexpected results in real fields. To overcome this week point, statistical question classifiers should be equipped with supplementary modules that rule-developers can easily edit rules or grammars for correct classification.

3. A Named Entity Dictionary

The hybrid system uses a named entity dictionary, which is called PLO dictionary. Using the PLO dictionary, the proposed system converts lexicons into semantic markers. We constructed the PLO dictionary with 477,596 entries. The PLO dictionary contains three kinds of entry words as follows:

- Proper nouns such as the names of people, countries, cities, organizations, and etc.
- Common nouns such as jobs, positions, hobbies, and etc.
- Unit nouns such as km, m, cm, kg, g, mg, and etc.

Table 1 shows a part of the PLO dictionary. As shown in Table 1, an entry of the dictionary consists of a lemma and semantic markers. The semantic markers mean semantic categories that the lemma is associated with. In a sense, the

Table 1 A part of the PLO dictionary

Lemma	Semantic mark
cm	(%length_unit)
dollar	(@money_unit)
Gardner	(@city|@person)
gangsa	(@position)
gangui	(%lecture)
ingan	(%person)
kg	(%weight_unit)
New York	(@city)
sogangdaehakgyo	(@organization|@building)

semantic markers are similar to sense codes in WordNet[13].

The 9th entry, "*sogangdaehakgyo (@organization| @building)*", means that *sogangdaehakgyo* (*Sogang University*) is the name of an organization or the name of a building. The 6th entry, "*ingan (%person)*", means that *ingan* (*human*) is semantically similar to *person*. "*X (@Y)*" implies that *X* is the hyponym of *Y*, and "*X (%Y)*" implies that *X* is the synonym of *Y*.

4. Hybrid Query Classification

To take advantages of the rule-based and statistical approaches, we propose a hybrid system for classifying question types. The hybrid system consists of three sub-modules: a rule-based classifier, a statistical classifier and a hybrid merger. The classification processes of the hybrid system are as follows. First, the preprocessor converts an input sentence into two different forms; lexico-syntactic patterns and semantic patterns. The rule-based classifier uses lexico-syntactic patterns, and the statistical classifier uses semantic patterns. Second, each classifier turns out the result individually. Finally, the hybrid system merges the outputs and select one according to some heuristic rules.

4.1 Semantic Category

We classify users' queries into 105 semantic categories. We think that the 105 semantic categories are frequently questioned in practical IR/QA systems.

Table 2 A part of 105 semantic categories

1st layer	2nd layer		
animal	bird	fish	mammal
	person	reptile	
location	address	building	city
	continent	country	state
	town		
date	day	month	season
	weekday	year	
time	hour	minute	second
organization	company	department	family
	group	laboratory	school
	team		

As shown in Table 2, the semantic categories consist of 2 layers. The semantic categories in the first layer have broader meanings than those in the second layer. To define the 105 categories, we referred to the categories of QA systems in TREC [1][2][3][4][5][6][7][8] and analyzed users' query logs that are collected by a commercial IR system[14].

4.2 Matching Lexico-syntactic Patterns

For matching a user's query with handcrafted lexico-syntactic patterns, the rule-based classifier converts the query into a suitable form, using the PLO dictionary. For example, the query "*yahukoriaui sajangeun nuguingayo?* (*Who is the president of Yahoo Korea?*)" is translated into "*yahukoria j %person j %who jp ef sf (%who auxiliary verb %person preposition Yahoo Korea symbol)*". In the example, *%person* and *%who* are the semantic markers. The content words that are not listed on the PLO dictionary keep their lexical forms. The functional words (*e.g.* auxiliary verb, preposition) are converted into POS's. After conversion, the rule based classifier matches the converted query against one of lexico-syntactic patterns, and classifies the query into the one of 105 semantic categories. When two or more patterns are matched, the classifier selects the first matched category. Table 3 shows some lexico-syntactic patterns for *person* and *tel_num* categories. The above sample query matches the first pattern in Table 3.

