Question Answering Technology for Pinpointing Answers to a Wide Range of Questions

Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, and Nozomi Kobayashi

Abstract

Question answering technology that provides pinpoint answers to a wide range of questions is expected to make speech interfaces more intelligent. This article describes question answering technology and reviews past and current approaches to it at NTT.

1. Introduction

The sheer number of documents that have been created on the Internet makes it impossible for one to read them all. Therefore, technologies that help us find pertinent information efficiently are becoming more and more important. One such technology is the web search engine. Web search engines search for documents in response to keywords provided by users, present the documents, and thereby facilitate our information access. Although web search engines are useful, they only return documents; that is, users still need to read through the returned documents for the information they need. It is easy to imagine that the amount of information we handle will increase at a pace faster than ever before, making it too time-consuming to even look through the returned documents. There is therefore an imminent need for technologies that make our information access more efficient. In this article, we describe question answering technology, which represents a major step forward in meeting this need. We also describe NTT's past and current approaches to question answering technologies.

2. Question answering technology

Question answering technology provides pinpoint answers to natural language questions. Systems based on this technology are called *question answering systems*. Users can obtain pinpoint answers to their questions just by asking the system; answers are presented immediately, and there is no need to read any documents. Question answering systems deal with a wide variety of questions. There are two types of systems depending on the type of questions they answer. One is a factoid question answering system, which answers factual questions by using words or short phrases. The other is a non-factoid question answering system, which provides answers in sentence or paragraph form. For example, the former system answers *Everest* to *What is the highest mountain in the world?* and *George Washington* to *Who was the first president of the United States?* The latter, for example, provides sentential answers to definition questions (e.g., *What is optical fiber?*), why-questions (e.g., *Why is the sky blue?*), and how-questions (e.g., *How do I cook delicious dumplings?*). If all questions that could be asked by users were known in advance, it would be possible to prepare the answers in advance and provide the correct answers when needed. However, in reality, user questions are too diverse to predict. Therefore, question answering systems mimic how humans find answers; that is, they interpret a question, search for relevant documents, and find answers.

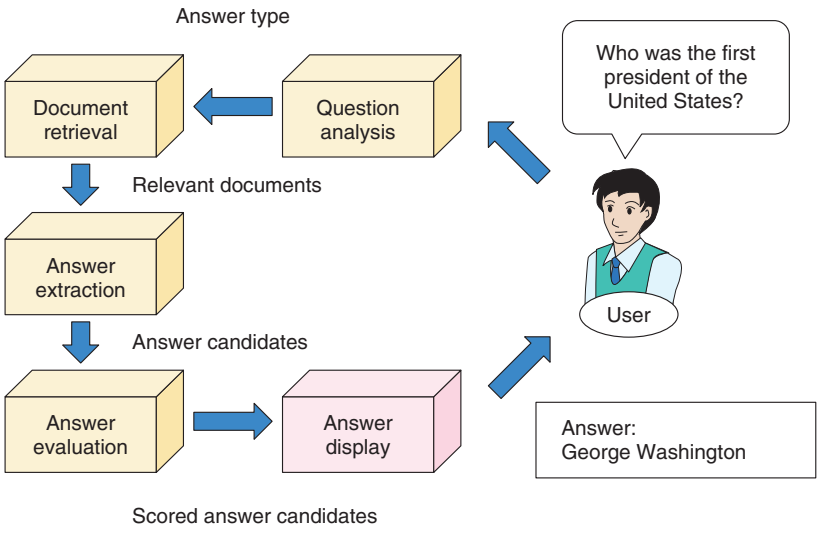


Fig. 1. General architecture of a question answering system.

3. Architecture of a question answering system

The general architecture of a question answering system is shown in **Fig. 1**. It has four modules: question analysis, document retrieval, answer extraction, and answer evaluation [1]. Note that although this article deals with a factoid question answering system, the architecture is almost the same for a non-factoid question answering system.

First, the question analysis module analyzes a question and determines its answer type. An answer type represents the type of information requested by a question: a person's name, place name, and a numerical expression are some of the possible answer types. For *Who was the first president of the United States?*, the answer type is *person* (person's name). When the granularity of answer types becomes fine-grained, it is possible to grasp the requested information with pinpoint precision, although there is a trade-off between the number of answer types and the accuracy of automatic answer-type classification. The information retrieval module uses a search engine to retrieve relevant documents on the basis of keywords in the question. Since the question answering system searches for answers only in those retrieved documents, the accuracy of document retrieval is very important. The answer extraction module extracts from the retrieved documents all answer candidates matching the answer type. When the answer type is *person*, all person names in the retrieved documents are extracted. Named entity recognition technology [2] is used for this extraction. Finally, the answer evaluation module evaluates the appropriateness of candidate answers by using such information as how they appear in the documents, and assigns scores to the candidate answers. Finally, highly scoring candidate answers are presented to the user.

4. Question answering systems at NTT

Research on question answering began around 1999. Around that time, an evaluation workshop on question answering was held at the Text Retrieval Conference (TREC), and researchers from all over the world started competing with one another to create question answering systems. At around the same time, NTT also started researching question answering systems and since then has developed a number of systems.

