5　言語基盤・情報分析技術

5 Language Infrastructure and Information Analysis Technology

5-1　情報分析技術の概要

5-1 Information Analysis Technologies at NICT

鳥澤健太郎
TORISAWA Kentaro

要旨

NICT では平成 18 年度から平成 22 年度の第 2 期中期目標期間及び、平成 23 年度から平成 27 年度に至る第 3 期中期目標期間に大量のテキスト情報などを自動的に分析する情報分析技術の開発に取り組んでいる。こうした研究開発の成果は様々な観点から Web 上の情報を分析し、また取得されたその妥当性を検証する材料を提供することを可能とし、一般にむけて公開されている Web 上の情報分析システム WISDOM や、高度な言語情報の解析を目指して、高度言語情報融合フォーラム（ALAGIN）などで公開されている言語資源、ツールなどとして社会で利用可能となりつつある。本稿ではそうした技術の概要、狙いについて述べる。

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.

[キーワード]
情報分析，自然言語処理，Word Wide Web，テキストマイニング，質問応答
Information analysis, Natural language processing, World Wide Web, Text mining, Question answering

1　はじめに

いわゆるインターネット上の情報爆発には収束の気配はなく、インターネット上に存在するいわゆる Big Data から価値を創出することは全世界的に課題と見なされている。こうした観点から、NICT においてもインターネット上の大量の情報を分析する技術、方法論の研究が平成 18 年度より進められてきた。その具体的成果としては、一般公開されている情報分析システム WISDOM や、高度言語情報融合（ALAGIN フォーラム）などにおいて公開されている各種の言語資源、ツール、サービス、さらには音声質問応答システム「一休」などが挙げられる。これらの具体的成果の内容については、本特集の他の論文が解説を行っているので、本稿では深く解説することはしない。

本稿ではむしろ、こうした技術の背景にあるものや、今後の研究開発の方向性について述べる。

2 テキストを深く解析する技術

NICT の情報分析技術の特徴は、「テキスト、文書をより深く解析する」ということである。ほとんどの検索エンジンを初めとして、インターネット上の情報にアクセスする手段の多くはいわゆるキーワード検索をベースとしている。これは、大量の Web ページ等から、ユーザの指定したキーワードを含む文書を選び出し、ランク付けした上でユーザに提示するものである。これらの技術は多くの場合、単語の意味を一定程度考慮しつつ、例えば、「Apple」と検索した場合、「アップル」を含む文書も提示するなどのいわゆる「クエリー拡張」という処理も行うが、むしろ技術の焦点はそうしたキーワードを含む文書をランキングする技術にある。こうした技術の代表として良く取り沙汰されるのが、Google の PageRank と呼ばれる技術である。しかしながら、この PageRank は Web ページに特有の機能であるリンクを利用するものであり、同じ Web ページに書かれているテキストを深く処理する訳ではない。

一方で、NICT の目指す情報分析技術は、テキストの表す意味、内容をより深く分析することを狙う。例えば、情報分析システム WISDOM では与えられたキーワードに関して、そのキーワードが指す対象を肯定的に評価している情報、否定的に評価している情報などを分類した上で列挙しているが、これは単にキーワードが文章に含まれているかどうかだけではなくて、文の文法的構造を解析し、肯定的あるいは否定的に評価をしているフレーズを機械学習によって特定しているのである。また、同様に、文書の構造等の分析によって、その文書を発信している発信者が、匿名なのか、企業なのかなどを示す発信者情報も自動的に取得できるようになっており、例えば、医療組織では「否定的意見」が多いが、企業からの発信では「肯定的意見」が多いといった、社会におけるキーワードの受容が簡単に分析できることになる。実際に企業サイトで大変肯定的な評価が発信されている食品に関連して、医療関係サイトで死亡事故（正確には類似品による死亡事故）がおきているといった事例も見つかっている。これはつまり、ある対象の肯定的評価、否定的評価を合わせて提示することで、そうした情報の信頼性を判断する手がかりを提供していることになる。WISDOM の詳細については、本特集 **5-3**「情報分析システム WISDOM」を参照されたい。

