ISSN 1883-1966
国立情報学研究所ニュース
No.48 June 2010

NII Today

National Institute of Informatics News

特集 言語 ―言語を「知」として生かすために―

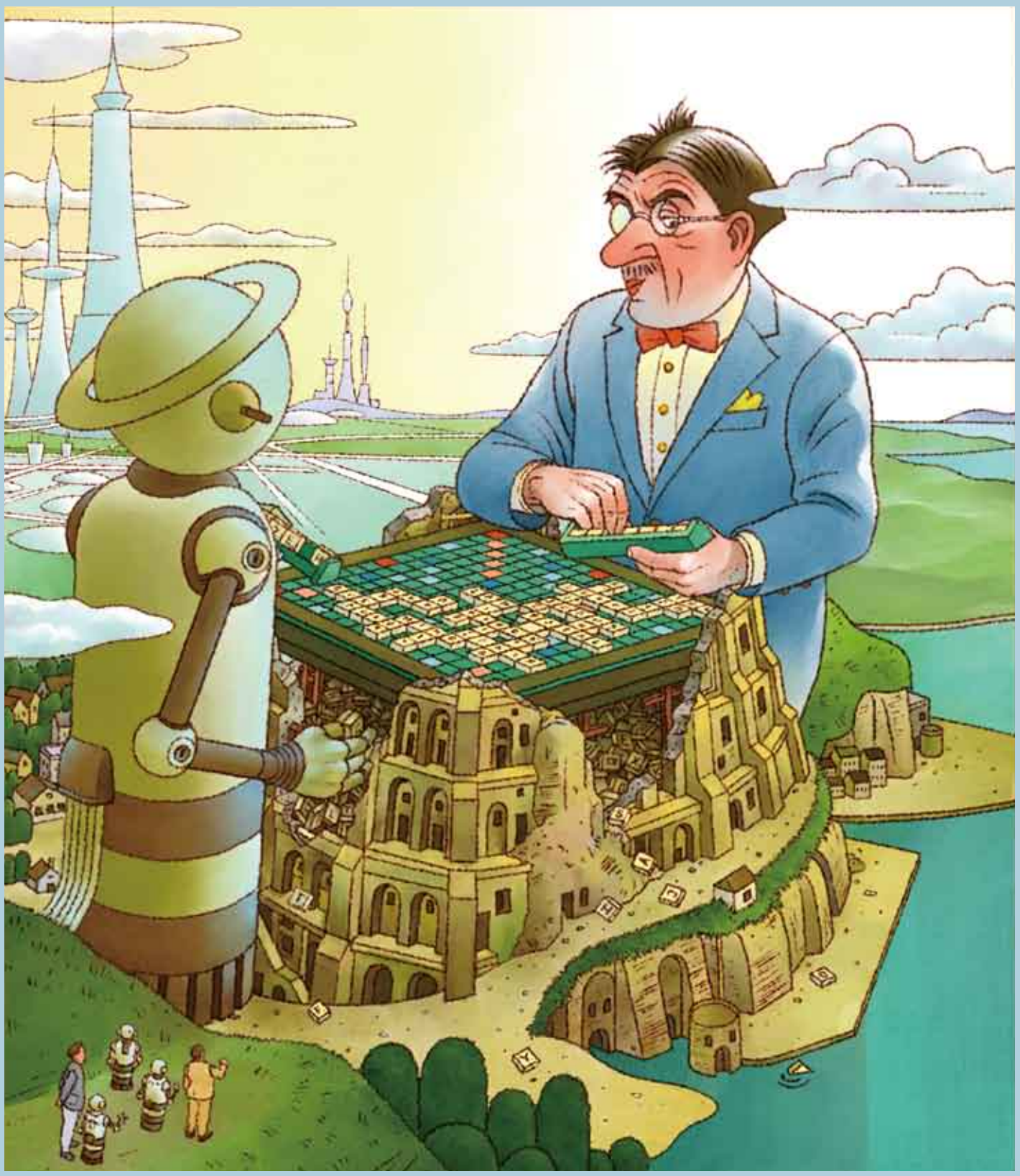
コンピュータがことばを読む

本当に必要な情報を、誰もが見つけられる時代をつくる

NTCIR が目指す情報検索の姿

情報が、感染症の広がりを食い止める

いくつもの顔を持った「言語」の魅力とは



［特集］
言語

言語を「知」として生かすために

人間が生活する上でなくてはならない、ことば（言語）。
言語はコンピュータの進化とともに、コミュニケーションの手段のみならず
それ自体が価値を持つ、「情報」あるいは「知」としてとらえられるようになった。
高度な処理能力を持ったコンピュータを用いて「言語」と向き合う、研究の最前線を紹介する。

NII Interview

コンピュータがことばを読む

相澤彰子
Akiko Aizawa
国立情報学研究所
コンテンツ科学研究系 教授

コンピュータは「ことば」を通じて現実を理解する

池谷　相澤教授は、言語処理がご専門の研究者の中でも、特に誤りなどが多く含まれる実データの処理に力を注がれていると聞いています。そんな“データの達人”がとらえる「ことば」について、今日はいろいろおうかがいしたいと思います。

相澤　最近の言語処理は本当に大量のデータを扱うようになってきました。私の仕事は主に雑多で大量のことばから、どんな価値ある情報を獲得できるかという挑戦ですが、情報を本当に処理しようと思ったら、コンピュータも、労力も、たくさん要るというのは、日々実感されるところです。

池谷　コンピュータが「ことば」を扱うというとき、私たち人間にとっての「ことば」と、どんな違いがあるのでしょうか？

相澤　一般に言語とは人間のコミュニケーションの道具だと考えられていますが、情報処理的な観点からいうと、まず電子化してコンピュータに取り込んではじめて、言語処理の対象になります。言語処理とは「コンピュータが読むとはどういうことか」を追究する学問である、とも言えますね。

