# 5-2 Speech-based Question Answering System "Ikkyu"

**DE SAEGER Stijn, GOTO Jun, and VARGA István**

This paper introduces the speech-based question answering system "Ikkyu" developed by the NICT Information Analysis Laboratory. Ikkyu is a next-generation information system that caters to users' various information needs, in the form of spoken natural language questions posed via smartphone. Ranging from causes of the Japanese deflation to preventive measures for strokes, Ikkyu exhaustively covers answers explicitly contained in our 600 million page Japanese Web archive, and furthermore is able to generate answer hypotheses that are not written explicitly but can be derived by combining seemingly unrelated information obtained from distinct documents. This system aims to provide a new search platform that enhances human decision making abilities by providing pinpoint information and relevant suggestions to questions people ask on a whim, thereby broadening their awareness of the various options available to them.

***Keywords***
Question answering, Knowledge acquisition, Big data, NLP, Speech recognition

## 1 Introduction

In the current situation where the volume of information on the Internet is growing exponentially, it has become evident that information search models that provide users with a large amount of documents that match search keywords have reached their limits. Since many users only check Web pages that appear higher up in search engine rankings, it cannot be denied that knowledge, or decisions made from that knowledge, can be biased by search engines. In this situation, it depends on chance which information can be found. Based on this, we broadly feel that it would be extremely difficult for current search engines to provide the information necessary for appropriate decision-making.

Conscious of this problem, we have developed a speech-based question answering system "Ikkyu" to find answers from the Web. Ikkyu performs semantic analysis of a Japanese Web archive of 600 million documents. To respond to a variety of questions given in text or speech without restricting the domain of questions, Ikkyu looks for answers in semantically analyzed Web pages and presents answers to meet various users' information needs. Ikkyu can exhaustively handle answers that are explicitly contained in the 600 million Web documents, for example from the causes of Japanese deflation to preventive measures for cerebral infarction. It can also generate, as a hypothesis, answers that are not explicitly contained in the Web documents by combining apparently different pieces of information. Ikkyu has the following characteristics that existing search engines lack.

1. Reduction of search errors

Ikkyu considers various paraphrases of the user's question and finds answers through its advanced semantic analysis of a large volume of Web documents. Therefore, unlike simple keyword matching, it can exhaustively find information that has the same content but is

written in different expressions.

2. Answers not explicitly written on the Web are provided by inference as hypotheses

At present, all useful world knowledge is not explicitly represented on the Web. By combining fragmented information on the Web, Ikkyu can create, as a hypothesis, new knowledge not explicitly written on the Web, and provide this as answers to users.

3. Answer presentation with an emphasis on easy discovery

To prevent overlooking information in the answer ranking, Ikkyu presents answers not in a list, which cannot be easily scanned, but in a word-cloud form. (See the right figure in Fig. 3.) As will be explained in Subsection **2.1**, the distance of an answer from the center of the word cloud indicates the certainty of the answer. The distance from other answers shows their semantic similarity, and answers that have similar meanings are displayed close to one another.

If we have a sudden random idea and give it to Ikkyu as input, we would be able to find unexpected but useful knowledge from a large amount of information beyond any one individual's grasp. An example is the question "What causes deflation?" Ikkyu may give not only common-sense answers such as "restructuring" or "imported products" but also unexpected ones. For example, Ikkyu provided a major company's name as a cause of Japanese deflation. The basis for the answer is given in Fig. 1. It may look unreliable as it was extracted from a blog, but Nikkei Newspaper published an article with similar content after we found the answer. This means that the answer was to some extent acknowledged by society. The information source from which the answer was derived and the question that we gave to Ikkyu are apparently quite different from each other as they have no common words besides "deflation". It is therefore difficult to reach this answer just by performing keyword matching. (The mechanism by which Ikkyu can find this kind of answer will be described in Section **2**). As this example indicates, exhaustively finding unknown but useful answers is extremely important. It can increase options for user's thoughts and actions and can even support appropriate decision-making from a broader perspective. This is exactly the final goal of the Ikkyu project.

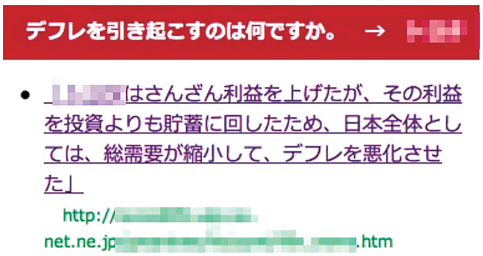



**Fig.1** *Example: the original sentence from which Ikkyu extracted an unexpected but valuable answer regarding the cause of the Japanese deflation*

This paper is organized as follows. In Section **2**, the core technology of this system is introduced. Approaches to respond to a larger variety of questions are described in Section **3**. Section **4** is devoted to the introduction of speech input and related technologies in Ikkyu. In Section **5**, we clarify the position of Ikkyu in the research field of question answering technologies, which have attracted attention in recent years. Conclusions are given in Section **6**.

## 2 Core technology of Ikkyu

In this section we introduce the core technology of Ikkyu. Ikkyu regards the question answering process as a relation extraction problem and solves it by a real-time adaptation of the semantic relation acquisition method proposed in [1] and [2]. For example, to the question "What is Paris famous for?" Ikkyu finds a semantic relation among the nouns, expressed in a language pattern "X is famous for Y." It then finds, as an answer, a noun that is in the same semantic relation with the noun "Paris." We describe a detailed flow of the

question answering process in the following.

