R&D

質問文で検索し
欲しい答えがすぐ見つかる

「なぜ」に答える質問応答システム

膨大な情報に紛れて求める情報にたどり着けない

インターネット上にあふれる情報が膨大になればなるほど、自分の知りたい情報がそこに含まれている可能性は高くなる。その一方で、求める情報になかなかたどり着けないというジレンマが生じているのも確かだ。膨大な情報の中から求める情報を効率的に見つけ出す新たな検索方法への期待は日に日に高まっている。

NTTコミュニケーション科学基礎研究所では、検索キーワードではなく、「アメリカを初めて発見したのは誰か」「エッフェル塔の高さは何mか」のように、文章の形で質問を入力し、質問内容にぴったり合う回答が得られる質問応答システムの研究を既に2000年から開始。日本における質問応答システム研究の主導的な立場を果たしてきており、その技術的蓄積に基づいて「SAIQA（サイカ）」という高性能な日本語質問応答システムを実現している。

このような従来型の質問応答技術は基礎研究の段階を脱してきており、gooなどの商用ポータルサイトで「自然文検索」として導入実験が行われている（http://qa.wpedia.search.goo.ne.jp/）。

しかし、従来型の質問応答システムが有効なのは、答えが固有名詞や数値表現のような簡単な言葉で答えられるものに限られ、原因や理由を問う高度な質問には対応しきれていない。同研究所協創情報研究部知識処理研究グループは、将来を見据えた基礎研究として、さらにその研究領域を拡大。「なぜ」「どうして」という原因や理由を問うものにも対応可能な質問応答システムをめざし、新たな挑戦を開始した。

原因・理由表現の抽出方法と回答候補の採点基準作りがポイント

従来の固有名詞や数値表現で答えられるものと違い、原因や理由は、多様な表現で表されることから、回答を自動で探し出すことは難しいとされ、これまで実用レベルでの研究はほとんど行われてこなかった。

質問応答システムの基本フローは、「質問の意味を解析」→「回答候補を指定された大量のテキストデータから抽出」→「より正解に近いと思われるものを上位に表示する」というもので、フロー自体には特別な変化はない。

「問題は、回答候補をいかに抽出するかだった。例えば『アメリカを発見した人は誰か』という質問であれば、『誰』という言葉から、聞かれているのは人名であると分かり、テキストデータ中の記事から人名と思われる部分をピックアップして回答候補とすることができる。一方、『熟年離婚の原因は何ですか』『洞爺湖がサミットの開催地に選ばれたのはなぜ？』というような質問の場合、その原因・理由を表しているとと思われる部分を探し出さなければならないが、明示的な手がかりのみに頼っていては不十分であり、検出が極めて難しい。そこでまず、原因・理由を表していると思われる表現を収集することから始めた」（同グループの磯崎秀樹主幹研究員）

原因表現の代表的なものとしては、「～によって」「～のため」のような言葉が挙げられるが、実はこれらの表現が使われているものは原因・理由を表している文章全体の半分程度に過ぎない。

例えば前述の質問「洞爺湖がサミットの開催地に選ばれたのはなぜ？」に対する回答には、「洞爺湖と有珠山をはじめとする変化に富んだ自然豊かな洞爺湖地域は、恵まれた自然環境と自然災害の深い傷跡をあわせ持つ、環境を主要なテーマとするサミットの開催地にふさわしい場所と言えるでしょう」というものがあるが、この文章のどこにも「～によって」「～のため」のような、私たちが原因・理由表現として思いつくものは含まれていない。

普段、意識して使っていない原因・理由表現を洗い出すには人力では無理がある。そこで、意味的なタグが付与された約20万文に上るテキストデータ（コーパス）を解析し、理由が記されている文章の特徴を自動的に選び取るという方法を採用。約700項目の原因・理由表現を抽出した。

これで、回答候補を探し出すための手がかりとなる原因・理由表現の「パターン集」は完成したが、次なる問題は回答候補にどうやって優先順位を付けるのか、つまり回答候補の採点方法をどうするかということだった。

