情報工学科　人工知能研究室

自然言語処理と機械学習

― 言葉の 処理の研究―

二宮崇准教授

―自然言語処理―

自然言語処理という言葉にまずなじみがないかもしれません。自然言語とは英語や日本語などの人間が日常用いる言語のことで、自然言語処理とはそれらの言語を処理・解析することによって、文書から知識を獲得したり、より便利なユーザーインターフェースやサービスを提供するための処理技術や研究分野のことを指します。自然言語処理の応用に近いところでは、仮名漢字変換や自動翻訳の研究があって、こういう研究を思い浮かべると自然言語処理というイメージがつかみやすいかもしれません。プログラミング言語の研究と区別をするために「言語処理」ではなくて「自然言語処理」という言葉が使われています。

我々が日常知識を得たり意志を伝えたり、意見を交換するときには自然言語を用います。人間の大きな知識の源である書籍はもちろん、コンピュータやインターネットの世界でも、膨大なウェブページやブログ、ウィキペディア、個々人の蓄積するメール、電子ドキュメント、それらは全て自然言語で記述されています。インターネット上で人間が新たな知識を得たり、新たな情報を発信する時にもやはり自然言語を用います。これらのメディアへのアクセスを支援する知的な言語処理を実現したり、何らかの知識や情報をこれらのメディアから抽出できたら面白いと思いませんか？自然言語処理は自然言語で記述されたメディアを処理し、知的サービスやユーザーインターフェースを提供することを目的にしています。

―自然言語処理の基礎技術―

自然言語処理には基礎から応用までの様々な研究があります。基礎的な技術としては、形態素解析、統語解析、照応解析、固有名解析があげられます。形態素解析は、ある文が与えられたときにその文を単語に分割し、その品詞をあてる解析のことです。たとえば、「太郎は花子を愛している。」という文があったときに、「太郎/固有名詞　は/係助詞　花子/固有名詞　を/格助詞　愛し/動詞五段サ行連用形　て/接続助詞　いる/動詞一段基本形　。/句点」と解析するのが形態素解析です。統語解析と呼ばれる解析は、名詞句や動詞句などの句を解析したり、単語の係り受けを解析します。照応解析は「それ」や「あれ」といった指示代名詞がどの単語を指しているのか解析する技術です。固有名解析は、「太郎」が人名であることを解析したり、「愛媛大学」が組織であることを解析します。その他の基礎技術には、意味の研究で、同義語を集める研究や、同じ意味の表現を集める「言い換え」の研究などがあります。

上に書いた基礎技術は主に言語学と機械学習と呼ばれる研究分野の研究成果や技術で実現されています。言語学は上の例をみてもわかるように、まず、単語という単位をどう設定するか、品詞体系をどう定義するか、といった問題設定そのものに深く関わってきます。また、統語解析には文法理論が大きく関わってきます。次に、機械学習という分野および技術ですが、これは簡単に述べると、あるデータに対するラベルの予測問題を数学的にきれいに解く技術の研究分野と言えます。たとえば、上

情報工学科　人工知能研究室

自然言語処理と機械学習

― 言葉の 処理の研究―

二宮崇准教授

の例だと、「太郎」という単語（＝データ）に割り当てられる品詞（＝ラベル）を予測する、ということになります。自然言語処理ではサポートベクターマシンやロジスティック回帰の一種である条件付き確率場がよく用いられています。また、最近ではオンライン学習と呼ばれる機械学習の手法がよく研究され自然言語処理にも用いられつつあります。

―自然言語処理の応用研究―

自然言語処理の応用は多岐に渡ります。古くは仮名漢字変換の技術があり、ジャストシステム、Microsoft、Googleの仮名漢字変換ソフトはみなさんも普段から使っているのでご存じだと思います。大きな応用研究には機械翻訳があります。これは自然言語処理研究者にとっては長年の夢の一つであり、英語やドイツ語などいろんな言語の文章を母国語に自動的に翻訳するソフトを実現する研究です。この研究は一つの研究分野といってもいいぐらい大きくかつ多岐に渡り、辞書の構築、対訳文書の作成、機械翻訳アルゴリズム、二言語間の単語の対応付け、評価手法など、様々な基礎技術からなっています。

　上記以外の有名な応用研究としては、文書分類、自動要約、評判分析、自動校正、読解/翻訳支援、質問応答などがあります。文書分類は、たくさんの文書を、経済、スポーツ、政治などジャンルに自動的に分類する技術の研究で、投稿サイトの自動タグ付け、ニュース記事の自動分類、メールの自動分類など応用は非常に多くあります。自動要約は長い文章を自動的に要約する技術の研究で、評判分析は、ある商品のコメント文からその商品の評判が良いか悪いかを判定する技術で、例えば、大量にあるブログから評判の良い商品を発見する、といったことが出来ます。質問応答は、自然言語の文による質問に自然言語の文で自動的に回答を与える技術の研究です。自動校正は、文章の文法誤りを自動的に検出する研究で、メールソフトやワープロソフトに組み込まれて実用的に用いられています。読解支援は、例えば、英語の文章を読むことを容易にするための技術で、単語訳をポップアップで表示させたり、単語訳を文章に埋め込むといったことがあげられます。

―研究テーマ―

形態素解析、言語学的な文法を用いた統語解析、機械翻訳、機械学習を主なテーマとして考えていますが、基本的にはここまで述べてきたような自然言語を処理する研究テーマや機械学習のテーマなら何でも良いと考えています。自然言語のデータは現在大量に存在しています。大量のウェブページや、ブログ、ウィキペディアがあったら、こんな面白い情報が抽出できるんじゃないか、こんな楽しいサービスが提供できるんじゃないか、といったアイデアや好奇心からスタートする研究が面白いと思っています。

二宮 崇

職位: 准教授
学位: 博士(理学)
専門: 自然言語処理 (特に構文解析)
　　計算言語学
　　機械学習
連絡: ninomiya@cs.ehime-u.ac.jp