池谷　もし人間ならば、読めばふつうは理解する。しかし、「コンピュータが読む」となると、人と同じにはいかないでしょうね？

相澤　はい。コンピュータにとって「読める」ということは、要するにそこから何らかの情報を獲得し、活用するということです。例えば、人間のようにはことばを理解していなくても、人と自然な会話ができるロボットがいれば、ことばを活用できていると言えるでしょう？　また言語というのは“やりとり”ですから、例えば私たちが情報を求めて検索エンジンに向かうとき、実は検索エンジンの方でも私たちから情報を得ています。「スカイツリー　高さ」と質問すれば、「スカイツリー」には「高さ」という属性があることが分かる。そういった質問が何百万、何千万とあれば、コンピュータは相当量の知識を得ることになります。

池谷　ことばを通じて、コンピュータが現実の「スカイツリー」を学んでいくわけですね？

あるものと別なものが「同じ」と判断することの大切さ

相澤　ええ。私たちがふだん交わすことばは、例えば「犬という動物は賢い」というような一般的な事柄ではなく、「昨日銀座へ行って○○に会った」というように１つ１つ個別的な事柄であることがほとんどです。コンピュータがこのような「事実」を集めると、コンピュータの中にあるバーチャルな世界は、私たちの世界にとても近いものになる可能性がある。ただし、少ない手がかりから正確な「事実」をつかみ出すのは、実はコンピュータはあまり得意ではないのです。ところが大量に集めた言語データから膨大な数の「事実」を取り出し、つき合わせて同じものをまとめたり、矛盾を調べたりする――こういった作業ならば、まさにコンピュータの能力が生かせます。

池谷　統計的な手法によって、さまざまな「事実」を切り出していくわけですね？

相澤　はい。私は、知的なものの本質は、何と何が同じであるという判断をするところにあ

ると思っています。この意味でコンピュータが切り出した「事実」が、現実世界にある対象物と一致しているのかどうか、という問題はとりわけ重要です。なかでも地名・人名などの「固有名」は、現実世界にぴたりと対応する具体物が存在しますから、言語の世界と現実の世界を結びつけるポインタの役割を担っています。このような部分を解決しないと、やはり言語というのは分からない。そこでこれについては、これまでもNIIの学術コンテンツ基盤高度化のプロジェクトで取り組んできました。またうまく成功すれば、人工知能の分野にも貢献できると考えています。

池谷　ことばというデータを介して現実世界、人、バーチャルな世界、コンピュータがつながっていく様相が、少しずつ見えてきました――もしかすると、このあたりが言語を扱う面白さなのでしょうか？

Web上の大量のことばを分析すると見えてくるもの

相澤　そうですね。ことばは社会を測るツールにもなるし、人間の脳の中をのぞくツールにもなる。言語を解析することで、人間の知識や社会的通念といったものが見えてくるのが、面白さだと思います。そこでこのようなことばの使い方を、私は「言語センサー」と呼んでいます。たくさんのことばを集めて人間の価値観を「感知」し、測定していこうというわけです。

池谷　うーん、面白いと同時に、すごく難しい問題のようにも思えてきました。

相澤　その通りです。何しろ意味をとらえるというのは永遠の課題ともいうべき難しい問題で、何千万、何億という文書を集めてきて巨大な量の計算をして、やっと、一般的なことばの文脈がおぼろげに分かってくる、それくらいのチャレンジ性を持っています。

池谷　例えば最近、人々がWeb上で“つぶやく”ようになりました。すると、人々が発信する大量のことばが、Web上に載ってきています。このような情報から、今後どんなことが分かる可能性があるでしょうか？

相澤　Webの出現によって、人と人がコンピュータを介して結びつくようになってきていますね。これまでは価値というと、ある程度画一的に価格という指標だけで測られていた面がありましたが、いまの人々が発信することばの中には、使い心地や安心感といった価格以外のさまざまな価値観があります。このような情報を集約することによって俯瞰的な傾向をとらえたり、あるいは逆にある種の多様性を見出したりすることができますね。

池谷　一般ユーザの立場から見ても、ことばを通じて、人々のさまざまな小さな思いのようなものが、大量にWebに流れ込んでいるように感じます。

相澤　そうですね。私たちの活動を記録するいろんな手段が出てきて、あらゆるモノに、それを手にした人々の行動記録や会話が刻まれるようになってきていると言ってもいいでしょう。記録というのはずっと残りますから、10年、100年後に街角に立てば、そこでふと、過去の人々が交わした会話を聞けるようになるかもしれない。人間社会やそれが担う知識がいったいどこへ行くのか、興味は尽きません。

池谷瑠絵
Rue Ikeya
サイエンス・コミュニケーター

インタビュアーの一言

相澤研究室では最近、コンピュータに向かう被験者の瞳の動きをとらえ「人が読む」行為の解明にも取り組んでいるという。コンピュータだけではなく、「人が読む」行為もまだまだ未知なる部分が多いのだ。人類の歴史全体からするとコンピュータを手にしたのはごく短期間であって「発展途中の今、計算する手間を惜しんではいけない」と相澤教授は言うが、人類が最初に文字を刻みつけた有史以来の人とことばの関わりが、これからどう変化していくのか。相澤教授の研究に、ますます目が離せない。

［特集］
言語

言語を「知」として生かすために

本当に必要な情報を、誰もが見つけられる時代をつくる

NTCIRが目指す情報検索の姿

Google、Yahoo!などの普及により、
検索エンジンで情報を探すという行為が、特別なことではなくなった今、
利用者にとって最適な情報とは何なのか。
情報アクセス技術のさらなる発展を目指し、ワークショップスタイルで活動するプロジェクト、
NTCIRに携わる3名の研究者に話をうかがった。

