

第1章

異分野融合研究のためのテキストマイニング

内田諭・大賀哲・中藤哲也

コンピュータおよび情報技術の発展で多くの言語データが利用可能となった。インターネット上には大量のテキスト情報が溢れ、研究利用を目的とした言語コーパスも多く公開されている。テキストマイニング(text mining)は、このような大量の言語データから意味のある情報を取り出すことを指す。「テキスト」(文章)から「マイニング」((情報の)発掘)を行う、ということである。近年では、より分析的な側面を強調して「テキストアナリティクス」(text analytics)という言葉で表されることもある(那須川 2018、坂地 2019 など)。

テキストマイニングを行うには、(1)データの収集、(2)分析・集計・可視化、(3)結果の解釈というプロセスを踏む(図1)。

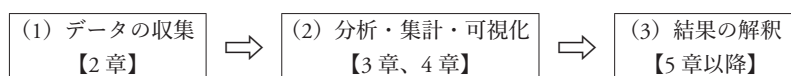


図1 テキストマイニングのプロセス

(1)の段階ではデータをどのように整理するか、という点がポイントになる。特に、データ分析を正しく行うためにはデータのクリーニングが必要で、文字情報の扱い方を理解しておく必要がある(2章参照)。

(2)の段階では、収集したデータをコンピュータで分析し、集計したり可視化したりする。最も単純な分析方法は、テキストごとの単語頻度の集計である。しかしながら、日本語を扱う場合は(英語の場合と異なり)単語間にスペースがないため、単純な頻度集計を行うに際してもソフトウェアを利用し

た形態素解析が必要となる(2章参照)。また、より高度な分析を行うためには、単語をあるカテゴリーに分類してそのカテゴリーの頻度を文書間で比較するということが必要となってくる。例えば、「権利」、「差別」、「平等」などの単語を『人権』というカテゴリーに分類し、「雇用」、「従業員」、「職場」などを『労働』に分類することで、特定の文書における『人権』や『労働』に関する単語の出現が多いか少ないかなどが明らかになる(コーディングと呼ばれる。詳しくは3章参照)。このような処理は、一定のプログラミングなどに関する技術が必要であるが、テキストマイニングツールを使うことで手軽に実行できる。本書の4章ではフリーで使える優れたツールを紹介し、読者が自分のデータで形態素解析や可視化が実施できるよう、操作方法などを解説している。

テキストマイニングを実施する際には、実施計画を入念に立てることが重要である。目的によって収集するデータの種類も異なり、データの整理の仕方やコーディングの方法も異なってくる。また、有効な統計手法や可視化の方法についても目的によって変わってくる。小林(2017)は、テキストマイニングのタイプとして、「仮説発見型」と「仮説検証型」があることを指摘している。前者の場合はターゲットとなるデータを決めてそれをある程度一般的な形で整理し(年代別、ジェンダー別など)、クロス集計や可視化などの結果から何か隠れた傾向がないかを発見するようなスタンスを指す。これが本来の意味でのテキスト「マイニング」であると言えるだろう。一方、後者の場合、より明確な目的を持って実験を行うもので、例えばある話題についてTwitterの投稿と新聞記事で使われている形容詞の種類を比較し、Twitterのほうが口語的で感情的なものが多い、といった仮説を検証するような場合である。この場合、分析の対象を絞ることができるため、データ収集のプロセスを省力化できる(一方、予想が外れた場合などにはデータを取り直す手間がかかることもある)。

以上の(1)データ収集、(2)分析・集計・可視化については非常に優れたガイドブックが多く出版されているが(石田2017、石田・小林2013、牛澤2018、小林2017、松村・三浦2014など)、(3)の結果の解釈の部分について、学術分野を横断した研究事例をまとめたものは非常に限られている。言語学分野の研究事例として岸江・田畑(編)(2014)などがあるが、テキストマイニングの研究は、言語学のみならず、より広い分野に応用可能である。テキ