### 2.1 Question answering algorithm and processing flow

The question answering algorithm of Ikkyu, shown in Fig. 2, consists of the following steps. A more detailed explanation of steps 3 and 4, which are of technical importance for the present method, is given in Subsection **2.2**.

1. Speech recognition of questions

A question in the form of text or speech is given to Ikkyu. Questions given from smart phones are converted to text using a speech recognition module specialized for question sentences. The speech recognition module will be introduced in Section **4**.

2. Extraction of lexico-syntactic pattern

Next, a rule-based syntactic transformations turn the question sentence into an affirmative sentence. Then, a syntactic analysis of this affirmative sentence is performed, and the lexico-syntactic pattern indicating the semantic relation between the words is extracted from the syntactic tree. The lexico-syntactic pattern consists of the words that lie on the path of dependency relations between two words in the syntactic tree. For example, the question in Fig. 2, "What can we fish in Kawazu river?" is converted first to "what can be fished in Kawazu river?" and then to an affirmative sentence, "what can be fished in Kawazu river." From the sentence, a lexico-syntactic pattern such as "In X, Y can be fished." or "Y can be fished in X." is extracted (X = Kawazu river, Y = what). In what follows, the lexico-syntactic pattern that is extracted from a question sentence is called a "query pattern."

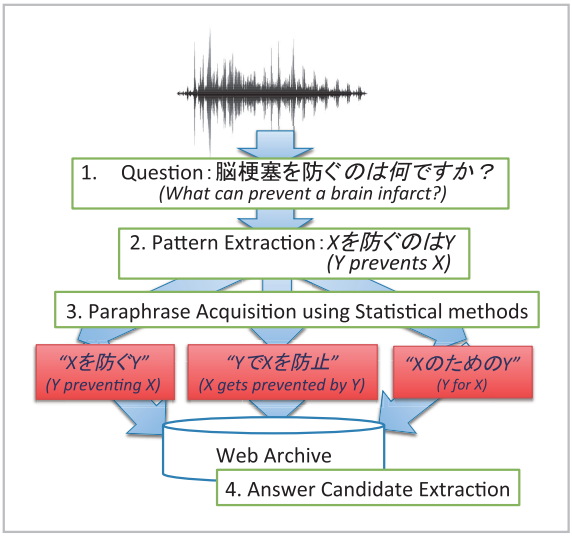



**Fig.2** *Overview of the question answering algorithm*

3. Acquisition of paraphrase patterns

The next step is an expansion process of the query pattern extracted from the question sentence. This step is a key process to exhaustively recognize possible answers on the Web. The query pattern is first expanded through basic syntactic transformations. Then lexico-syntactic patterns that can be considered paraphrases of the query pattern extracted from the question sentence are automatically obtained based on the context similarity between lexico-syntactic patterns, which is calculated from the statistical data acquired from large Web archives. The query pattern is thus expanded into dozens to hundreds of paraphrase patterns. These paraphrase patterns are sometimes superficially quite different from the query patterns. For example, as extended patterns of a query pattern, "X causes Y," paraphrase patterns such as "X is the cause of Y", "X induces Y", "Y by X", and "X brings about Y" can be acquired.

4. Extraction of potential answers

The above pattern set is used to extract potential answers from a Web corpus, which are then ranked and shown to users. The ranking of the potential answers considers a combination of various sub-scores such as the semantic class of the words (described in Subsection **2.3**), the paraphrase score (which presents the semantic similarity of the query pattern and the lexico-syntactic pattern from which the potential answer is obtained), and the relevance between the potential answer and the pattern from which the answer is acquired [1][2]. Figure 3 gives an example of how these an-


**Fig.3** *Ikkyu's answer display (left: smartphone, right: PC browser)*

swers are shown in a web browser or on a smart phone. In a browser (right of Fig. 3), answers are presented not in a list, which cannot be viewed at a glance, but in a word cloud form ("cloud" of answers) in order to find useful answers easily. This display form has the following two characteristics. The distance of an answer from the center of the word cloud indicates its relative score. Highly reliable answers are displayed in the vicinity of the center. The distance between the answers on the screen represents the semantic similarity between them. Words that are semantically similar are shown close to each other by the word cloud's display algorithm. On smart phones, on the other hand, answers are displayed in a list, as the phone screens are too small to show answers in such an advanced display style (left in Fig. 3).

## 2.2 Automatic paraphrase acquisition

In this section we introduce the paraphrase pattern recognition algorithm of Ikkyu. One of the features of Ikkyu is that it can recognize various paraphrase expressions of users' question when it extracts potential answers from a Web corpus. Ikkyu can exhaustively extract answers found in paraphrased expressions that are superficially quite different from the original question and thus cannot be acquired by simple keyword matching. The automatic acquisition of these paraphrases is based on our semantic relation acquirement method proposed in [1] and [2], which uses class-dependent paraphrases. Concretely, Ikkyu's paraphrase recognition is a real-time implementation of the pattern learning algorithm given in [1] and [2].

Paraphrases of a lexico-syntactic pattern can be obtained by gathering word pairs that instantiate a pattern's variables from the Web corpus, and calculating the relative overlap between the word pairs co-occurring with each pattern. For example, let us consider two lexico-syntactic patterns: "Y is cured by X." and "Y is medically treated with X." If there are many word pairs of X and Y that the two patterns have in common (e.g. steroid drug and atopy), it is likely that these patterns are paraphrase expressions of each other. The hypothesis that words that appear in a similar context have similar meanings is well known in linguistics and is called the "distributional hypothesis" (Harris [3]).

