WISDOM X, DISAANA and D-SUMM: Large-scale NLP Systems for Analyzing Textual Big Data

Junta Mizuno Masahiro Tanaka Kiyonori Ohtake Jong-Hoon Oh
Julien Kloetzer Chikara Hashimoto Kentaro Torisawa
Data-driven Intelligent System Research Center (DIRECT), NICT / Kyoto, Japan
{junta-m, mtnk, kiyonori.ohtake, rovellia, julien, ch, torisawa}@nict.go.jp

Abstract

We demonstrate our large-scale NLP systems: WISDOM X, DISAANA, and D-SUMM. WISDOM X provides numerous possible answers including unpredictable ones to widely diverse natural language questions to provide deep insights about a broad range of issues. DISAANA and D-SUMM enable us to assess the damage caused by large-scale disasters in real time using Twitter as an information source.

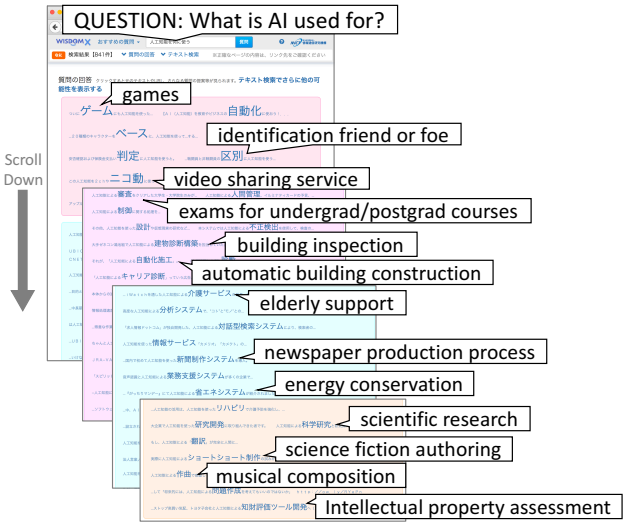
1 Introduction

This paper describes three large-scale NLP systems we have developed at NICT: WISDOM X, DISAANA, and D-SUMM. The first system, WISDOM X[1], is an open-domain question-answering (QA) system for Japanese using 4-billion web pages as an information source. It was designed to enable users to obtain a wide and deep perspective on a broad range of issues. The range of questions that humans can pose is unlimited and web texts are a valuable information source for compiling a comprehensive list of answers. Such answers are expected to include *unknown unknowns* in the infamous words of Donald Rumsfeld, which are things that "we don't know we don't know" (Torisawa et al., 2010). For instance, even though global warming is a severe and widely discussed problem that might result in devastating *unknown unknowns* for many people in the future, no exhaustive list of answers has been compiled to the question: "What happens if global warming worsens?" Although many documents available on the web actually describe the possible consequences of global warming, only a few can be discovered using commercial search engines because they merely provide a huge number of documents that users have to read. By contrast, WISDOM X, for instance, provides hundreds of answers to the question, and furthermore suggests new questions related to the first question to have deeper knowledge related to the issue.

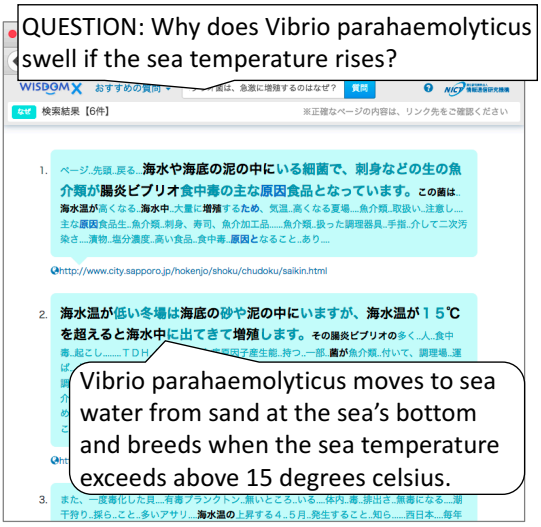
Our other two systems, DISAANA[2] and D-SUMM[3], were developed to help disaster victims and rescue workers in the aftermath of large-scale disasters. One lesson from the 2011 Great East Japan Earthquake was that a large-scale disaster can destroy a wide range of infrastructure, disrupt lives, and cause many unpredictable situations. Immediately after the disaster, much useful information was transmitted into cyberspace, especially for such social media as Twitter. Nevertheless, because most people were overwhelmed by the huge amount of information, they were unable to make proper decisions and much confusion ensued. DISAANA provides a list of answers to questions such as "What is in short supply in City X?" and displays locations related to each answer on a map (e.g., locations where food is in short supply) in real time using Twitter as an information source. D-SUMM summarizes the disaster reports from a specified area in a compact format and enables rescue workers to quickly grasp the disaster situations from a *macro* perspective. In the 2016 Kumamoto Earthquake (M7.0), DISAANA was actually used by the Japanese government[4] and provided a wide range of useful information, including such unpredictable one as the shortage of *Halal foods*[5].