この問題解決にあたっては、まず1000問の質問と、それに対する模擬回答を作成。その回答を解析することで、より正解に近いものの条件を洗い出し、その条件に合致する度合いに応じたポイントを各表現に自動的に付与していった。

こうして、ポイントの高い表現が多く含まれているものほど上位に表示される「採点基準」を作り上げていった。

多様な質問への応答を実現し知的欲求を満たしたい

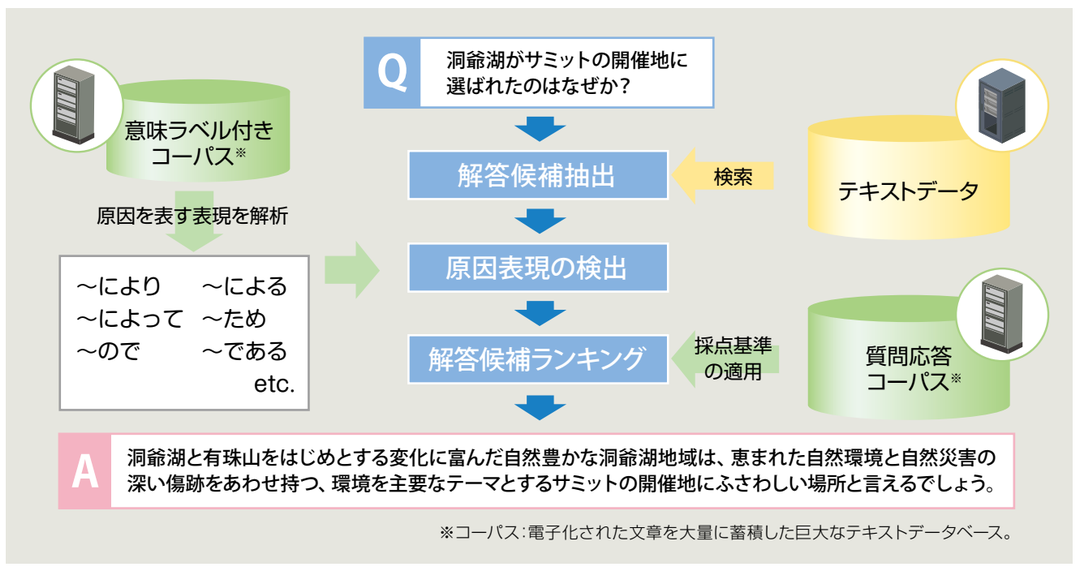
「なぜ」に答える質問応答システムによる回答の精度は、回答候補を探し出す手がかりとなる「パターン集」と、優先順位を決定する「採点基準」の二つの精度にかかっていると言っても過言ではない。

「思い込みや処理数の限界から見落としが生じることを避けるため、コンピュータ解析による自動化を徹底したことで、精度の大幅な向上を実現することができた。試行錯誤の結果、現在では、『正解』と見なせるものが、ほぼ上位3位までに表示することが可能となっており、実用化もそう遠いことではない」（同グループの東中竜一郎研究員）

「なぜ」という質問は、人間にとって最も自然な疑問の形である。また、「なぜ」に対する回答が、さらなる疑問を生み出すことも多い。「なぜ」を繰り返し、検索を繰り返すことは、人間の知的欲求を満たし、さらなる〝知〟を生み出す。

「学ぶ動物」である人間にとって、この「なぜ」に答える質問応答システムを待つ市場ニーズは大きいだろう。Webで「検索する」から、Webで「学ぶ」へ。より多様で高度な質問に答えられるシステム開発が進めば、Webに存在する膨大な情報の輝きは今以上に増してくるはずだ。

■「『なぜ』に答える質問応答システム」の基本フロー



※コーパス：電子化された文章を大量に蓄積した巨大なテキストデータベース。

■ 質問とその回答例

Webから回答を抽出