On the other hand, when the semantic classes of words co-occurring in a lexico-syntactic pattern are restricted, the pattern is called a class-dependent lexico-syntactic pattern. Lexico-syntactic patterns' ambiguity can be greatly reduced by placing semantic class restrictions on the pattern's variables. Let us consider a pattern "X for Y" as an example. It

would indicate a medical treatment relationship between X and Y if the words X and Y are of a certain semantic class, where X is a disease name and Y is a drug, such as "X: drug for Y: disease name." In this case the pattern can be considered as a paraphrase of the above-mentioned pattern "Y: disease name is cured by X: drug." On the other hand, the pattern "X for Y" would represent a semantic relation of means or tools if it is given as "Y: tool for X: work." If the words that co-occur in the calculation of pattern similarity are restricted to a certain semantic class in the above-mentioned manner, the ambiguity of patterns is significantly reduced. We can thus handle highly frequent, ambiguous patterns to increase the coverage of the acquired relation instances (word pairs).

These semantic classes are automatically acquired by the word clustering method proposed in [4] as well as in [1] and [2]. Under this method, co-occurrence frequencies of dependency relations between nouns and verbs are extracted from a large Web corpus, and this data is used to derive the posterior probability distribution to a hidden class of nouns. If the membership probability of a noun in a hidden class is above 0.2, the hidden class is regarded as a semantic class of the noun. Ikkyu currently uses this clustering method to classify 1 million nouns into 500 classes. These semantic classes are utilized not only for paraphrase recognition but also, for example, for the inference of a semantic class of promising potential answers.

### 2.2 Answering with hypothesis generated by inference

In Subsection **2.1** we described how Ikkyu acquires paraphrases to extract answers exhaustively from the Web. This process makes it possible to extract answers in superficially different expressions from the questions of users. However, no matter how large the volume of Web documents is, there will always be useful knowledge that is not expressed in a single sentence. Therefore, to find answers that people would be able to infer, Ikkyu seeks answers by combining information acquired from two different Web pages. When the answer extraction algorithm based on automatic paraphrase acquisition fails to find certain answers, they can instead be found as hypotheses using the inference process proposed in [5]. For example, Ikkyu automatically discovered the inference rule that "if X is the cause of Y and if Z prevents X, then Z is likely to prevent Y." Then, from the sentences "dark chocolate prevents arteriosclerosis" and "arteriosclerosis is a cause of cerebral infarction" found on different Web pages, Ikkyu generates, according to the inference rule, a hypothesis about the desirable effect of dark chocolate that "Dark chocolate works to prevent cerebral infarction." This side effect was not well known in the Web archives from which Ikkyu extracted the information data, but is now described on many Web pages.

The inference process of Ikkyu consists of two phases. One is for the automatic learning of the inference rules and another is an inference phase for applying the rules. We briefly describe each phase below.

1. Automatic learning of inference rules

For the automatic learning of inference rules, specific semantic relation instances such as "cause and effect" and "prevention" are prepared as seed data using the semantic relation acquisition method described in [1] and [2]. These instances consist of word pairs. We focus on the word pairs that include a common word. For example, suppose the seed data for prevention relations contains two word pairs: "coffee and sleepiness" and "caffeine and sleepiness." Then we assume that there exists some semantic relation between coffee and caffeine., and extract the lexico-syntactic patterns that may describe this relation from the corpus. From these lexico-syntactic patterns, many candidates for an inference rule about the "prevention" relationship, such as "if a prevention relation (between A and B) holds and 'C contains A', then a prevention relation (between C and B) holds," are automatically created as shown in Fig. 4. These inference

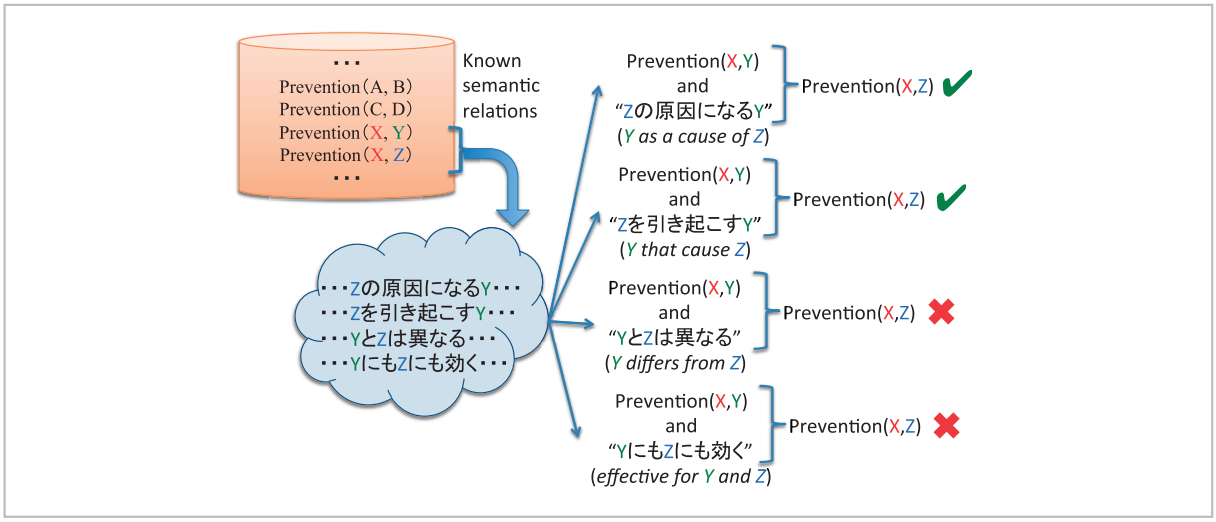



**Fig.4** *Example: the automatic acquisition of inference rules*

rule candidates are automatically evaluated and scored according to how correctly the seed data for the input semantic relation is reproduced. Inference rules of higher scores are considered reliable.