NTCIRとは？

パソコンやインターネットが広く普及した今日、「人間とコンピュータが将棋やチェスの対決をする」という類いのニュースを当たり前に耳にするようになった。そして今や、「コンピュータがクイズ番組に出演する」時代だ。アメリカの『ジョパディ!』という人気クイズ番組に、回答者として参加するコンピュータシステムの研究開発プロジェクトがIBMで進行中だ。情報検索や言語処理の技術を組み合わせた、質問応答のシステムを搭載し、いわゆる「ひっかけ問題」などにも対応できるような質問応答技術を研究しているという。

このような情報アクセス技術の発展を目的とし、国際的に活動しているワークショップスタイルのプロジェクトがNTCIR（エンティサイル）だ。「情報検索」ではなく、「情報アクセス（access）」という言葉を使うのには理由がある。NTCIRが目指すのは、「利用者が膨大な情報の集積から『新たな価値をうみだす』ことを支援する」ためのシステムなのだ。文書検索の技術も含め、文書中の情報を活用するための技術（質問応答・要約・意見分析・動向分析など）や、利用者が適切な質問を探すのを支援する技術を研究しているのである。

1997年に立ち上がった当プロジェクトは、情報アクセス技術に関するいくつかの研究部門を設定し、それぞれの部門を、研究者が「オーガナイザ」として企画運営するというものだ。研究部門の選定にも、ワークショップスタイルを重要視するプロジェクトの理念が垣間見える。運営サイドが一方的に研究部門を立ち上げることはせず、その分野に関わる研究者から研究候補案を募った上で、内容や実現可能性、国際的な研究動向や技術動向、社会的意義などの観点で委員会が審議をして決定するという。1年半を1サイクルとして活動しており、現在NTCIR-8（8サイクル目）が進行中である。1サイクルのプロセスの概要は次の通りだ。まずオーガナイザが研究部門の目的と評価方法を提案し、参加希望の研究者も交えた議論により評価方法やデータを決定する。その後、オーガナイザから共通の検索対象の文書データと検索に使用する質問データのセットが配布される。参加者はこのデータセットを用いて検索を実行することで、自らが開発した検索システムの検証を行う。その結果を集め、人手で判定して正解を作成する。あらかじめ作成してあった正解案が利用できるケースもあり、その場合には、多数の参加者で検討・検証し、もとの正解案の信頼性や妥当性を高めていくことになる。最後に検証結果を論文としてまとめ、成果を報告し合い、1サイクルが終了となる。はじめに配布された文書データと質問データ、それに正解を加えたセットを「テストコレクション」と呼ぶ。NTCIRに参加した研究者が、これを繰り返し使用するのはもちろん、NTCIRに参加していない研究者にも公開することで、研究を効率的に進めているのである。情報アクセス技術の効果を検証するには、実験において多数の質問と利用者が必要となる。しかし、アイディアが生まれるたびに検証が必要な、研究の初期段階において、大人数の利用者を集めて長時間の実験をすることは難しい。テストコレクションを用いることで、研究アイディアの有効性を研究室内の実験で、すぐに、しかも繰り返し検証することが可能になり、研究の展開速度が格段に上がるのだ。

半世紀にわたる情報検索システムの歩み

NTCIRの立ち上げ時からその活動に深く携わってきたのが、NIIの神門典子教授だ。「コンピュータシステムによる情報検索の研究が始まったのは、1950年代のことでした」と、研究の歴史を以下のように説明する。

情報検索の研究は、始まってほどなく、商用での実用化を目指す流れと、検索アルゴリズムなどの理論を研究する流れとに二分化する。当時商用のシステムで行っていた検索は、検索クエリと完全に一致するものを探し出すエグザクトマッチの手法を用いて、論文のタイトルや抄録などを対象に行うシンプルな検索がほとんどだった。しかし、ハードウェアの開発技術の向上により、取り扱うデータが格段に増えたことをきっかけに、検索の仕組みを見直さざるを得

神門典子
国立情報学研究所
情報社会相関研究系 教授

藤井 敦
東京工業大学 准教授

武田浩一
日本アイ・ビー・エム株式会社
東京基礎研究所 主席研究員

ない状況に陥ってしまう。具体的には、1つの単語で検索すると膨大な件数のデータがヒットしてしまい、一方で複数の単語で検索すると絞り込みすぎて1件もヒットしない、という状況である。これを打開するために、検索アルゴリズムなどを研究していたもう1つの流れに白羽の矢が立った。「商用のシステムが採用していたエグザクトマッチに対し、彼らが研究していたのはベストマッチという手法でした。これは、利用者の情報要求にもっとも『レレバント（適合する）』な順に情報を提供することを目的とした手法です」

検索アルゴリズムの研究成果を実用化するためには、大規模なテストコレクション上での、技術や手法の相互比較による検証が不可欠であった。これが、評価ワークショップによるオープンな研究の始まりである。1992年の米国におけるTREC（トレック）を皮切りに、日本のNTCIR、ヨーロッパのCLEF（クレ）と、研究拠点となるプロジェクトが立ち上がり、研究成果の技術移転、緊密な研究交流を続けることで、評価ワークショップが新たな研究課題を提案する場となっていった。

ベストマッチ検索という手法

では、レレバントな情報を検索するための手法、ベストマッチとは一体どのような検索手法なのだろう。「ベストマッチとは、検索システムが、利用者にとって最もレレバントだと判断したものから順に検索結果をランキングする検索方式です。検索対象の文書と質問の中に含まれる、単語や文字列の出現頻度、出現のパターン、文書の長さなどを評価項目とし、それらの類似性を計算するための数学的なモデル、『検索モデル』（ベクトル空間型、確率型、言語モデルなど）を作成します。Webのリンクやクリックログといった利用履歴、さまざまな経験則なども用います。それらを使い、検索対象の文書と質問の類似性、および重要性を計算し、検索結果をランキングします。質問の文字列ではなく、その背後にある情報ニーズや意図に適合する情報を探すことを目的とした手法なのです」

