Information Processing and Management 50 (2014) 683–692

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Open domain question answering using Wikipedia-based knowledge model

Pum-Mo Ryu *, Myung-Gil Jang, Hyun-Ki Kim

Electronics and Telecommunications Research Institute, 138 Gajeongno, Yuseong-gu, Daejeon 305-700, Republic of Korea

ARTICLE INFO

Article history:
Received 8 September 2012
Received in revised form 25 April 2014
Accepted 28 April 2014
Available online 6 June 2014

Keywords:
Question-answering
Wikipedia
Semi-structured knowledge

ABSTRACT

This paper describes the use of Wikipedia as a rich knowledge source for a question answering (QA) system. We suggest multiple answer matching modules based on different types of semi-structured knowledge sources of Wikipedia, including article content, infoboxes, article structure, category structure, and definitions. These semi-structured knowledge sources each have their unique strengths in finding answers for specific question types, such as infoboxes for factoid questions, category structure for list questions, and definitions for descriptive questions. The answers extracted from multiple modules are merged using an answer merging strategy that reflects the specialized nature of the answer matching modules. Through an experiment, our system showed promising results, with a precision of 87.1%, a recall of 52.7%, and an F-measure of 65.6%, all of which are much higher than the results of a simple text analysis based system.

1. Introduction

The goal of a question answering (QA) system is to directly return answers, rather than documents containing answers, in response to a natural language question. The answers can be fact-based short answers, lists of instances, or descriptions about a particular topic. Many of the initial efforts in QA research, ignited by the QA track in TREC (Dang, Kelly & Lin, 2007; Voorhees, 2004), have focused on mining unstructured texts such as news sites and blogs. However, these systems show a relatively low performance, with at most 71% accuracy for factoid questions, 48% F-score for list questions, and 33% F-score for descriptive questions. On the other hand, specialized QA systems have relied on well-structured knowledge bases in specific domains (Demner-Fushman & Lin, 2007; Frank et al., 2007). Although these works achieved high accuracy, building large-scale, well-structured knowledge bases for a general domain QA is a very expensive task.

Wikipedia is a semi-structured and wide covering, rapidly growing knowledge source that has been built through a collaborative effort of volunteers. Wikipedia has become a stable and sufficiently large knowledge source for many knowledge-based engineering works (Bizer et al., 2009; Hoffart et al., 2013; Nastase & Strube, 2008; Suchanek et al., 2007). Furthermore, Wikipedia was applied to QA systems as a knowledge base (Ahn et al., 2004; Buscaldi & Rosso, 2006; Simmons, 2012). However, these systems utilized only parts of Wikipedia information. Thus, we developed an open-domain QA system that fully utilizes semi-structured Wikipedia knowledge model. The knowledge model can serve as certified information sources and enable a QA system to generate correct answers in a general domain. We exploit the category structure, article structure, infoboxes, definitions, redirection, and article contents of Wikipedia as knowledge sources for

* Corresponding author. Tel.: +82 42 860 5327; fax: +82 42 860 4889.
E-mail addresses: pmryu@etri.re.kr (P.-M. Ryu), mgjang@etri.re.kr (M.-G. Jang), hkk@etri.re.kr (H.-K. Kim).

http://dx.doi.org/10.1016/j.ipm.2014.04.007
0306-4573/

a QA system. We assume each knowledge source has its own strengths for answering different types of answer formats such as factoid, list, and description. For example, an infobox is effective in answering factoid questions, and the category structure is effective in answering lists of questions.

Well-organized knowledge bases do not guarantee high performance if the questions are in natural language instead of formal query. Mapping linguistic expression in questions to knowledge representation in knowledge-base is another hard task if we build full-featured knowledge from Wikipedia like YAGO (Hoffart et al., 2013). F-scores of QA systems in QALD task are about 50%, which are much lower than expected result (http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/). The task covers extracting answers from well-organized knowledge-bases for given natural language questions. So, our system is based on a conventional QA system, which consists of a question analysis module, document retrieval module, and answer matching module. To this end, the different types of knowledge sources are converted into text documents for the document retrieval module. Instead, specialized answer matching modules are developed for the knowledge types.

Section 2 describes the question analysis, while Section 3 describes our Wikipedia QA system. Section 4 describes the experiment used, and concluding remarks are given in Section 5.

2. Question analysis

To make full use of knowledge sources of Wikipedia for many types of questions, it is critical to analyze user questions in terms of the nature of the answers being sought. The availability of a question categorization scheme will help not only in analyzing an incoming user question but also in identifying QA capabilities and techniques to be developed in the future (Oh et al., 2011). To this end, we collected 600 questions from a commercial Korean website, Naver™ Manual QA Service (http://kin.naver.com). The questions were analyzed to characterize the types of questions and answers. The results of our analysis are shown in Table 1 for surface-level question type classes determined based on interrogative pronouns. These results are further divided into answer formats corresponding to the classes used in TREC (Voorhees, 2004). While TREC used a "definitional" type, we generalized it as a "descriptive" type, which includes "definitional," "reasons," and "methods" types (Oh et al., 2009).

A user question in natural language form is analyzed using multiple linguistic analysis techniques including POS tagging, chunking, and named entity tagging (Lee et al., 2006; Lee & Jang, 2011). The analyzed result of a question has three components including answer format (AF), answer theme (AT) and question target (QT). The AF has three possible values: factoid, list, and descriptive. They can be distinguished based on the surface-level description of questions. For example, "Where is the Nile River located" looks for a single factoid answer, whereas "Who are American politicians who have emigrated from Austria" requires a list of answers. A descriptive question needs an answer that contains definitional, causal, or method information about a key term, as in "What is X," "What is the cause of X," or "What is the method for X?". An AT is the class of the object or description sought by the question, such as PERSON, LOCATION, and DATE for a factoid; list answer format; and DEFINITION, REASON, and METHOD for a descriptive answer format. We used a total of 147 answer themes, which are organized into a hierarchical structure (Lee et al., 2006). A QT consists of two parts: object and property. The former is the main object or event that the question is about, whereas the latter is the property of interest that a question attempts to get at regarding the object. In "Where is the Nile River located," for example, the object is "Nile River" and the property is "be located." The key elements in detecting the question target are the predicate-argument structure or noun phrase structure in the dependency structure of the given question. When the property is not clear, it can remain empty.

Given a question q, we want to find a question analysis result $r = (af, at, qt)$ which most likely explains what the question means as follows;

$$\hat{r} \leftarrow \arg\max_{r} S_Q(r|q)$$

$$S_Q(r|q) = S_{AF}(af|q) \cdot S_{AT}(at|q) \cdot S_{QT}(qt|q)$$

Table 1
Distribution of surface-level question types.

Surface level	Answer format	# Questions	Example
Who	Factoid	44	Who is the president of Korea?
	List	50	Who are American politicians emigrated from Austria?
	Descriptive (definition)	72	Who is Barack Obama?
What/Which	Factoid	87	What is the name of the oldest aircraft?
	List	71	Who are an American politicians emigrated from Austria?
	Descriptive (definition)	84	What is the movie based on Facebook?
Where	Factoid	62	Where is the Nile river located?
When	Factoid	28	When did World War I end?
Why	Descriptive (reasons)	52	Why do typhoons occur?
How	Descriptive (methods)	50	How to fix a flat tire?
Total		600	