［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。

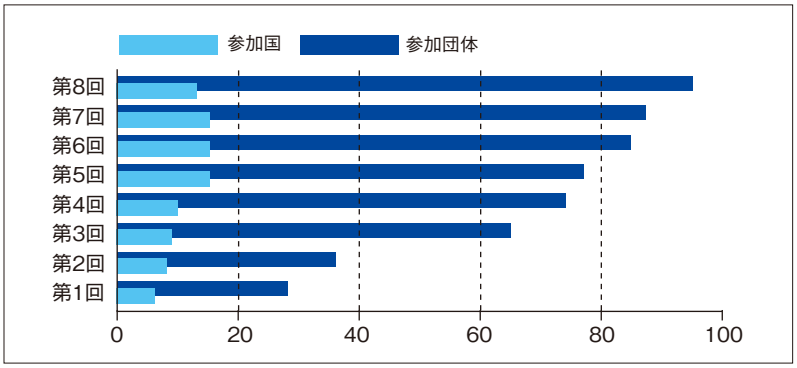

図1　NTCIRの参加国、参加団体の推移

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。