What Kind of Answers in Web Open Knowledge Resources?: for Quiz Questions in Education

Hyo-Jung Oh[1] and Yong Kim[2]

[1]*Graduate School of Archives and Records Management,*
Institute of Culture Convergence Archiving, Chonbuk National University
567 Baekje-daero, Deokjin-gu, Jeonju, Jeollabuk-do, 561-756, KOREA
First Author: ohj@jbnu.ac.kr
[2]*Dept. of Library and Information Science, Chonbuk National University*
567 Baekje-daero, Deokjin-gu, Jeonju, Jeollabuk-do, 561-756, KOREA
Corresponding Author: yk9118@jbnu.ac.kr

Abstract

Automatic Question Answering (QA) system aims an appropriate answer for a natural language question. There are several attempts to apply QA techniques to education field in terms of tutoring system. The ultimate goal of this paper is to discover reliable answer resources readily available knowledge in natural language within broad domains. To evaluate the usage of web open knowledge resources for QA in terms of educational purpose, we choose 'quiz' game as the target domain of television show. Based on experimental results of our QA system, we classified questions according to their clause structures and determined their complexity. We also analyzed correlation between question classes and answer knowledge resources.

1. Introduction

Question Answering (QA), asking for curious things, is one of the typical activities in education. Toward mimicking this human behavior, many research to develop automatic QA systems have been progressed. In 2011, a great challenging event related with QA was announced. A machine, called WATSON, compete against two humans in the game Jeopardy – which is a very famous quiz show in America. The result, WATSON had won and it is a very remarkable result in the computer science area. WATSON is an artificially intelligent computer system capable of answering questions posed in natural language developed in IBM's DeepQA project [1]. The sources of information for WATSON include encyclopedias, dictionaries, thesauri, newswire articles, and literary works. They also leveraged other kinds of semi-structured and structured content. They tried to identify and collect the answer evidence resources, which include databases, taxonomies, and ontologies, such as dbPedia [2] WordNet [3], and the YAGO [4] ontology.

In this paper, we discover reliable answer resources readily available knowledge in natural language within broad domains. In particular, we give attention to encyclopedic contents as sources to contain candidate answers and try to apply automatic QA system for educational applications.

2. Related Researches

2.1. Automatic Question Answering in Education

Automatic QA system aims an appropriate answer for a natural language question [5]. QA system is attractive as it is one of the most challenging in exploring new ways for computer science technology to affect education, business, and society [6]. We believe QA system ultimately can be a useful interface or instrument in tasks like acquiring knowledge or training skills.

Under the same premise, there are several attempts to apply QA techniques to education field in terms of tutoring system. Atlas-Andes [7] provides fifty-five separate directed lines of reasoning that teach the basic conceptual rules targeting all aspects of Newtonian Mechanics. AutoTutor [8] simulates a human tutor by holding a conversation with the learner in natural language. It has been applied to conceptual physics and presents a series of challenging questions that require approximately a paragraph of information to answer correctly. Recently Moré et al. [9] presented a system that helps tutors to answer questions asked by their students at an online university.

While these researches are dialog-based tutoring systems holding a conversation with the learner in natural language, they are restricted to narrow-specific domains such as physics or mechanics. Furthermore, manually constructed answers are likely to be limited by their poor adaptability and low coverage, even though they guarantee high accuracy. In e-learning area, Liu and Jia [10] designed an innovative method for answering students' or scholars' academic questions by automatically recommending e-learning resources in a cyber-infrastructure-enabled learning environment to enhance the learning experiences of students and scholars. Based on examining preliminary evaluation results and student feedback, they found that cyber learning resources can provide high-quality and straightforward answers for students' and scholars' questions concerning the content of academic publications.

2.2. Web Open Knowledge Resources

QA systems have proven to be helpful to users because they can provide succinct answers that do not require users to wade through a large number of documents. Trust is also an important area in question answering, since contradictory answers can be obtained from diverse sources in an answer to a question.