Table 3 Lexico-syntactic patterns

Semantic category	Lexico-syntactic patterns
person	%who (j|ef)?
	(%person|@person) j? (sf)* S
	(%person|@person) j? %ident j? (sf)* S
	(%person|@person) j? (%about)? @req
	(%person|@person) j? (%ident)? @req
	(%person|@person) jp ef (sf) S*
	%which (%person|@person)
tel_num	*(%tel_num|@tel_num) (%num)? j? (sf)*S*
	(%tel_num|@tel_num) (%num)? j? %what
	(%tel_num|@tel_num) j? (%about)? @req
	(%tel_num|@tel_num) j? (%what_num)

4.3 Applying Statistical Rules

For learning statistical rules, we select a decision tree method because it is very fast and efficient with a good generalization capability. Among the various decision tree learning-algorithm, we choose C4.5 algorithm to train the statistical query classifier. C4.5 algorithm generates a decision tree by finding a feature that yields the maximum information gain[9]. In the decision tree, a node is generated with a set of rules corresponding to the feature. This process is repeated for all other features in succession until no further information gain is obtainable. In testing, a pattern is repeatedly compared with a node of a decision tree starting from the root and following appropriate branches based on the condition and feature value until a terminal node is reached. The pattern is then presumed to belong to the class that the terminal node represents.

To construct input patterns, the statistical classifier approximates a user's query to a suitable form[15][16]. We call this form as a semantic pattern. Generally, the semantic pattern includes semantic features like the semantic markers of keywords as well as syntactic features like the type of a main verb. Table 4 shows a composition of the semantic pattern.

Table 4 A composition of the semantic pattern

Semantic feature	Values	Notes
Interrogative	NULL, what, what_num, when, which, where, who, why, how, how_much	The type of an interrogative
Main-verb	NULL, pv, pa, pv_define, a_method, be	The type of a main verb
POS1	NULL, ncn, ncp, pv, pa, j, jp, ep, ef, etc (total 43 kinds)	The POS of the first focus word
POS2	NULL, ncn, ncp, pv, pa, j, jp, ep, ef, etc (total 43 kinds)	The POS of the second focus word
Sem-mark1	NULL, none, %person, @person, %city, %country, %company, etc (total 245 kinds)	The semantic marker of the first focus word
Sem-mark2	NULL, none, %person, @person, %city, %country, %company, etc (total 245 kinds)	The semantic marker of the second focus word

As shown in Table 4, the semantic pattern consists of 6 semantic features; the type of an interrogative, the type of a main verb, POS's of two focused words, and semantic markers of the two focus words. In Table 4, *NULL* means that the feature does not exist in a sentence, and *none* means that a value of the feature does not exist. To extract focus words from a query, the statistical classifier applies heuristic rules to the query, as shown in Algorithm 1. Then, it transforms the focus words into semantic markers after looking up the PLO dictionary. For example, if a user inputs the sentence "*yeoreum banghaki myeot wole sijakhapnikka?* (*What month will the summer vacation begin on?*)", the statistical classifier sets the values of *Sem-mark1* and *Sem-mark2* to *%month* and *none* after extracting *wol* (*month*) and *sijakha* (*begin*) from the sentence.

1. Count the number of specific interrogatives such as *eoneu* (*which*), *museun* (*what*), and *myeot* (*how many*).
2. If the number is 2, select the head nouns of the specific interrogatives as the focus words.
3. If the number is 1, select the head noun of the specific interrogative and the last content word in the query as the focus words. If there are no any other content words except the head noun, select the head noun and *NULL*.
4. If the number is 0, select the last 2 content words as the focus words. If the number of content words is 1, select the content word and *NULL* as the focus words.

Algorithm 1 Extraction of focus words

Figure 1 shows an example to extract the semantic patterns from an input sentence. As shown in Figure 1, if a focus word has a multi-semantic marker, the statistical classifier generates multiple semantic patterns.