2. Inference by application of inference rules

A new instance of the target semantic relation is generated as a hypothesis by applying the automatically learned inference rules to the Web corpus. As shown in Fig. 5, the hypotheses generated from reliable inference rules are assumed to be accurate and the reliability of the hypotheses is evaluated by calculating the total score of the inference rules that generated this hypotheses. For example in Fig. 5, a hypothesis that "dark chocolate" and "strokes" are in a "prevention" relation is generated by multiple inference rules such as "prevention relation (between X = dark chocolate and Y = arteriosclerosis) < Y = arteriosclerosis causes Z = cerebral infarction > → prevention relation (between X = dark chocolate and Z = cerebral infarction)", "prevention relation (between Y = polyphenol and Z = cerebral infarction) < X = dark chocolate contains Y = polyphenol > → prevention relation (between X = dark chocolate and Z = cerebral infarction)." This hypothesis is considered highly reliable. The above relation between dark chocolate and cerebral infarction was not widely described in the Web corpus that we used in 2007 and the relation was not mentioned explicitly in any single sentence in our corpus. Therefore the relation could not be extracted even by using the lexico-syntactic patterns of the current version of Ikkyu. On the other hand, this relation attracted a lot of attention in the media and now many documents that describe this relation can be found with Google and other Web search engines. In other words one could say that our method found, in advance, the relation between dark chocolate and cerebral infarction.

Although the inference based semantic relation acquisition method is still at an initial stage, useful answers that are not explicitly described in a single sentence can already be provided to users as hypotheses, as the above dark chocolate example shows. However, cur-

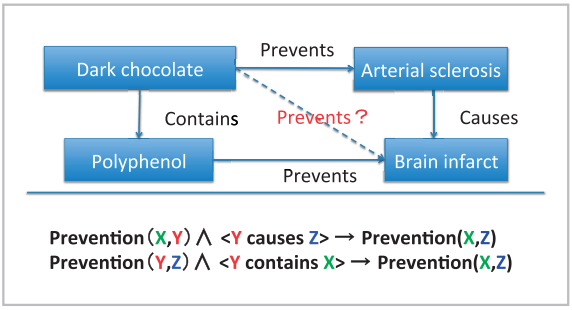



**Fig.5** *Example: inference through application of the learned inference rules*

rent hypothesis generation technology is effective only in a limited target and can be applied, for example, only to the relations between words. We are currently conducting research to widen the applicable range. One possible way is to acquire useful information for users, not from the relation between words, but from the semantic relation between phrases. Some results in this approach have been reported recently [6].

## 3 Respond to varied questions

In this section we describe how the core technology of Ikkyu based on automatic paraphrase recognition technology is used to handle a broader range of questions.

### 3.1 Questions including multiple query patterns

As far as the above-mentioned method (called core system of Ikkyu) is used straightforwardly, answerable questions are restricted to the ones that have a predicate relation between a noun ("deflation") and an interrogative word ("what"), such as "what causes deflation?" Under such limitations, searching a pattern or noun in a large amount of documents is not heavy work and answers can be immediately extracted even from 100 million Web documents. On the other hand, it is difficult to perform a high-speed search of answers for questions that do not have simple semantic relations. In what follows we explain a method of obtaining answers to complicated questions at a high speed by using the core system of Ikkyu.

The fundamental idea is as follows. When a question expressing a semantic relation between multiple nouns is given, it is broken into sub-questions which Ikkyu's core system can answer, and the answers to the sub-questions are then integrated to obtain the answer the question intended. The flow of this process is shown in Fig. 6.

(1) Generation of sub-questions

Using syntactic analysis, a question sentence is broken into smaller question sentences that contain only a single query pattern. All lexico-syntactic patterns consisting of dependency relation paths between nouns and interrogative words are generated as sub-questions. For example, a question "What products does Japan import from China?" can be broken into two sub-questions: (A) "What products does Japan import?" and (B) "What products are imported from China?" (See Fig. 7).

(2) Acquisition of answers to sub-questions

Ikkyu's core system described in Section **2** is used to acquire answers to sub-questions. First, query patterns for obtaining basic answers are created by breaking down the question sentence into sub-questions. Next, paraphrase patterns are generated from the query patterns extracted from the sub-questions. Finally, answers to these sub-questions are searched using these extended patterns. An example of answers to sub-questions is shown in Fig. 8.

