5 Language Infrastructure and Information Analysis Technology

5-1 Information Analysis Technologies at NICT

TORISAWA Kentaro

We have conducted research on information analysis technologies, which enables us to automatically analyze a huge amount of information available on the Web since 2006. Our research achievements include the information analysis service WISDOM, as well as several language resources and tools available from the ALAGIN forum. The former allows users to analyze information on the Web from several perspectives and provides users the insight necessary to help them assess the credibility of information obtained from the Web. The latter are indispensible resources for a deep analysis of textual information. In this paper we give an overview of these research activities and describe the underlying aims for which we developed them.

Keywords
Information analysis, Natural language processing, World Wide Web, Text mining, Question answering

1 Introduction

The so-called information explosion shows no sign of ending, accumulating the ever-increasing information called Big Data on the Internet. To create value out of this "Big Data" is now a global challenge. To this end, we have pursued research on technologies and methods to analyze a huge amount of information on the Internet since 2006. Our research achievements include the information analysis system WISDOM, several language resources, tools, web services available from Advanced LAnGuage INformation Forum (ALAGIN), and the speech-based question answering system "Ikkyu". We are not going to present the details of each achievement in this paper since they are presented in other papers in this special issue. Instead, we will present what lies in the background of those technologies and the future direction of our research and development.

2 Technologies for deeper text analysis

What makes NICT's information analysis technologies unique is their ability to analyze texts and documents at a deeper level. Many of the methods used in accessing information on the Web including those used by most search engines are based on what is called keyword search. Keyword search systems retrieve documents that contain the keyword(s) given by a user and rank them according to a certain standard to provide them for the user. Certainly, search engines take the meaning of a keyword into consideration to some extent. For instance, they use the technique called "query expansion" to achieve this. For example, when the word "Apple" is inputted as a keyword, the system retrieves a documents containing not only the word "Apple" but also "アップル/apple". However, the core part of their searching technology is not the query ex-

pansion technique but the ranking technology. One of the most typical examples of the ranking technologies is PageRank developed by Google. PageRank analysis is based on the link function characteristically possessed by Web pages without a deeper analysis of the contents of a Web page. On the other hand, in NICT, we aim at developing information analysis technologies that achieve a deeper analysis of the meaning and contents of what is written in a text. One such example is the information analysis system WISDOM. WISDOM does not provide the search results simply because the keyword is contained in a document. It provides information regarding the input keyword by organizing positive and negative evaluations of what the keyword denotes by analyzing the syntactic structures of candidate sentences and identifying the phrases that give either positive or negative evaluations using a machine learning technique. Moreover, WISDOM identifies the originator of the document, e.g. a company, a university or an anonymous user. By using these two functionalities, the way a keyword is accepted in the society is easily analyzed by looking at the evaluations and opinions about it in a certain field such as its bad reputation in the medical field and good reputation in the business field. Actually we found that an accident death caused by a food product was reported by a medical school at a university while the same product were receiving good evaluations on business related sites (to be precise, the product that caused the accident was not exactly identical but similar to the one introduced on business sites). In short, WISDOM provides the users with clues for judging the credibility of the information by showing them both positive and negative evaluations of a certain object. For the details of WISDOM, please see the Chapter **5-3** "Development of the Information Analysis System WISDOM" in this special issue.

The speech-based question answering system "Ikkyu" runs on a smartphone and provides users with list of answers to a spoken question. For example, when a user asks a smartphone "What causes deflation?" with a voice, it retrieves a list of answers to that question out of hundreds of millions of Web pages. A common search engine would give you just a bunch of documents containing the keywords such as "deflation" and "cause", making you read a large amount of documents and identify the answers yourself. In addition, keyword search has another problem that the information given by a user is just a set of words. A search engine cannot judge whether what the user wants to know is "the cause of deflation" or "what can be caused by deflation" and thus ends up retrieving large amounts of documents that the users otherwise would not have to read. On the other hand, Ikkyu provides straightforward answers in the forms of words or phrases. Therefore, even for a question that may have a huge amount of answers, like the one asking for the cause of deflation, each answer is very short, enabling users to get an overall image of the issue or to find interesting related cases. For example, as an answer to the question asking the cause of deflation, Ikkyu identified the name of a very famous major Japanese company. Seemingly, the answer did not make any sense, but as we tracked the link provided by Ikkyu to identify the documents where Ikkyu had extracted the answer, we found a logical ground for its answer in a sentence stating "… has allocated a huge amount of profits to their internal reserves to stop their capital from flowing out to the market" (see the video demonstration of this operation on http://www2.nict.go.jp/univ-com/info_analysis/). After we had found the answer and its grounds, an article in an economic magazine named the Japanese company as the causer of deflation on the same grounds based on the statistics that the total amount of internal reserves held by Japanese firms had reached 200 trillion yen. Actually, the Web page where Ikkyu had extracted the answer had been written by an anonymous person who seemed not to be an economist. This is a very interesting incident because it implicates the complexity of the modern economy and a very high level of ability possessed by some

non-professionals living in the so-called Internet society to gather and understand information. It is not coincidental that Ikkyu found such unexpected information since it has originally been developed based on the same concept as the one for its predecessor "TORISHIKI-KAI", the Web search support system containing a conceptual dictionary whose development concept is to help users "to find unexpected but useful information" [1] [2].

