■ **相澤彰子**　コンテンツ科学研究系　教授

【タイトル】言語テキストから“知”を生み出す

【本文】

人間にとって、言葉は不可欠な道具です。たがいに意思を伝え合うコミュニケーションの手段として、また記録を伝えるためにも必要です。逆にいえば、言葉の使われ方を解析すると人間の活動を垣間見ることができます。私の興味は、実にその点にあります。

大量のテキストを整理する

私は、コンピューターを使って、書き言葉である“言語テキスト（以下テキスト）”から、“知識”を引き出したいと考えています。ポイントは、“大量”のテキストを処理することです。コンピューターに文字列の意味はわかりません。しかし大量のテキストがあれば、繰り返し出現する文字列について、統計的に頻度や傾向を計算し、結びつきが強い文字列の“かたまり”を抽出できます。これが “知識”のもとです。この仕組みは、テキストからの辞書やシソーラスの構築、テキストの分類や検索ナビゲーションの技術へつなげることができます。

少ない規則性から情報をふやす

テキストの中にあらわれる文字列の“かたまり”を実世界の対象に対応づけたり、あるいは“かたまり”同士の関係を抽出したりすることで、さらに情報を増やすことができます。

例えば「情報研」および「ＮＩＩ」は、いずれも「国立情報学研究所」を指しますが、文字列が異なります。このため、コンピューターが同一と判定するためには、新たなルールが必要です。しかし、照合を要するデータは数百万件から数千万件以上にもなります。そのつど照合ルールを導入していたのでは、追いつきません。そこで具体的なルールではなく、与えられたテキストから、自動的に候補となる組み合わせや判定ルールを得ることを検討しています。このような学習が成功すれば、「読むコンピューター」の実現に、一歩近づくことができるでしょう。

情報を伝える“単位”をみつける

新聞記事を分析した際、複数の記事にまたがり完全に一致する長い文字列（フレーズ）が存在することに、興味をもちました。このような文字列はコピーされ、使い回され、場所を越え、時間を越え、伝わっていきます。情報を伝える“単位”といえそうです。これは、リチャード・ドーキンス氏が提唱するミーム（文化を伝える自己複製）論を彷彿させます。フレーズの特性や、伝播の仕方を詳しく調べれば、テキストを対象とする新しい科学に発展できるかもしれません。今後の研究に期待が膨らみます。

（取材・構成　那須川真澄）