(3) Integration of answers to sub-questions

The potential answers acquired from sub-questions are integrated to obtain an answer to the original question. The simplest way is to take the intersection of all answers to the sub-questions. However, even when some sub-

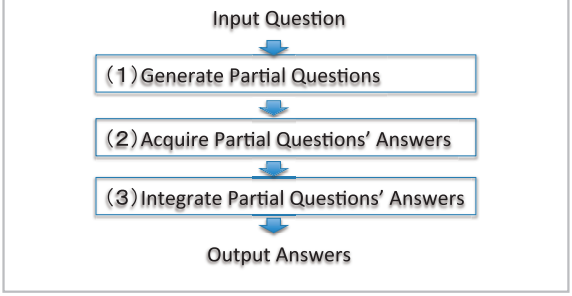



**Fig.6** *Processing flow for questions containing multiple query patterns*

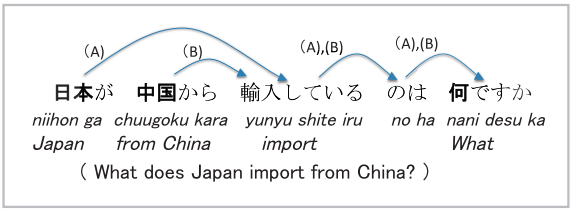
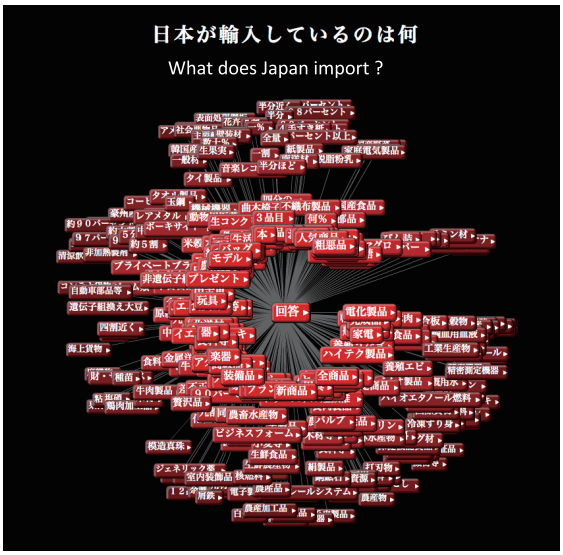
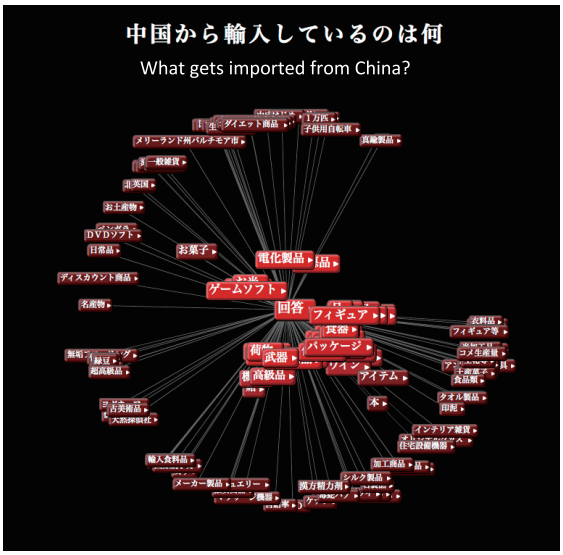



**Fig.7** *Dividing the question into sub-queries using syntactic analysis*




Partial questions (A)




Partial questions (B)

**Fig.8** *Example answers for the partial questions*

questions have the same answer, they might be obtained from contexts in different documents and may not always be correct as an answer to the original question. For example, suppose we find a sentence "Japan imports product A" in Document 1 and another sentence "China exports product B" in Document 2. These may indicate that Japan and China have a trade relationship with each other, but it may be the case that they do not trade directly with each other for political or other reasons.

Therefore we set a priority to each sub-question's answer by identifying the document and sentence from which the answer is obtained. If some sub-questions have the same answer and it is obtained from the same sentence, the priority of the answer is set to the highest. For example in Fig. 8, when both of the sub-questions (A) "What products does Japan import?" and (B) "What products are imported from China?" have the same answer "electric appliances," its priority is highest if it is found in the same sentence, second highest if in the same document, and third highest if found in different documents. Even when sub-questions do not have the same answer, their answers are added to the final answer list if a noun contained in a sub-question appears around the sentence from which the answer of another sub-question is obtained. For example, even when an answer "rare metal" is included only in the answer list of the sub-question (A), it is added to the list of final answers if the noun "Japan" given as a query pattern parameter of the sub-question (B), appears in the vicinity of the sentence from which the answer "rare metal" is acquired. The answers to the sub-questions are thus integrated to answer a complex question that contains multiple query patterns. Figure 9 shows the answer result.

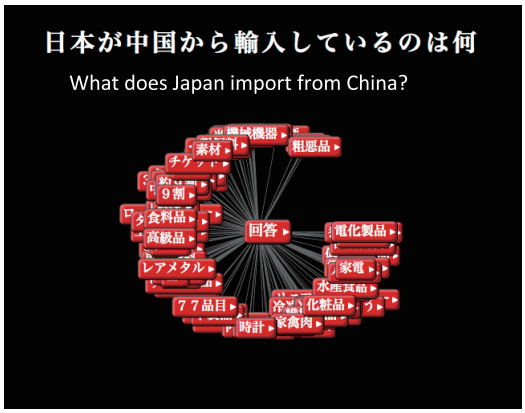



**Fig.9** *Example answers for the integrated partial questions*

### 3.2 Answer filtering by subject terms

When a question like "What does Japan import?" does not restrict the answer category, the answers given by Ikkyu will contain a variety of nouns from "shiitake mushroom" to "enriched uranium." However, if one wants to restrict answers to a certain category, one would ask a question like "What **food** does Japan import?" to restrict the answer category. In order to give answers to the questions that restrict the answer category, an answer filtering process is employed. The answer filtering process uses the category words (hereafter called subject terms), such as "food" in the above question. We explain the process in this section.