On training time, the statistical classifier uses semantic patterns as input patterns and uses semantic categories as output patterns. After training a decision tree, the statistical classifier promptly determines semantic categories of users' queries by traversing the decision tree. If a query has multiple semantic patterns, it selects a semantic category that has the maximum value.

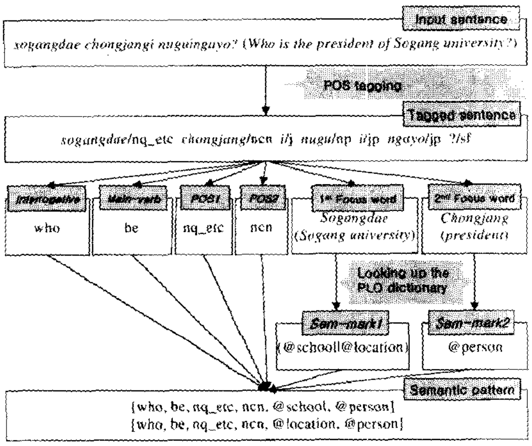
Input sentence
sogangdae chongjangi nuguinguyo? (Who is the president of Sogang university?)
POS tagging
Tagged sentence
sogangdae/nq_etc chongjang/ncn i/j nugu/np i/jp ngayo/jp ?/sf
Interrogative
who
Main-verb
be
POS1
nq_etc
POS2
ncn
1st Focus word
Sogangdae (Sogang university)
2nd Focus word
Chongjang (president)
Looking up the PLO dictionary
Sem-mark1
(@school|@location)
Sem-mark2
@person
Semantic pattern
{who, be, nq_etc, ncn, @school, @person}
{who, be, nq_etc, ncn, @location, @person}

Fig 1 An example to extract the semantic pattern

4.4 Hybrid System

The hybrid system selects a semantic category by merging the results of the rule-based classifier and the statistical classifier, as shown in Algorithm 2. By using Algorithm 2, the hybrid system performs well in specific domains because the handcrafted lexico-syntactic patterns guarantee fast and precise responses. More, the hybrid system robustly operates in any domains because the system is backed by the statistical method.

1. If the rule based classifier fails to return a semantic category, select the output of the statistical classifier.
2. If both classifiers return semantic categories which are similar in a wide meaning but included in different layers, select one in the second layer. For example, if *year* and *date* are the outputs, select *year* as the output of the hybrid system.
3. If both classifiers return quite different semantic categories, select the output of the rule-based classifier.

Algorithm 2 Merging results of the classifiers

5. Experiment

5.1 The Experiment Data

To experiment on the hybrid system, we collected users' queries from real web sites such as *www.sogang.ac.kr* and *korea.internet.com*. We manually annotated the queries using the 105 semantic categories. We call the corpus AQUS (Annotated QUery Set). AQUS for the experiment consists of 78 semantic categories with 7,726 queries. 27 categories are excluded from the experiment because there are no queries in 27 categories. Figure 2 shows the distribution ratio of the semantic categories in AQUS. As shown in Figure 2, AQUS includes a lot of explanation-seeking queries. The ratio of the fact-seeking queries to the explanation-seeking queries is 2.54 to 1.

	person	desc.	method	URL	tel_num
Num.	1100	1096	870	571	523
	loc.	dept.	doc.	date	etc.
Num.	440	383	270	233	2240

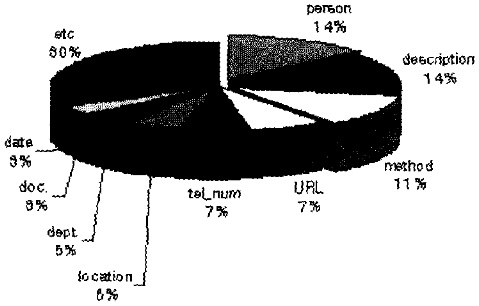

Fig 2 The distribution ratio of the semantic categories in AQUS

To experiment on the rule-based classifier, we manually constructed 580 lexico-syntactic patterns for the 78 semantic categories. The patterns are carefully generated to cover the categories. However, we excluded lexical-only patterns. We believe that it is impossible to build rules that cover all queries with lexical-level rules in real fields. For example, to classify the query like "*Who is the president of Sogang university?*", we did not construct the lexicon-only pattern like "*(Who is)? the president (of .+ university)? \??*".

