「ことば」を工学する

精密工学研究所
奥村・高村 研究室

奥村 学 教授　1962年京都府生まれ。東京工業大学大学院理工学研究科情報工学専攻博士課程修了。2009年より東京工業大学精密工学研究所教授。

高村 大也 准教授　1974年静岡県生まれ。奈良先端科学技術大学院大学情報科学研究科自然言語処理学専攻博士課程修了。2010年より東京工業大学精密工学研究所准教授。

奥村・高村研究室では、人間が使用する「ことば」をコンピュータ上で処理し、さまざまな応用を試みる自然言語処理という学問分野について研究している。本稿では、テキストに秘められた感情的な側面を汲み取ることができるようなシステム開発についての研究、そして現在大きな注目を浴びているSNSを対象にした研究を主に紹介する。

現代社会における情報

インターネットが私たちの生活の中に取り入れられるようになってから、約20年が経った。普及当初から現在に至るまで、情報を媒介するコミュニケーションツールとして電子掲示板やブログといったものが主に利用されてきた。特にここ数年では、TwitterやFacebookといったSNSの登場によって、私たちはより多くの情報をより速くやり取りできるようになり、今でもそのさらなる進歩は続いている。

情報社会と称される現代では、以前に比べ情報に対する関心が高まっている。このような社会では、自ら情報を得ることが重要であり、その手段としてインターネットを活用することはごく一般的なことであるだろう。しかし、必要な情報が断片的に存在しているとき、インターネット上に飛び交う膨大なデータの中から自分の欲する情報だけを効率よく抜き出すことは難しい。

この問題を解決する手段の一つとして、自動要約という技術がある。これは長い文章から重要な要素を取り出し、コンピュータが自動で要約を生成する技術である。実際に自動要約技術は多岐にわたって応用されており、身近な例としてインターネットの検索エンジンが挙げられる。

この自動要約技術には自然言語処理という分野の技術が応用されている。自然言語とは、コンピュータの動作のために用いられるプログラミン

グ言語に対して、人間が意志疎通を行うための一般的な手段として用いる言語のことを指し、日本語や英語をはじめとした多くの言語が自然言語に該当する。つまり自然言語処理とは、人間のことばをコンピュータに理解させる研究分野であるのだ。また、自然言語処理は情報工学や言語学などと関連して長らく研究が続けられており、その応用例は数知れない。

奥村・高村研究室では自然言語処理を研究テーマに掲げ、いろいろな特性をもったテキストを対象にさまざまな観点からアプローチしている。本稿では先生方の数ある研究の中からいくつかを紹介していく。

形態素解析とは

テキストから得られる情報は、単に文面から得られる情報そのものだけでなく、書き手の主観的な感情を反映した副次的な情報を含んでいる場合がある。このような情報は、私たちが感覚的に理解しているものであるため、具体化しにくい。しかし、ネット上には口コミやレビューといった形で、こういった主観的感情が大きく寄与しているテキストが多数存在している。そのため、感情が秘められた大量のテキストを分析することができるようになれば、その恩恵は非常に大きいものだろう。例えば、ある製品について書かれた膨大なレビューを一つひとつ解析することで、企業はその製品に対する改善策を打ち出すために有用な情報を得ることができる。

しかし、これらの作業を人の手で行おうとすると、時間的、経済的に膨大なコストがかかってしまい、現実的な手段とは言えない。そこで奥村・高村研究室では、膨大なテキストをコンピュータによって自動で処理し、それらを「ポジティブ（肯定的）な文」「ネガティブ（否定的）な文」「ニュートラル（中立的）な文」という3つの極性に分類するシステムの開発を試みた。では、実際にどのようなシステムであるのか、順に説明していこう。

私たちが普段何気なく読み書きしている文はコンピュータにとってはただの文字の羅列にすぎない。そのため、まずはコンピュータに文字の羅列を文として理解させる過程が必要である。コンピュータ上で言語を処理する際にまず欠かせないのが、文を小さい単位に区切り、理解しやすくする形態素解析を行うことだ。

形態素とは言語において、意味を有する最小の単位のことで、本稿では単語のことを指すと思ってもらえればよい。形態素解析で行うことは、文を単語に区切り、それと同時に品詞を定めるということである。英語のように、単語同士がスペースによって区切られている言語については形態素解析を容易に行うことができるが、日本語は単語同士が区切られることなく連続的であるため、他の言語より複雑な形態素解析技術が必要となる。