[1] publicly available at http://wisdom-nict.jp (in Japanese)
[2] publicly available at http://disaana.jp (in Japanese)
[3] publicly available at http://disaana.jp/d-summ (in Japanese)
[4] "Analyzing tweets to comprehend necessities," Yomiuri Shimbun Evening edition, p.1, 2016, May 11.
[5] The Muslim population in Japan is quite small.

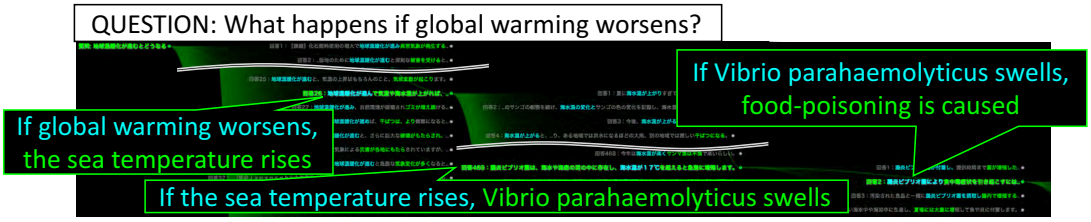
Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 263–267, Osaka, Japan, December 11-17 2016.



(a) Results from the factoid QA system.



(b) Results from the why-QA system.



(c) Results from the what-happens-if QA system.

Figure 1: Example screenshots of WISDOM X.

In the following, we provide an overview of WISDOM X, DISAANA, and D-SUMM.

2 WISDOM X: Information Analysis System

WISDOM X, which discovers answers to given questions from about 4-billion web pages by several kinds of deep semantic processing, consists of four QA systems, each of which deals with different types of questions: factoid (e.g., What prevents global warming?), why-type (Oh et al., 2012; Oh et al., 2013; Oh et al., 2016) (e.g., Why did the global warming worsen?), what-happens-if-type (Hashimoto et al., 2012; Hashimoto et al., 2014) (e.g., What happens if global warming worsens?), and definition type (e.g., What is global warming?). It also has a functionality that suggests questions to users. These QA systems use a large-scale knowledge base for entailment recognition(Saeger et al., 2009; Hashimoto et al., 2009; Saeger et al., 2011; Hashimoto et al., 2011; Kloetzer et al., 2013; Sano et al., 2014; Kloetzer et al., 2015) and semantic noun clusters (Kazama and Torisawa, 2008). We also developed a middleware RaSC (Tanaka et al., 2016) to efficiently run various NLP tools on hundreds of computation nodes.

We designed WISDOM X to provide a wide range of pin-point answers, e.g., a noun phrase for factoid questions and a sentence for what-happens-if-type questions. This feature constitutes a major difference from commercial search engines, which merely give web pages for a given query and rely on human effort to ascertain pin-point answers. In addition, WISDOM X can provide numerous answers to a given question. For instance, the current version of WISDOM X provides around 800 answers to the question, "What is AI used for?" (Figure 1a). Since all the answers are presented as noun phrases, it is relatively easy to find useful or interesting answers from them. It would be extremely difficult to find 800 answers for the same question by reading the documents provided by search engines. This feature of WISDOM X is expected to be useful for the discovery of relatively unknown ideas in AI applications, for instance, and for the creation of novel and innovative ideas using such unknown but already written ideas as hints. WISDOM X also enables us to search for relatively unknown future risks, such as the undesirable side effects of the Tokyo Olympic games in 2020.

