大規模情報分析システム WISDOM X, DISAANA, D-SUMM

水野 淳太　田仲 正弘　大竹 清敬　呉 鍾勲　Julien Kloetzer　橋本 力　鳥澤 健太郎
国立研究開発法人 情報通信研究機構 (NICT) データ駆動知能システム研究センター (DIRECT)
{junta-m, mtnk, kiyonori.ohtake, rovellia, julien, ch, torisawa}@nict.go.jp

1 はじめに

本稿では、NICT にて開発、一般公開している、自然言語処理技術を用いた 3 つの大規模情報分析システム WISDOM X、DISAANA、D-SUMM[1]の概要を述べ、その特徴をまとめる。

まず、WISDOM X は、約 40 億件のウェブ文書を情報源としたオープンドメイン質問応答システム（以下、QA システム）である。従来の検索エンジンとは異なり、「地球温暖化によって何が起きる？」といった自然文の質問を入力として、多種多様な回答（名詞句や動詞句）と、その抽出元のウェブページへのリンクを返す。本システムは、ユーザが考え得る様々な質問に対して、予想だにしなかった回答を含めて幅広に提示することができる。例えば、「地球温暖化」は将来的に壊滅的な被害をもたらす可能性のある深刻な問題であるため、個人では想定しきれないほど、想定される様々な被害がウェブ上に発信されている。しかしながら、従来のウェブ検索エンジンは、入力されたキーワードを含むユーザが読むべき文書への URL を羅列するだけであったため、多種多様な情報を効率良く把握することは困難であった。一方で、WISDOM X では、「地球温暖化が進むとどうなる？」といった簡単な質問を入力すると、ウェブ文書の情報をもとに、その質問の回答を瞬時に数百件提示することができる。こうした機能は、一個人の持つ知識を拡張、代替し、思考の幅を拡大するものと考えることができる。WISDOM X は、質問に対して多くて数百件の回答を提示するが、「回答」にフォーカスして提示するため、回答の全体像を把握することは容易である。また、それぞれの回答に関して、さらに深掘りする質問を提案する機能も備えており、より深い知識を得ることが可能である。

DISAANA と D-SUMM は、WISDOM X の技術をベースとし、災害発生時に被災者と救援者を支援するために開発した。2011 年の東日本大震災では、その発災直後から、大量の有用な情報が、Twitter などのソーシャルメディア上に発信された。しかしながら、あまりにも多くの情報が発信されため、効率的な情報取得が難しく、適切な意思決定への貢献は限定的であった他、デマ拡散なども発生した。DISAANA は、「○○市で何が不足している？」といった質問を入力すると、Twitter を情報源として、その回答を場所と紐付けて一覧化することができる。この結果、「ハラール食が不足している」といった事前に想定しづらい情報を含む、有用な情報を効率良く発見することができる。また、「どこで救助を求めているか」といった質問を入力すると、救助を求めているツイートを瞬時に発見することもできる。特に、後者に関して言えば、災害時には「救助」というキーワードを含んでいるものの、「早く救助してあげてください」といった、当事者以外からのツイートが大量に発信される。DISAANA はそうした情報の中から、詳細な地名を含むことやその他の言語的手がかりによって、当事者からの救助要請である可能性が高いツイートだけを発見、提示することができる。さらに、もう一つの機能として、県や市などの自治体名を指定するだけで、その自治体内で発生している災害・トラブル情報を一覧化することもできる。D-SUMM は、そうした DISAANA の出力を災害やトラブルの種類に基づいてまとめ上げ、地域ごとに分類、要約してコンパクトな形で提示するシステムである。2016 年の熊本地震においては、首相官邸（内閣官房）が DISAANA を実際に活用し、被災地でのニーズの分析を行っており、その有用性が報道されている。[2]