Lin [11] conducted a user study to investigate the effective interface design in the QA area and discovered that users prefer paragraph-sized chunks of text as the answer to their questions and the reliability of source is important. The user required much context to accept an answer from a less trustworthy source than from a more trustworthy one and accepted nearly 98% of the answers for trusted sources. In other contexts, detecting opinions is useful when no single ground truth can be provided in answer to a question, and instead multiple perspectives are summarized as the answer provided [12]. Lin [13] also evaluated the quality of answer patterns and lists of relevant documents currently employed in automatic question answering evaluation, and concluded that they are not suitable for post-hoc experimentation.

There are various knowledge resources in which we can use to find answers, for example, web documents, electronic newspaper, and encyclopedic contents such as Wikipedia [14]. In encyclopedia, there is a lot of recorded information like Guinness Book. The Web, while nearly infinite in content, is not a complete repository of useful information [15]. Contrary to some web documents which contain indefinite information in terms of uncertainty of web, encyclopedic contents contain the facts which were already approved or occurred in the past. It is true that encyclopedia provides much descriptive information and it appears in particular

forms. As the result of analyzing the log from Korean web search engine, most questions can find the answer in an encyclopedia except questions from gossips or scandals about entertainers [16].

In this paper, we will try to distinguish characteristics of web open knowledge resources corresponding to question and answer classes. We'll also classify questions according to their clause structures and reveal correlation between questions and answer sources.

3. Automatic Question Answering System for Quiz

ETRI has developed an automatic QA system, named *WiseQA* [17]. Figure 1 illustrates system overview. WiseQA used highly sophisticated linguistic analysis and utilized a lot of web open knowledge resources collected from bigdata. To answer given questions, WiseQA generates answer candidate hypotheses and verifies optimal answers among them based on continues self-learning framework.

WiseQA advocates "generate & type" approach, which means that the system analyzes expected answer type from incoming questions, mapping it into a fixed set of known types, and restricts answer candidates retrieved from the corpus to those that match this answer type. It brings several advantages. First, more flexible as QA system does not break down if type is not detected or meaningful. Second, much wider type coverage is possible using variety of sources and methods for type inference. We built WiseWordNet, a linked resource for type inference, to overcome limitation of available resources.
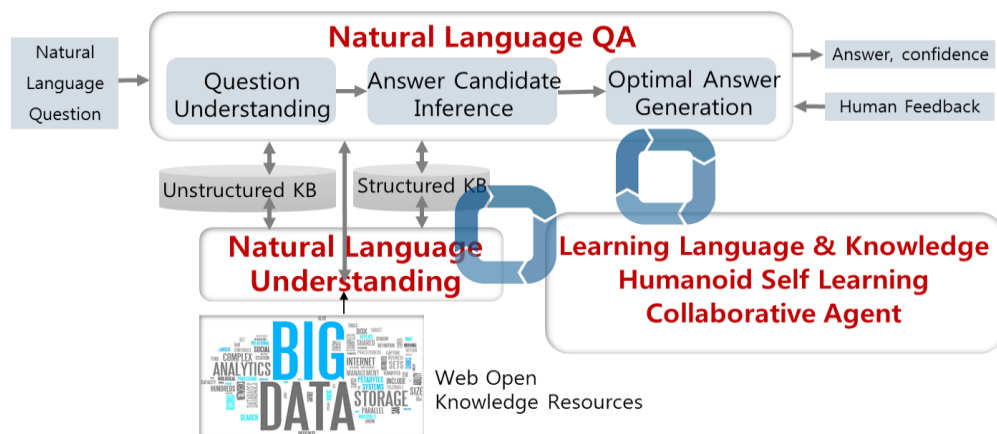

Figure 1. System Overview of WiseQA

3. Characteristics Analysis of Quiz Questions

3.1. Question Classes in Quiz

To analyze the usage of web open knowledge resources for QA in terms of educational purpose, we choose 'quiz' game as the target domain of television show. The show, originally called "Janghak Quiz" in Korea, has been seen as a respectable thing for a Korean high school student to participate in. It is a popular show among high schoolers and was aired by the broadcaster MBC in 1973, but was taken over by EBS in 1997. So far, 1,700 episodes have been taped and up to 11,000 people have participated as contestants. For the detailed survey, we collected over 1,000 questions in this show from May to July in 2014. Among them, we selected 586 questions because we focused on the 'short answers' or factoid questions, where the correct response is a single word or a short phrase from the answer sentence.