The first question answering system developed at NTT was SAIQA (System for Advanced Interactive Question Answering) in 2001 [3]. SAIQA was a factoid question answering system that obtained accurate answers thanks to its early use of machine learning techniques for answer-type classification and named entity recognition. Machine learning is a process of statistically learning criteria for judgment from a large amount of training data. In 2004, SAIQA achieved the best performance at the evaluation workshop Question Answering Challenge (QAC) in Japan. SAIQA used a database of newspaper articles for document retrieval.

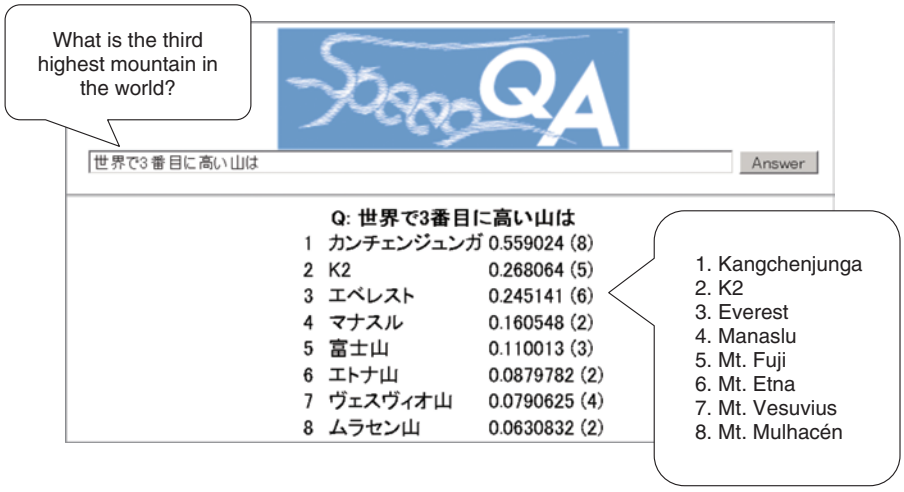


Fig. 2. Screenshot of SpeedQA. The numbers indicate scores of candidate answers and their frequencies (in parentheses) in retrieved documents.

In 2004, Web Answers was developed [1]. This system, as its name implies, retrieves documents from the web. The system was made public on the Internet so that anyone could use it. It answered not only factoid questions but also definition and reputation questions (e.g., *What is the reputation of X?*), which made it popular among users for its wide coverage. In 2007, we developed a non-factoid question answering system called NAZEQA in order to provide answers to why-questions [4]. Conventional approaches to why-questions had used cue words such as *node* and *tame* (corresponding to *because* in English) to find answer sentences. However, it was also pointed out in the literature that relying only on such cue words resulted in limited coverage of answers. Therefore, we statistically mined causal expressions from a large number of documents and used them to detect answer sentences, which achieved accurate answers to why-questions. In 2012, we developed SpeedQA, our most recent system. A screenshot is shown in **Fig. 2**. This system is the culmination of our experience in researching question answering systems at NTT. For example, it uses machine learning in almost all of its modules. It uses the Internet for document retrieval and can answer factoid as well as non-factoid questions. Non-factoid questions it can deal with are definition, reputation, and also why-questions. This system, minus some of its functions, has been integrated into the knowledge Q&A service of NTT DOCOMO's Shabette-Concier* voice-agent service [5].

5. Recent advances: Answer-type classification and timeliness detection in SpeedQA

5.1 Answer-type classification

In developing SpeedQA, we concentrated in particular on answer-type classification because we wanted to provide pinpoint answers. The answer analysis module of SpeedQA first determines whether a question type is factoid or non-factoid. Then, for a factoid question, it further classifies its answer type by referring to a taxonomy of over 100 answer types. Even with machine learning, classification into such fine-grained answer types is not an easy task. Therefore, we used a large-scale Japanese thesaurus created at NTT [6] and put a special focus on noun suffixes and counter suffixes. In this way, we succeeded in finding useful information for answer-type classification and achieved high accuracy.

5.2 Timeliness detection

To find pinpoint answers, in addition to accurately classifying answer types, it is also important to recognize user intentions, that is, to determine what users really want to know. Consider the question *Who won the gold medal?* On the surface, this question looks like an easy one for which the system can simply present past gold medalists. However, if it is posed during the Olympic Games, it would be reasonable to present only the gold medalists in the current

* Shabette-Concier is a trademark of NTT DOCOMO Inc.

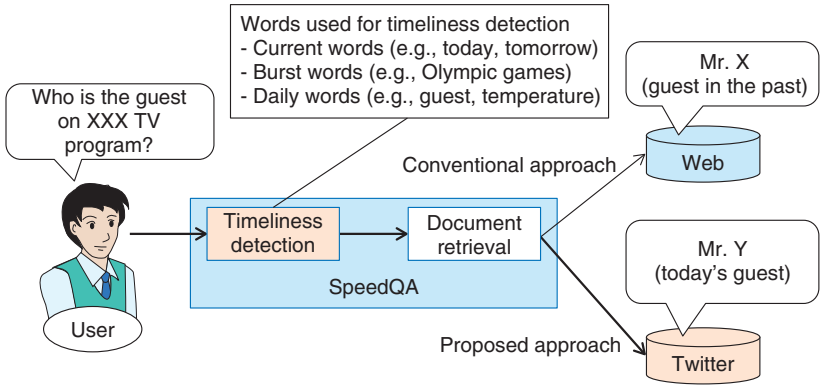


Fig. 3. Flow of timeliness detection.