#### *3.2.1 Subject term extension by synonym dictionary*

Subject terms have the role of specifying the range of answers that users want to have for a question. It is obvious that answers like "towels" or "electric appliances" would not be suitable to the question "What food does Japan import?" Therefore, if syntactic analysis of the (Japanese) question sentence detects a noun that directly depends on an interrogative word, we select that noun as a subject term. In the above question, since the noun "food" has a direct dependency on the interrogative word "what", we acquire "food" as a subject term. Next we acquire synonyms for the subject term from a synonym dictionary (Language Resources A-9: Example base of fundamental semantic relations) released by the Advanced Language Information Forum, ALAGIN (www.alagin.jp). For example, for the word "食べ物" (meaning "food" in Japanese), we can acquire notational variant nouns, "食べもの" and "たべもの," and a synonym "食物." For details of the synonym dictionary, see "**5-5** Fundamental Language Resources" in this special issue.

#### *3.2.2 Answer filtering by subject terms*

Answers are filtered by these subject terms extended by the synonym dictionary. The answer filtering process uses context similarity and a hypernym-hyponym dictionary. By using the subject term and the Alagin language resources to restrict the range of answers, we can obtain appropriate answers. At present, only the answers that do not meet the following filtering conditions are deleted from the answer list in order to prevent excessive filtering.

(1) Answer filtering by context similarity

This filtering uses the context similarity that utilizes the probabilistic clustering of dependency relations obtained from 600 million Web documents. Semantic similarity of words calculated based on their distributional profile is called context similarity. In the present study we use the Context-based Similar Word Dictionary developed according to the method proposed in [7] and released by the Advanced Language Information Forum (ALAGIN). (See "**5-5** Fundamental Language Resources" in this special issue.) For example, words of high context similarity to "food" include snacks, alcohol, fish, meat, wine, coffee, beer, chocolate, banana, mushroom, and so on. On the other hand, the context similarity of "towel" or "electric appliance" is relatively low as they are less connected to "food." Context similarity can thus be used to eliminate answers that are used in a different context from the subject term and thus less relevant to the term.

(2) Answer filtering by hypernym-hyponym dictionary

This filtering uses a hypernym-hyponym dictionary acquired from Wikipedia (ALAGIN's language resource A-4: Hypernym hierarchy data, [8]). For example, we can obtain hyponyms of "food," such as "fruit", "mushrooms", "fish", "seafood", "Japanese sake", and "cake," and furthermore hyponyms of "fruit," such as "cherry", "fig" and "earl's melon." By recursively selecting hyponyms of the subject term, we can thus acquire only answers that are in the answer category intended by the user.

## 4 Ikkyu's speech interface

Ikkyu accepts questions from a variety of fields ranging from economy, health, and philosophy, to hobbies, sightseeing, cartoon films,

and cooking. In order to use smart phones as input devices, the construction of a language model that can perform accurate speech recognition of open domain questions is an important challenge. In this section, we introduce the speech frontend to Ikkyu's core QA system described in Section **2**.

Most previous studies of language model construction assume a manually constructed corpus that restricts the domains of target applications (e.g. to travel or healthcare) as well as the input sentence's style, and construct a language model by adding similar sentences from the Web to the corpus [9]-[11]. On the other hand, Ikkyu's input questions must contain a query pattern in each question, as described in Section **2** (Hereafter, this question format is simply called the input style.). However Ikkyu is open domain, meaning that the domain of the questions is not restricted (e.g. to sightseeing or medical care). To create a language model for Ikkyu, we manually collected question sentences in the required style and put them together into a seed corpus. Similar sentences were then automatically collected from Web documents to build a new corpus. We then constructed an open-domain language model from this corpus.

It is known that the above method is useful to some extent for the development of a language model for domain-restricted speech recognition. On the other hand, it was not clear whether this method would be useful for Ikkyu, which restricts question styles but not domains. A problem that we can expect first is that, since the vocabulary of the seed corpus is considerably smaller than that used in open domain, i.e. the entire Web, the corpus would cover only a limited vocabulary even after it automatically collects similar sentences to the seed corpus from the Web. To address this issue, the seed corpus is extended by automatically replacing nouns in the seed corpus with semantically similar ones [7] in order to make the vocabulary larger. As a result, a large number of question sentences that have wider vocabulary and meet the style required by Ikkyu could be efficiently collected from the Web corpus. Then, a low-cost, high-performance language model can be obtained by applying the existing domain adaptation method from [9] to this seed corpus.

In what follows we explain this method in more detail. The domain adaptation method [9] calculates the perplexity of the sentences from the Web based on n-grams obtained from the seed corpus, and collects Web sentences that have a similar tendency to the seed corpus. To create a language model, we manually build a seed corpus of 500 sentences that meet Ikkyu's style and cover various topics. Then, using this seed corpus and the Web as input, the following steps are taken.

1. As in Section **3**, the nouns contained in the sentences of the seed corpus are replaced with the k most similar words, using the "Context-based similar words database" released by ALAGIN. The newly obtained sentences are added to the seed corpus.
2. A learning corpus is constructed by applying the method given in [9] to the extended seed corpus and Web corpus.
3. A speech recognition language model is created from the learning corpus by using existing tools [12].