2 情報分析システム WISDOM X

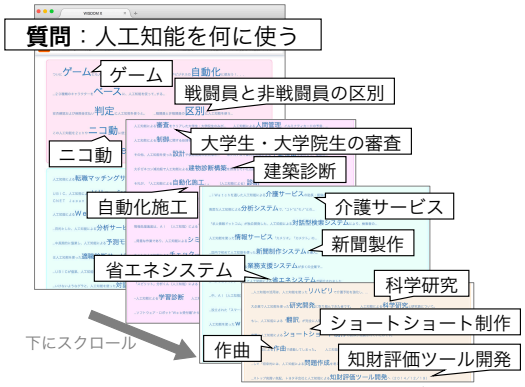
WISDOM X は、約 40 億件のウェブ文書に対して、種々の言語解析を行い、回答を抽出するシステムであり、ファクトイド型 QA システム（例えば、「何で地球温暖化を防ぐ？」）、なぜ型 QA システム（例えば、「なぜ地球温暖化が起きる？」）[1, 2, 3]、どうなる型 QA システム（例えば、「地球温暖化が進むとどうなる？」）[4, 5]、定義型 QA システム（例えば、「地球温暖化とは何？」）の 4 つの QA システムからなる。さらに、ユーザに質問を提案する機能を持っている。これらのシステムは、約 3 億件のエントリを含む大規模な含意関係知識 [6, 7, 8, 9, 10, 11, 12] と、名詞の意味クラスタ [13] を利用している。さらに、本システムの構築にあたっては、種々の言語解析器を数百台の計算機上で効率良く運用するために、ミドルウェアである RaSC [3] [14, 15] も開発した。

WISDOM X では、ファクトイド型質問やどうなる型質問に対して、名詞句や動詞句などのピンポイントな回答を得られる。これは、従来の検索エンジンと大きく異なる点である。検索エンジンは、入力されたキーワードに関連する文書の URL を提供するだけであり、回答部分を特定するにはユーザによる労力が必要となる。それに対して、WISDOM X では、例えば「人工知能を何に使う？」という質問を入力すると、約 800 件の回答が得られる。図 1a にその一部を示す。回答は、同

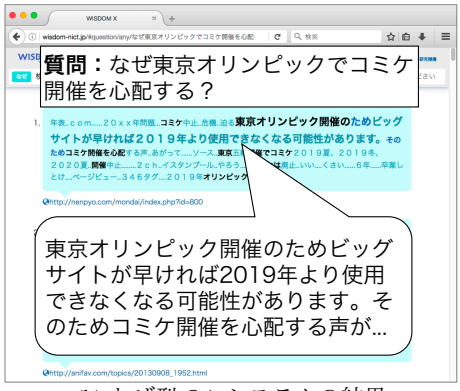
[1] それぞれ、http://wisdom-nict.jp、http://disaana.jp、http://disaana.jp/d-summ で一般公開中。

[2] 「つぶやき分析 ニーズ把握」読売新聞夕刊 1 面 2016/5/11、「被災情報ツイッターで集約」西日本新聞 2016/6/12

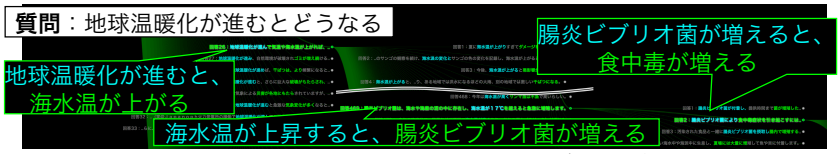
[3] https://alaginrc.nict.go.jp/rasc/ja/



