

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、たくさん要するというのは、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始めて、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が情報を得ています。「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、「昨日銀座へ行って〇〇に会った」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、私たちの世界にとっ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

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象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ト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、人間の脳の中のを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のでしょうか？

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、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、興味は尽きませぬ。



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。