AIのある未来へ

「自然言語処理」という言葉を聞いたことがあるだろうか。コンピューターによる処理技術の一つだ。聞いたことがない人でも、パソコンの日本語入力ソフト（かな漢字変換）やスマートフォンの音声認識など日常でよく触れている技術である。

そんな自然言語処理の「いろはのい」を、九州工業大情報工学部（福岡県飯塚市）の嶋田和孝准教授＝写真＝に語ってもらった。新聞の未来を握る技術でもあった。

コンピューターの自然言語処理

新聞の未来も変える

九州工業大・嶋田和孝准教授に聞く

自然言語処理とは、コンピューターに自然言語を理解させ、意味を捉えたりする処理技術です。

自然言語とは日本語や英語など人間が日常使っている言語のことです。それに「自然」を付けているのは、対になる言葉の「人工」があるから。人工言語はコンピューターのプログラミング言語です。

自然言語と人工言語との大きな違いは曖昧性の有無です。曖昧性とは単語や文の意味や構造の解釈がはっきりと決まらず、複数の考え方ができることを指します。人間なら、行間を読み取って意図をくみ取ることができます。「愛する花子を太郎は」という文があります。これは文法が誤っており、文章として成立していない非文です。しかし人間は、非文であっても意図はちゃんと理解できます。

一方、人工言語には、そういった曖昧性がありません。曖昧性なく解釈が一つになるように、厳密に文法を定義したものがプログラミング言語なのです。

モデル化して計算

基本的な自然言語処理の流れを説明します。古くからある手法ですが、文を単語に分割して品詞を推定する「形態素解析」、主語述語など単語間の構文関係を決める「構文解析」、単語と文の意味を考える「意味解析」、文脈まで広げて処理をする「文脈解析」という段階を踏みます。

しかし現実には難しいものです。例えば「このひとことでげんきになった（この一言で元気になった）」という文を、自然言語処理のかな漢字変換をするとします。コンピューターは辞書を引いて「この」という単語に「個の」「子の」などの字を充て正解を導こうとしますが、組み合わせの数は膨大になります。

そこでコンピューターによる機械学習で、どんな組み合わせが起きやすいのか、数値でモデル化して計算します。文頭に「個の」が来る文はほとんどなく「この」が圧倒的に多い、ということが分かるようになります。

古くは人がルールを書いて、コンピューターにひたすら覚えさせていました。しかし限界があります。グーグルを使って調べることが「ググる」という動詞として定着するなど言葉は日々変化するため、人間が毎回メンテナンスすることは現実的ではありません。そこで深層学習（ディープラーニング）のニューラルネットワーク（NN、☆1）など、機械学習の進んだ手法を使って処理を行うようになりました。テキスト（文書）データの量も格段に増えたため、自然言語処理でこれまでできなかったことの幾つかが、できるようになりました。

意図の通りに動作

言葉の意味をコンピューター上で表現することについて考えます。人は単語の意味を、辞書に載っている語釈で理解します。しかしコンピューターは辞書の語釈だけでは言語の処理ができません。「単語の意味は周りの単語で決まる」という概念で、工学的に意味をとらえているのです。

ある単語が文中に出たとき、その文中に別の単語が頻繁に出現することを共起といいます。共起の度合いを測定すれば工学的に意味は量れるはずだ、という考え方です。

英語の「ｂａｎｋ」は銀行と堤防の意味があります。ｂａｎｋを含んだ文で、その周りにｄｅｐｏｓｉｔ（預金）があれば銀行で、ｒｉｖｅｒ（川）があれば堤防、とコンピューターは考えます。

ｗｏｒｄ２ｖｅｃ（ワード・ツー・ベック）というツールで、単語をベクトル化（☆2）することが流行しました。特定の単語の周囲にどんな単語が出てくるか、ＮＮで学習し算出したものです。単語をベクトル化すると、単語同士の意味の演算ができるようになります。有名な例ですが「Ｋｉｎｇ－Ｍａｎ＋Ｗｏｍａｎ＝Ｑｕｅｅｎ」という式があります。ＫｉｎｇとＱｕｅｅｎが似たところにあり、ＭａｎとＷｏｍａｎが似た方向にあるベクトルの状態を表します。