形態素解析を行う際の基本的な動作は、あらかじめ用意された辞書データにコンピュータがアクセスし、解析したい文に含まれる単語を探っていくことである。コンピュータは文頭から順に文を探っていき、辞書データにある単語を見つけ出していく。この作業を文末まで繰り返していくと、文がすべて単語に区切られた形ができるのだ。

ここで、「彼女は図書館にいます」という例文をとって考えてみよう（**図1**）。この例文で考えてみ

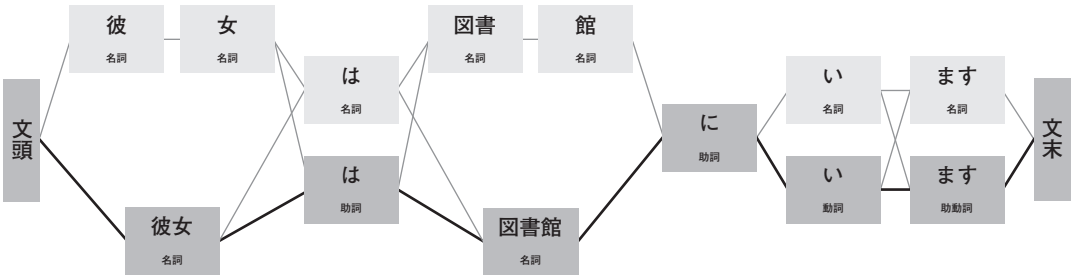
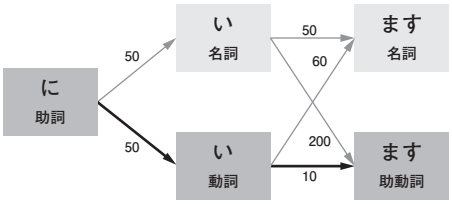


図1　形態素解析のグラフ構造

文に部分的に一致する単語を辞書データから抜き出していく。この例文だけでも2^5=32通り以上の分割パターンができることがわかる。

ると、最初の「彼」という文字に注目するだけで「彼」「彼女」という名詞を考えることができ、それに続く「は」という文字に注目すれば名詞の「葉」や助詞の「は」などについて考えることができる。この例で見てもわかるように、起こりうるすべての分割パターンについて考えていくと、最終的にその分割パターンの総数は膨大な数になってしまうことになる。

そこで、この分割パターンから、正しく区切ることができているものを絞り込むためにコスト最小法という手法が用いられる。この方法は、形態素解析で得られたそれぞれの分割パターンについてどの程度自然な文であるかを点数化し、順位付けするものである（**図2**）。この点数の根拠となるものは、品詞同士の繋がりがどの程度自然であるかということだ。例えば、一般的に名詞と助詞が文の中において隣同士で使われることは多いが、名詞と助動詞、という組み合わせは一部を除けば使われることのない組み合わせである。このように、それぞれの品詞同士の繋がりやすさをコストという形を用いて表していく。すなわち、よく使われる品詞の組み合わせに対しては低いコストをつけ、あまり使われることのない品詞の組み合わせに対しては高いコストをつける。そして、これらのコストの合計を形態素解析で得られた分割パターンごとに割り出していく。その中で最もコストの低いものを正しい解析が行われた結果として判断するのだ。



に（助詞）＋い（動詞）＋ます（助動詞）＝ 60
に（助詞）＋い（名詞）＋ます（名詞）　＝ 100
に（助詞）＋い（動詞）＋ます（名詞）　＝ 110
に（助詞）＋い（名詞）＋ます（助動詞）＝ 250

図2　コスト最小法の一例
単語同士の繋がりにコストが設けられ、合計コストが算出される。

極性分類と機械学習

前述の方法にしたがってテキストが解析されることで、文字の羅列が単語で区切られた文として認識されるようになる。これに続く手順として、本題のテキストの極性分類が行われることになる。

一般に、コンピュータで単語の感情極性を分類することはそれほど難しいことではない。例えば、「美味しい」「素敵だ」といった単語は、辞書データを引用することですぐにポジティブな表現であることがわかるが、逆に「まずい」「醜い」といった単語は、いずれもネガティブな意味を含んでいることがわかる。しかし、複数の語から構成されるもの、すなわち文やフレーズの形になっているものを同じように分類するとどうだろうか。「背が高い」という文は良い意味で使われることが多いが、「背」や「高い」といった単語それ自体は必ずしもポジティブな意味を表すものではない。また「事故によるけが人がいない」という文については、「事故」「けが人」といったネガティブな意味を含んでいるにもかかわらず、文そのものはポジティブであり、極性の反転が起きている。このように、複数の語を対象に評価を行う際にはさまざまな難題がつきまとう。