(a) ファクトイド型 QA システムの結果



(b) なぜ型 QA システムの結果



(c) どうなる型 QA システムを用いた仮説生成

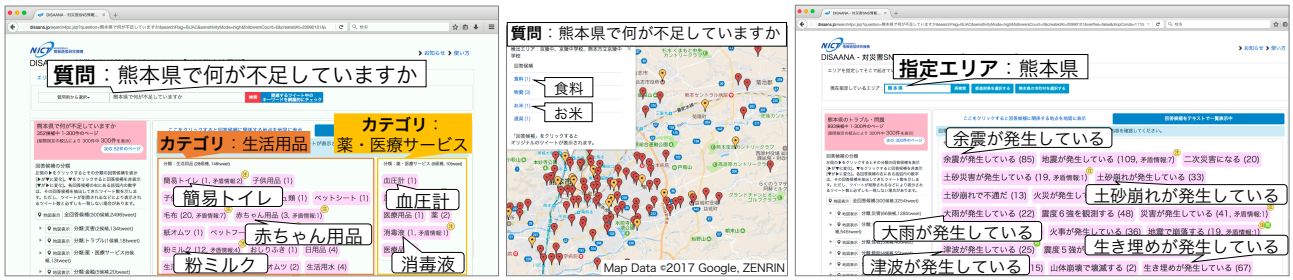
図 1: WISDOM X のスクリーンショット

一のものは一つにまとめられて表示される他、全てハイライトされた名詞句とその前後の短い文字列からなるため、有用あるいは興味深い回答を素早く発見することができる。一方で、同量の回答を、従来の検索エンジンを用いて発見することは、ほとんど不可能であろう。本機能で得られた回答は、新製品のヒントや小説のネタなど、新たなアイディアのヒントに繋がることを期待している。また、WISDOM X では将来的なリスクを発見することもできる。例えば、2020 年に開催予定の東京オリンピックで心配するべき事柄は、「東京オリンピックで何を心配すべきか?」という質問によって得ることができる。回答には、「資材高騰」や「テロ」といった想定しやすいリスクだけでなく、「コミケ開催」や「区の水泳大会の開催」といった、容易には想定できないリスクも含まれている。また、各々の回答をクリックすると、さらに情報を深掘りする質問等が提示されるが、その中には例えば「なぜ東京オリンピックでコミケ開催を心配する?」といった、なぜ型質問もあり、それをクリックすることで、図 1b に示すように、回答のより深い根拠を発見することも可能である。こうした機能は情報の信ぴょう性を判断する材料を提供するものと考えることもできる。

WISDOM X では、どうなる型 QA システムによって仮説を生成することもできる。生成された仮説は、情報源のウェブ文書にも含まれていないことがしばしばある。図 1c に仮説生成の例を示す。まず、どうなる型 QA システムに「地球温暖化が進むとどうなる?」を入力すると、「海水温が上昇する」といった回答が得られる。その回答をクリックすると、システムは他の質問として「海水温が上昇するとどうなる?」を提案し、その回答として「腸炎ビブリオ菌が増える」が得られる。この手順を繰り返すことで、最終的に「地球温暖化が進むと海水温が上昇し、腸炎ビブリオ菌が増え、食中毒が増える」という仮説を生成することができる。これはつまるところ因果関係の連鎖であり、連鎖全体を見渡すと、「地球温暖化が進むと食中毒が増える」という仮説が提示されたものと考えることができる。本仮説が最初に見つかったのは、2007 年までに収集した約 6 億件のウェブ文書からであるが、これらの中には、本仮説全体が記載された文書は存在しなかった。一方で、Baker-Austin ら [16] は、2013 年に、この仮説が部分的に事実であることを報告した。これはつまり、権威ある論文誌で事実として報告された仮説を、部分的とはいえ、ウェブ文書の情報を組み合わせで作ることができたということになる。

前にも述べたように、WISDOM X は様々な質問を提案する。上の仮説生成の途上では、なぜ型質問の「なぜ海水温が上昇すると腸炎ビブリオ菌が増える?」や、定義型質問の「腸炎ビブリオ菌とは何か?」といった質問が提案される。一つ目の質問は、仮説中の因果関係（この例では「海水温が上昇する」と「腸炎ビブリオ菌が増える」の因果関係）について、その根拠を聞く質問になっており、その回答を閲覧することで、より信頼性の高い情報を選んで仮説を生成することができる。さらに、WISDOM X に質問ではなくキーワードを入力すると、そのキーワードに関連し、回答が得られる質問を提案する。例えば、「スマートフォン」と入力すると、「スマートフォンによって何が解決する?」といった質問が約 500 個得られる。ユーザは、これらの質問と回答を閲覧することで、入力したキーワードに関する深い知識を得ることができる。