Table 1. Question Class Distribution of the Quiz Show

Question Class		# of questions		Examples
Definition		151	25.8%	*All proteins are made of _____ acids.* (Answer: amino)
Factoid		395	67.4%	*Where was Napoleon I born?* (Answer: Corsica in France)
Linguistic Knowledge	Four character idioms	10	1.7%	*The act to Crack down on cornering the market.* (Answer: 買占賣惜 매점매석)
	Origins	21	3.6%	*The unit of frequency originated from German physicist's name.* (Answer: Hertz)
	Proverbs	6	1.0%	*There is a saying in English, 'to lose by a ____.'.* (Answer: Hair)
	Synonym/ Dialects	3	0.5%	*What is meaning of 'MoolKuruck 물구럭' in Jeju dialect?* (Answer: Octopus)
All		586	100%	

Table 1 is distribution of question classes. As expected, the majority of questions (93.2%) are factoid questions and definitional questions (67.4% and 25.8% respectively). Factoid question has a typical form of "What is *X*?" like "*Sakura cheese from Hokkaido is a soft cheese flavored with leaves from this fruit tree. What is it*? (Answer: cherry)". Especially most of definitional questions in the quiz show form 'FITB(fill-in-the-blank)' style which find appropriate title words given the question text. For example, "*All proteins are made of _____ acids*". The answer of this question in blank is 'amino'.

The others questions (6.8%) are related with linguistic knowledge and they are likely to find potential answers from our dictionaries. For an example question, "*The unit of frequency originated from German physicist's name*" is looking for the source of scientific terminology, "Hertz". Some questions require four character idioms, such as "Mejummesuk: 買占賣惜(매점매석)" means "*The act to Crack down on cornering the market*". Some questions related local brogue because the Korean language has many dialects and their pronunciations are very different.

3.2. Question Complexity in Quiz

We distinguished questions according to their clause structures: simple, compound, complex, and other mixed sentences. A simple sentence consists of only one clause likes "*This French protestant group was inspired by John Calvin in the 1530s*" whereas a compound sentence consists of two or more independent clauses. For example, "*Where a person is born has invented penicillin?*" consists two sentences "*Where the person is born?*" and "*He was born when penicillin was invented.*" A complex sentence has at least one independent clause plus at least one dependent clause.

Meanwhile, many questions in quiz consist of several sentences which indicated the same answer. The competitions should find common candidates for multiple questions. The example in last column of Table 2 is a typical sample question in Jeopardy! Show. Both sentences, "*It's largest airport is named for a world war II hero.*" and "*It's second largest, for a world war II battle*" are looking for the answer, "Toronto".

Table 2. Question Complexity Distribution of the Quiz Show

Question Complexity	# of questions		Examples
Simple sentence	96	16.4%	*This French protestant group was inspired by John Calvin in the 1530s* (Answer: Huguenot)
Compound sentence	14	2.4%	*Where was penicillin invented?* (Answer: Scotland Ayrshire)
Complex sentence	138	23.5%	*This movie is based on the chaos theory, the sensitive dependence on initial conditions in which a small change in one state of a deterministic nonlinear system can result in large differences in a later state* (Answer: Butterfly effect)
Mixed	338	57.7%	*It's largest airport is named for a world war II hero; It's second largest, for a world war II battle* (Answer: Toronto)
All	586	100%	

Table 2 is complexity distribution of quiz questions. As mentioned before, the largest portion (57.7%) is mixed sentences because of nature of the quiz domain. Complex sentence questions take charge of 23.5%. Simple and compound sentence questions are only 16.4% and 2.4%, respectively. The later section will discuss answer characteristics according to quiz question types.