ストデータは社会学、政治学、教育学、看護学など様々な分野で蓄積されているものの、「そのデータをどのように料理すればよいのか」ということがわからず、日の目を見ないことも多い。しかしながら、このようなテキストデータを分析することで、これまで明らかになっていなかった様々な事象が発見され、当該分野のみならず、他分野への示唆を得ることができる可能性がある。テキストマイニングの結果の解釈には、他分野の知見が必要となる場合が大いに想定され、その意味でテキストマイニングを使った研究は、自然と異分野融合的な性質を帯びることが多い。

本書の実践編(5章以降)では、様々な分野のテキストマイニングの実践例を紹介する。5章はテキストマイニング的なアプローチを言語学の研究に実践した事例で、「構文」についてコーパスデータから実態を明らかにしようとするものである。「構文」は「形式と意味の対応付け」を指し、昨今の言語学(特に認知言語学)では広く議論されているが、「二重目的語構文」や「使役移動構文」など特定の構文に関する研究が多い。本章では、言語データからより広範な構文をマイニングするための試みが示されている。独自の旅行課題遂行会話データベースから、パターンマッチングの手法を用いて場面に応じた構文の抽出を試みる。

6章は情報学の分野からの研究事例で、機関(大学や研究所など)に所属している研究者の論文などが公開されている「機関リポジトリ」を対象として、研究者同士のつながりを可視化するという試みである。トピックモデルという手法を用いて、収録されている論文のトピックを推定し、それによって研究者のリンクを辿るというアプローチが用いられている。分野横断的な論文を対象としているという点で学際的な研究事例であると同時に、この研究の成果が研究者同士のつながりを生み出し、共同研究のきっかけになる可能性があるという点で、異分野融合研究を推進するための貴重な試みとなっている。

7章では図書館情報学での研究事例を数多く紹介している。「図書館」に関わる諸問題を扱うこの分野では、自然とテキストを対象とした研究が多く、テキストマイニングは一般的な手法として広く用いられてきており、様々な研究の蓄積がある。その中には図書の自動分類、図書の推薦、文献の生産・流通・利用などを統計的に分析するビブリオメトリックス(計量書誌学)などが含まれており、本章ではこれらの具体的な研究事例が示されている。

8章は政治学分野の研究事例である。有権者の政治的態度の一貫性に国際政治学の理論的枠組み(リアリズム・リベラリズム・コンストラクティビズム)がどの程度反映されているか(この枠組みが一般有権者の分析にも有用であるか)ということ、アンケート調査を通して明らかにするものである。この論考では、北朝鮮ミサイル問題や日本の核武装の問題など、前述の3つの理論的枠組み(3つのism)によって違いが出ると考えられるトピックについて、客観的な設問と自由回答記述によってそれぞれの立場の特徴を明らかにすることを試みる。コーディングルールを用いてテキストマイニングを実施することで、それぞれの立場のキーワードを抽出し、有権者の回答の裏にある政治的態度を明らかにしている。

9章は教育現場における評価の手法の1つとして、テキストマイニングを応用した研究事例である。教育は言語を介して行われることが多いため、テキストデータに溢れている。例えば、「感想文」などのようなものもテキストデータであり、また議論の内容について文字起こしを行えばそれもテキストデータとして扱うことができる。しかしながら、このようなデータは、客観的に評価を行うことが難しい。本章では、対馬の水産業に関わる外部指導者を活用した協働学習の事例を取り上げ、外部指導者と小学生の発話をテキスト化し、ネットワーク分析によって議論の内容を明らかにしている。これにより、外部指導者から提供された知識がどのように議論に取り入れられているかなどについて可視化することができ、協働学習の評価に客観的な要素を取り入れることが可能となることを示している。

10章は、社会学におけるテキストマイニングの活用事例である。社会に大きなインパクトを残す出来事は、テレビや新聞などのメディアで大きく報じられることが多い。特に新聞の場合、情報が文字化され、またアーカイブとして過去の記事を検索することができ、テキストマイニングを行う上で重要な資源となる。本章では、1986年のチェルノブイリ原発事故に関する新聞記事の計量テキスト分析を行ったものである。1986年、1996年、そして東日本大震災が起きた2011年の新聞記事を時系列で比較し、それぞれの時代に広く共有された情報をテキストマイニングの手法を用いて客観的に示している。