また、音声質問応答システム「一休」では、ユーザがスマートフォンに対して音声で「デフレを引き起こすのは何ですか？」といった質問を行うと、億単位の Web ページから得られる情報を基にその質問に対する回答をリストアップする。検索エンジンで同様の情報を得ようとしても、「デフレ」、「原因」といったキーワードを入力し、表示された膨大な文書を自ら読んで、具体的な原因を特定するより他ない。また、「デフレ」「原因」というキーワードでは、「デフレの原因」なのか、「デフレが原因となる別の事象」なのか、ユーザの意図が伝えきれず、結果として、読まなければいけない文書もさらに増える、といった問題も生じる。一方で「一休」が提示する回答は質問に対する端的な回答となっている単語ないしはフレーズであり、例えば、デフレの場合のように、様々な原因が考えられる場合には、大量の回答が表示されるが、各々の回答が非常に短いため、それら全体を概観し、問題の全体像や興味深い事例を把握することが容易である。例えば、一休はデフレの原因として、日本を代表するある大企業の名称を回答として提示した。一見ナンセンスに見える回答であったが、システムが回答を抽出した文書を一休の提示したリンクから辿ったところ、一応、「巨額の利益を内部留保に回し、資金が市場に出回るのを妨げた」という論理的な根拠も提示されていた。（この状況を示すデモビデオが http://www2.nict.go.jp/univ-com/info_analysis/ にあるので参照されたい。）我々がこの回答を発見した後、同主旨のロジックと日本企業の内部留保の総額が 200 兆円にのぼるというデータに基づき、デフレの原因としてその企業を挙げる記事が経済雑誌に実際に掲載されたことは、回答が抽出されたページが一般人の書いた匿名のブログであったことも合わせて、ネット社会の一部における一般人の情報の受容・理解の高度さや、現代における経済の複雑さを示唆するようで非常に興味深い。また、一休のもともとの開発の意図は、そ

の前から開発されてきた概念辞書、検索支援システム「鳥式改」のコンセプトに基づくもので、「意外でありながら有用な情報の発見」を支援することが狙いであったことを補足しておく[1][2]。

一休が文書ではなく、質問の回答を端的にリストアップできるのは、やはり「テキストを深く解析する」ことによる。具体的には、テキスト中から「X が Y を引き起こす」「X が Y を悪化させる」「X による Y」といったパターンを、変数 X、Y にマッチさせる名詞の対、例えば、「グローバライゼーション」と「デフレ」、ある企業名と「デフレ」といったものを抽出し、前もって一種のデータベースに保存しており、また、「X が Y を引き起こす」「X が Y を悪化させる」「X による Y」といったパターンがほぼ同義であることを自動的に認識しているからである。これにより、「何がデフレを引き起こしますか？」といった質問への回答を例えば、一見かけ離れた「グローバライゼーションによるデフレ」といった表現から抽出することが可能となる。まず、こうした情報の抽出を行うためには、意味をなすパターンを特定するために、やはり文の文法的構造を認識することが必要である。例えば、「ハウスダストが例えばアトピーを引き起こします」といった文からも「X が Y を引き起こす」というパターンが抽出できてほしいが、その場合「例えば」というような表現は重要でない情報としてパターンから削除しても良い、ということが認識できる必要がある。こうした処理に文法的構造の認識、いわゆる構文解析は必要である。また、「X が Y を引き起こす」と「X による Y」という表現が同義であると述べたが、これは一般的には正しくない。例えば、X = Apple、Y = iPhone と仮定した「Apple による iPhone」という表現は「Apple が iPhone を引き起こす」と言い換えることは出来ない。同義性がいえるのは、X、Y に来る名詞が特定のタイプの場合だけである。例えば、X がホルムアルデヒドのような化学物質、Y がアトピーのような病名の場合は非常に高い確率で同義であるといってよいであろう。一休ではこのような単語のタイプ、意味的分類を自動的に計算し、考慮した上でパターン間の同義性を自動認識しており、これ自体深くテキストを分析している事例と考えることができるが、こうした意味分類でもやはり、テキストの文法構造の認識が重要である。一休のさらなる詳細については、本特集の **5-2**「音声質問応答システム一休」を参照されたい。また、やはりここで詳細に立ち入ることはできないが、上述した単語の自動的な意味分類の計算結果やパターンの同義性の認識結果などは、高度言語情報融合フォーラム（ALAGIN）において言語資源として公開されている。これらの詳細については、本特集 **5-5**「基盤的言語資源」、ならびに **8-1**「高度言語情報融合フォーラム（ALAGIN）」を参照されたい。

また、一休は現在「Why 型質問」への回答が出来るようになるよう、拡張が進められている[3]。この Why 型質問への回答は、単語ではなく、文章でなされるべきものであり、米国においてクイズショーの人間のチャンピオンに勝ったことで一躍有名になった IBM の Watson でも現状答えられない難しいタスクである。一休も、未だ全体的な精度は高くないが、例えば、現在は「ガダルカナル島で米軍に負けたのはなぜですか？」といった質問に対して「兵力の逐次投入」「前線と基地の間の距離」等に言及しつつ、歴史的経緯を解説した文章を回答することなどが出来ている。本稿では詳細に立ち入るスペースはないが、このタスクでは WISDOM で使われた技術、上で述べた様々な一休で使われてきた技術、すなわち様々な深いテキストの分析処理を統合して回答を行う。つまり、テキストの深い解析をすることによって、Why 型質問への回答のような難しいタスクを行うシステムも実現可能になりつつあるということである。