4. Answers in Web Open Knowledge Resources for Quiz Questions

4.1. Web Open Knowledge Resources for Quiz

WiseQA consists of four major open knowledge resources in Korean [18]: Wikipedia(kr)[1], Korean standard unabridged dictionary[2], Doosan[tm] EnCyber[3], and Korean proverb dictionary[4]. Table 3 describes the characteristics of our knowledge resources. Wikipedia is the biggest free encyclopedia that anyone can edit and currently contains 4,899,771 articles in English whereas Korean Wikipedia contains 318,913 articles. Figure 2 illustrates a sample page of Wikipedia titled "Bruno Mars". The main pages in Wikipedia consist of contents and various template including infobox and images.

Table 3. Korean Web Open Knowledge Resources in WiseQA

Resources	# of entries	Specification
Korean Wikipedia	318,913	Main Pages, section and subsections, infobox, images, *etc*.
Korean standard unabridged dictionary	509,076 (Nouns: 335,057, Verbs: 68,394)	Head words, definition, idioms, and examples.
Doosan[tm] EnCyber	351,440	Main Pages, infobox, in/out links, category, etc.
Korean Proverb Dictionary	Proverbs: 8,877	Title Head word, Proverbs,

---

[1] https://ko.wikipedia.org/wiki
[2] http://stdweb2.korean.go.kr/main.jsp
[3] http://www.doopedia.co.kr/, or http://dic.naver.com/
[4] http://www.natmal.com/

| | Examples: 171,448 | Examples, and four character idioms |

Figure 2. An Example of Wikepedia Contents (headword: Bruno Mars)

Korean standard unabridged dictionary has been built by NIKL (the National Institute of the Korean Language). It consists of 335,057 and 68,394 definitions for nouns and verbs respectively. Doosan[tm] Encyclopedia is provided via Naver[tm] portal services. It consists of 351,440 articles and their page structures are similar with Wikipedia. Korean Proverb Dictionary has been developed by Natmal© and contains 8,877 proverbs and 171,488 word example s including four character idioms.

4.2. Answer Source Distribution for Quiz Questions

We tagged sources in which answers were firstly founded for all questions by manual. Figure 3 depicts the distribution of answers from our web knowledge resources. Among 586 questions, 110 questions (18.8%) fail to find an answer within our target resources. Predictably, the most important source of answers is revealed: the first ranked is Korean Wikipedia (59.6%) and the second ranked resource is Doosan EnCyber (13.5%).
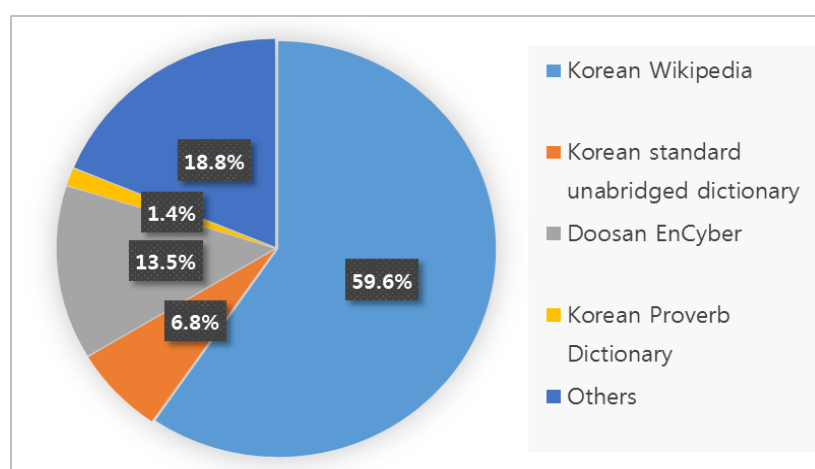
Figure 3. Answer Sources for All Questions

Figure 4 show the distribution of 476 (586-110) correct answers automatically submitted by WiseQA as a Venn diagram. A large portion of answers (23+58+3+1=85%) are founded in mutually more than two resources. Even 58% of answers were extracted from major three resources. It means our web open resources are enough to contain potential answers for quiz questions.