Webだけに留まらない情報検索システムの活躍

情報検索と聞くと、ついGoogleやYahoo!などの検索エンジンのことを連想しがちだが、情報検索技術の活躍はWeb上だけには留まらない。日本アイ・ビー・エム株式会社に在籍し、機械翻訳やテキストマイニング（※）の研究を専門とするNTCIRのタスクオーガナイザの1人、武田浩一氏はこう語る。「オフィスに蓄積されている情報の大半は、データベースのようには構造化されていない（テキストや画像などの）情報であり、また、ホワイトワーカーの仕事の最大30%が情報の検索や分析に費やされていると言われています。仕事を効率的に行うためにも、利用者が探したいと思っている情報を、量的にも質的にも扱いやすい状態にして提供できる、つまりレレバントな情報を提供できるソリューションの開発が求められているのです」

「特許」から「Yahoo!知恵袋」まで、幅広く研究を展開

1997年にスタートしたNTCIR。8サイクル目に入った現在、17カ国の研究者が参加する国際的な研究プロジェクトへと発展している（図1参照）。特徴的な研究をいくつか紹介していこう。

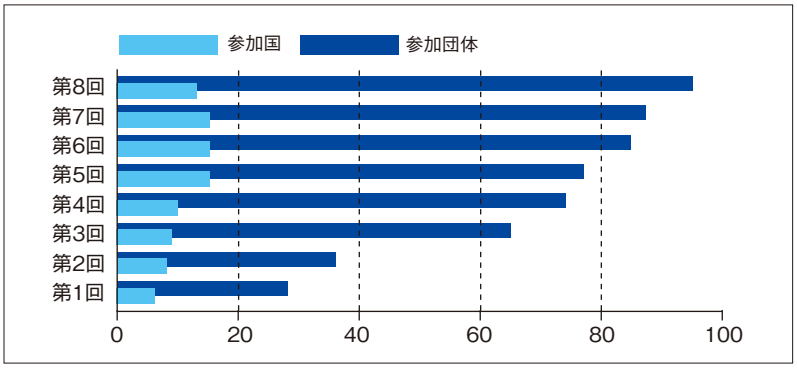


図1　NTCIRの参加国、参加団体の推移

NTCIRが初期から研究してきた分野に、複数の言語にまたがった検索を行う、言語横断検索がある。その研究の草分け的な存在として、黎明期に研究部門に参加した東京工業大学の藤井敦准教授は、当時の言語横断検索システムの概要を、以下のように説明する。「英語の論文を日本語の質問で検索する言語横断型システムの例を紹介しましょう。このシステムで検索を行う場合、日本語の質問を英語に翻訳する、英語の論文を日本語に翻訳する、という2つの検索方法が考えられます。後者は、システムの処理として負荷が高いものなので、前者を選択するのが一般的でしょう。ただし、より検索結果の精度を高めるために、以下のようにもう1つ段階を踏みます。英語に翻訳した質問で検索した上位何件か、ここでは仮に1000件とします。この1000件の論文を日本語に翻訳し、もともとの日本語の質問で再度検索するのです。この手法はその後、バイディレクショナル（bi-directional）と呼ばれるようになり、現在では、言語横断検索において主流の手法となっています」

2001年より始まった特許情報の検索も、国際的に、NTCIRがいち早く注目した研究テーマの1つだ。はじめはいくつもの課題に直面したが、特許の情報サービスを提供している企業や、日本知的財産協会の知的財産情報検索委員会の方々との共同研究やバックアップがあり、課題を解決してこれたのだという。結果、言語の横断や、文書の種別の横断、技術動向を一覧する「特許マップ」の自動生成など、さまざまな研究が実現し、実用化した研究成果も少なくない。

大量の文書から、答えそのものを引き出す質問応答も、NTCIRが力を入れてきた研究部門の1つだ。これまでの研究テーマは、「日本の首相はだれ？」といった簡単な事実をたずねる質問への応答から、「情報検索ってなに？」といった質問のような、定義や関係の説明など複雑な答えが要求される対話型の質問応答、答えがない質問への「答えがない」という応答、さらには複数言語間での質問応答など多岐にわたる。文書の検索、質問応答、要約は、別々の技術分野として発展してきたが、利用者にとってレレバントな情報を適切なかたちで提供するという目的のもと、これらの技術分野が融合していくとNTCIRでは考えている。

また、現在進行中である、Yahoo!知恵袋の検索の研究も大変ユニークなものと言えるだろう。ご存知の方も多いと思うが、Yahoo!知恵袋は、利用者が投稿した質問に、回答可能な利用者が答えを書き込むWebサービスで、優れた回答には「ベストアンサー」という評価がつく。NTCIRでは、利用者の評価とは別に、システムが自動的にベストアンサーを決定する仕組みを研究している。評価の客観性を高めるために、Yahoo!知恵袋で実際に質問者が選んだベストアンサーに加えて、新たに複数の判定者によるベストアンサーの選定を行ったそうだ。「多くの人が選んだベストアンサーを解析してみると、導入として質問者への同意の文章が入っている、根拠となるページのURLが入っているなど、いくつかの傾向が見えてきました。それは、同じ内容の答えでも、文章の書き方で評価が分かれるということを意味しています。そこから、コミュニケーションスタイルや表現方法を考慮することが、利用者の必要としているかたちでの情報提供につながることが分かりました」と、神門教授は言う。この研究成果はYahoo!知恵袋のサービスの向上はもちろん、さまざまなコミュニケーションへの応用が可能だ。例えば企業の問い合わせ窓口を、システムにより自動応答化する、ということも不可能な話ではない。「Yahoo!知恵袋のような情報は利用者のプライバシーに関わるものなので、研究で使用させていただくことは難しいのですが、Yahoo!知恵袋の立ち上げに関わった方がNTCIRの活動をご存知で、色々な手続きを経た上で提供してくださったのです」