WISDOM X が出力する回答は、その確信度によってソートされている。例えば、なぜ型 QA システムでは、その中核である機械学習器 [2] のスコアによってソートされる。さらに、ファクトイド型 QA システムでは、意味的に類似する回答をまとめあげて、有用な回答を見つけやすくしている。なお、回答の類似性は、教師無しクラスタリング [13] によって計算している。



(a) QA モードの結果　(b) QA モードの地図表示　(c) エリア検索モードの結果

図 2: DISAANA のスクリーンショット

図 3: D-SUMM のスクリーンショット

3 被災状況を把握するための DISAANA と D-SUMM

DISAANA は、ツイートをリアルタイムに解析し[4]、災害関連情報を発見するためのシステムである。DISAANA には、QA モードとエリア検索モードの 2 つのモードがある。QA モードでは、図 2a に示すように、「熊本県で何が不足している？」という質問を入力すると、熊本県で不足している物資を一覧化することができる。回答は、閲覧性を向上させるために、「生活用品」や「薬・医療サービス」といった意味的なカテゴリに分類して表示する。図 2b は、これらの回答を、地図上に表示したものである。地図上には、不足物資があると報告のあった場所に対してピンを立てており、そのピンから、不足物資を閲覧することができる。

エリア検索モード（図 2c）では、迅速に被災状況を把握するために、質問を入力するのではなく県や市などの自治体（この例では熊本県）を指定するだけで、「生き埋めが発生している」や「電話が繋がらない」といった災害・トラブル情報を一覧化できる。本モードは、Varga ら [17] による、要望・対策の自動抽出技術を利用して実現した。本モードの回答には、後述の地名データベースを利用することで、ユーザが指定した自治体だけでなく、その下位区分[5]で発生した情報も含まれる。例えば、「熊本県」と指定したとき、「熊本県」が陽に記述されていない「益城町で電話が繋がらなくなっている」といったツイートから、益城町は熊本県の一部であるという関係を参照しつつ、「電話が繋がらない」というトラブルを抽出する。この機能は、従来の検索エンジンでは実現されていない。

DISAANA を実現するために、Wikipedia や電話帳等を利用し、数百万エントリからなる地名データベースを構築した。このデータベースには、「熊本県」に対する「益城町」といった部分全体関係と、地名やランドマークに対する緯度経度情報が含まれており、ツイート中に表れる地名を認識し、地図上に表示するために利用される。ツイートは投稿時に位置情報を付与することができるが、我々はこの情報を利用していない。その理由は二つあり、一つ目は、位置情報が付与されているツイートが非常に少ないためである。二つ目は、ツイートを投稿する場所と、投稿内容の地名は、しばしば異なるためである。これは、被災情報を避難先で投稿するケースや、被災者から電話などで得た情報を別の場所で投稿するケースなどによって生じる。

DISAANA のエリア検索モードには、特に大規模災害発生時において、非常に多くの回答が得られてしまい、状況の把握が困難になるという問題がある。また、指定された自治体内で発生している災害・トラブル情報を、その自治体内のどこで発生しているかを考慮することなくリスト表示するため、複数の下位区分について情報を収集するためには、複数回検索する必要がある。これらの問題を解消するために、D-SUMM を開発した（図 3）。一つ目の問題に対して、D-SUMM は、半自動的に整備した数千万エントリからなる災害オントロジーを用いて、類似情報をまとめる。例えば、「ビルが倒壊」と「家が壊れた」は、「建造物が崩れる」という回答にまとめられる。二つ目の問題に対して、D-SUMM では、災害・トラブル情報を、指定された自治体の下位区分ごとに分類し、整理された形で提示する。下位区分は、画面上では災害・トラブル情報の量（ツイート数）の降順でソートされるため、被害が重大な地域

