機械学習による自然言語処理
言葉を扱う技術とビッグデータの接点

人間の言葉をコンピューターで処理する技術を自然言語処理と呼びます。機械翻訳、質問応答、テキスト・マイニング、文書分類などの基盤技術である自然言語処理が、実はビッグデータと機械学習によって成り立っていることをご紹介します。

1. 自然言語処理の応用例

自然言語処理（Natural Language Processing：NLP）とは、自然言語をコンピューターで処理する技術の総称です。なお自然言語とは、コンピューター言語と対比して、日本語や英語といった人間が日常的に使っている言語のことを指します。

人間の知的活動の結果の多くは自然言語として記録されます。そのため、自然言語を扱うNLPには幅広い応用先があります（**図1**）。例えば、Web上のサービスとして提供されている「機械翻訳」や、2011年にクイズ番組でコンピューターが人間のチャンピオン2人を負かした「質問応答」についてはご存じの読者の方も多いでしょう。また、テキストデータから知識を発見するための「テキスト・マイニング」も応用先の一つです。これらの応用と同様、ビジネスの現場でよく活用されている「文書分類」を以下で紹介します。

文書分類は、文や文書を入力して、一つまたは複数のカテゴリを自動で付与する処理です。文書分類技術の活用例として、医療保険でのコード化が挙げられます。米国では病院や医師から保険業者に送る領収書に、病名などを表す国際疾病分類コードを記載する必要があり、そのコード化にはコンピューターによる補助コードシステム（Computer-Assisted Coding）が利用されています。診断書（文書）を文書分類システムに入力することにより、十数万あるコードから候補となるコードを絞り込むことができ、入力者の負荷やミスを軽減することに役立っています[1]。また日本の保険会社でも、コンピューター

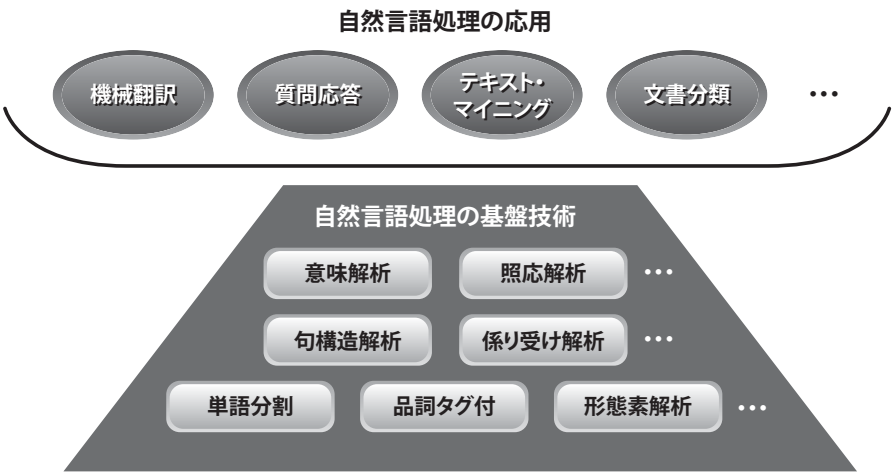

図1. 自然言語処理の基盤技術と応用

による診断書の自動コード化を保険金や給付金の請求内容の確認に役立てています[2]。

文書分類システムでは文書から抽出したキーワードやフレーズなどを根拠としてコードを予測します。意外に思われるかもしれませんが、診断書に書かれる病名や手術名は“揺れ”が大きく、例えば「悪性腫瘍摘除手術」が「ガン摘出術」と書かれていたりします。この揺れを機械が判断するのは容易ではありません。また、手術部位が別の欄に書かれているため、一つのキーワードだけでは部位によって異なるコードが判断できないこともあります。そのため、機械学習を使ってキーワードを組み合わせた分類ルールを導き出す必要があります。さらに保険の場合には、契約情報などの定型情報とテキストを入力して複合的に判断することもあります。例えば保険請求者の性別も併用することで、男性・女性に特有の疾病コードや出産関係の手術コードの認識をより正確に行うことができます。