In our evaluation experiments we used the speech recognition device ATRASR [12], the vocabulary size of the language model of the proposed method was 410 thousand words, the word error rate is 15.49%, and the sentence error rate is 54.73%. These rates are 3.25 points (word error rate) and 4.28 points (sentence error rate) lower than those of a baseline language model that is constructed with sentences extracted randomly from the Web corpus. Table 1 shows examples of question sentences that were correctly recognized. We thus found that, using the speech recognition language model constructed by this method, we could obtain a highly accurate speech-based question answering system. Table 1 also contains more complicated questions than the ones that the core system of Ikkyu can answer. For details, see [13].

**Table 1** *Examples of correctly recognized question sentences*

| | |
|---|---|
| はやぶさは何年ぶりに地球に帰還した？ | *After how many years did Hayabusa (Japanese space probe) return to the Earth?* |
| 最近発売されたソニーの学習リモコンの型番は？ | *What's the model ID of Sony's recent universal remote control device?* |
| 板付遺跡はどこにありますか | *Where are the Itazuke ruins?* |
| 東京ディズニーランドの最寄り駅はどこですか | *Which station is closest to Tokyo Disneyland?* |
| 5 月の誕生石を教えて下さい | *Tell me the birthstones of May.* |
| 熱中症の初期症状は？ | *(What is) the first symptom of hyperthermia?* |
| 国勢調査は何年おきに実施される？ | *How long is the interval (in year) between each national census?* |
| ステロイドの副作用にはどんな物がありますか | *What are the side effects of steroids?* |
| かいけつゾロリの作者はだれ？ | *Who is the author of Kaiketsu Zorori (cartoon)?* |
| ウインブルドンで優勝した人はだれ？ | *Who is the champion at Wimbledon?* |
| ルイ 14 世の業績は何ですか | *What are the achievements of Louis XIV?* |
| 日本で iPhone はどれ位売れていますか | *How many iPhones have been sold in Japan?* |
| ポストモダンとは何ですか | *What is postmodern?* |
| Java の最新バージョンは？ | *What is the latest version of Java?* |

## 5 Ikkyu's position in question answering research

Information access methods such as search engines and question answering systems have made remarkable progress in recent years. For example, IBM's Watson [14] has attracted much attention as a question answering system.

Watson came to prominence by its overwhelming victory over the human champion of a TV quiz show, Jeopardy. It however requires a supercomputer. It is also said that Jeopardy's questions needed to be adjusted. On the other hand, Ikkyu is a simple system that runs on a single server for the real-time extraction of answers from hundreds of millions of Web pages, not taking data updates into account. Also, the adjustment of questions of specific types is not necessary. Taking advantage of the characteristics of Ikkyu, we are currently in the process of incorporating it as a sub-module of NICT's information analysis system WISDOM (www.wisdom-nict.jp), in order to answer various questions from users based on billions of Web pages. We are also planning to release it as an information system that can be used for support and reconstruction in case of a disaster. At the Resilient ICT Research Center of NICT, the system would extract information about isolated disaster sites, necessary goods, supplied goods, support and so on from the large amount of internet information that is generated when a disaster occurs.

As described in the beginning of this paper, the final goal of the development of Ikkyu is to find unexpected but useful information from users' whimsical questions. For this goal, Ikkyu must be able to find potential answers as hypotheses to questions that may or may not have a definite answer, or a function to list all answers that it can find. This is in sharp contrast to Watson, which aims to provide a single correct answer to questions that have a unique answer that is just difficult for people to find. Ikkyu has been developed with this functionality in mind. To get closer to this goal, research is now in progress to strengthen the hypothesis generation method, and to develop a mechanism for recommending more useful questions and answers based on a users' history of making questions and viewing their answers. We are also conducting research to handle questions that require full sentence answers such as "why" and "how" questions. Even Watson cannot answer these types of questions at this

moment. In fact, a prototype "why" question answering system is currently included in Ikkyu. For example, given the question "Why did Japan lose to the US forces at Guadalcanal?" Ikkyu displays paragraphs from Web pages that mention these reasons, such as the sequential deployment of forces or the distance from the base. Furthermore, as we briefly described in this paper, it has become possible to acquire and accumulate a large amount of global knowledge from Web pages. An example of this global knowledge is a causal relation between sentences or phrases such as "depreciation of yen" ⇒ "increasing exports." In the future, we will develop a mechanism that utilizes global knowledge, shows more useful hypotheses, and recommends more useful questions and answers.