ワークショップスタイルの魅力

このように、NTCIRの研究はたくさんの人に支えられて成り立っている。それは、外部組織との連携に限った話ではない。神門教授は、連携の意義を次のように語る。「みんなで集まって同じ課題に取り組むワークショップスタイルであることが、NTCIRの1番の魅力だと思います。1つの組織で研究をする場合、アイディアや試せる方法は数が限られてしまうし、客観的な評価をすることも難しいです。プロジェクト期間の1年半の間に、何度かラウンドテーブルミーティングという意見交換の場を設けているのですが、とくに成果報告会でのラウンドテーブルミーティングは研究者の間で、すごく話が盛り上がるのです。同じ課題に取り組んでいるという共通点があるから、テーマに対するアプローチ方法や実験のノウハウなど、論文に書ききれない細かい話でも共感度が高いのでしょうね」


前回のNTCIR-7成果報告会の様子

さらに、ワークショップスタイルは、技術的な連携の部分でも大きなメリットとなるようだ。質問応答システムを例に挙げよう。これは、情報の収集、解析・抽出、集約・提示、というようにシステムをフェーズごとに分割することが可能、つまり、機能ごとにモジュールを分割することが可能なシステムである。このような場合、1つの組織ですべての開発をするより、別々の組織が得意とするモジュールを開発し、それらを組み合わせた方が優れたシステムになる、というケースが多々あるというのだ。

情報アクセス技術研究のさらなる発展にむけて

今年の6月にNTCIR-8が終了し、一区切りとなるが、今後のNTCIRの活動はどのように展開していくのだろうか。「他国と比較して、日本は情報検索分野の研究者層が薄いため、その部分の強化が必要だと思います」と語るのは、自ら学生を指導する立場にある藤井准教授だ。「教育現場に身を置く者として、検索システムの設計や開発ができる人材だけでなく、テストコレクションを使いこなし、システムを適切に評価できる人材も育てていきたいです。検索システムを評価することは、設計や開発と同じくらい重要で難しいのです。また、検索システムを評価することと学生の成績評価にはある程度の共通点があるのではないか、と考えています。テストコレクションも学生に解かせる試験問題も、公正な評価基準をベースに、問題とそれを解くための材料、そして正解を、適切な難易度で多様に作成する必要があります。また、それによって評価された学生もシステムも、社会に出て実用的なタスクをこなせるレベルに達していなければならないのです。学生とシステム、どちらに関しても、社会の役に立つということを意識して育成していきたいと思います」。一方、企業人として武田氏は、「NTCIRの研究内容の実社会への応用という部分に力を入れて取り組んでいきたいと思います。今までは、検索、翻訳、テキストマイニングなどモジュール単体での研究が中心でしたが、今後はそれらを組み合わせることで、より効率的なシステムが構築されるようになるでしょう」と抱負を述べた。最後に、神門教授が今後の目標を2点語ってくれた。「1点目は、探索的な検索（Exploratory Search）です。Webのサーチエンジンでは、例えばNIIの地図や明日の天気など、事実確認や答えが用意されていることを知っていて質問をすることがあります。しかし、一方で、探索者自身が何を探したいのかが明確ではなかったり、探索者自身があまり分かっていないことを探したり、探索のゴールが明確ではないことを調べたりというような、インタラクティブに調べながら学んでいくケースも現実にはたくさんあります。このようなインタラクティブな探索的検索と情報活用を可能にし、検索システムが導き出す答えを利用者の要求にもっともっと近づけていきたいのです。例えば、子どもが入る幼稚園を探しているお母さんがいるとします。いい幼稚園に入れたいと思うのが親心ですが、はじめてのことであれば、どのような観点の『いい』があるか分からないと思うのです。そういう立場の人には、観点を選択肢として提示してあげる必要があります。また、すごく漠然とした調べものをする場合、例えば「大学受験」について調べたいと思っている高校生がいるとします。『大学受験』というキーワードだけで、学部別の大学ランキングの一覧表や、海外の大学を受験するために必要な手続きの一連の流れ、卒業生の進路の割合がわかるグラフなどを見ることができたら、次のアクションを起こしやすくなるでしょう。このように、検索に必要な観点を提示したり、検索した情報を分類・集約し、加工して見せたりすることで、利用者が探索的に情報を探し、学び、調べていくことができるようなシステムをつくっていきたいのです。これは、膨大な情報の集積から、利用者が『新たな価値をうみだす』のを支援する技術であり、NIIが目指している『情報から知を紡ぎだす。』を情報アクセス研究という立場から追求していくものです。その研究基盤として、インタラクティブな情報アクセス技術の評価手法の確立が必要で、現在国際的にも研究が非常に盛り上がりつつあるところです。

2点目は、1点目の実現にも大きく関わることなのですが、NTCIRを本当の意味でコミュニティとして機能させていくということです。より多くの研究者が、研究部門のオーガナイザとして、参加者として、自分の取り組んでいる研究をオープンに展開し発展させていくために、あるいは、学生や若手研究者を育成するために、NTCIRという場を活用してくださるといいなと思います。このような自発的なコミュニティの動きを先導するのではなく、サポートするのがNTCIRのあるべき姿だと思います。そうすることで、ワークショップスタイルの効果がより強く発揮され、各自の専門性を生かした連携が生まれ、よりよいシステムが構築できると考えているからです」