そのほか、コールセンターのお客様の声に商品コードや問い合わせ種別コードを付与するような場面でも、文書分類が活用されています。分類を自動化することで、お客様の声を商品や問い合わせ種別ごとに集計して、より迅速に商品開発や不具合対応に生かせるようになります。文書分類はNLPのシンプルな応用ですが、文書に何らかの専門的なカテゴリを付与する業務は多くの業種で見られるため、今後も使われる場面が増えていくと予想されます。

2. 自然言語処理の基盤技術

NLP応用として取り上げた文書分類では、文書の中からキーワードやフレーズを抜き出す処理を前提としています。この処理の中では、

a) 単語の文法的役割を識別する「品詞タグ付」処理（名詞・動詞・助詞など）を行って助詞などの意味的な役割の少ない語をキーワードから除外する処理

b) 単語同士の関係を判断する「係り受け解析」を行って、主語と述語の組を抽出する処理

などを行っています。また、日本語のように空白で単語が区切られていない言語では、「単語分割」処理も行う必要があります。

図2は「肺は悪性腫瘍の疑い」という日本語に対して、単語分割、品詞タグ付、係り受け解析を順に適用する過程を示しています。単語分割が行われていないと、単語への品詞タグ付や単語間の係り受け解析ができません。そのためNLPの処理は、複数の解析器を順番に適用していくパイプライン処理が一般的です。

これらの品詞タグ付、係り受け解析、単語分割などは比較的応用に依存しないNLPの基礎技術であり、さまざまなNLPの応用に活用されています。

3. 自然言語処理とビッグデータ

Web上のデータをはじめとした、「ビッグデータ」と呼ばれる大規模データから知識を引き出す技術が注目さ

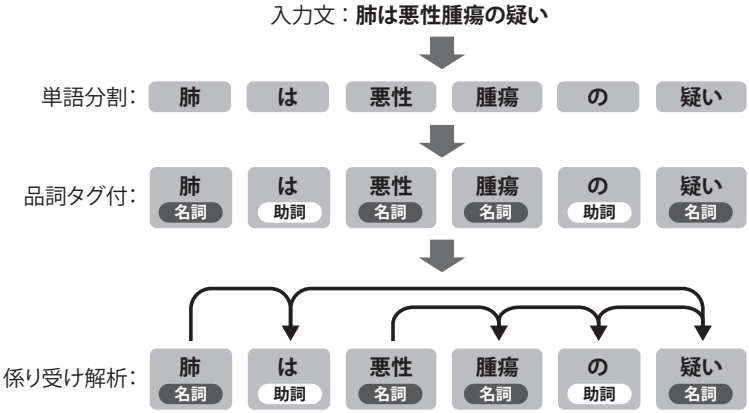

図2. 自然言語処理の基礎技術

れています。この流れは、データの背後に隠れた規則性を高精度で発見する「機械学習」技術の発展に支えられています。

実は、2節にあげたNLPの基盤技術や文書分類などの応用システムも機械学習を用いて構築されています[3]。NLPの基盤システムは、もともと文法によって記述されていましたが、例外的な現象が多く人手で記述しつくすのは困難でした。しかも時代とともに言葉の使われ方は変わり、新しい単語が生まれるとともに単語の組み合わせや使われ方にも新しいものが生まれます。例えば、スイカという単語は、フルーツの西瓜の意味で他の食品を表わす単語と一緒に文書に現れていましたが、現在では電子マネーのSuicaの意味で「駅」「支払」といった単語と一緒に文書に現れるようになっています。

そのため、21世紀に入ってからのNLPは機械学習に基づく手法が主流になりました。複数の望ましい解析結果（例えば品詞付きの文）を事例として与えることで、機械学習は自動的に規則性を見つけ、高性能な解析器を構築してくれます。新しいデータを追加して学習しなおすことで新しい表現に対応できることもメリットの一つです。

