K-QARD: A Practical Korean Question Answering Framework for Restricted Domain

Young-In Song, HooJung Chung, Kyoung-Soo Han, JooYoung Lee, Hae-Chang Rim
Dept. of Computer Science & Engineering
Korea University
Seongbuk-gu, Seoul 136-701, Korea
{song, hjchung, kshan, jylee rim}@nlp.korea.ac.kr

Jae-Won Lee
Computing Lab.
Samsung Advanced Institute of Technology
Nongseo-ri, Giheung-eup,
Yongin-si, Gyeonggi-do 449-712, Korea
jwonlee@samsung.com

Abstract

We present a Korean question answering framework for restricted domains, called K-QARD. K-QARD is developed to achieve domain portability and robustness, and the framework is successfully applied to build question answering systems for several domains.

1 Introduction

K-QARD is a framework for implementing a fully automated question answering system including the Web information extraction (IE). The goal of the framework is to provide a practical environment for the restricted domain question answering (QA) system with the following requirements:

- **Domain portability**: Domain adaptation of QA systems based on the framework should be possible with minimum human efforts.
- **Robustness**: The framework has to provide methodologies to ensure robust performance for various expressions of a question.

For the domain portability, K-QARD is designed as a domain-independent architecture and it keeps all domain-dependent elements in external resources. In addition, the framework tries to employ various techniques for reducing the human effort, such as simplifying rules based on linguistic information and machine learning approaches.

Our effort for the robustness is focused the question analysis. Instead of using a technique for deep understanding of the question, the question analysis component of K-QARD tries to extract only essential information for answering using the information extraction technique with linguistic information. Such approach is helpful for not only the robustness but also the domain portability because it generally requires smaller size of hand-crafted rules than a complex semantic grammar.

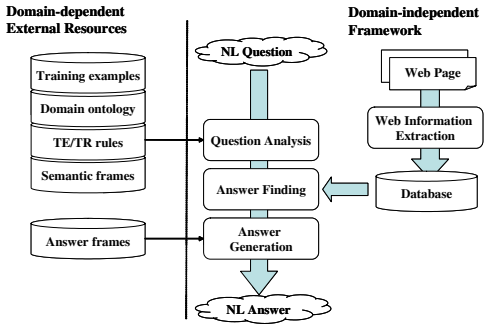


Figure 1: Architecture of K-QARD

K-QARD uses the structural information automatically extracted from Web pages which include domain-specific information for question answering. It has the disavantage that the coverage of QA system is limited, but it can simplify the question answering process with robust performance.

2 Architecture of K-QARD

As shown in Figure 1, K-QARD has four major components: Web information extraction, question analysis, answer finding, and answer generation.

The Web information extraction (IE) component extracts the domain-specific information for question answering from Web pages and stores the information into the relational database. For the domain portability, the Web IE component is based on the automatic wrapper induction approach which can be learned from small size of training examples.

The question analysis component analyzes an

Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pages 29–32,
Sydney, July 2006.

input question, extracts important information using the IE approach, and matches the question with pre-defined semantic frames. The component outputs the best-matched frame whose slots are filled with the information extracted from the question.

In the answer finding component, K-QARD retrieves the answers from the database using the SQL generation script defined in each semantic frame. The SQL script dynamically generates SQL using the values of the frame slots.

The answer generation component provides the answer to the user as a natural language sentence or a table by using the generation rules and the answer frames which consist of canned texts.

3 Question Analysis

The key component for ensuring the robustness and domain portability is the question analysis because it naturally requires many domain-dependent resources and has responsibility to solve the problem caused by various ways of expressing a question. In K-QARD, a question is analyzed using the methods devised by the information extraction approach. This IE-based question analysis method consists of several steps:

1. **Natural language analysis**: Analyzing the syntactic structure of the user's question and also identifiying named-entities and some important words, such as domain-specific predicate or terms.

2. **Question focus recognition**: Finding the intention of the user's question using the question focus classifier. It is learned from the training examples based on decision tree(C4.5)(Quinlan, 1993).

3. **Template Element(TE) recognition**: Finding important concept for filling the slots of the semantic frame, namely template elements, using the rules, NE information, and ontology, etc.

4. **Template Relation(TR) recognition**: Finding the relation between TEs and a question focus based on TR rules, and syntactic information, etc.

Finally, the question analysis component selects the proper frame for the question and fills proper values of each slot of the selected frame.

Compared to other question analysis methods such as the complex semantic grammar(Martin et al., 1996), our approach has several advantages. First, it shows robust performance for the variation of a question because IE-based approach does not require the understanding of the entire sentence. It is sufficient to identify and process only the important concepts. Second, it also enhances the portability of the QA systems. This method is based on the divide-and-conquer strategy and uses only limited context information. By virture of these characteristics, the question analysis can be processed by using a small number of simple rules.