[4]DISAANA で利用可能なツイートは、日本語による全ての投稿の約 10%であり、投稿されたら即時に解析され検索可能になる。利用日を含めて間近 4 日間のツイートが解析対象となる。

[5]例えば、最初に指定したのが「熊本県」であれば、「熊本市」「益城町」が下位区分になる

を一目で発見することが可能である。また、D-SUMM を使うと、DISAANA のようにキーボードから質問を入力しなくても、重大な被災情報を簡単に発見することができる。例えば「熊本県」で発生した救助要請のツイートは、1. D-SUMM で「熊本県」を指定して要約された情報を提示し、2. 各市町村区の「救助」カテゴリをクリック、3. 情報抽出元となったツイートをチェックという 3 ステップを踏むことで、即座に発見することができる。

また、発災時に生じる重大な問題としては、いわゆるデマの流布がある。例えば、東日本大震災では、「イソジンを飲んで被曝を防ぐ」というデマが広まった。DISAANA および D-SUMM は、回答を検索すると同時に、回答と矛盾する内容を含むツイートも検索する。矛盾ツイートが見つかると、回答の情報抽出源のツイートとあわせてユーザに提示することで、その回答がデマである可能性があることを示す。このような矛盾情報は、回答が誤情報であるかをユーザが検証するための重要な手がかりとなる [18]。矛盾ツイートは、否定や推量といったモダリティの認識 [19] と、述語間の矛盾関係知識 [10] を用いることで実現した。例えば、「石油コンビナートで何が発生している？」という質問に対して、「酸性雨が降る」といったツイートから、「酸性雨」が回答の一つとして得られる。同時に、「発生していないもの」も検索する。その結果、「酸性雨が降るというのはデマ（つまり酸性雨は発生していない）」というツイートを発見することができる。

4 おわりに

本稿では、3 つの大規模情報分析システム WISDOM X、DISAANA、D-SUMM についてその概要と特徴を述べた。今後は、より高精度の仮説生成 [20] や、照応解析 [21] といった深い言語解析の導入、深層学習による各解析器の性能向上 [22, 23] などを考えている。

謝辞：本研究の一部は、総合科学技術・イノベーション会議の SIP（戦略的イノベーション創造プログラム「レジリエントな防災・減災機能の強化」（管理法人：JST）によって実施された。

参考文献

[1] 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 EMNLP-CoNLL 2012*, pp. 368–378, 2012.

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

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

[4] 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 EMNLP-CoNLL 2012*, pp. 619–630, 2012.

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

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

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

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

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

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

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

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

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

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

[15] Masahiro Tanaka, Kenjiro Taura, and Kentaro Torisawa. Autonomic resource management for program orchestration in large-scale data analysis. In *Proceedings of IPDPS 2017*, 2017 (to appear).

[16] Craig Baker-Austin, Joaquin A. Trinanes, Nick G. H. Taylor, Rachel Hartnell, Anja Siitonen, and Jaime Martinez-Urtaza. Emerging Vibrio risk at high latitudes in response to ocean warming. *Nature Climate Change*, pp. 3:73–77, 2013.

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

[18] 村上浩司, 松吉俊, 隅田飛鳥, 森田啓, 佐尾ちとせ, 増田祥子, 松本裕治, 乾健太郎. 言論マップ生成課題：言説間の類似・対立の構造を捉えるために. 情報処理学会研究報告 2008-NL-186, pp. 55–60, 2008.

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

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

[21] Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, Canasai Kruengkrai, and Julien Kloetzer. Intra-sentential subject zero anaphora resolution using multi-column convolutional neural network. In *Proceedings of EMNLP 2016*, pp. 1244–1254, 2016.

[22] Jong-Hoon Oh, Kentaro Torisawa, Canasai Kruengkrai, Ryu Iida, and Julien Kloetzer. Multi-column convolutional neural networks with causality-attention for why-question answering. In *Proceedings of WSDM 2017*, 2017 (to appear).

[23] Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *Proceedings of AAAI-17*, 2017 (to appear).