Concept-based Question Answering System

Yu-Hwan Kang*, Seung-Eun Shin, Young-Min Ahn, Young-Hoon Seo
Department of Computer Engineering
Chungbuk National University, Cheongju, Korea

ABSTRACT

In this paper, we describe a concept-based question-answering system in which concept rather than keyword itself makes an important role on both question analysis and answer extraction. Our idea is that concepts occurred in same type of questions are similar, and if a question is analyzed according to those concepts then we can extract more accurate answer because we know the semantic role of each word or phrase in question. Concept frame is defined for each type of question, and it is composed of important concepts in that question type. Currently the number of question type is 79 including 34 types for person, 14 types for location, and so on. We experiment this concept-based approach about questions which require person's name as their answer. Experimental results show that our system has high accuracy in answer extraction. Also, this concept-based approach can be used in combination with conventional approaches.

***Keywords**: Concept-based QA system, Answer Extraction, Question-Answering System, IR System, Natural Language Processing*

1. INTRODUCTION

IR(Information retrieval) system offers relevant documents to user for a given word or question. But considerable work is often needed to find information that he/she really wants from great amount of documents extracted by the system.

QA(Question-answering) system is one that offers an answer instead of large volume of documents for user's various natural language questions[1]. Therefore, users can reduce time to find answer and be served a more user-friendly interface because QA system accepts an input sentence written in natural language as a question. Many researches about QA system has been centered upon AAAI[2] and TREC[3].

General QA system is composed of three components. Question analysis component extracts question type, answer type, keywords, and so on from question. Passage retrieval component searches documents relevant to question using conventional IR method and then extracts likely passage to contain answer in searched documents. Finally, answer extraction component extracts named entity related with answer type from that passage as answer.

The studies for question analysis mostly extract answer type corresponding to question type using named entities, answer type taxonomies, and ontologies such as WordNet[4,5,6,7,8,9]. The number of answer types varies widely from single digits to a few thousands. The subdivided classification of answer type helps to extract more accurate answer by reducing the number of answer candidates in phase of answer extraction.

The studies for answer extraction are largely divided into keyword approaches and language analysis approaches. Keyword approaches[4,5,7] use keywords' frequency and position in passages, an expansion of keywords, and so on. These approaches are efficient in speed, but it is difficult to find correct answer because only keywords and corresponding limited knowledge are used. Language analysis approaches[6,8,10,11] use natural language processing technologies such as syntax analysis. Approaches using linguistic knowledge can extract more accurate answer than keyword approaches.

Approaches using syntactic patterns for answer extraction have a limit to extract answer because there are many answers which can not be found by syntactic information obtained from question, and because syntactic patterns for answer extraction can hardly be derived. Therefore, we need semantic information as well as syntactic information for more accurate answer extraction.

In this paper, we propose a concept-based approach which uses semantic information to overcome above problems for question-answering system. We classify questions to 79 types currently, and define a concept frame which contains important concepts for that question type. For example, concept frame for "writer" question type contains concepts such as title of the book, nationality of the writer, time to write the book, and other

Corresponding author. E-mail: eric@ chungbuk.ac.kr

Manuscript received Sep. 28, 2005 ; accepted Mar. 13, 2006

This work was supported by the Korea Research Foundation Grant funded by Korean Government (MOEHRD).(D00607)

This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

concepts. In question analysis, words in question are analyzed and are assigned to appropriate concept slot in concept frame. This concept-based question analysis makes the system to know the role of each word in question. So, we can extract answers or documents containing answers more accurately using answer patterns with both syntactic information and concepts. Additionally, our approach can be applied to any language and can be combined to any conventional question-answering approaches.

The composition of this paper is same as following. Section 2 explains about our approach. Section 3 discusses about experiment of performance and result for proposed method. Finally, section 4 describes about conclusion and further works.

2. CONCEPT-BASED QUESTION ANSWERING SYSTEM

An overall architecture of our system is shown in Fig. 1. Our system is largely divided into three phases, question analysis, document search, and answer extraction. Input question is analyzed through morphological analysis and concept analysis. The result of question analysis is a concept frame which contains question type, answer type, concepts filled with keywords of question, and expanded keywords. Keywords in concept frame are presented to the IR engine for document retrieval. Answer extraction receives relevant documents extracted by general IR engine and extracts probable paragraphs containing answer through morphological analysis, concept analysis, and named entity recognition for them. And it extracts answer candidates from paragraphs, calculates score of each candidate, and presents up to 5 answers.