In the following subsections, we will describe each component of our question analyzer in K-QARD.

3.1 Natural language analysis

The natural language analyzer in K-QARD identifies morphemes, tags part-of-speeches to them, and analyzes dependency relations between the morphemes. A stochastic part-of-speech tagger and dependency parser(Chung and Rim, 2004) for the Korean language are trained on a general domain corpus and are used for the analyzer. Then, several domain-specific named entities, such as a TV program name, and general named entities, such as a date, in the question are recognized using our dictionary and pattern-based named entity tagger(Lee et al., 2004). Finally some important words, such as domain-specific predicates, terminologies or interrogatives, are replaced by the proper concept names in the ontology. The manually constructed ontology includes two different types of information: domain-specific and general domain words.

The role of this analyzer is to analyze user's question and transform it to the more generalized representation form. So, the task of the question focus recognition and the TE/TR recognition can be simplified because of the generalized linguistic information without decreasing the performance of the question analyzer.

One of possible defects of using such linguistic information is the loss of the robustness caused by the error of the NLP components. However, our IE-based approach for question analysis uses the very restricted and essential contextual information in each step and can avoid such a risk successfully.

The example of the analysis process of this

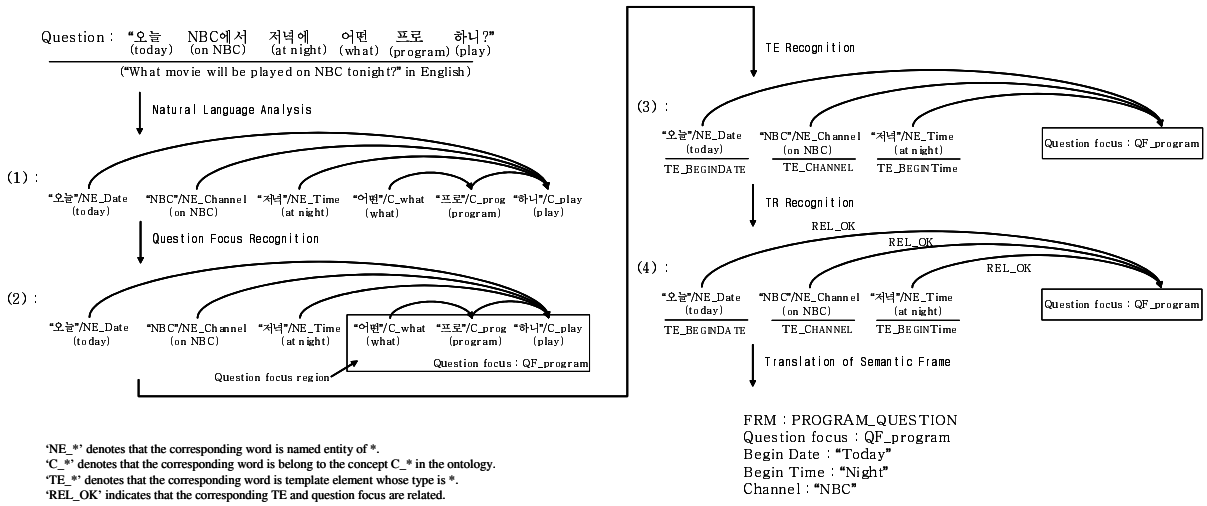


Figure 2: Example of Question Analysis Process in K-QARD

component is shown in Figure 2-(1).

3.2 Question focus recognition

We define a question focus as a type of information that a user wants to know. For example, in the question *"What movies will be shown on TV tonight?"*, the question focus is a program title, or titles. For another example, the question focus is a current rainfall in a question *"San Francisco is raining now, is it raining in Los Angeles too?"*.

To find the question focus, we define *question focus region*, a part of a question that may contain clues for deciding the question focus. The question focus region is identified with a set of simple rules which consider the characteristic of the Korean interrogatives. Generally, the question focus region has a fixed pattern that is typically used in interrogative questions(Akiba et al., 2002). Thus a small number of simple rules is enough to cover the most of question focus region pattern. Figure 2-(2) shows the part recognized as a question focus region in the sample question.

