こころの測り方

テキストを機械が要約するために

大阪大学大学院人間科学研究科 助教
村中誠司（むらなか　せいじ）

読者のみなさんは，これまでに数多の文書を読み，それらがどのような話題の文書であったかを心に留めたり，記録したり，誰かと共有したりしたことがあるかと思います。文書を読んだとき，その文書の特徴をつかみ，「どのような話題であるか」を判断する方法の一つとして，キーワードを拾い，その出現頻度から推論することが挙げられるかと思います。「晴れ」「雨」「くもり」「前線」といったキーワードがよく出現した文書であれば，これは天気について話しているのだろうと想像すると思います。他方，「マスク」「三密」「アルコール」といったキーワードが頻出していれば，現状を踏まえると新型コロナウイルス感染拡大やその感染対策についての文書であるのかなと想像する方が多いかと思います。

上記のような，文書の特徴をつかみ，話題を推論する過程を，コンピュータにさせることは可能でしょうか。情報科学では自然言語処理という技術の開発が進んでおり，機械学習技術を活用することで実現しつつあります。本稿では，この技術の一つである，トピックモデルについて紹介します。まずは，文書の特徴のつかみ方として，シンプルな手順で算出できる特徴量を紹介し，この特徴量を確率モデルに入力して得られるトピックモデルについて紹介したいと思います。

テキスト解析について本稿で述べるにあたり，数ある自然言語処理の技術の中でトピックモデルを選んだ理由は，筆者が最近扱った方法だからです[1]。その経験から，非常に効率的に，「良い感じに」文書を要約してくれることがわかりました。また，トピックモデルは非常に拡張性が高く，さまざまな文書の形に対応できるため，心理学研究と非常に相性が良いと考え，本稿でご紹介することにしました。

文書の特徴

文書の特徴の表し方の一つとして，冒頭ではキーワードの出現頻度を挙げました。この方法は，自然言語処理においてはバッグオブワーズ（Bag of Words; 以下，BoWと表記）と呼ばれる特徴量として幅広く使用されています。この特徴量は，まず全ての文書集合で出現する単語のリスト（これを語彙といいます）を作成して，次に文書ごとに単語をカウントしていきます。この文書ごとの単語のカウントデータがBoWです。順序や単語の意味は考慮しておらず，名前の通り「言葉のカバン」です。文書ごとのBoWを積み重ねていくと，語彙数×文書数の表ができます。これを検索語・文書行列（Term-Document matrix）といいます。BoWは算出方法が簡単でありながら，前述のキーワードの出現頻度の観点からうまく特徴を表せる強みがあります。

BoWには弱点もあります。第一に，単純な出現回数を算出するため，どの文書にでも存在するような単語は高い値を示してしまいます。この弱点については，ストップワードという，解析に関係ないと判断した単語をあらかじめ除去する前処理を行うなどして対応をします。第二に，多くの場合に共起すると期待される単語がたまたまその文書では現れなかったとき，その文書の特徴が人の感覚から離れてしまうことです。例えば，「晴れ」「くもり」「前線」といった単語が出て来ていたとき，「雨」という単語も共起すると多くの人が期待すると思いますが，もしたまたま確認したこの文書の中で一度も出現しなかったのであれば，BoW上は「雨」は0となります。

次節で紹介するトピックモデルでは，確率を通した単語の出現頻度を得ることができるため，「雨」の期待度を確率的に見積もることで上記の第二の弱点を補うことができます。

トピックモデルとは

トピックモデルとは，トピックとしてまとめられる潜在変数に基づいて観測された単語を生成する統計的なアプローチです。ここでのトピックとは，同じ文書で現れやすい語彙の集合を指します[2]。トピックモデルはおおまかに以下の手順で行います。①語彙の辞書を作成する，②文書ごとに，単語の出現をカウントし，BoWを作成し，検索語・文書行列を得る，③モデル訓練し，パラメータを推定する，④結果を可視化する，の4つの手順です。日本語の場合は，英語のように文章が空白文字で区切られていないため，①の前に形態素解析という，品詞情報を参照しながら文章を分割していく技術を使って文章の分かち書きをします。