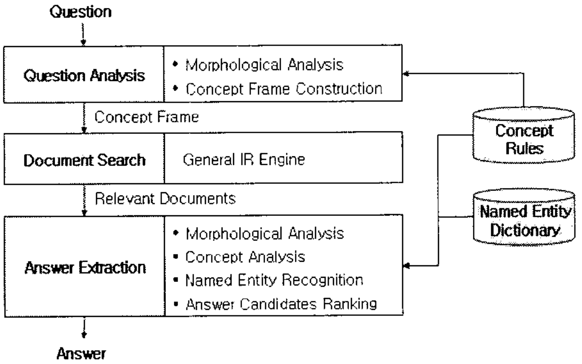


Fig. 1. The overall architecture of our system

2.1 Question Analysis

Question analysis phase determines question type and answer type of given question, and constructs concept frame for the question. We currently classified 79 question types including 34 types for person, 14 types for location, and so on. The answer type is almost similar with question type in our system because we classified question type in detail. Table 1 shows part of the question types.

Table 1. The lists of question type for each category

Category	Question type
Person	writer, politician, family, entertainer, athlete, developer, ...
Location	country, continent, ocean, river, mountain, ...
...	...

Concept frame is defined for each type of question, and is composed of concepts which can be occurred in common with that type of questions. When we use these concepts for question analysis and answer extraction we can make rules more precise and compact, so we can extract answers that are very difficult to extract by syntactic information only. For example, for the question "Who is the eldest son of Taisho?" we can extract correct answer from below paragraph if we know relation that Hirohito is eldest son of Taisho and Taisho is father of him and if we use concept-based rule.

> gwiguk hoo hirohitonun aberjee taishoga jeongsinjilhwaneuro euntoehaja seopjeongeuro immyongtoeeotda.(After coming home, Hirohito became a regent because his father Taisho retired with a mental disease

The followings show lists of concepts in concept frames for writer, politician, and family.

- •writer : title of book, nationality, genre, pen name, description about writer, description about book time, period
- •politician : duties, nationality, position, election, event(Event_X, Event_Y), qualification, time, period, adverb information
- •family : basis person, description about basis person, relation, description about relation, opposite relation

Table 2. An example of a question of writer type and result concept frame

Question	1991nyeon soseol 「gaemi」 rul chulgahan purangseui jakganun? (Who is the writer of France that publish a novel 「The ant」 in 1991?)	
Concept Frame	title of book	gaemi(The ant)
	Nationality	purangs(France)
	Time	1991nyeon, 91nyeon(year of 1991)

Table 3. An example of a question of politician type and result concept frame

Question	1884nyeon gabsinjeongbyeoneul ileukin chosunhukieui jeongchiganun? (Who is the politician of the latter term of Chosun who raises the gabsinjeongbyeon in 1884?)	
Concept Frame	Position	jeongchiga(politician)
	Event_X	Gabsinjeongbyeon
	Event_Y	ileuki, baksangsiki, yagiha, ... (raise an event)
	Period	chosunhuki, chosun huki, ...
	Time	1884nyeon, 84nyeon(the year of 1884)

In politician, Event_X and Event_Y are the object and the main verb of sentence respectively. Time means a particular point of minutes, days, years, and so on. Period means stage, interval of time, and so on. In family, basis person means person's name in question. And relation means relation between basis person and target person to be extracted as an answer. Table 2 and 3 shows the result of question analysis about questions of writer and politician.

Followings are detailed steps in question analysis.

- •Step 1 : extract keywords from input question.
- •Step 2 : expand keywords using synonym dictionary.
- •Step 3 : construct a concept frame using concept rules for question analysis.

We extract keywords from input question, remove stop words from them, and expand keywords using synonym dictionary. In case of compound noun, the module adds spacing information in order to match both compound noun itself and unit words consisting of it. In case of a year expression, the module expands it to all four digits form and two digits form. Finally, the module composes concept frame using concept rules for question analysis.

The concept rules for question analysis are constructed by using common concepts and syntactic information of questions of same type. The followings are examples of concept rules for question analysis for writer and politician.

•An example of concept rules for writer
- [title of book]+co+etm [genre]+jc <write>+etm [writer]+jx
- [title of book]+jc <write>+etm [writer]+jx

•An example of concept rules for politician
- [position]+jc <elect>+etm [position]+jx
- [Event_X]+jc [Event_Y]+etm [position]+jx

In above rules, components enclosed with '[' and ']' are concepts, and component enclosed with '<' and '>' is verb in question. 'co' is copular, 'etm' is adnominal ending, 'jx' is auxiliary postposition. 'jc' is case postposition which is further divided to jc_subj for subject case, jc_obj for object case, and so on.