Finally let us show a specific example of useful hypotheses. The Chinese government stopped rare earth exportation to Japan as a result of a territorial dispute in 2010. Expecting that the exportation of other materials would also be stopped, we used Ikkyu to survey the raw materials Japan depends on China for, the products for which these raw materials are used, and the companies that manufacture these products. As a result, we found that Japan imports tungsten from China, and a famous Japanese multinational that uses tungsten to manufacture ultra-hard tools. We then generated the hypothesis that if China would stop the export of tungsten to Japan, this Japanese company would have a problem in the production of ultra-hard tools, and introduced this scenario as an example in a research report. A week later, Nikkei Newspaper published an article with the title of "tungsten follows rare earth", stating that the Chinese government declared that it would increase the price of tungsten. The article also included an interview with a representative of that same Japanese company. This indicated that to a certain extent our hypothesis may be confirmed in reality. However, manual work was necessary to make such an advanced hypothesis. Ikkyu could acquire the partial information required to create this hypothesis such as "Japan depends on China to import tungsten", "tungsten is used to manufacture ultra-hard tools", and "Company X manufactures ultra-hard tools?" but currently human intervention is necessary to integrate this partial information into a well-formed hypothesis. In the future we intend to conduct further research on the automatic creation of hypotheses without any human intervention. It is quite difficult to automatically generate a hypothesis, such as the tungsten story, that exactly mirrors events in the real world, but it may be feasible to show at least many probable hypotheses to users. We would also like to study the automatic generation of various measures that users can take against undesirable hypothesis. For example, companies depending on ultra-hard tools may negotiate with other manufacturers to find another route of buying alternative tools. We think that the next step of our research would be to develop Ikkyu not only as simple question answering system but also as a strategic adviser or assistant for users.

## 6 Conclusions

This paper described "Ikkyu," a speech-based question answering system based on the Web, developed by the Information Analysis Laboratory of Universal Communication Research Institute. Ikkyu performs a semantic analysis of a large amount of information beyond any individual's grasp, and generates unknown but potentially useful hypotheses as answers by flexibly combining the acquired information to meet the various information needs of users.

At present the information explosion shows no sign of abating. In this situation, the improvement of access to useful information and knowledge can directly improve the quality of the appropriate decision making of individuals and society. On the other hand, it has become clear that current search engines and other systems that simply list up a large amount of Web documents that match some query keywords cannot provide the type of information required for effective decision mak-

ing. With Ikkyu, we would like to contribute to the improvement of this decision making process and to the development of efficient information gathering.

## *References*


1 Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata, "Large Scale Relation Acquisition using Class Dependent Patterns," in Proceedings of the IEEE International Conference on Data Mining (ICDM'09), pp. 764–769, Miami, Florida, USA, Dec. 2009.

2 De Saeger Stijn, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata, "The Large-scale Semantic Relation Acquisition based on Pattern Learning using Semantic Class of Words," NPL 2010 (16th annual meeting of The Association for Natural Language Processing).

3 Zellig Harris, "Distributional Structure. In Word 10(23)," pp. 142–146, 1954.

4 Jun'ichi Kazama and Kentaro Torisawa, "Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations," In ACL08-HLT: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 407–415, 2008.

5 Masaaki Tsuchida, Kentaro Torisawa, Stijn De Saeger, Jong Hoon Oh, Jun'ichi Kazama, Chikara Hashimoto, and Hayato Ohwada, "Toward Finding Semantic Relations not Written in a Single Sentence: An Inference Method using Auto-Discovered Rules," In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011), Chiang Mai, Thailand, Nov. 2011.

6 Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama, "Excitatory or Inhibitory: A New Semantic Orientation Extracts Contradiction and Causality from the Web," Proceedings of EMNLP-CoNLL 2012: Conference on Empirical Methods in Natural Language Processing and Natural Language Learning, 2012.

7 Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa, "A Bayesian Method for Robust Estimation of Distributional Similarities," In Proceedings of ACL 2010, pp. 247–256.

8 Ichiro Yamada, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata, Stijn De Saeger, Francis Bond, and Asuka Sumida, "Hypernym Discovery Based on Distributional Similarity and Hierarchical Structures," EMNLP'09, 2009.

9 Teruhisa Misu and Tatsuya Kawahara, "A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts," In Proceedings of Interspeech 2006, pp. 9–13.

10 Ruhi Sarikaya, Agustin Gravano, and Yuking Gao, "Rapid Language Model Development Using External Resources for New Spoken Dialog Domains," In Proceedings of ICASSP 2005, Vol. I, pp. 573–576.

11 Mathias Creutz, Sami Virpioja, and Anna Kovaleva, "Web augmentation of language models for continuous speech recognition of SMS text messages," In Proceedings of the 12th Conference of the European Chapter of the ACL, pp. 157–165.

12 Shigeki Matsuda, Takatoshi Jitsuhiro, Konstantin Markov, Satoshi Nakamura. 2006. "ATR Parallel Decoding Based Speech Recognition System Robust to Noise and Speaking Styles," IEEE Transactions on Information and Systems Vol. E89-D(3), pp. 989–997.

13 Istvan Varga, Kiyonori Ohtake, Kentaro Torisawa, Stijn De Saeger, Teruhisa Misu, Shigeki Matsuda, and Jun'ichi Kazama, "Similarity Based Language Model Construction for Voice Activated Open-Domain Question Answering," In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011), Chiang Mai, Thailand, Nov. 2011.

14 Ferrucci et al., "IBM Research Report: Towards the Open Advancement of Question Answering Systems," http://domino.watson.ibm.com/library/CyberDig.nsf/papers/D12791EAA13BB952852575A1004A055C/$File/rc24789.pdf

(Accepted June 14, 2012)

***DE SAEGER Stijn**, Ph.D.*

*Senior Researcher, Information Analysis Laboratory, Universal Communication Research Institute*

*Natural Language Processing, Knowledge Acquisition*

***GOTO Jun***

*Research Expert, Information Analysis Laboratory, Universal Communication Research Institute*

*Natural Language Processing, Information Extraction*

***VARGA István**, Dr. Eng.*

*Researcher, Information Analysis Laboratory, Universal Communication Research Institute*

*Natural Language Processing, Information Extraction*