It is Ikkyu's ability to "deeply analyze the contents of texts" to enable Ikkyu to retrieve not documents but a list of straightforward answers. To be concrete, Ikkyu extracts pairs of nouns that can fill the variable slots X and Y in such patterns as "X causes Y," "X aggravates Y" and "Y by X" (e.g. "globalization" + "deflation" or the name of a company + "deflation") from texts, and stores them as a kind of database. Not only storing them, but Ikkyu automatically identifies that the patterns "X causes Y," "X aggravates Y" and "Y by X" are all synonymous, denoting the same situation. Thus, Ikkyu can identify the phrase "deflation by globalization" which has no common words other than "deflation" with the question "What causes deflation?" as a source of answer for the question. In order to successfully extract such information, the system has to be able to recognize the syntactic structure of a sentence to identify the patterns that make sense. For example, it is desirable for Ikkyu to identify a pattern "X causes Y" in a sentence "House dust, for example, causes allergy," but to do this, Ikkyu has to know that the expression "for example" is not important therefore can be cut out of the pattern. Thus, identification of grammatical structures, i.e. syntactic analysis, is required for such task. Although I wrote that "X causes Y" and "Y by X" are synonymous, this is not the right way to describe them in more general cases. For example, when X is "Apple" and Y is "iPhone", the expressions "iPhone by Apple" and "Apple causes iPhone" are not interchangeable. They can be said to be synonymous only when the noun pairs to fill the X and Y slots fall in certain types. Examples of such noun pairs include a chemical substance such as formaldehyde as X and a name of disease such as atopic dermatitis as Y. Ikkyu automatically judges the synonymity between phrase patterns based on the results of automatic identification of the types and semantic categories of given words, which can be said to be another instance of Ikkyu's deep analysis. Such semantic categorization also requires the ability to identify grammatical structures of texts. For further details about Ikkyu, please see the Chapter **5-2** "Speech-based Question Answering System 'Ikkyu'" in this special issue. Note that the technologies and languages resources as dictionaries are distributed through a consortium named the ALAGIN. For their details, please see the Chapter **5-5** "Fundamental Language Resources" and the Chapter **8-1** "Advanced Language Information Forum (ALAGIN)" in this special issue.

Ikkyu is now being expanded so that it can answer questions with more complex forms, such as so-called "why questions", [3]. The answer to a why question should be a sentence, not a word. IBM Watson has become very famous for its victory over a human champion in a US TV quiz show, but to the best of our knowledge, even Watson does not have the ability to answer the right answer to question demanding sentences as their answers. For example, to the question "Why did we lose the battle of Guadalcanal to the United States?", current Ikkyu succeeded in retrieving the right sentence that explained the historical context of the battle by referring to such phrases as "successive deployment of small manpower" and "the distance between the front and the base". Although Ikkyu's overall performance level for such complex questions cannot be said to be very high, we expect that it will achieve a reasonable performance in the near future. Unfortunately, this article is not going to present further details about the "why-question" answering. The task requires integration of the deep analysis technologies developed for WISDOM and Ikkyu.

3 Future tasks

The technologies we have described so far could only be imagined 20 years ago, when I was a postgraduate student studying natural language processing. Actually I could not have imagined that twenty years later, I would be one of the developers of the system that can answer to real-world questions such as "what causes deflation" or "Why did we lose the battle of Guadalcanal to the United States?" through a mobile phone. The enabler of such unexpected technological evolution is the explosion of the Web, which provides us with huge amount of textual data, and the advanced statistical analysis and machine learning technologies.

Nonetheless, I do think we are still yet to achieve everything that can be achieved from these enablers. The systems using analysis technologies presented in this paper effectively functions only when the user gives them appropriate questions or queries. Let us consider the example "Sendai Plain" which is the name of a place in east Japan which was heavily damaged by tsunami caused by the 2011 Great East Japan Earthquake. After the earthquake we asked Ikkyu for the place hit by a tsunami in the past. It could successfully give the answer "Sendai Plain" using Web pages created before the earthquake. The information source was the Web site run by a research institute and their geological survey found that a tsunami had hit Sendai Plain about 1,000 years ago. This is the tsunami that Ikkyu found on the Web.