Moreover, WISDOM X enables us to create valuable hypotheses, which are not described in our information source, i.e., 4-billion web pages. Figure 1c portrays the process of hypothesis creation. Initially, a user poses a question, "What happens if global warming worsens?" and one answer is that "the sea temperature will rise." If the user clicks on that answer, the system suggests another question, "What happens if the sea temperature rises?" and the answer includes "Vibrio parahaemolyticus swells." By repeating this process, the user can create the following hypothesis: "if global warming worsens, the sea temperature rises and an increase of food poisoning will be caused by Vibrio parahaemolyticus." This is actually a chain of causalities. Although we were unable to find any web pages that describe the entire hypothesis in our web archive, Baker-Austin et al. (2013) partly confirmed it.

In the above hypothesis creation process, the question suggestion played an important role. WISDOM X suggests other types of questions as well, including "Why does Vibrio parahaemolyticus swell if the sea temperature rises?" (Figure 1b) and "What is Vibrio parahaemolyticus?" The first question can be regarded as a question asking for textual support for the causality between the rise of the sea temperature and Vibrio parahaemolyticus. Such questions can be used to identify highly reliable answers among those provided by WISDOM X. In addition, when a user gives a keyword instead of a question as a query, WISDOM X lists answerable questions related to it. For example, when a user inputs "smartphone," WISDOM X suggests roughly five hundred questions, such as "What can smartphones resolve?"

WISDOM X sorts the answers according to their confidence scores, whose computation varies depending on the question type. For instance, the scores of answers to the why-type questions are provided by a supervised classifiers (Oh et al., 2013). In addition, semantically similar answers to factoid questions are grouped together as far as possible to help users to find answers that are valuable to them. The semantic similarities are computed using unsupervised word clustering (Kazama and Torisawa, 2008).

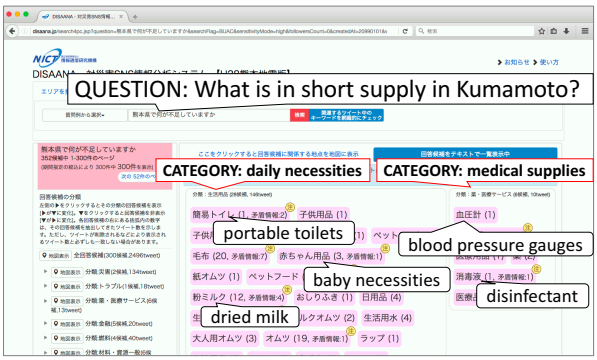
3 DISAANA and D-SUMM: Disaster Information Analyzer and Summarizer

DISAANA analyzes tweets in real time, discovers disaster-related information, and presents it in organized formats. It has two modes: QA and problem-listing. In the QA mode, for example, a user can enumerate goods in short supply in Kumamoto merely by asking, "What is in short supply in Kumamoto?" (Figure 2a). The answers are classified by such semantic categories as *medical supplies* for readability. A user can also enumerate them on a map (Figure 2c). In the problem-listing mode, a user can obtain a list of problems, such as "people were buried alive," which are occurring in a specified area (prefecture, city or town) without questions by using Varga et al. (2013)'s method (Figure 2b).

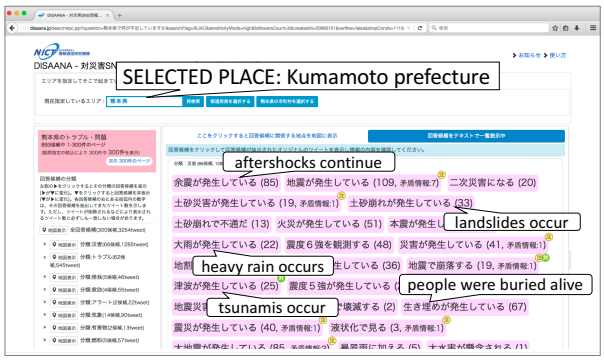
We constructed a million-scale location DB, which includes *part-of* relations between locations (e.g., Mashiki town is *part-of* Kumamoto prefecture) and each latitude and longitude of locations, to identify locations in tweets and display them related to the answers on a map. We did not use geotags attached to tweets because only a small fraction of them are actually geotagged due to privacy issues and the reported locations are often different from the locations from which users post tweets. When an area is specified in queries, the answers and problems related to the subparts are also presented to users. This is yet another function that has not been provided by commercial search engines.