For training and testing the statistical classifier, we divided AQUS into a test set with 772 queries (10% of the number of total queries) and a training set with 6,954 queries (90% of the number of total queries) according to distribution ratio of the semantic categories. Then, we performed 10-fold cross validation. The classifier generated 10,500 semantic patterns (1.51 semantic patterns per query) in training set and constructed 1,482 semantic patterns (1.92 semantic patterns per query) in the test set.

5.2 Analysis of Experiment Results

We evaluated four systems; baseline system, stand-alone rule-based system, stand-alone statistical system and hybrid system. The baseline system determines semantic categories according to specific interrogatives and semantic markers of focus words, as shown in Algorithm 3.

1. If a query includes a specific interrogative such as *who*, *when* and *where*, the baseline system classifies the query according to the interrogative.
2. If a query does not include any interrogatives, the baseline system classifies the query according to the semantic marker of the last focus word in the query. If a focus word has several semantic markers, the baseline system selects the first one.

Algorithm 3 The baseline system

For example, if a user asks "*yahu sajangeun nugujyo? (Who is the president of Yahoo?)*", the baseline system select *person*. If a user asks "*yahu sajangeun? (The president of Yahoo?)*", the baseline system checks the semantic marker of *sajang* (*president*) and classifies the query into *person* since the semantic marker is *@person*.

To evaluate the performances of the systems, we calculated the precision rate and the miss rate with the test set, as shown in Table 5. The miss rate is the ratio of the cases that a system fails to classify input queries because of insufficient linguistic knowledge like lexico-syntactic patterns. Only the rule-based classifier has miss rate because handcrafted patterns cannot cover all the users' queries. Precision-1 is the precision rate when missed queries are regarded as false classification. To practical QA systems, precision-1 is more meaningful than precision. As shown in Table 5, the hybrid system significantly surpasses the precision rate of the statistical system and eliminates the miss rate of the rule-based system. It is more difficult to construct lexico-syntactic patterns of explanation-seeking queries. Therefore, the rule-based system missed most of the explanation-seeking queries because of insufficient lexico-syntactic patterns. As a result, the hybrid system obtains high precision because the system chooses the results of the statistical system when the rule-based system missed queries. For example, when a query is "*DC moteoui gujo (The structure of DC motor)*", the rule-based system missed the query. However, the statistical system returned *description* that is the correct semantic category.

Table 5 The average precision on 10 fold cross validation

	Precision	Miss rate	Precision 1
Baseline system	0.56	0.00	0.56
Rule-based system	0.86	0.14	0.74
Statistical system	0.75	0.00	0.75
Hybrid system	0.80	0.00	0.80

6. Conclusion

We proposed a hybrid system that efficiently classifies users' queries into predefined semantic categories. The hybrid system combined two different sub-modules; the rule-based classifier and the statistical classifier. By adopting the rule based classifier, the hybrid system can easily add new rules and yield higher precision rate than the underlying statistical classifier. By adopting the statistical classifier, it can easily shift to other domains with good precision rate and reduce time for constructing handcrafted patterns. Furthermore, the hybrid system can guarantee robustness of question classification owing to the statistical classifier. In the future, we will study on finding new methods that combine the rule-based classifier with the statistical classifier. As one of the methods, we will consider machine learning methods such as a neural network, C4.5, and SVM.

References

[1] Clarke C. L. A., Cormack G. V., Kisman D. I. E. and Lynam T. R., "Question Answering by

Passage Selection (MultiText Experiments for TREC-9," In *Proceedings of the Ninth Text REtrieval Conference*, Gaithersburg, Maryland, 2000.

[2] Iittycheriah A., Franz M., Zhu W. and Ratnaparkhi A., "IBM's Statistical Question Answering System," In *Proceedings of the Ninth Text REtrieval Conference*, Gaithersburg, Maryland, 2000.