機械学習の対象としてのNLPは、当初から“大規模な問題”であることが知られていました。1990年代に作られた品詞タグ付や係り受け解析のベンチマーク・データ[4]でさえ、100万単語以上のテキストデータで、単語の種類も2万以上になります。各単語を処理単位とする問題（例えば品詞タグ付）では、データ数が100万以上になります。また、機械学習の入力データはベクトルで表現しますが、NLPではベクトルの各要素は単語や単語の組み合わせを示すように設計します（**図3**）。単語の種類が数万あるため、その組み合わせを変数にすると、頻度などで足切りしても数十万から数億の変数を扱う問題になります。過去においては100〜1000変数の問題でも大規模問題と考えられていたことを考慮すると、当時からNLPは桁違いに大規模な問題であったことが分かります。

一方で、100万単語ではさまざまな表現をカバーできないことが知られています。Web上のテキストが大量に入手できるようになったため、2014年時点では10億〜1000億単語のテキストデータを扱うことも一般的になってきています。

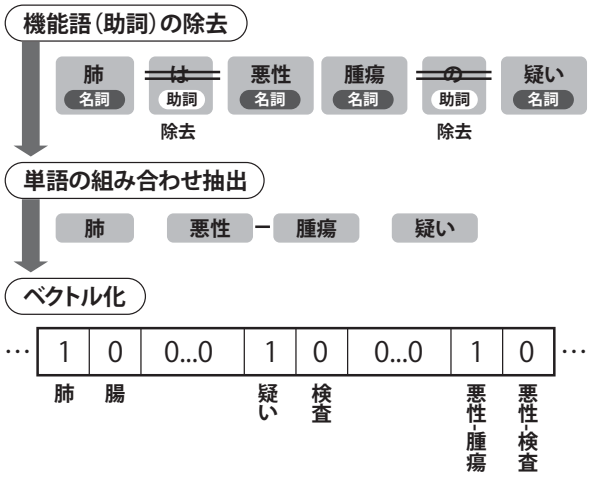

図3. 文のベクトル表現

4. 自然言語処理を支える機械学習技術

大規模問題に対処するため、NLPでは積極的に新しい機械学習技術を取り入れてきました。新しい技術の例として「オンライン学習」と「スパース正則化」を紹介します。

オンライン学習は、一つずつデータを見るごとに逐次的にモデルを更新していく学習手法です。従来の機械学習（バッチ学習）では何度も全データを参照する必要がありましたが、オンライン学習は全データを数回参照するだけでも高精度を実現できます。また全データを主記憶上に保持する必要もありません。オンライン学習のおかげで、従来の手法では扱えなかった大量のテキストデータを使って学習することが可能になっています。

一方のスパース正則化は、予測に必要な変数を比較的低い計算コストで見つけ出すことができる手法です。単語の組み合わせなどを機械的にベクトルの要素にして学習させると自動的に分類規則を見つけてくれることが機械学習のメリットですが、データを増やせば増やすほど扱うべき変数も増えてしまいます。スパース正則化によって予測に貢献しない変数を考慮する必要がなくなるため、膨大な変数を扱う必要のあった解析エンジンの主記憶使用量を一桁以上減らすことが可能になりました。

これらの機械学習の最新技術を使って構築されたNLP基盤はIBMのテキスト・マイニング製品[5]にも取り入れられており、高性能の自然言語処理を実現しています。

5. 自然言語処理における機械学習の今後

多様で多量なデータが増え続けている現在、自然言語処理の中で機械学習が重要な役割を占め続けることは間違いないでしょう。最近の研究では、

queen = king – man + woman

（王様から男性要素を引いて女性要素を足すと女王になる）

といった単語間の類推規則を、単語が出てくる文脈を大規模データから学習だけで獲得できたことが報告されています[6]。ある種の一般常識が大規模テキストデータを機械的に処理するだけで得られることは驚くべきことです。

この結果には広い応用先があると考えられており、自然言語処理の基盤技術の一つとなりつつあります。例えば、「悪性腫瘍」と「ガン（癌）」が同じ病気を示していることは先にあげた文書分類では重要ですが、数に限りがある診断書データではそうした類義関係が学習できないことがあります。診断書以外のデータでは必要なカテゴリ（コード）を予測するシステムは学習できませんが、幅広いテキストデータを用いて類義関係を学習しておくことで、診断書データの不足を補える可能性があります（**図4**）。