One problem with DISAANA is that it often provides too many answers, which are difficult to grasp instantly. To address this problem, we developed D-SUMM (Figure 3), which summarizes the list of problems in a specified area provided by DISAANA. Similar problem reports such as "buildings collapsed" and "houses were demolished" were merged into a single problem report. In addition, the reports were classified according to their subparts of a specified area.

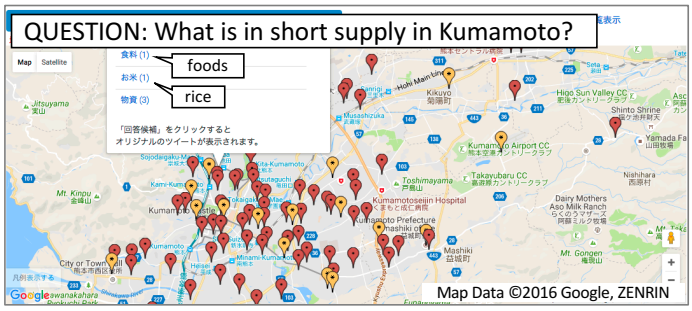
Another important issue is false rumors. In past disaster situations, numerous false rumors were spread widely on Twitter (e.g., "Drinking iodine protects against radiation" during the 2011 Great East Japan Earthquake). DISAANA and D-SUMM give an alert to users by retrieving not only answers but also information that contradicts the answers by using a modality analyzer (Mizuno et al., 2015) and contradictory patterns (Kloetzer et al., 2013). For example, when "acid rain" is one of the answers to the question, "What happened in a petrochemical complex?" and there is a tweet that contradicts the answer such as "Acid rain in the petrochemical complex is a false rumor," DISAANA presents it alongside the original tweet: the source of the answer. By examining such contradictory information, users can notice



(a) Results from the QA mode.



(b) Results from the problem-listing mode.



(c) Results from the QA mode located on a map.

Figure 2: Example screenshots of DISAANA.

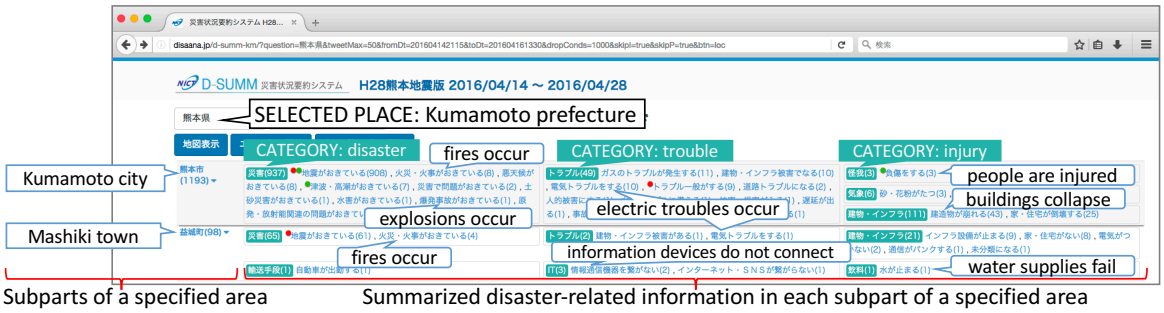


Figure 3: Example screenshot of D-SUMM.

the possibility of false rumors.

4 Conclusion

In this paper, we introduced three systems: WISDOM X, DISAANA and D-SUMM. We are going to add more intelligent functionality to the systems, such as more advanced reasoning mechanisms (Hashimoto et al., 2015) and such highly accurate linguistic analysis tools as anaphora resolution (Iida et al., 2016).