Olympics. As shown by this example, there are questions whose answers vary depending on when they are posed, and they need to be treated accordingly to provide pinpoint answers. To this end, we developed a method of timeliness detection that detects whether a question is asking about timely events or not. If it is detected as being timely, the document retrieval module is switched from a web search engine to a real-time search engine (a search engine for Twitter) for timely information. The flow of timeliness detection is shown in **Fig. 3**. When questions contain time-related words such as *today* or *now* (called *current words*), it is easy to detect their timeliness. When questions contain *burst words* (words whose occurrence frequencies have shown a sudden increase in a short time span), it is also reasonable to determine that they are timely questions. The problem is when a question does not contain such words. We collected and analyzed many questions that asked for timely information and discovered that words such as *guest*, *starting player*, *game*, and *temperature*, whose referents (entities being referred to) vary day by day, are used frequently. We named such words *daily words* and developed a technique for automatically acquiring them. By using this technique, we were able to obtain many daily words, and it became possible to successfully detect the timeliness of questions containing such words.

6. Conclusion

This article described question answering technology and past and current approaches to this technology at NTT. We acknowledge that our latest system SpeedQA still has a lot of room for improvement. We plan to improve our algorithms further to achieve more pinpoint answers.

References

[1] H. Isozaki, R. Higashinaka, M. Nagata, and T. Kato, "Question Answering Systems," Corona Publishing Co. Ltd., 2009 (in Japanese).
[2] K. Sadamitsu, R. Higashinaka, T. Hirano, and T. Izumi, "Knowledge Extraction from Text for Intelligent Responses," NTT Technical Review, Vol. 11, No. 7, 2013.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa3.html
[3] E. Maeda, H. Isozaki, Y. Sasaki, H. Kazawa, T. Hirao, and J. Suzuki, "Question answering system: SAIQA—A "Learned Computer" that answers any questions," NTT R&D, Vol. 52, No. 2, pp. 122–133, 2003 (in Japanese).
[4] R. Higashinaka and H. Isozaki, "Automatically acquiring causal expression patterns from relation-annotated corpora to improve question answering for why-questions," ACM Transactions on Asian Language Information Processing, Vol. 7, No. 2, 2008.
[5] W. Uchida, C. Morita, and T. Yoshimura, "Knowledge Q&A: Direct Answers to Natural Questions," NTT DOCOMO Technical Journal, Vol. 14, No. 4, pp. 4–9, 2013.
http://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/rd/technical_journal/new/vol14_4_004en.pdf
[6] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi, "Goi-Taikei—A Japanese Lexicon," Iwanami Shoten, 1997.

Ryuichiro Higashinaka

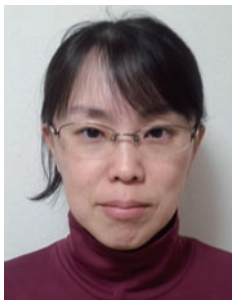
Senior Research Scientist, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.A. degree in environmental information, the Master of Media and Governance, and the Ph.D. degree from Keio University, Kanagawa, in 1999, 2001, and 2008, respectively. He joined NTT in 2001. His research interests include building question answering systems and spoken dialogue systems. From Nov. 2004 to Mar. 2006, he was a visiting researcher at the University of Sheffield in the UK. From 2006 to 2008, he was a part-time lecturer at Osaka Electro-Communication University. Since 2010, he has been a part-time lecturer at Keio University. He is a member of the Japanese Society for Artificial Intelligence, the Information Processing Society of Japan (IPSJ), and the Association for Natural Language Processing (NLP).

Kugatsu Sadamitsu

Researcher, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

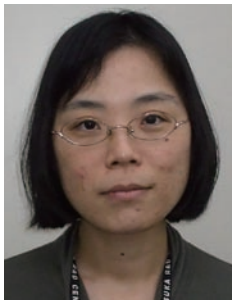
He received the B.E., M.E., and Ph.D. degrees in engineering from Tsukuba University, Ibaraki, in 2004, 2006, and 2009, respectively. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2009. His current research interests include natural language processing and machine learning. He is a member of IPSJ and NLP.



Kuniko Saito

Senior Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

She received the B.S. and M.S. degrees in chemistry from the University of Tokyo in 1996 and 1998, respectively. She joined NTT Information and Communication Systems Laboratories in 1998 and then moved to NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories). Her research focuses on part-of-speech tagging, named entity recognition, and term extraction. She is a member of IPSJ and NLP.



Nozomi Kobayashi

Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

She received the M.E. and Dr.Eng. degrees in information science from Nara Institute of Science and Technology in 2004 and 2007, respectively. She joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2007. Her current research interests include information extraction. She is a member of IPSJ and NLP.