It is rather well known in Japan that places hit by a tsunami tend to be hit repeatedly. If this information had been spread among larger numbers of people living in the area, the people would have asked the local government to take better safety measures against tsunamis. One problem is that, it is extremely difficult to formulate the question "Where did tsunami hit in the past?" in the first place when searching the potential risks related to a given area such as Sendai Plain. First, there are huge variations of potential risks related to a certain area. Tsunami is just one of them. Secondly, it is a well known fact that tsunami seems to hit the same place many times but recognizing the tsunami as one of the clear dangers considering that fact and investigating the past tsunami demands a relatively complex reasoning. Considering that the tsunami is just one of the many potential risks and that conducting such reasoning for many of those risks requires many trials-and-errors and is time-consuming, most people are unlikely to do this. In a sense, one may be able to infer that the geological survey that revealed the tsunami occurred 1,000 years ago was truly valuable because of the unlikeliness.

Thus, if there were a system that can automatically do the inference based on general knowledge including the ones on tsunami in order to formulate the question "Where did tsunami hit in the past?" and could recognize that the answer suggests a real danger, it would be valuable in the sense that it can provide us with valuable information that many people would not reach.

In order to realize such systems, we are now working on the development of automatic systems to retrieve general knowledge from a large amount of Web pages which existing systems do not possess. When such systems are implemented, it will be able to provide "information about tsunamis in neighboring areas in the past" if it receives the task regarding "safety investigation of nuclear power plants" by using general knowledge such as "tsunamis tend to hit the same place." In other words, it is a system that can guess and decide what a user wants before the user tells so to the system and provide the user with a broader range of information, which in the end would lead the user to make better decisions. Figure 1 illustrates a network of large amounts of phrases denoting events that are causally related. It was acquired from Web pages by using our automatic acquisition procedure of causal knowledge [5]. The network includes chains of events that can be judged to be causally related based on the acquired knowledge, such as "have less traffic ⇒ alleviate traffic congestion

⇒ reduce air pollution/prevent traffic accidents" and "slip into inflation ⇒ have a weaker yen ⇒ export increases/have a stronger dollar". Although the system is still incomplete in terms of its precision and coverage, it can extract and generate millions of causally related pairs. We believe that in the future, we will have a system that can guess and decide what a user needs before the user tells so to the system, i.e. a system that can infer from general knowledge found on the Web. In other words, it is a technology to enable "the Web that thinks".

Moreover, we are expanding "Ikkyu" to develop a system that can promptly provide lists of important information in case of disaster by using information available on the Internet, especially those found among twitter comments or posted on local government or volunteer group websites, such as the names of isolated areas or hospitals offering specific treatments like dialysis and announcements regarding offering to provide assistance and relief supplies or requests for them. In addition, to check the spread of false information which caused some problems in the aftermath of the great earthquake, we plan to provide information that contradicts to certain information about specific products (e.g. "You may have heard the announcement that *** is giving out free contact lenses, but it is a lie"). We plan to make the system available to the public by the end of FY 2014.

4 Conclusion

This paper presented an overview of NICT's researches on information analysis technology. It should be noted that our deeper text analysis has contributed to the implemen-