2.2 Answer Extraction

Paragraph extraction phase includes morphological analyzer, concept analyzer, and named entity recognizer. Concept rules for answer extraction are used to extract paragraphs containing answers from relevant documents extracted by general IR system.

After morphological analysis for extracted documents, concept analyzer attaches concept tags to the same words that are in concept frame. Named entity recognizer finds named entities that correspond to answer type and attaches answer tag to them. The following shows the process of paragraph extraction.

Finally, concept rules for answer extraction are applied to each sentence in document. The followings are example of concept rules for answer extraction for writer, politician and family.

•Example of concept rules for answer extraction

for writer :
- [target person]+jc [*] [tile of book]+jc <write>
- [target person]+jx [title of book]+jm <synonym of writer>
- [target person]+jc [*] <write>+etm [title of book]

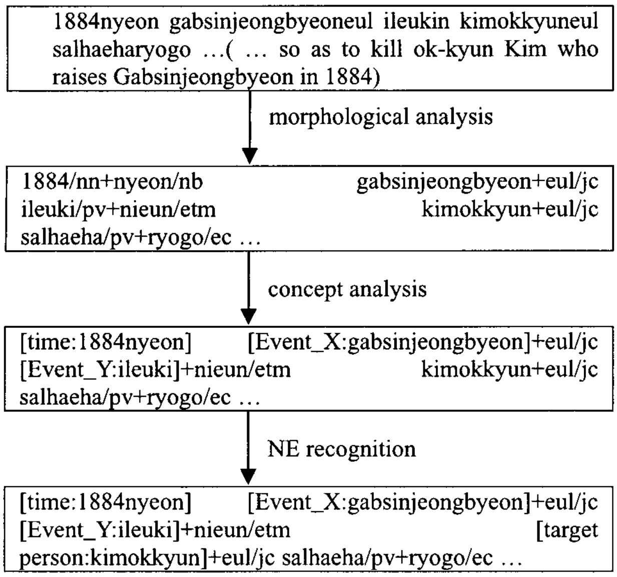
for politician :
- [time] [Event_X]+jc [Event_Y]+etm [target person]
- [order] [position] [L-] [*] [target person]+jc [*] <elect>

for family :
- [target person] [basis person]+[*]+jm [relation]+!etm
- [basis person]+[*]+jm [relation]+etm [target person]

The concept rule basically extracts a sentence as a paragraph. But a paragraph may have 5 sentences in maximum. '[L-]' or '[L+]' means that paragraph is extracted in several sentences. '[*]' means that zero or more words can be skipped over when the rule is applied to sentence. If any paragraph is not extracted by rules, then the system extracts answer paragraphs using keyword approach.

Answer candidates can be extracted directly from answer paragraphs because named entities that are related with answer type were already tagged. Answer candidates are ranked by differentiated weights against applied rule because each rule has different precedence.

We sorts answer candidates primarily as the precedence of applied rule and readjusts order using frequency of answer candidates.



1884nyeon gabsinjeongbyeoneul ileukin kimokkyuneul salhaeharyogo ...(... so as to kill ok-kyun Kim who raises Gabsinjeongbyeon in 1884)

1884/nn+nyeon/nb gabsinjeongbyeon+eul/jc ileuki/pv+nieun/etm kimokkyun+eul/jc salhaeha/pv+ryogo/ec ...

[time:1884nyeon] [Event_X:gabsinjeongbyeon]+eul/jc [Event_Y:ileuki]+nieun/etm kimokkyun+eul/jc salhaeha/pv+ryogo/ec ...

[time:1884nyeon] [Event_X:gabsinjeongbyeon]+eul/jc [Event_Y:ileuki]+nieun/etm [target person:kimokkyun]+eul/jc salhaeha/pv+ryogo/ec ...

Fig. 2. The process of document analysis for answer extraction

3. EXPERIMENT AND ANALYSIS

In experiment, we evaluated the number of correct answers that are ranked in top five about questions which require person's name as their answer. The number of questions used in experiment is 20. These are selected from our training set of questions. The types of question used in experiment are writer, politician, family, awardee, and entertainer. And we use common IR system and collect 15 documents for each question as the test documents. The following table shows the number of questions for each question type and the number of correct answers that are ranked in top five.