応用編で「西日本新聞－福岡＋東京」をやると、最初に東京新聞が出てきました。その次に出てくるのが中京新聞。今はない新聞ですが、何となく意味があるっぽい結果が出たような気になります。

実際はコンピューターは意味が何なのか、分かっていません。「言葉の意味が分からないコンピューターに、何ができる？」と思う人もいるでしょう。でも工学的には必ずしも人間のように意味を理解して解答を導く必要はないかもしれません。飛行機が鳥のように羽ばたいて飛ぶ仕組みで動いているわけではないように、人間のように意味が理解できなくても人間が意図したように動作すれば十分に役立つものが実現できます。

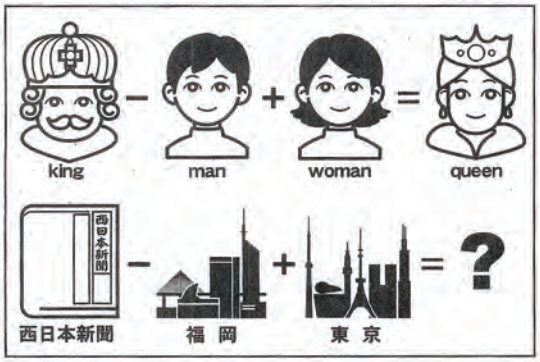
自然言語処理が新聞づくりにどう生かせるか考えてみましょう。利用できる技術は少なくありません。文章を要約したり、株価のグラフや天気図などを文章で説明したり、画像に何が写っているかを示すキャプションを作ったりする研究が進んでいます。朝日新聞は昨年、人工知能（ＡＩ）による文章の自動校正システムを開発したと発表しました。10月の当欄で紹介された「スペクティ」のような情報分析の補助となる速報サービスも広まっています。

基本的にコンピューターが得意なことは①大量のデータを高速に処理する②大量のデータから何かを予測する③誰がやっても同じ結果になるような定型的な作業の自動化（新聞でいえば野球やサッカーの試合結果の記事作成）―です。一方で「何を書くべきか」と考えること、インタビューなど人からの情報を抽出し整理することはできません。できることとできないことがあり、できることの中心は「新聞づくりの手助け」なのです。記者が取材や執筆により注力できる環境づくりの補助となるイメージでしょう。

ビジネスチャンス

新聞社の強みは、データベース化した過去の記事など豊富なテキストデータを所有していることです。データがないとコンピューターは何もできません。しかも所有したテキストデータは、ＳＮＳ上の文章などと比べるととても良質です。正しい用字用語で事象を的確に伝え、しかも裏付けのあるファクト（事実）を発信しています。正しくてきれいな文章がたくさんあるということは、文章の自動校正システムなどのモデル構築に役立つはずです。

記事執筆にとどまらず、新聞社のデータを使い、言葉を処理するためのツールを構築することが可能となってきます。それは新聞社にしかないビジネスチャンスではないでしょうか。



しまだ・かずたか　1974年生まれ、大分市出身。大分大大学院で博士（工学）を取得。2012年から現職。専門は知能情報学（自然言語処理）で、情報要約や人間同士の会話理解などの研究に取り組む。

注釈

☆1　コンピューターによる機械学習の一つ。人間の脳神経回路のように多層的に情報処理を行い、コンピューター自身がデータに含まれる特徴を捉え、より正確で効率的な判断をする。

☆2　コンピューターによって単語を数値化する一手法。ベクトルとは大きさだけでなく向きも持った量のことで、意味が似た単語同士は数値同士も近い値となる。

※「ＡＩのある未来へ」は原則、毎月最終日曜日掲載です。