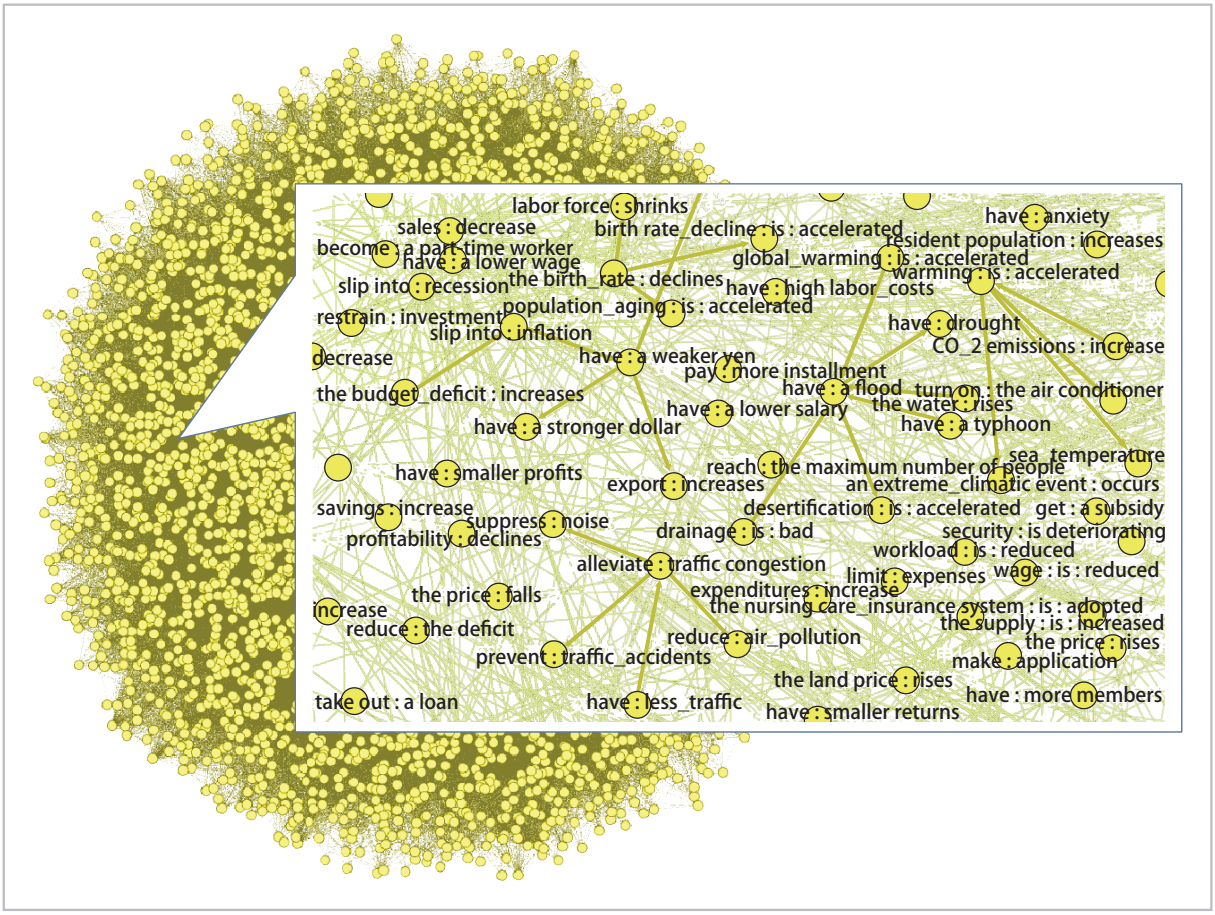

Fig.1 *Network of phrases to denote causally related events acquired on the Web (The original expressions are Japanese and are translated into English by a human translator)*

tation of various high-level information analysis systems. Those systems have already enabled the discovery of information including the unexpected but useful ones and the analysis of information from several viewpoints that cannot be found elsewhere. We plan to further promote our research on deeper analysis technology to develop systems to support various decision making processes not just by providing superficial information on the Web but by making "the Web to think" and providing users the information and the hypotheses derived by high-level automatic inference based on the general knowledge acquired from the Web.

Acknowledgement

I am deeply grateful to the members of Information Analysis Laboratory who have always been willingly giving us their helpful opinions and the members of the development team of the information analysis system WISDOM.

References

1 Kentaro Torisawa, Stijn de Saeger, Jun'ichi Kazama, Asuka Sumida, Daisuke Noguchi, Yasunari Kakizawa, Masaki Murata, Kow Kuroda, and Ichiro Yamada, "Organizing the Web's Information Explosion to Discover Unknown Unknowns," in New Generation Computing (Special Issue on Information Explosion), Vol. 28(3), pp. 217–236, July 2010.

2 TORISAWA Kentaro, NAKAGAWA Hiroshi, KUROHASHI Sadao, INUI Kentaro, YOSHIOKA Masaharu, FUJII Atsushi, and KITSUREGAWA Masaru, "Beyond Keyword Search: Info-plosion Search for Mining Valuable Unknowns(<Special Features> Creating Vital Information Technologies for the Info-plosion Era)," Journal of Information Processing (JIP), Special issue of "Info-plosion", Vol. 49, No. 8, pp. 12–18, 2008.

3 Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Takuya Kawada, Stijn De Saeger, Jun'ichi Kazama, and Yiou Wang, "Why Question Answering using Sentiment Analysis and Word Classes," In Proceedings of Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLPCoNLL 2012), Jeju, Korea, July 2012. (To appear)

4 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), pp. 902–910, Chiang Mai, Thailand, Nov. 2011.

5 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," In Proceedings of Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP- CoNLL 2012), Jeju, Korea, July 2012. (To appear)

(Accepted June 14, 2012)

***Kentaro Torisawa**, Ph.D.*

Director, Information Analysis Laboratory, Universal Communication Research Institute

Computational Linguistics, Knowledge Acquisition, Web Mining