Table 4. The number of correct answers that are ranked in top five

Question Type	Number of Questions	Number of Correct Answers
writer	5	4.2
politician	4	3
family	3	3.7
awardee	4	2.8
entertainer	4	2.3
Total	20	3.2

The average number of correct answers among top 5 answer candidates was 3.2. In case of writer and family, many correct answers were ranked in high position because those question types contain many answers in documents and their syntactic patterns are somewhat regular. So, many answer paragraphs could be matched with concept rules for answer extraction. On the other hand, the rest types contain a few answers in documents relatively, and their syntactic patterns are so complex as not to describe by pattern. But, answers extracted by concept rules were mostly correct.

4. CONCLUSION AND FURTHER WORKS

In this paper, we proposed a concept-based approach for question-answering system. We classified questions into small groups corresponding to semantic similarity of their answers. We called the group to question type. A concept frame was defined for each type of question by concepts commonly occurred in that type of questions. Question corpus was used to define the concept frames and their concepts. The number of question type defined until now is 79, but it would be increased as progress of research.

We defined and used concept-based rules for both question analysis and answer extraction. Experimental results showed that correct answers were extracted by these concept-based rules even to paragraphs including answer which was so complex as not to derive any syntactic pattern. Other advantages of this concept-based approach are that it can be applied to any language and can be used together with any other conventional question answering approaches in way concept-based approach would be used only to defined concept frames and other approaches to undefined question types.

REFERENCES

[1] Ellen M. Voorhees, "The TREC question answering track", **Natural Language Engineering**, Vol. 7, 2001.

[2] AAAI Fall Symposium on Question Answering, http://www.aaai.org/Press/Reports/Symposia/Fall/fs-99-02.html

[3] TREC(Text Retrieval Conference) Overview. http://trec,nist.gov/overview.html

[4] S. Abney, M. Colins, A. Singhal, "Answer Extraction", **In 6th Applied Natural Language Processing Conference**, 2000.

[5] A. Ittycheriah, M. Franz, W. Zhu, A. Ratnaparkhi, "IBM's Statistical Question Answering System", **In 9th Text Retrieval Conference**, 2000, pp.229-334.

[6] S. Haragagiu, M. Pasca, S. Maiorano, "Experiments with open-domain with open-domain textual question answering", **In COLLING-2000**, 2000, pp.292-298.

[7] G.G Lee, J. Seo, S. Lee, H. Jung, B. Cho, C. Lee, B. Kwak, J. Cha, D. Kim, J. Ann, H. Kim, K. Kim, "SiteQ:Engineering High Performance QA System Using Lexico-Semantic Pattern Matching and Shallow NLP", **In 10th Text Retrieval Conference**, 2001, pp.437-446.

[8] C. Cardie, "Examining the Role of Statistical and Linguistic Knowledge Sources in a General-Knowledge Question-Answering System", **In 6th Applied Natural Language Processing Conference**, 2000, pp.180-187.

[9] M. Pasca and S. Harabagui, "High Performance Question / Answer", **In 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**, 2001, pp.366-374.

[10] S. Buchholz, W. Daelemans, "Complex Answers: a case study using a WWW question answering system", **Natural Language Engineering**, Vol. 7, 2001, pp.301-323.

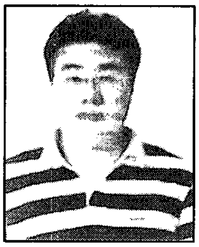
[11] Valdo Kesel, "Question Answering using Unification-based Grammar", **Advanced in Artificial Intelligence, AI 2001**, Vol. LNAI 2056 of Lecture Notes in Computer Science, 2001.

Yu-Hwan Kang

He received the B.E. and M.E. degrees in computer engineering from Chungbuk National University in 1998 and 2000, respectively. Currently, he is a Ph.D. candidate in computer engineering at Chungbuk National University. His research interests include machine translation, information retrieval and natural language processing.

Seung-Eun Shin

He received the B.E. and M.E. and ph.D degrees in computer engineering from Chungbuk National University in 1999, 2001 and 2006,1 respectively. His research interests include information retrieval and natural language processing.



Young-Min Ahn

He received the B.E. and M.E. degrees in computer engineering from Chungbuk National University in 2000 and 2002, respectively. Currently, he is a Ph.D. candidate in computer engineering at Chungbuk National University. His research interests include morphological analysis, 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.