[3] Moldovan D., Harabagiu S., Pasca M., Mihalcea R., Goodrum R., Girju R. and Rus V., "LASSO-A tool for Surfing the Answer Net," In *Proceedings of the Eighth Text REtrieval Conference*, Gaithersburg, Maryland, 1999.

[4] Ferret O., Grau B., Illouz G., and Jacquemin C., "QALC the Question-Answering program of the Language and Cognition group at LIMSI CNRS," In *Proceedings of the Eighth Text REtrieval Conference*, Gaithersburg, Maryland, 1999.

[5] Hull D.A., "Xerox TREC-8 Question Answering Track Report," In *Proceedings of the Eighth Text REtrieval Conference*, Gaithersburg, Maryland, 1999.

[6] Prager J., Radev D., Brown E., and Coden A., "The Use of Predictive Annotation for Question Answering in TREC8," In *Proceedings of the Eighth Text REtrieval Conference*, Gaithersburg, Maryland, 1999.

[7] Srihari R., and Li W., "Information Extraction Supported Question Answering," In *Proceedings of the Eighth Text REtrieval Conference*, Gaithersburg, Maryland, 1999.

[8] Kupiec J., "MURAX: A Robust Linguistic Approach for Question Answering Using an On-line Encyclopedia," In *Proceedings of SIGIR'93*, 1993.

[9] Quinlan R. J., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.

[10] Ittycheriah A., Franz M., Zhu W. and Ratnaparkhi A., "Question Answering Using Maximum Entropy Components," In *Proceedings of NAACL*, 2001.

[11] Hermjakob U., "Parsing and Question Classification Classification for Question Answering," In *Proceedings of the ACL Workshop Open-Domain Question Answering*, pp. 17-22, 2001.

[12] Mann G. S., "A Statistical Method for Short Answer Extraction," In *Proceedings of the ACL Workshop Open-Domain Question Answering*, pp. 13-30, 2001.

[13] Miller G., *WordNet: An on line lexical database*, International Journal of Lexicography, Vol. 3(4), 1990.

[14] diquest, *http://www.diquest.com*.

[15] Kim H. and Seo J., "Automatic Extraction of a Syntactic Pattern for an Analysis of Speech Act: A Neural Network Model," In *Proceedings of ICONIP-2000*, Korea, 2000.

[16] Kim H., Cho J., and Seo J., "Fuzzy Trigram Model for Speech Act Analysis of Utterances in Dialogues," In *Proceedings of Conference on FUZZ-IEEE99*, Vol. 2, pp. 598-602, Seoul, Korea, 1999.

김 학 수

1996년 건국대학교 전자계산학과 학사. 1998년 서강대학교 컴퓨터학과 석사. 1998년~현재 서강대학교 컴퓨터학과 박사과정. 2001년~현재 다이퀘스트 연구소 선임연구원. 관심분야는 자연어 처리, 대화 시스템, 생략 및 대용어 처리, 화행 분석, 정보 검색, 질의 응답 시스템

안 영 훈

2000년 서강대학교 컴퓨터학과 학사. 2002년 서강대학교 컴퓨터학과 석사. 2000년~현재 다이퀘스트 연구소 선임연구원. 관심 분야는 자연어 처리, 질의 의도 분석, 정보 검색

서 정 연

1981년 서강대학교 수학과 학사. 1985년 미국 Univ. of Texas, Austin 전산학과 석사. 1990년 미국 Univ. of Texas, Austin 전산학과 박사. 1990년~1991년 미국 Texas Austin, UniSQL Inc. Senior Researcher. 1991년 한국과학기술원 인공지능 연구센터 선임연구원. 1991년~1995년 한국과학기술원 전산학과 조교수. 1995년~1996년 서강대학교 전산학과 조교수. 1996년~현재 서강대학교 컴퓨터학과 부교수. 관심분야는 한국어정보처리, 자연언어처리, 기계번역, 대화처리