このような難題を解決するために、奥村・高村研究室では、機械学習という手法を用いてテキストを統計的に処理することを考えた。機械学習とは、コンピュータがデータの中からルールを自動的に獲得することができるようなプログラムを与えることで、人間における学習能力をコンピュータ上で実現させる技術である。

極性分類で用いられる機械学習の手法は教師あり学習と呼ばれる。教師あり学習とは、人間が例題となるような訓練データを与えることによって、コンピュータがそれを統計的に解析し、自らルールを獲得する学習法である。例えば、文Aは「ポジティブ」、文Bは「ネガティブ」、文Cは「ニュートラル」、文Dは「ポジティブ」、……といったように、事前に人の手で感情極性が割り振られた訓練データをコンピュータの手本となるように与える。するとコンピュータはその訓練データからテキストの傾向や特徴を自動的に学習し、「○○という単語が使われるときはポジティブ極性をもつ確

率が高い」「××という単語が△△という単語と組み合わせて使われるときはネガティブ極性をもつ確率が高い」といったように、統計的な解析を行うことで確率モデルを生成する。そうすることで、極性分類が定められていない新しいデータに出会ったとき、訓練データから得た確率モデルにしたがって、文がどの極性をもつのかを算出できるようになるのだ。このように、機械学習を用いると、ある程度の量の訓練データを用意するだけで膨大なデータを処理することができるようになるのである。

また、コンピュータが獲得した確率モデルにしたがって極性分類のなされたテキストを、新たに訓練データとして活用することで、コンピュータの極性分類の性能がさらに向上するのではないかと奥村・高村研究室では考えた（**図３**）。分類がなされたテキストのうち、確実性が高いものだけを取り出し、それまでの訓練データに追加する。そして、それを用いて再びコンピュータに学習させ、同様にテキストの極性分類を行うのだ。実際、データを新しく追加する以前に行なった極性分類と比べて、分類性能がわずかに上がっているのがわかった。この操作を何度も繰り返すことによって、より正確な極性分類を行うことができるようになったのだ。

現在では、顔文字やネットスラングを含んだテキストや、SNSなどで多く見られるくだけた表現なども対象にして極性分類を行なっており、あらゆるテキストに対して極性分類が行えるように改良を続けている。

SNSを対象にした研究

奥村・高村研究室が現在力を入れている研究の一つとして、SNSを対象にした研究がある。その中でも比較的ユーザの多いTwitterに焦点を当て、奥村・高村研究室はTwitter上で投稿される多数のスポーツ実況を抽出し、その要約をする自動スポーツ速報生成に取り組んでいる。これは、自動要約の一種としても考えられるが、対象とするテキストがSNS上に寄せられるテキストであること、そしてその要約が速報であることから、生成される速報はいち早く発信され、状況を正確に表すことができているものでなければならない。ゆえに、従来の自動要約とは異なった技術が必要となるのだ。

この研究において最初に行われるのが、あるメインイベント（スポーツの実況においては1つの試合を指す）中に起きるいくつかのサブイベントの検出である。TVで生中継されているあるサッカーの試合をメインイベントとすれば、「選手がシュートしてゴールを決める」「ゴールキーパーが選手のシュートをはじく」「前半終了」といったものがサブイベントの例として挙げられる。そして、上に挙げたようなサブイベントが起きたときTwitterではそれに関連した投稿、すなわちtweetが急激に増えることが予想される。このように一時的にtweet数が上昇する現象をバースト現象と呼ぶ。バースト現象が検出されたとき、あるサブイベントが起こったのだと考え、そのサブイベントに対応した速報生成を行うのである。