Acknowledgments

This work was partially supported by the Council for Science, Technology and Innovation (CSTI) through the Cross-ministerial Strategic Innovation Promotion Program (SIP), titled “Enhancement of societal resiliency against natural disasters” (Funding agency: JST).

References

Craig Baker-Austin, Joaquin A. Trinanes, Nick G. H. Taylor, Rachel Hartnell, Anja Siitonen, and Jaime Martinez-Urtaza. 2013. Emerging Vibrio risk at high latitudes in response to ocean warming. *Nature Climate Change*,

pages 3:73–77.

Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Stijn De Saeger, Masaki Murata, and Jun'ichi Kazama. 2009. Large-scale verb entailment acquisition from the web. In *Proceedings of EMNLP 2009*, pages 1172–1181.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Sager, Jun'ichi Kazama, and Sadao Kurohashi. 2011. Extracting paraphrases from definition sentences on the web. In *Proceedings of ACL-HLT 2011*, pages 1087–1097.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of EMNLP-CoNLL 2012*, pages 619–630.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of ACL 2014*, pages 987–997.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, and Jong-Hoon Oh. 2015. Generating event causality hypotheses through semantic relations. In *Proceedings of AAAI-15*, pages 2396–2403.

Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, Canasai Kruengkrai, and Julien Kloetzer. 2016. Intra-sentential subject zero anaphora resolution using multi-column convolutional neural network. In *Proceedings of EMNLP 2016 (to appear)*.

Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of ACL08: HLT*, pages 407–415.

Julien Kloetzer, Stijn De Saeger, Kentaro Torisawa, Chikara Hashimoto, Jong-Hoon Oh, and Kiyonori Ohtake. 2013. Two-stage method for large-scale acquisition of contradiction pattern pairs using entailment. In *Proceedings of EMNLP 2013*, pages 693–703.

Julien Kloetzer, Kentaro Torisawa, Chikara Hashimoto, and Jong-Hoon Oh. 2015. Large-scale acquisition of entailment pattern pairs by exploiting transitivity. In *Proceedings of EMNLP 2015*, pages 1649–1655.

Junta Mizuno, Canasai Kruengkrai, Kiyonori Ohtake, Chikara Hashimoto, Kentaro Torisawa, and Julien Kloetzer. 2015. Recognizing complex negation on Twitter. In *Proceedings of PACLIC 2015*, pages 544–552.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Takuya Kawada, Stijn De Saeger, Jun'ichi Kazama, and Yiou Wang. 2012. Why question answering using sentiment analysis and word classes. In *Proceedings of EMNLP-CoNLL 2012*, pages 368–378.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra- and inter-sentential causal relations. In *Proceedings of ACL 2013*, pages 1733–1743.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. 2016. A semi-supervised learning approach to why-question answering. In *Proceedings of AAAI-16*, pages 3022–3029.

Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata. 2009. Large scale relation acquisition using class dependent patterns. In *Proceedings of ICDM'09*, pages 764–769.

Stijn De Saeger, Kentaro Torisawa, Masaaki Tsuchida, Jun'ichi Kazama, Chikara Hashimoto, Ichiro Yamada, Jong-Hoon Oh, István Varga, and Yulan Yan. 2011. Relation acquisition using word classes and partial patterns. In *Proceedings of EMNLP 2011*, pages 825–835.

Motoki Sano, Kentaro Torisawa, Julien Kloetzer, Chikara Hashimoto, István Varga, and Jong-Hoon Oh. 2014. Million-scale derivation of semantic relations from a manually constructed predicate taxonomy. In *Proceedings of COLING 2014*, pages 1423–1434.

Masahiro Tanaka, Kenjiro Taura, and Kentaro Torisawa. 2016. Low latency and resource-aware program composition for large-scale data analysis. In *Proceedings of CCGrid 2016*, pages 325–330.

Kentaro Torisawa, Stijn de Saeger, Jun'ichi Kazama, Asuka Sumida, Daisuke Noguchi, Yasunari Kakizawa, Masaki Murata, Kow Kuroda, and Ichiro Yamada. 2010. Organizing the web's information explosion to discover unknown unknowns. *New Generation Computing*, 28(3):217–236.

István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of ACL 2013*, pages 1619–1629.