3.5.3　ユニバーサルコミュニケーション研究所　情報分析研究室

室長　　鳥澤健太郎　ほか 11 名

ネットの情報を深く分析する

【概　要】

インターネット上には膨大な情報が存在し、多くの人々が検索エンジンを用いてそうした情報にアクセスしている。一方で、情報の間には様々なつながりがあり、本来はそうしたつながりをきちんと見ていくことで初めて本当の意味で情報を活用することが可能になる。例えば、ある出来事を示す情報があったとして、その出来事の帰結（その出来事が起きると次になにが起きるのか？）や原因（なぜそれが起きたのか？）が分かれば、将来の潜在的リスクやチャンスを知ることにつながり、意思決定をする際に有効である。また、最初に見つけた情報と、それらの帰結、原因といった情報のつながりはインターネット上で明示的に書かれているとは限らず、情報システムが「考えて」、仮説として原因や帰結をユーザに提供する必要もある。例えば、地球温暖化の潜在的な帰結には膨大な可能性があり、その中にはいまだ誰も検討していないが、将来には現実的な脅威となり得るようなものもあるであろう。

こうした将来のリスクを前もって調べ、それによるダメージを軽減するためには、先に述べたような「仮説」の生成が必要となる。本中長期目標期間中ではこうした情報の分析、仮説の生成を大量の Web 文書をもとに行えるシステムである大規模 Web 情報分析システム WISDOM X を開発し、平成 26 年度にインターネット上で一般に公開した。また、耐災害 ICT 研究センター 情報配信基盤研究室と共同でツイッター上の災害関連情報の分析を行う対災害 SNS 情報分析システム DISAANA の開発、公開、自治体と連携しての実証実験を実施した。また、同様に情報配信基盤研究室と連携し、内閣府の戦略的イノベーション創造プログラム（SIP）の支援を受けて大規模災害時の膨大な災害関連情報を A4 サイズの用紙 1 枚程度に要約するシステムである D-SUMM の研究開発も行い、順調に進捗している。（DISAANA、D-SUMM に関しては、**3.11.3** 情報配信基盤研究室の項で詳述する。）

【平成 27 年度の成果】

平成 27 年度は平成 26 年度に公開した WISDOM X の拡張、改良を行い、また、次期中長期計画をにらんだ新規な課題の研究に取り組んだ。

【WISDOM X の分析対象 Web ページの増加】公開開始時点の平成 27 年 3 月 31 日、WISDOM X の分析対象、つまり、質問への回答等で情報源として利用する Web 文書は、約 10 億件であったが、その後、平成 27 年 11 月に、それを 40 億件へと増大させた。平成 27 年 3 月末現在公開しているバージョンも 40 億件の Web 文書を分析対象とし、日々ページを更新している。

【WISDOM X に音声・画像の取り扱いを可能とさせる拡張】WISDOM X に拡張を加え、動画中の人の発言を音声認識によってテキスト化し、そのテキストを質問応答で使うことを可能とした。また、質問の回答が、Wikipedia の記事の見出しになっている単語である場合には、その Wikipedia の画像を回答とともに提供できるように拡張を行った（図 1 参照）。ただし、現在、動画、画像は著作権等で問題があるケースも多いため、公

図 1　WISDOM X で出力された画像の回答

開版ではそれらの機能は提供しておらず、NICT 内でのデモンストレーションでの利用に限定している。

【WISDOM X のオープン化に必要なミドルウェアの改良】 WISDOM X の大規模情報分析機能を支える処理基盤、ミドルウェア RaSC を大幅に機能強化した。具体的には、多数の計算機上で分散並列実行される分析プログラムのプロファイルを自動的に収集しつつ、各分析プログラムの並列実行数を自動的にチューニングする機構を実現した。これにより、大規模情報分析における技術やノウハウを十分に持たない他組織においても、WISDOM X が必要とする分析環境を、簡単かつより少ない計算機資源で構築できる。また、仮想マシンの管理機能を強化し、計算機環境を問わず、様々な分析プログラムをより小さなオーバヘッドで稼動させられるようにすると共に、従来数十分を要していたような、新しい計算機環境への分析プログラムの設置を、数秒～数十秒程度と大幅に高速化した。なお、RaSC に関しては、Amazon EC2 を用いた技術チュートリアルを電子情報通信学会で行っており、普及活動も進めている。