トピックモデルには，いくつか拡張版が開発されています。その

うちの一つである構造的トピックモデルによる研究活用例を，次節でご紹介します。構造的トピックモデルは，トピックモデルに対して，トピック間の相関を考慮の上，文書のトピック出現確率を説明する共変量を設定でき，さらに限られたデータ量でも予測精度が向上するよう拡張されたモデルです。

トピックモデルの活用例

トピックモデルは文書の分類や要約などに活用されています。村中・竹林（2021）[1]では，遠隔心理支援に関する研究動向を明らかにするために，構造的トピックモデルによるアブストラクトの解析を実施しています。トピックを構成する単語をワードクラウドで表し，文書集合内でのトピックの出現確率をランク付けしました。さらに，本研究では論文の出版年を共変量としてモデルに組み込み，トピックの出現確率の経年変化を推定し，トピックの「流行り廃り」を可視化しました。その結果，578件のアブストラクトから12のトピックが抽出されました。その後の解析からモバイルアプリを活用したうつや不安への支援に関する検討が急速に伸びていることから，まずはこのような研究課題から優先されることがわかり，その他支援者へのサポートや予防的介入についても徐々に報告数が増加しており，今後検討する必要があることを提案しました。トピックモデルを文献のレビューに加えて行うことで，研究の方向性がよりクリアになるので，ある程度知見が蓄積された研究から新たな研究課題を抽出したり，卒論・修論などで研究テーマを考えたりする上で非常に参考になる情報が得られると期待しています。

テキスト解析の課題と展望

テキスト解析は，「テキストマイニング」とも称されながら，これまでに，幅広い領域で応用されてきました。精神医学領域にも応用されており，その方法は多様です[3]。方法論としてのアプローチの多様さは，妥当な方法を選択する上で障壁ともなり得ることですが，ソフトウェアのバージョンなどの開発環境がそろっており，プログラム，学習済みモデル，入力データがあれば追試ができるため，方法論としての議論はむしろしやすいと考えます。

また，テキスト解析は前処理が大変です。大量の文書データを相手にするので，手作業では到底追いつかないため，プログラムで処理することがほとんどだと思います。プログラムを組む上で，前処理を行うためのルールを設定したとしても，文書は人が作ったものですから，往々にして例外的なデータが出てきます。この例外を一つひとつ確認し，対処しながら進めていくのはなかなか骨が折れます。

今回は（構造的）トピックモデルについて紹介しました。本稿のタイトルに示すテキストの要約において，非常に有効な手法だと筆者は考えます。また，上述の通り，トピックモデルはその拡張性の高さから，多くの拡張版モデルが開発されています。時系列に対応したダイナミックトピックモデルは，カウンセリングなどの時間に対応したテキストデータに有効かもしれませんし，今回のような“マニュアルな”トピック数の決定をせずに，階層ディリクレ過程トピックモデルを使えば，最適なトピック数をあわせて推定することも可能です。

言わずもがな，ことばを扱う心理学研究において，自然言語処理技術は非常に強力なツールでありますが，質問紙の得点を集計し要約するように，テキスト解析においても要約が心理学への応用の第一歩だと考えます。トピックモデルはことばを要約する優れたモデルであるため，心理学研究との相性が良く，今後も応用例が蓄積されていくだろうと確信しています。

文　献

1 村中誠司・竹林由武 (2021).「遠隔心理支援（Telepsychology）におけるこれまでの検討課題：Structural Topic Model によるアブストラクト解析」『認知行動療法研究』47, 127–138. https://doi.org/10.24468/jjbct.20-026

2 岩田具治 (2015).『トピックモデル』講談社

3 Abbe, A., Grouin, C., Zweigenbaum, P., & Falissard, B. (2016). Text mining applications in psychiatry: A systematic literature review. *International Journal of Methods in Psychiatric Research, 25*(2), 86–100. https://doi.org/10.1002/mpr.1481

Profile — 村中誠司
国立精神・神経医療研究センター 認知行動療法センター 客員研究員を兼職。専門は臨床心理学。主著に『遠隔心理支援スキルガイド：どこへでもつながる援助』（共著，誠信書房）など。