7G-05

文脈解析に向けた作者別概念辞書の構築

秋岡明香†
明治大学 総合数理学部

1. はじめに

本研究では、J-POP における作詞家の語彙選択や文章組み立てにおける意図を分析し、可視化することを目指す。作者の作品や意図を構造化することは、作品の理解を深める助けとなるだけでなく、一般的な文章の文脈理解に役立つ。さらに、人工知能による表現や創作に個性を与えたり、ソフトウエアにおける人間らしさを提供したりといった応用も考えられる。

これまでの研究で、J-POP 歌詞に現れる語彙群は作詞家ごとに有意な差があり、語彙群の差と作品や作詞家への印象は、強い相関があることがわかった[1]。これにより、作詞家特有の語彙群と一般的な概念辞書の差分は、作詞家の意図が反映され個性に繋がる部分であると推定できる。そこで本稿では、作詞家ごとの語彙群を構造化し、一般的な概念辞書との比較を行う。

なお、形態素解析を超えた文学作品の分析研究には、トピックモデルによる推理小説の研究[2]、N-gram を用いた近代短歌の研究[3]、夏目漱石作品の研究[4]などがある。これらの研究では、主に作者ごとの文体に注目している。

2. 作者別概念辞書の可能性

作風に大きな違いがあると広く認知されている作詞家 2 名（作者 A および B）について、曲ごとに出現単語群を生成し、TF-IDF[5]を計算した後に、k-means[6]で各曲を 2 クラスタへ分類した結果を図 1 に示す。なお、図 1 では UMAP[7]で特徴量を 2 次元に削減しており、各色は各作者を表す。一方で、同じ作品群について、LDA[8]による 2 クラスタへの分類を試みたが、収束しなかった。

これにより、トピックモデルでの作者識別は困難であるが、単語選択や使用頻度により作者を識別できる可能性があることがわかる。つまり、作者ごとに重点的に用いる概念やカテゴリに特徴があり、作者別の概念辞書が作者ごとの特徴を捉えている可能性がある。

図 1　対象作詞家 2 名の使用語彙と頻度の比較

3. 作者別概念辞書の構築

本稿では、作者別概念辞書の構築起点として、日本語ワードネット[9]を利用する。日本語ワードネットは、英語概念辞書である Princeton WordNet[10]の synset に対して日本語を付与したものである。日本語と英語の synset の違いや、全エントリのうち 5%程度にエラーが含まれるなどの問題が指摘されているが、57,000 以上の概念と 93,000 以上の単語を収録しており、網羅性を重視した構成である。

最初に、作者ごとに歌詞データを収集し、曲単位で形態素解析・クリーニングを行い、曲ごとの単語群を生成する。次に、曲ごとに得た単語群から作者ごとの単語群を生成する。最後に、この単語群を日本語ワードネットにマッピングすることで、特定の作者に特化した概念辞書を得る。

4. 結果と考察

作者 A および B の各単語群について、出現頻度が 100 回以上の単語について、日本語ワードネットへのマッピングを試みた。作者 A は出現単語群のうち約 1%である 57 単語が、作者 B は出現単語群のうち約 0.4%である 25 単語が対象である。このうち 17 単語は、両者に共通した頻出単語であった。表 1 に、両者に共通した頻出単語、A のみ

Lexical and conceptual database of lyrics organized by songwriters.
† Sayaka Akioka, School of Interdisciplinary Mathematical Sciences, Meiji University.