【文脈処理技術の研究開発】 文脈処理の主要な問題である、省略された主語等を復元する省略解析技術に関して、平成 26 年度に研究開発した、一文中に現れる述語のペア（例：「血栓を溶かし、脳梗塞を予防する」という文中の 2 つの述語、「溶かす」と「予防する」）の間で、主語が共有されるかどうかを自動で判断する手法の適用方法を拡張し、新しい省略解析手法を研究開発した。また、その出力を既存の省略解析システムに統合することで最終的な性能向上を実現した。この手法では、同一文内で主語を共有する述語のネットワークを構築し、そのネットワーク内で主語を伝播することで、省略の復元を行う。例えば、「政府は 50 人を被災地に（X が）派遣することを（X が）決め、準備を進めている」（「X が」が省略された主語を表す）に出現する「派遣する」「決める」「進める」の 3 つの述語について、主語を共有するかどうかを判定する技術を適用、3 つの述語が同じ主語を持つことを認識することでこの 3 つの述語の主語共有に関するネットワークを構築し、さらに、そのネットワークを通じて主語「政府」を「派遣する」や「決める」の省略された主語の位置に伝播することで、省略の復元を行う。

【WISDOM X の質問サジェスト機能の強化及び質問応答の精度向上】 WISDOM X は、単語で回答可能な質問（例：「地球温暖化を防ぐのは何か？」）に答える「なに」型質問応答機能、（複数の）文で表現される理由や原因を問う質問（例：「なぜ日本はデフレに陥ったか？」）に回答する「なぜ」型質問応答機能、文あるいは文の連鎖で表現される未来シナリオを問う質問（例：「人工知能が進化するとどうなる？」）に回答する「どうなる」型質問応答機能等、複数の質問応答機能から構成されている。平成 26 年度までは、「なに」型質問応答及び、「どうなる」型質問応答の 2 つに関して、キーワードを入力すると回答可能な質問を列挙する質問サジェスト機構を開発、公開版で稼働していた。平成 27 年度は新たに「なぜ」型質問応答に対して質問サジェスト機構を開発した。これは「近年世界中で CO_2 が大量に排出されている。その結果、地球温暖化が進行している。」といった文章に書かれた因果関係を自動的に特定し、その結果部（地球温暖化が進行している。）を質問（なぜ地球温暖化が進行しているか？）へと変換し、もとの文章に類似した文章がその質問の回答として提示するものである。問題は、自動的に特定された因果関係の結果部は必ずしも質問に変換ができないことである。例えば、「地球温暖化が話題になっている。近年 CO_2 が大量に排出されているため、クローズアップされているのである。」といった文章からは、因果関係として原因：「CO_2 が大量に排出されている」→結果：「クローズアップされている」といったものが認識される。この場合、結果部から質問を生成すると、「クローズアップされているのはなぜ？」といった意味をなさない質問が生成されてしまう。こうした問題に対処するために質問として適切な言語技術を開発して対応した。また、この質問サジェスト機構を利用し、「なぜ」型質問応答の精度を半教師あり学習で向上させることに成功した。これは、システムが自ら提示した回答の正否を回答が含まれている文書とは異なる文書の情報をもとに検証し、その検証結果を自らが賢くなるために利用する手法ととらえることができる。つまり、システムが自律的に賢くなる技術であり、こうしたアプローチは次期中長期計画でも大きな柱となるものと考えている。

【より広範な仮説生成に向けた推論規則の自動獲得】 「X がポリフェノールを含む」→「X が脳梗塞を防ぐ」といった変数を含むテキスト間の推論規則を獲得する技術を開発した。これは、これまでに開発してきた「ポリフェノールを含む→脳梗塞に効く」といった因果関係を自動獲得する技術と、上で述べた文脈処理技術を組み合わせることで開発できたものである。こうした推論規則は例えば「何が脳梗塞に効くか？」といった質問に「赤ワインが脳梗塞に効果的である」といった直接回答を表す表現がなかったとしても、「赤ワインがポリフェノールを含む」といったテキストから仮説としての回答を見つけ出すことを可能にする。次期中長期計画では、こうした推論規則のより複雑なものを獲得し、さらにはそれらを複数組み合わせた推論を行うことで、よりユーザにとって有用な仮説を提示できる技術を開発する予定である。