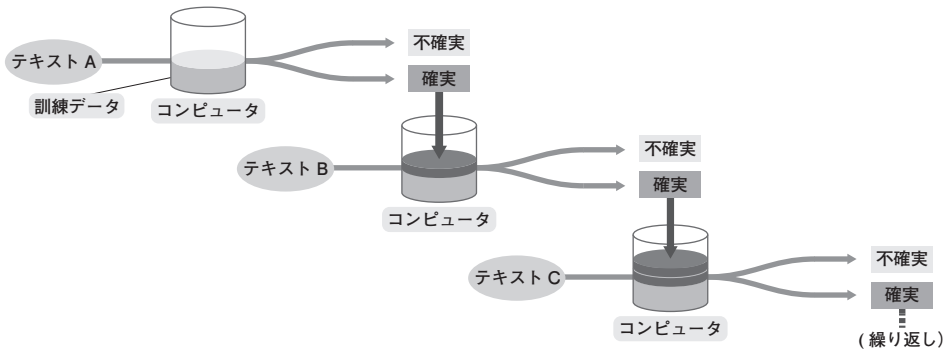


図３　機械学習のモデル
コンピュータが何度も学習を繰り返していくことで、分類の精度が上がっていく。

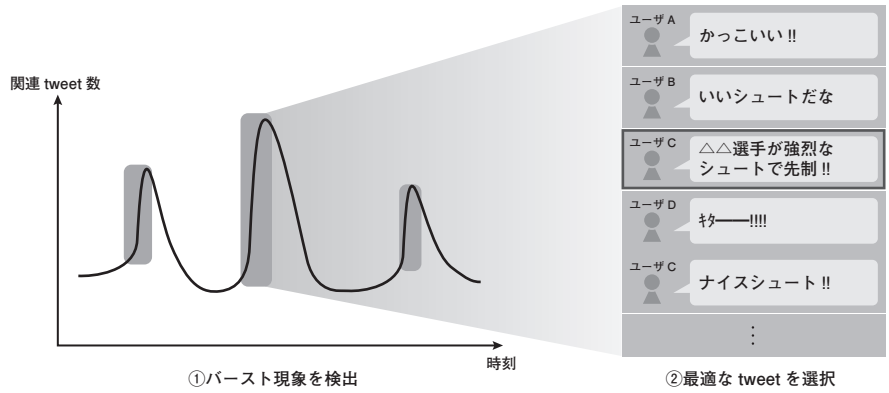


図４　バースト現象と自動スポーツ速報生成

バースト現象が検出されると、その際に投稿された多数のtweetから最も適切なものが1つ選択される。

次に、各サブイベントに対し、それぞれの要約が生成されていく。この研究では、バースト現象が観測されている間に投稿される膨大な数のtweetの中から、最も状況を詳しく説明しているtweetを手を加えずに1つ選出し、それをそのサブイベントにおける要約として定めることにした（**図４**）。そのため、いかにしてよりよい実況をしているtweetやユーザを探し出すかがこの研究での焦点となる。そこで奥村・高村研究室では、要約として使用するtweetを選び出す指標として、tweetがどれだけキーワードとなるような用語を含んでいるかを「tweetスコア」というもので表し、また、どのユーザがどれだけそのイベントに関連したtweetを多く投稿しているかを「ユーザスコア」で表した。そして、これらの2つのスコアを組み合わせ、より高い点数を獲得したtweetを速報要約としてピックアップすることによって、自動で速報を生成できるようにした。

奥村・高村研究室で行われている研究は数多くあり、特に今回紹介したようなインターネット上のテキストを対象にした研究は、自然言語処理において今最も注目が集まっている分野の一つである。それは、インターネットが現代の私たちの生活の中において重要な位置を占め、多くの人が興味、関心を示すからにほかならない。ブログや電子掲示板に始まり、今ではSNSというツールを媒介して見知らぬ人とも気軽にコミュニケーションをとることができる。奥村・高村研究室では、ブログ形式のウェブサイトが世間一般に浸透する前から、その潜在能力の高さにいち早く気が付き、率先してブログを対象にした研究を始めていた。

ブログやSNSのようなコミュニケーションツールが身近なものとなるにつれ、私たちはことばのもつ力を改めて認識させられるのではないだろうか。時として、人間が発信したことばは計り知れないほど大きな影響力をもつことがあり、それはインターネット上で交わされていることばについても例外ではない。

奥村・高村研究室では、インターネットを通じてことばのもつ力を生かすことができるような研究に精力的に取り組んでいる。そして、情報の海とも称されるインターネット上に溢れる大量のことばの中から、科学の力を用いることによって人の役に立つような面白いコンテンツを見つけ出そうと日々研究を続けている。

執筆者より

本稿を執筆するにあたって行なった取材では、自然言語処理に関連した先生方のさまざまな研究や取り組みについてお話を伺いました。先生方が説明してくださった研究はどれも興味深く、私たち学部生がなかなか触れることのない生の研究を身近に感じることができました。

最後になりますが、大変お忙しい中、快く取材を引き受けてくださった奥村・高村研究室のみなさまに心より御礼申し上げます。

（沖野 亮太）