開かれた研究スタイルからは、きっとこれまでの予想を超える新たな技術、新たな価値が生まれることであろう。

（取材・構成　工藤 拓也）

※テキストマイニング：大量のテキストデータから単語や（人名、地名などの）固有表現、感情表現、主語/目的語+述語といった係受け表現などの多様な情報を抽出し、その出現頻度/パターンやそれらの相関関係を分析することで、一見しただけでは気がつかない知見の獲得や、文書の組織化や報告書の作成などを支援するための手法。

information

NTCIR-8の成果を問う場としてワークショップ成果報告会が開催されます

第8回　NTCIR ワークショップ成果報告会

テーマ
「情報アクセス技術の評価：情報検索、質問応答、言語横断情報アクセス」

2010年6月15～18日／学術総合センター

主催：NTCIR実行委員会　後援：国立情報学研究所
使用言語：英語

http://research.nii.ac.jp/ntcir/ntcir-ws8/meeting/

招待講演は、クイズ番組に挑戦する質問応答システムの研究開発プロジェクトDeepQAについてです。どなたでもご参加いただけます。前回のNTCIR-7には、17カ国から200名以上の研究者が出席し、活発な議論と意見交換をしました。出席者の約半数はNTCIRの研究部門参加者、残りの半数は積極的な議論のみへの参加者でした。多数のご参加、お待ちしています。

情報が、感染症の広がりを食い止める

**「テキストマイニング」という言葉をご存知だろうか。
体系化されていないテキストデータの中の、
言葉の意味やそれらの相関関係を分析することで、
特定の目的に対するそのテキストの有用性を評価する言語処理の手法のことである。
このテキストマイニングを用いて感染症問題の解決に取り組むプロジェクト、
BioCasterが各国の研究機関の連携によって進められているという。
その活動の最前線を紹介しよう。**

Webテキストの利用価値

感染症と聞いて、まず思い浮かべるのは昨年の新型インフルエンザの大流行だ。流行情報に関して重要なのは、発生地域をできるだけ正確に知ることである。

そんな中、情報システムによって感染症の拡大を防ぐことを目的としたプロジェクト、BioCasterが注目を集めている。Web上に存在する、感染症に関するさまざまな情報をシステムで自動的に収集・解析し、警告および対策や治療のための参考情報としてWeb上で公開すること。それが、プロジェクトのメインの活動だ。膨大な量のテキストから感染症関連の必要な情報を抽出するために、システムにテキストマイニングの手法が使われているところが特徴である。

「鳥インフルエンザが流行したときに、自然言語処理分野における私の研究が、感染症の監視システムというかたちで社会に貢献できるのではないかと考えました」と、BioCasterの研究責任者であるNIIのコリアー・ナイジェル准教授は言う。

BioCasterの概略はこうだ。まず、Web上から収集してきた大量のテキストデータから感染症関連の情報だけを抽出する。その後、症例、病原体、日時、場所などの重要概念を認識し、構造化されたイベント情報を取り出す。それらの緊急度に応じて文書にランクを付け、Web上に公開し、専門家に感染症対策のために活用してもらう。以下、フェーズごとの機能を詳しく紹介していく。

コリアー・ナイジェル
Nigel Collier
情報学プリンシプル研究系
准教授

ニュースレポートから Twitterまで幅広く情報を収集

BioCasterのシステムによる感染症監視は、情報の収集からはじまる。ニュースレポートや公的機関から発表される文書はもちろん、公のメーリングリストでのやりとりに至るまで、Web上に公開されたさまざまな情報を集めてくるのだ。誰もが個人単位で情報発信ができる、というのがWebの大きな特徴であるが、コリアー准教授はその点を最大限に生かそうとしている。「ブログ、TwitterやFacebookなどのソーシャルメディアも今後は情報収集の対象として考えていく必要があると思います。というのも、地域によっては、公的機関が

情報を提供するためのインフラがほとんど整備されていないためです。確かに、個人が発信する情報ですから、情報の信憑性は高いとは言えません。しかし、そこから必要な情報を抽出する仕組みを開発するのは、私にとって大変やりがいがあることなのです」

多言語情報に対応する

「英語、タイ語、ベトナム語、日本語の4カ国語からスタートし、現在12カ国語対応に向けて動いています」と、コリアー准教授はBioCasterで扱う言語の多様化について語る。さまざまな言語で書かれた情報は、どの言語のものも、分類の前にシステム内でいったん英語に翻訳される。この過程には感染症分野特有の問題がつきまとう。感染症は、1つの決まった名前で呼ばれることがほとんどない。豚インフルエンザがそのいい例で、「Swine Flu」「Swine Influenza」「Pig Flu」など、英語だけでもこれだけの呼び名がある。そのため、英語に翻訳する際、英語圏で慣用的には使われない単語に訳されるという問題が頻発してしまうのだ。単語同士の組み合わせを1つ1つシステムに登録することにより解消していくことになるが、地道な作業が必要なことは想像に難くない。この過程は、まだ人手を介さなくてはならない部分なのである。

即時性の向上が課題

英語に訳された情報は、テキストマイニングの手法で分類されていく。BioCasterのシステムでは、「感染症の名前」「発生している地域」など、情報を分類する際にキーとなる語句に注目し、それらの関係性などから1つのカテゴリーを特定していくという。現在、文書の分類という観点では、人間の専門家に比べて70%くらいの正確性で情報のカテゴリーを特定することが可能であるそうだが、「解決すべき課題はまだたくさんある」という。「感染症の流行拡大を防ぐには即時性がまだまだ低いのです。テキストマイニングというのは、情報がそろわなければできないことですが、感染症の発生当初は、ごく限られた情報しかWeb上に現れません。情報がそろってきた段階では、すでに感染症が広がっているということです。情報が少ない状態で、その感染症の発生を検知し、流行の兆しを予知することができるのがベストなのです」。分類された情報は、データベースに登録され、BioCasterのWebサイトから閲覧が可能となる。「世界中の公衆衛生に携わる専門家への情報提供を主目的としていますが、それと同時に一般の方の感染症への関心も高まれば嬉しいですね」