3　今後の研究

さて、これまでに述べてきたように、億単位の Web ページから必要な情報を質問に対する端的な回答のリストという形で取得し、その全体像を把握し、意外でありながら有用な回答を発見するなどの操作や、あるキーワードの肯定的／否定的評価、発信者の情報の分析のように、Web における傾向をある観点から分析することは可能になりつつある。

一方で現状技術の重要な課題として挙げられるのは、上述して来たような分析はあくまでユーザ

が適切な質問、クエリーを与えるなどの操作をして初めて効果を発揮するということである。例えば、一休は昨年の震災以前の Web ページから、「津波が過去に襲った場所」として仙台平野を提示することができた。これは、震災後に有名になった情報であり、ある研究所の Web ページで報告されていた地質調査の結果判明した約 1,000 年前の地震による津波のことである。もしこの情報が震災前により広く普及しており、それに基づいて市民がさらなる安全対策を要求することなどにより、より適切な安全対策、防災対策が取られた可能性もあったかもしれない。しかしながら、現状技術のみを利用した場合、震災前に仙台エリアの防災対策の調査、あるいは原子力発電所の安全性を調査していたユーザがこのような情報に接する可能性は高くない。つまり、そうした漠然とした安全性の調査というタスクから、「津波が過去に襲った場所」を尋ねる質問を質問応答システムに与えるという操作に至るまでは大分距離があるからである。まず、人間であれば、常識として備えている知識、すなわち「過去に津波が襲った場所は再度津波に襲われる可能性が高いこと」「仙台平野に隣接したエリアに原子力発電所が存在すること」を現状のシステムは備えておらず、例えば、「＊原子力発電所の安全性は確保されているか？」といった質問に対して、上述の仙台平野を襲った津波を関連情報として提供することは不可能である。

現在、我々は上述したような現状のシステムが持っていない常識的知識を大量の Web ページから自動獲得させる研究を行っており、最終的には、「津波は同じ場所を繰り返し襲う」といった常識的知識から、上述したように「原子力発電所の安全性調査」には、「近隣を過去に襲った津波の情報」を提供するといったシステムを構築したいと考えている。これはある意味でユーザの要求、意図を先回りして、より広範な情報を提供することになり、最終的にはユーザにより適切な意