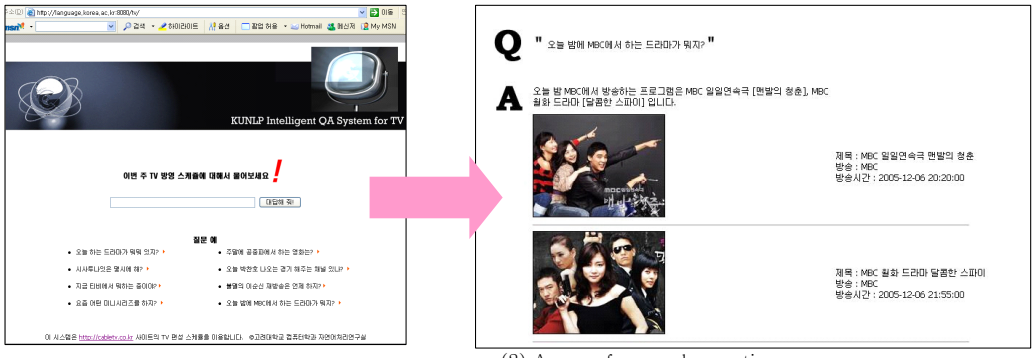
After recognizing the region, the actual focus of the question is determined with features extracted from the question focus region. For the detection, we build the question focus classifier using decision tree (C4.5) and several linguistic or domain-specific features such as the kind of the interrogative and the concept name of the predicate.

Dividing the focus recognition process into two parts helps to increase domain portability. While the second part of deciding the actual question focus is domain-dependent because every domain-application has its own set of question foci, the first part that recognizes the question focus region is domain-independent.

3.3 TE recognition

In the TE identification phase, pre-defined words, phrases, and named entities are identified as slot-filler candidates for appropriate slots, according to TE tagging rules. For instance, *movie* and *NBC* are tagged as Genre and Channel in the sample question *"Tell me the movie on NBC tonight."* (i.e. *movie* will be used to fill Genre slot and *NBC* will be used to fill Channel slot in a semantic frame). The hand-crafted TE tagging rules basically consider the surface form and the concept name (derived from domain ontologies) of a target word. The context surrounding the target word or word dependency information is also considered in some cases. In the example question of Figure 2, the date expression *'오늘(today)'*, time expression *'저녁(night)'* and the channel name *'MBC'* are selected as TE candidates.

In K-QARD, such identification is accomplished by a set of simple rules, which only examines the semantic type of each constituent word in the question, except the words in the question region. It is mainly because of our divide-and-conquer strategy motivated by IE. The result of this component may include some wrong template elements, which do not have any relation to the user's intention or the question focus. However, they are expected to be removed in the next component, the TR recognizer which examines the relation between the recognized TE and the question focus.



(1) Broadcast−domain QA system

(2) Answer for sample question, "What soap opera will be played on MBC tonight?"

Figure 3: Broadcast-domain QA System using K-QARD

3.4 TR recognition

In the TR recognition phase, all entities identified in the TE recognition phase are examined whether they have any relationships with the question focus region of the question. For example, in the question *"Is it raining in Los Angeles like in San Francisco?"*, both *Los Angeles* and *San Francisco* are identified as a TE. However, by the TR recognition, only *Los Angeles* is identified as a related entity with the question focus region.

Selectional restriction and dependency relations between TEs are mainly considered in TR tagging rules. Thus, the TR rules can be quite simplified. For example, many relations between the TEs and the question region can be simply identified by examining whether there is a syntactic dependency between them as shown in Figure 2-(4). Moreover, to make up for the errors in dependency parsing, lexico-semantic patterns are also encoded in the TR tagging rules.

4 Application of K-QARD

To evaluate the K-QARD framework, we built restricted domain question answering systems for the several domains: *weather, broadcast,* and *traffic*. For the adaptation of QA system to each domain, we rewrote the domain ontology consisting of about 150 concepts, about 30 TE/TR rules, and 7-23 semantic frames and answer templates. In addition, we learned the question focus classifier from training examples of about 100 questions for the each domain. All information for the question answering was automatically extracted using the Web IE module of K-QARD, which was also learned from training examples consisting of several annotated Web pages of the target Web site. It took about a half of week for two graduate students who clearly understood the framework to build each QA system. Figure 3 shows an example of QA system applied to the broadcast domain.

5 Conclusion

In this paper, we described the Korean question answering framework, namely K-QARD, for restricted domains. Specifically, this framework is designed to enhance the robustness and domain portability. To achieve this goal, we use the IE-based question analyzer using the generalized information acquired by several NLP components. We also showed the usability of K-QARD by successfully applying the framework to several domains.

References

T. Akiba, K. Itou, A. Fujii, and T Ishikawa. 2002. Towards speech-driven question answering: Experiments using the NTCIR-3 question answering collection. In *Proceedings of the Third NTCIR Workshop*.

H. Chung and H. Rim. 2004. Unlexicalized dependency parser for variable word order languages based on local contextual pattern. *Lecture Note in Computer Science*, (2945):112–123.

J. Lee, Y. Song, S. Kim, H. Chung, and H. Rim. 2004. Title recognition using lexical pattern and entity dictionary. In *Proceedings of the 1st Asia Information Retrieval Symposium (AIRS2004)*, pages 345–348.

P. Martin, F. Crabbe, S. Adams, E. Baatz, and N. Yankelovich. 1996. Speechacts: a spoken language framework. *IEEE Computer*, 7(29):33–40.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.