■BioCasterシステムによる感染症流行地域表示例（アジア地域中心）

国内外の組織との連携、今後の展望

BioCasterの活動はさまざまな組織とのつながりの上に成り立っている。国立感染症研究所、国立遺伝学研究所、岡山大学、ベトナム国立大学ホーチミン市校、タイのカセサート大学などと連携して、システムの性能向上に日夜勤しんでいる。「私たちのシステムには、まだ人の手を介さなければ成り立たない部分もあります。しかし、着実に成長していることも確かです。このシステムが、情報収集をサポートすることで、感染症の流行拡大を防ぐことが可能となり、社会がより安全になる。その実現のため、今後さらに力を入れて研究に取り組んでいきたいです」

BioCasterのシステムが、性能向上、利用範囲拡大により、世界の公衆衛生情報インフラとなること。それは、医療技術や情報技術の未熟さが原因で、感染症に苦しむ世界各地の人々を救うことにつながっている。

（取材・構成　工藤拓也）

いくつもの顔を持った「言語」の魅力とは

会話で、メールで、Webで、普段私たちがあまり疑問を感じずに使用している「言語」。
しかしそこには、知的好奇心をかき立てるさまざまな不思議があると研究者は言う。
「書き言葉」「話し言葉」さらには、Web上のテキストなど、
多面的な言語の魅力、さらには現在関心のある日本語などについて、
3名の研究者に語ってもらった。

Web検索から手話まで、多彩な言語研究の範囲

――言語研究の裾野は非常に広いと聞いています。まずは皆さんの研究対象についてお聞かせください。

内山　学術コンテンツサービス研究開発センターでは、学術論文や科学研究費補助金の研究成果をデータベース化し、一般に公開するサービスを提供しています。私はそこで、テキスト化したコンテンツから有益な情報を得るための研究をしています。学術論文には、専門用語をたくさん使用した難しい内容もあれば、比較的易しい内容のものもあります。この専門用語の難易度を判別できれば、検索ユーザーは自分のレベルにあった論文を絞り込むことができます。その基礎となる言語処理を研究しています。

坊農　私はもともと話し言葉やジェスチャーに興味を持っていました。ここ数年はその延長で日本手話の研究を始めています。NIIには、音声資源コンソーシアムという音声データを研究の資料として世の中に流通させるプロジェクトがあります。将来的には、書き言葉を持たない「手話」も言語研究の資源としてまとめたいと考えています。

阿辺川　僕は連想情報学研究開発センターに所属し、自然言語処理技術を用いた連想検索※システムの開発を行っています。現在は、当センターが開発に協力している図書検索サービスのWebcat Plusのリニューアルに携わっており、最終調整の真っ最中です。

――リニューアル後の特徴を教えてもらえますか？

阿辺川　ユーザーが「自然言語処理」という検索ワードを入力したとします。これまでの技術ですとコンピュータは「自然」「言語」「処理」の3つの言葉に分けて検索してしまったため、本来探している「自然言語処理」についての書籍が見つけにくいという問題がありました。それを1つの言葉としてコンピュータに認識させることで、ユーザーのリクエストにマッチした検索情報を提供できるようになります。

内山清子
Kiyoko Uchiyama
学術コンテンツサービス
研究開発センター
特任研究員

「言語」を究めようと思った原点

――皆さんの研究分野が広範囲に及んでいることが分かりました。ところでそもそもなぜこうした研究を志したのでしょうか？

内山　学生時代、実は英語が苦手で（笑）、あるとき日本語をすぐに英語に翻訳してくれる機械翻訳システムを使ってみました。ところが画面に表示されたのは、とんでもない英語だったのです。日本語入力の後に、一体コンピュータはどのような処理によって英語に翻訳しているのか、その仕組みを知りたいと思ったのが言語処理研究を志したきっかけです。また、日本語だったらうまくニュアンスを

伝えられることができるのに、英語ではそれができない。すると今度は「ニュアンス」って一体なんだろうと考え始めてしまう。現在の研究対象とは少し分野が外れていますが、書き言葉だけでは表現できないものにも、一研究者としては大変興味を持っています。

坊農　私はもともと人前で話すのが結構好きで、高校時代は演劇部に所属していたほどでした（笑）。大学時代には「演劇を支える言葉の表現って何だろう」と思い、言語学の勉強を始めました。私たちは向かい合ってしゃべるとき、音声だけでなく目線や身振り手振りなども使っていますよね。大学院では、音声と身体動作を使って人はどのように思考を伝えようとしているのか、という方向に興味が広がっていきました。

――「言語」と、身体動作などの「非言語」というのは、明確に分けることはできるのですか？

坊農　私自身その境界に興味があって研究を志したのですが、現在では構造主義的にすべての要素を分けることに疑問を感じています。やはり発話と身体動作を総合的に眺めることが必要なのではないかと考えています。

阿辺川　僕はお二人とはアプローチが異なり、もともと読書が趣味だったこともあり、ずっと書き言葉に興味がありました。そこで自然言語処理の研究を始めたのですが、まず痛感したのが、コンピュータに言語を理解させる難しさです。例えば「クロールで泳ぐ人を見た」という文の場合、「クロールで」は文法上「泳ぐ」に係っています。しかし、「双眼鏡で泳ぐ人を見た」となると「双眼鏡で」は「泳ぐ」に係らない。同じ「で」を使っているのに、係り先が違っているのです。こうした違いは人間なら自明ですが、一昔前のコンピュータではその違いが理解できなかった。そこでコンピュータにこうした知識や常識をどんどん教え込んでいきたい、というのがここ10年近く考えていることです。例えるなら、子どもにさまざまな経験をさせて成長させていく、ということでしょうか。

坊農　子ども、という言葉が出たので、私からも言語の獲得と育つ環境について一つ。例えば、ろう者のご夫婦に耳が聞こえる子どもがいる場合、子どもは第一言語として手話を獲得する場合があります。そして、幼稚園に通い始めた段階や両親以外の人とのコミュニケーションの中で第二言語として日本語を獲得します。このように言語の獲得は、育つ環境と大きく関連しています。