To analyze correlation between answer sources and question classes, we detected sources of answers for each question. Table 4 shows the distribution of answer sources corresponding of question classes. As same with Figure 3, major portion of correct answer sources is Korean Wikipedia. Especially answers for factoid questions can be found in Wikipedia. It is an interesting point is that the distribution of Doosan Encyber is same with Korean Wikipedia. As expected, answers for definitional question can be easily found in Korean Standard unabridged dictionary. The answers for questions about proverbs understandably can be produced by Korean Proverb Dictionary.
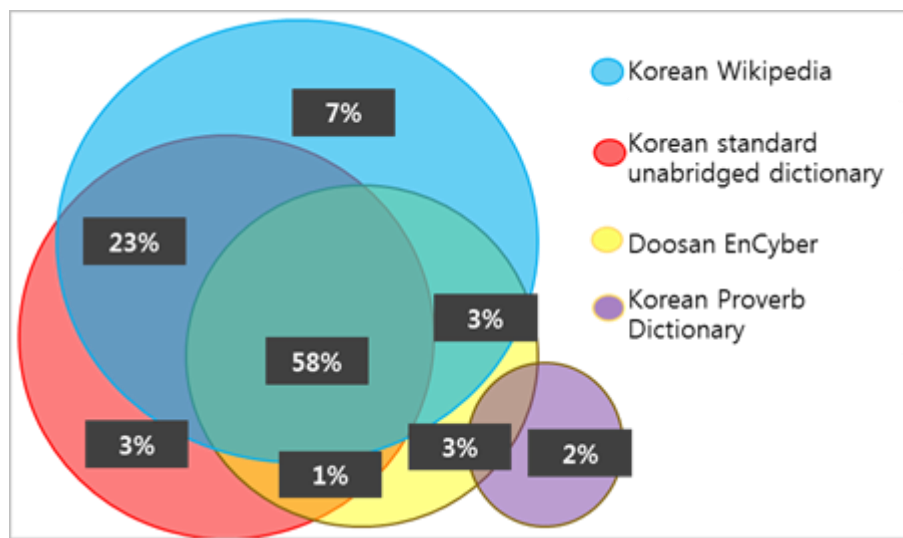

Figure 4. Answer Source Distribution

Table 4. Answer Source Distribution According to Question Classes

Question Class / Answer sources		Korean Wikipedia	Korean Standard unabridged dictionary	Doosan Encyber	Korean Proverb Dictionary
Definition		59	24	23	0
Factoid		273	9	51	0
Linguistic Knowledge	Four character idioms	3	4	1	3
	Origins	13	2	3	0
	Proverbs	0	0	1	5
	Synonym/ Dialects	1	1	0	0
Sum		349	40	*79*	*8*
All		*476*			

5. Conclusions

In this paper, we built automatic question answering system based on web resources for education purpose. We revealed the characteristics of four major Korean web open knowledge resources. We also defined question classes based on real questions collected from real TV quiz show and experimental results were used to verify our assumptions. In the future, expansion of various other resources, such as YAGO or WordNet, and improvement of accuracy of our automatic QA system will be attempted.