Web上の画像や動画と、それを説明するテキストのペアから言葉の意味を学習するマルチモーダルな試みも増えてきています[7][8][9]。例えば、ビーチ写真には空・海・砂浜が含まれていることが多いため、画像情報を通じて「ビーチ」と「空」「海」「砂浜」との関係を学習することが可能になります。また、動作を表す言葉は動画も併用して学習することが合理的です。人は五感を通じて知識を習得するため、NLP研究においても複数種類の入力を考慮することが注目を集めています。

さらに、fMRIなど脳の活動信号と言葉の関係の研究も進んできています[10][11]。テキストの世界にとどまらず多様なデータを扱う際には、数学的に抽象化した世界で統合することで可能な機械学習の強みがますます発揮されることでしょう。

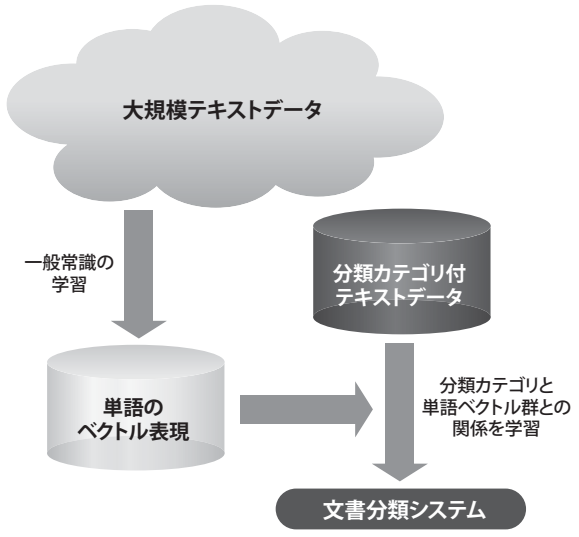

図4. 大規模データから取得した一般常識の利用

[参考文献]

[1] Josh Pollatsek, Dealing With ICD-10 — Computer-Assisted Coding Can Help Manage Transition, Radiology Today, http://www.radiologytoday.net/archive/rt0312p9.shtml

[2] 日本IBM, お客様導入事例: 明治安田生命保険相互会社, http://www-06.ibm.com/jp/solutions/casestudies/20130408meijiyasuda.html

[3] 高村 大也, 言語処理のための機械学習入門, コロナ社, 2010.

[4] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz, “Building a Large Annotated Corpus of English: The Penn Treebank”, Computational Linguistics, 1993.

[5] 日本IBM, 非構造化情報の自然言語分析, http://www-03.ibm.com/software/products/ja/category/content-analytics

[6] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig, “Linguistic regularities in continuous space word representations”, Association for Computational Linguistics,, 2013.

[7] Nitish Srivastava and Ruslan Salakhutdinov, “Multimodal Learning with deep Boltzmann Machines”, Advances in Neural Information Processing Systems, 2012.

[8] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, Andrew Y. Ng, Transactions of the Association for Computational Linguistics, 2014.

[9] Haonan Yu and Jeffrey Mark Siskind, “Grounded Language Learning from Video Described with Sentences”, Association for Computational Linguistics, 2013.

[10] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. “Predicting human brain activity associated with the meanings of nouns,” .Science, 2008

[11] Alona Fyshe, Partha Talukdar, Brian Murphy, and Tom Mitchell, “Interpretable Semantic Vectors from a Joint Model of Brain- and Text-based Meaning”, Association for Computational Linguistics, 2014.

日本アイ・ビー・エム株式会社
東京基礎研究所
スタッフ・リサーチャー

坪井 祐太
Yuta Tsuboi

国際基督教大学教養学部、奈良先端科学技術大学院大学情報科学研究科を経て、2002年にIBM東京基礎研究所に入所。博士（工学）。テキスト・マイニングの研究・開発に従事。2009年度人工知能学会現場イノベーション賞（金賞）、平成21年度情報処理学会論文賞を受賞。