ろう者の親は自らのアイデンティティとして手話の使用を理解してほしい、けれども聞こえる子どもは日々音声としての日本語にも触れていく。その状況をどう乗り越えていくか、迷っているご家族は少なくありません。

言語のプロとして、世の中と向き合う

阿辺川　今のお話しで坊農先生は「耳が聞こえる子ども」とおっしゃいました。私はかつて日本語の連体修飾について研究していたので、これは「耳が聞こえる」が「子ども」に係っている言語構造だな、と無意識に分析してしまいます。他人の文章を読んでいても「この表現はおかしい」と気になることもしばしば…。

内山　私も同じですよ。研究対象である専門用語はいくつかの単語が結合した「複合語」が多く、そのためテレビを見ていると「事業仕分作業」という複合語が気になって仕方ない。「仕分」も「作業」も、動詞由来の名詞なのですが、結合する規則や順番をつい考えてしまいます。

坊農　私が最近気になるのは、携帯電話とパソコンの画面の大きさです。パソコンの画面は大きいので文章を推敲できますが、携帯だと画面が小さくて推敲が難しい。携帯メールでのコミュニケーションに慣れ親しんだ若者とそうでない世代では、書き言葉に対してもスタンスが違ってくるのでしょうね。

阿辺川　最近の学生の中には、卒論を携帯メールで教授に送る人もいるぐらいですから。

内山　坊農　信じられない！

―― 3人とも研究者ならではの立場から日本語や現代社会と向かい合っていますね。では最後に皆さんの研究を社会にどう還元していけるか、お聞かせいただけますか？

阿辺川　研究成果をいかにして実社会へ応用できるか、そこに主眼を置いて研究してきました。これからも大勢の方に使っていただけるシステムを作っていきたいです。

坊農　日本人はハローやニイハオなど近隣国の挨拶は言えても、ほとんどの人が日本手話の「こんにちは」を知らない。最初にお話した「日本手話」のデータベース化を通じて、「日本手話」という日本におけるもう一つの言語のあり方を人々に伝え、「言語とはそもそも何なのか」を考えることが私の仕事かなと思います。

内山　学生と接していると、検索エンジンには興味あるのに、検索エンジンなどに使われる技術の一つである自然言語処理には興味を持ってくれない。隠れたところで人々の暮らしに役立っている自然言語処理について、もっと分かりやすく伝えていきたいですね。

（取材・構成　升國義浩）

※連想検索：人間がある言葉から無意識にいくつもの関連単語を思い浮かべるように、検索キーワードから関連性の高い単語を抽出し、それを含む図書などの情報を探し出す検索方法。

NII ESSAY

うわさ話とヒトの群れ

小林哲郎
（国立情報学研究所 情報社会相関研究系 助教）

うわさ話が楽しいのはなぜだろう。われわれの日常会話の大部分は他人のうわさ話で占められているという研究結果は少なくない。そこまでヒトがうわさ話好きなのには何かわけがあるに違いない。

言語は毛づくろいの代替物？

人類学者のロビン・ダンバーは、うわさ話は群れを維持するために必要であり、そもそも言葉はうわさ話をするために生み出されたという大胆な仮説を唱えている（Dunbar, 1996）。言語を持たない類人猿は、必要以上に長時間の毛づくろいをすることによって群れの結束を維持している。毛づくろいには、衛生的な機能以外に、「誰が信頼できるのか」「誰がお返しをしなかったのか」など、群れの協力関係を維持するために必要な情報を得て覚えておくという社会的機能があるからだ。ところが、大脳新皮質の進化に伴って群れの規模が大きくなると、多くのメンバーと毛づくろいすることは時間的に不可能になる。そこで、人類は毛づくろいの代替物として言語を生みだしたのではないか。「誰が狩りの最中にサボっていたか」「誰が獲物の分配で信頼に足る行動をしたか」といった情報をうわさ話として流すことによって、毛づくろいと同じ社会的機能を効率的に果たすことができる。つまり、そもそも人類の言語は、大規模な群れの維持に必要なうわさ話をするために進化した可能性がある。

「うわさ話」が、人と人をつなぐ

こうして考えるとヒトのうわさ話好きにも理由がありそうだ。ヒトは常にうわさ話をして他人の評判を流し、また自分の評判を知っておくことによって、群れの中での協力関係を維持している。知らない他人のうわさ話を聞いてもちっとも面白くないのは、うわさ話が群れの維持と関係しているからだろう。ゴシップ（Gossip）の語源は"God Sib"、つまり「名付け親（ゴッドファーザー）」であり、そこから転じて家族ぐるみの付き合いを意味したという（川上, 1997）。つまり、ゴシップは縁戚関係のように信頼できる関係を示す言葉から生まれたということになる。ここにもうわさ話と群れの維持の接点がありそうだ。

さて、世の中ＳＮＳやツイッターが話題だ。ネットは、うわさ話が広まるスピードが上がり、範囲を拡大することでヒトの群れに何か本質的な変化を起こしているだろうか。進化し続けるネットサービスにおいても、案外、古い「群れの論理」が生き続けているかもしれない。

今月の表紙イラスト：今月のテーマは「言語」。中央には無数の言語が蓄積された「バベルの塔」。
そこから単語を掘り出して、人間とコンピュータがワードゲームのスクラブルで知恵比べ。果たして勝つのはどっち!?

情報から知を紡ぎだす。
NII

国立情報学研究所 ニュース（NII Today）第48号 平成22年6月

発行：大学共同利用機関法人 情報・システム研究機構 国立情報学研究所　http://www.nii.ac.jp/
〒101-8430 東京都千代田区一ツ橋2丁目1番2号 学術総合センター
編集長：東倉洋一　表紙画：小森 誠　写真撮影：谷口弘幸　制作：株式会社 商業デザインセンター
本誌についてのお問合せ：企画推進本部広報普及チーム TEL：03-4212-2131 FAX:03-4212-2150 e-mail：kouhou@nii.ac.jp