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

に頻出した単語、B のみに頻出した単語の例を示す。また、表中の太字は、日本語ワードネットに収録されていない単語であることを示す。

表 1　頻出単語の例

共通	**私**/**あなた**/**何**/心/愛/夢/見る/**君（きみ）**/忘れる/来る/時/今/遠い/明日
A のみ	人/**あたし**/**誰**/くれる/言う/**僕**/風/知る/泣く/女/それ/**どこ**/夜/**誰か**/変わる/笑う/雨/日/**おまえ**/わかる/流れる/呼ぶ/生きる/**みんな**/思う/者/行く/歌/涙/手/**ひとつ**/言葉/男/海
B のみ	恋/**Me**/遠い/二人/いつか/消える/空/見える

日本語ワードネットには、口語表現、一人称、二人称、三人称などは含まれない。また、今回の 100 回以上の頻出単語には該当しないが、作者独自の造語も当然含まれない。しかし、こうした表現に作者の意図や特徴が表れることは多く、無視できない存在であるため、概念辞書に的確に取り込む手段が必要である。

また、「君」と「きみ」などの表記の違いは、形態素解析などの段階で統一されてしまう場合も多いが、表記方法にも作者のこだわりは表れる。表 1 では「君」と「きみ」は同じ意味としてまとめているが、「君」は A の頻出単語、「きみ」は B の頻出単語である。このような明示的でない意味や概念の違いをどのように構造化するかも重要な課題である。

A および B のそれぞれに固有の頻出単語のうち、名詞に着目すると、A は「女」や「男」など、上位概念で person に属する語を含む。「あたし」などの日本語ワードネットに含まれない単語の中にも、person に属すると推測できる単語が複数存在する。また、「風」や「雨」は weather に属する単語である。これらの考察から、A は特定の概念に含まれる単語群を選んで使用する傾向があると考えられる。B については出現頻度が高い単語の種類が少ないこともあり、傾向を掴むことは難しい。動詞についても同様で、A では「笑う」と「泣く」など、express_emotion に属する単語が頻出している。しかし B については特徴を捉えることができなかった。

一方で形容詞については、文脈により意味が変化する場合があり、今回のような統計的な扱いの中で比較を行うことは難しい。しかし、形態素解析の段階で微妙な差異を考慮することも難しく、こうした繊細な意味の違いをどのように扱うかも課題である。

5. まとめ

本項では、J-POP 歌詞に現れる語彙群に作詞家ごとの有意な差があることに注目し、日本語ワードネットを用いて作詞家ごとの語彙群の違いを構造化することで、作詞家の意図や個性に繋がる特徴を得ることを試みた。日本語ワードネットのような一般的な概念辞書を用いることで、一定の特徴抽出を行うことは可能であるが、概念辞書自体の拡充や表記揺れへの対応など、課題が多いことも明らかになった。

参考文献

[1] 秋岡, “意味ベクトル表現を用いた J-POP 歌詞の文体分析”, 情報処理学会第 82 回全国大会講演論文集, 2020(1), 2020.

[2] 黒田, “19 世紀の推理小説：機械学習アプローチによる文体分析”, 情報処理学会研究報告, Vol. 2019-CH-119, No. 10, 2019.

[3] 村田, “N-gram 統計を用いた近代短歌テキストの分析”, 情報処理学会研究報告, Vol. 2019-CH-120, No.9, 2019.

[4] 土山, “文末表現の計量分析に基づく夏目漱石の小説の分類”, 情報処理学会研究報告, Vol. 2019-CH-120, No.6, 2019.

[5] G. Salton, et al., “Introduction to Modern Information Retrieval”, McGraw-Hill Inc., USA, 1983.

[6] J. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations”, Proc. 5th Berkeley Symp. on Math Statist. and Prob., Vol. 1, pp. 281 – 297, 1967.

[7] L. McInnes, et al., “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”, ArXiv e-prints 1802.03426, 2018

[8] D. M. Blei, et al., “Latent Dirichlet Allocation”, J. of Machine Learning Research, Vol. 3, 2003.

[9] F. Bond, et al., “Japanese SemCor: A Sense-tagged Corpus of Japanese”, The 6th International Conference of the Global WordNet Association (GWC-2012), 2012.

[10] Princeton University, “WordNet | A Lexical Database for English”, https://wordnet.princeton.edu, 2021 年 1 月 7 日取得.

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.