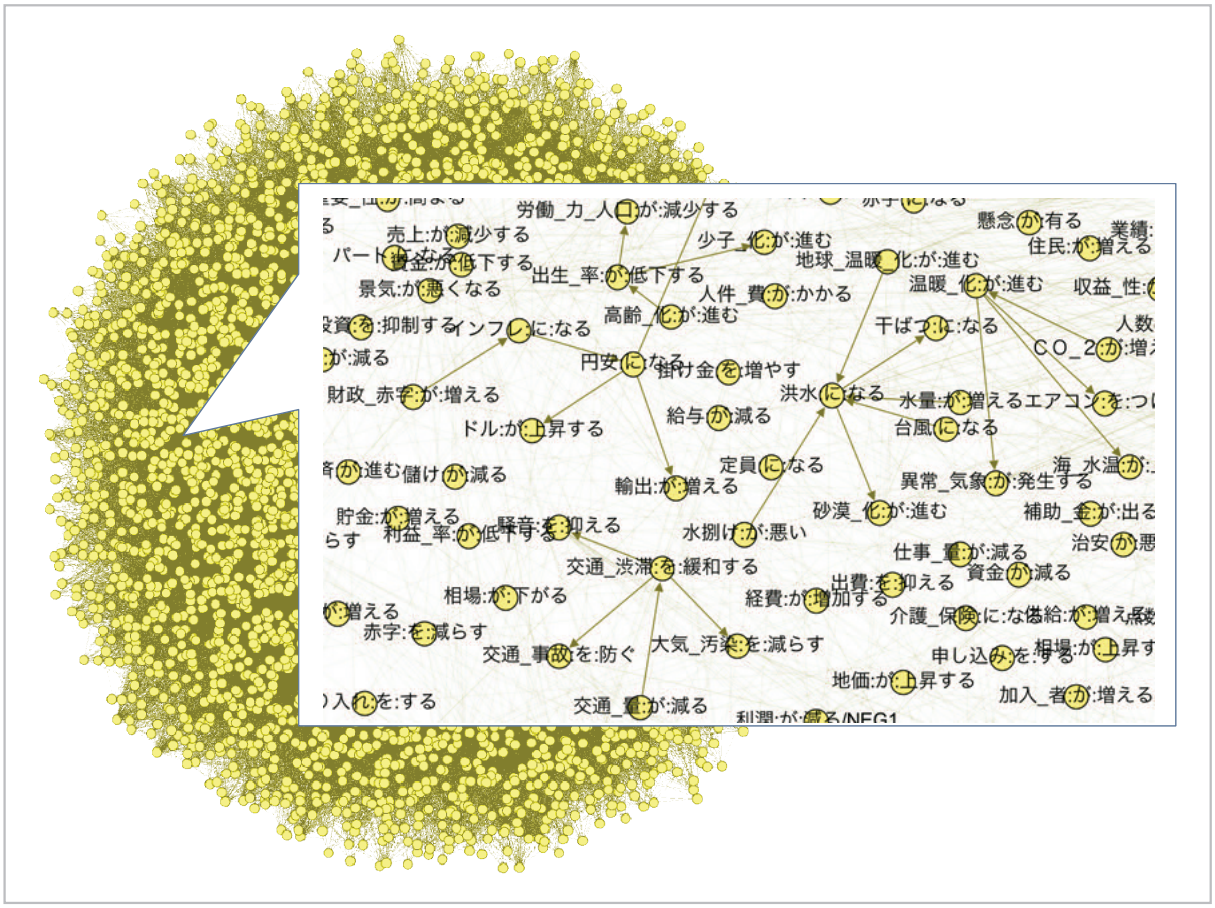

図 1 Web から抽出した因果関係のネットワーク

思決定を促すことになろう。図1はこうした常識的知識の自動獲得手法[5]の結果を示しているが、これはWebから抽出した大量のフレーズ間の因果関係のネットワークであり、例えば、「交通量が減る」⇒「交通渋滞を緩和する」⇒「大気汚染を減らす」「交通事故を防ぐ」であるとか、「インフレになる」⇒「円安になる」⇒「輸出が増える」、「ドルが上昇する」などの、いわば常識的な因果関係のチェーンを含んでいる。未だ精度に問題はあるものの、現在ではこうした因果関係を数百万個オーダーで抽出、生成することが可能であり、将来的にはこうした手法を拡張することで、上で述べたようなユーザの要求、意図を先回りできるシステムの開発も可能となろう。これはネット上の常識を基にシステムが推論を行うということであり、いわば「ネットが考える」技術であると言えよう。

また、上では震災に関連する例を挙げたが、一休を拡張することによって、災害時のネット情報、特にtwitter、地方公共団体、支援団体の情報等から、孤立している地点、支援・物資の提供のお知らせ、あるいは逆にリクエスト、透析など特定の治療が提供されている病院などの重要な情報を迅速にリストアップするシステムも開発中である。また、今回の震災時に問題となったデマ等を抑制するため、そうして得られた情報、例えば特定の物資の提供に矛盾する情報（例：「＊でコンタクトレンズを提供しているという情報はウソです。」）も合わせて提供する予定である。最終的には平成26年度までにこうしたシステムを一般公開することも計画している。

4　むすび

本稿では、NICTの情報分析技術の概要について述べてきた。重要な点は、NICTの情報分析技術においては文の構文解析やパターンの同義性の認識を含む、いわゆる深いテキストの分析が、様々な分析機能実現に貢献していること、ならびにそうした機能によって、意外でありながら有用な情報も含め、他では見つけにくい様々な情報の発見や、他では提供されない観点での情報の分析を可能にしていることである。今後は、こうした深い分析をさらに押し進め、ネット上の表面的な情報の提供にとどまらず、情報をもとに一種の推論を行い、その結果得られた仮説をユーザに提供し、様々な意思決定に資するシステムの開発も進めて行く予定である。

謝辞

本研究について常日頃から議論をさせていただいている情報分析研究室のメンバー、ならびに情報分析システムWISDOMの開発メンバーに深く感謝する。

参考文献

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　鳥澤健太郎，中川裕志，黒橋禎夫，乾健太郎，吉岡真治，藤井敦，喜連川優，“キーワードサーチを越える情報爆発サーチ―自然言語処理で価値ある未知をマイニング―，”情報処理学会学会誌「情報爆発」特集号，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 (EMNLP-CoNLL 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)

（平成 24 年 6 月 14 日 採録）

鳥澤健太郎（とりさわけんたろう）
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
博士（理学）
自然言語処理、知識獲得、Web マイニング
torisawa@nict.go.jp