11章は、看護学におけるテキストマイニングの研究事例で、生体肝移植ドナーの語りから、看護学への示唆を抽出する試みである。これまで看護学

生や看護師など、医療を実施する側の研究が多く行われてきており、看護学教育のカリキュラムの改善などに役立てられてきたが、患者やその家族などの「語り」を対象とした研究も増えてきている。本章では、生体肝移植のドナーとなった患者の語りを書き起こしたデータから、ネットワーク分析などのテキストマイニングの手法を用いてその内容を可視化している。医療現場のコミュニケーションには感情や人生哲学といった数値化できない抽象的な思考が現れることが多いが、本章の研究事例からテキストマイニングの手法によってこれらを客観化することができることが示されている。

12章は、社会調査におけるテキストマイニングについての論考で、「テキストマイニングを社会調査データにどのように活かすとおもしろい研究ができ、学術論文になるか」ということを論じている。人生について時系列に語ったデータ(ライフヒストリー法)を例に取り、テキストマイニングによって内容を客観的に要約することだけではなく、言葉の持つ曖昧性を分析することで有意義な研究となることを示している。ネットワーク分析などを利用することで、ある単語と共起する語が人生の前半と後半において異なるということを明らかにし、テキストに潜む意味の変化を示す。

13章は、デジタル・ヒューマニティーズ(以下DH)におけるデータベースの構築についてレポートしたものである。DHは人文学にデジタル技術を応用する研究分野であり、ITの発達に伴って2000年代以降に盛んになっている。この分野の基礎となるものは電子化された歴史的な書物や文学作品などのデータである。DHで用いられるデータには画像のイメージファイルやテキストデータなどが含まれる。本章では、約2,900点の仏典関連資料の集成である大正新脩大藏経をデータベース化したプロセスが記録されている。この資料には特殊な文字が含まれているため、それを電子化すること自体が困難なことであり、さらに検索しやすい形にすることは大きなチャレンジである。本章はこのような難題を伴うテキストデータの電子化をいかに進めたかについての貴重な報告となっている。

実践編の各章は、当該分野におけるテキストマイニングの先行事例を含むように構成されている。これらの文献レビューに加えて、本書で示す多分野にまたがる論考は、テキストマイニングを活用した異分野融合研究の手がかりとなるだろう。テキストマイニングは、分野ごとに形成された専門的知見の統合を促す可能性がある。テキストの定性的な内容に詳しい研究者とテ

キストの定量的な分析を行う研究者が共同研究を組織することで、従来は意識されなかった要素が可視化される。また、そうした共同研究を通じて、複数の分野のテキストを比較・統合・再構築することが可能になる。テキストマイニングは、個別分野ごとの知見を統合し、知を再構築するポテンシャルを持っているのである。

参考文献

- 石田基広(2017)『Rによるテキストマイニング入門』第2版。東京：森北出版
- 石田基広・小林雄一郎(2013)『Rで学ぶ日本語テキストマイニング』東京：ひつじ書房
- 牛澤賢二(2018)『やってみよう テキストマイニングー自由回答アンケートの分析に挑戦!』東京：朝倉書店
- 岸江信介・田畑智司(編)(2014)『テキストマイニングによる言語研究』東京：ひつじ書房
- 小林雄一郎(2017)『Rによるやさしいテキストマイニング』東京：オーム社
- 坂地泰紀(2019)「テキストマイニングからテキストアナリティクスへ」『情報・システムソサイエティ誌』24(2), 4-5.
- 那須川哲哉(2018)「テキストアナリティクスと特許情報分析」『情報の科学と技術』68(7), 326-331.
- 松村真宏・三浦麻子(2014)『人文・社会科学のためのテキストマイニング』改訂新版。東京：誠信書房