Acknowledgements

References

[1] D. A. Ferrucci, A. Levas, S. Bagchi, D. Gondek and E. T. Mueller, "Watson: beyond jeopardy!", Artificial Intelligence, vol. 199-200, (2013), pp. 93-105.
[2] "DBpedia Knowledge Base", http://wiki.dbpedia.org/about.
[3] G. A. Miller, "WordNet: A Lexical Database for English", Communications of the ACM, vol. 38, no. 11, (1995), pp. 39-41.
[4] "YAGO: A High-Quality Knowledge Base", www.mpi-inf.mpg.de/yago-naga/yago.
[5] E. M. Voorhees, "Overview of TREC 2003 QA Track", Proceedings of the TREC-12, NIST, (2003), pp. 1-13.
[6] D. A. Ferrucci, E. Brown and J. Chu-Carroll, "Building Watson: An Overview of the DeepQA project," AI magazine, Association for the Advancement of Artificial Intelligence, (2010), pp. 59-79.
[7] C. Rosé, P. Jordan, M. Ringenberg, S. Siler and A. Weinstein, "Interactive Conceptual Tutoring in Atlas-Andes," Proceedings of the 10th Artificial Intelligent in Education Conference, (2001), pp. 1-9.
[8] A. Graesser, P. Chipman, B. Haynes and A. Olney, "AutoTutor: An intelligent tutoring system with mixed-initiative dialogue," IEEE Transactions on Education, vol. 48, no. 4, (2005), pp. 612-618.
[9] J. Moré, S. Climent and M. Coll-Florit, "An Answering System for Questions Asked by Students in an e-Learning Context", [online article], Universities and Knowledge Society Journal (RUSC), vol. 9, no. 2, (2012), pp. 229-239.
[10] X. Liu and H. Jia, "Answering academic questions for education by recommending cyberlearning resources", Journal of the American Society for Information Science and Technology, vol. 64, no. 8, (2013), pp. 1707-1722.
[11] J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, D., B. Katz, and D. R. Karger, "What Makes a Good Answer? The Role of Context in Question Answering", Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction, (2003), pp. 25-32.
[12] C. Cardie, J. Wiebe, J., T. Wilson and D. Litman, "Low-Level Annotations and Summary Representations of Opinions for Multi-perspective Question Answering", New Directions in Question Answering, AAAI Press/MIT Press, (2004).
[13] J. Lin, "Evaluation of Resources for Question Answering Evaluation", Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, New York, (2005).
[14] B. Katz, M. Bilotti, S. Felshin, A. Fernandes, W. Hildebrandt, R. Katzir and Ö. Uzuner, "Answering Multiple Questions on a Topic from Heterogeneous Resources", Proceedings of the 13th Text REtreival Conference (TREC-13), (2004).
[15] W. Hilderbrandt, B. Katz and J. Lin, "Answering Definition Question using Multiple Knowledge Sources", Proceedings of HLT/NAACL, (2004), pp. 49-56.
[16] H. J. Oh, K. Y. Sung and S. H. Myaeng, "Compositional Question Answering: a Divide and Conquer Approach", Information Processing & Management, vol. 47, issue 6, (2011), pp. 808-824.
[17] P. M. Ryu, "Natural Language Question Understanding for WiseQA", Proceedings of the 2nd international conf. on BIGComp, Workshop, vol. 2, (2015).
[18] H. J. Oh and Y. Kim, "Web Open Resources for Automatic Question Answering System", Advanced Science and Technology Letters, vol. 103, (2015), pp. 49-52.

Authors

Hyo-Jung Oh received the BS and the MS degrees in computer science from Chungnam National University in 1998 and 2000, respectively. She received the PhD degree in engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2008. During 2000 – 2015, she was a principle researcher in Electronics and Telecommunications Research Institute (ETRI). Currently she is an assistant professor of Chonbuk National University. Her research interests include machine learning, question answering, and knowledge mining.

Yong Kim received the MS degrees in information science from University of North Texas, USA in 1996 and the MS degrees in computer science from Chungnam National University in 2000. He received the PhD degree in information science from Yonsei University in 2006. During 1996 – 2008, he was a principle researcher in KT R&D center. Currently he is an associate professor of Chonbuk National University. His research interests include digital library, data mining and records archiving system.