6H-01

意味ベクトル表現を用いた J-POP 歌詞の文体分析

秋岡明香†
明治大学 総合数理学部†

1. はじめに

自然言語処理の進化により、文書解析手法が変化している。従来の統計的手法に加えて、語句の意味や文脈を加味した文書解析が可能になり始めた。特に、機械学習で大量の文書を解析し、単語の意味や文書の特徴を表すベクトルを生成する手法が注目を集めている。単語や文書をベクトル化することで、類似度計算が明確になる、単語の意味を複数単語のベクトル演算で表現できるといった利点がある。

本研究では、これらの意味ベクトルを用いた言語モデルを J-POP 歌詞に適用し、文体分析を行なうことを目指す。多くの作品を、人手で抜けなく均質に解析することは難しい。本研究は、精読的分析を補助し、同時に従来の文体分析とは異なる観点を提供する。そのために、本稿では、意味ベクトルを用いた複数の言語モデルを文体分析の観点から比較する。形態素解析を超えた文体分析の既存研究には、トピックモデルによる推理小説の研究[1]、N-gram を用いた近代短歌の研究[2]、夏目漱石作品の研究[3]などがある。

2. 特定作詞家に注目した言語モデル比較

単語ごとのベクトル（単語ベクトル）を提供する Word2Vec[4]と fastText[5]、さらに文書ごとのベクトル（文書ベクトル）を提供する Doc2Vec[6]を用いて、特定の作詞家の楽曲群を分類する。各言語モデルから得ることができる文体情報を比較することが目的である。

分類を行なう上で、Doc2Vec では文書ベクトルの類似度を特徴量とした。Word2Vec および fastText では、単語ベクトルから擬似文書ベクトルを生成し、類似度を求めた。ある曲の擬似文書ベクトルは、その曲に出現する単語の単語ベクトルの相加平均とする。学習データは、曲名をラベルとし、歌詞を形態素解析・クリーニングして得た曲ごとの単語群とする。今回は約 450 曲を用いた。この学習データでモデルを学習させ、学習済みモデルに再び同じデータを入力して特徴量を得る。どの言語モデルも、学習パラメータの最適化が不可欠である。しかし最適化指標は既知でなく、以下の 2 点を満たすことを条件とした（最適化条件）。

1. 同タイトルのバージョン違いの楽曲群すべてが、オリジナルと同じクラスタに含まれること。
2. ある楽曲とその返歌が同じクラスタに含まれること。

比較のために、LDA[7]で同楽曲群を分類した結果を図 1 に示す。図 2 は Word2Vec、図 3 は fastText、図 4 は Doc2Vec を用い、k-means[8]で分類した結果である。いずれもクラスタ数は 10 であり、クラスタごとに色分けした曲名をプロットした。図 2〜図 4 は、UMAP[9]で 2 次元に次元削減した図である。どのモデルも実行時の乱数要素を完全に排除できないため結果が毎回異なるが、大きな傾向は変わらない。

LDA は最適化条件 2 を満たすことが極端に困難であった。対象 2 曲を比較すると、モチーフは明確で、共通する語句が少なくない。他の言語モデルは学習パラメータの最適化でこの条件をクリアした。また、LDA は明確な分類を行なうが、各クラスタの意味づけは難しかった。

図 2 が示すように、Word2Vec は極端に曲数が少ないクラスタを形成する傾向があった。少曲数クラスタには 2 種類の傾向があった。まず、特定の 1 曲が頻繁に単独クラスタを形成した。この楽曲には、他の楽曲とは全く異なる分野の単語を多用している、ある単語が繰り返され他の語句の繰り返しはほぼ無い、などの際立った特徴がある。さらに、最適化条件 1 の楽曲群を単独クラスタとする頻度が高かった。Word2Vec は連続単語集合モデルを使用している。作詞において語の並びは重要であると推測できるため、歌詞の文体分析と Word2Vec の親和性が高かったと考えられる。

fastText は分類の再現性が高かった。他の言語モデルでは、モデル構築後、特徴量計算時に最適化条件 2 を満たさない現象が頻発した。fastText も同様だが、その頻度は低かった。また、特定の語句を極端に繰り返す楽曲群でクラスタを形成する傾向があった。ただし、同傾向の楽曲群を網羅的に単独クラスタにすることはなく、異なる傾向をもつ楽曲群も同じクラスタに混在した。繰り返し以外の評価軸を含むと推測できるが、その詳細は明らかできなかった。

図 4 が示すように、Doc2Vec は各クラスタの意味づけが困難であった。いずれの言語モデルも LDA と比較してクラスタ境界が曖昧であるが、Doc2Vec は特にその傾向が強かった。最適化条件の見直しや、Doc2Vec が生成する単語ベクトルと Word2Vec や fastText が生成する単語ベクトルの比較など、詳細な調査が必要である。

3. 対比される作詞家との文体比較

2 章で注目した作詞家には、楽曲の方向性などの違いから、

Stylistic analysis on J-POP lyrics by meaning vectors
† Sayaka Akioka, School of Interdisciplinary Mathematical Sciences, Meiji University

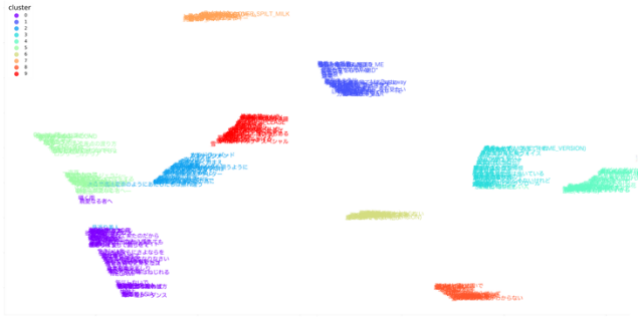

図 1　LDA による 10 クラスタへの分類

図 2　Word2Vec 単語ベクトルの相加平均を特徴量とした K-means によるクラスタリング

図 3　fastText 単語ベクトルの相加平均を特徴量とした K-means によるクラスタリング

図 4　Doc2Vec 文書ベクトルを特徴量とした K-means によるクラスタリング

頻繁に対比される作詞家が存在する。言語モデルが二者の文体を区別するか検討するため、先述の 4 言語モデルで 2 値分類を行なった。モデル構築時のパラメータや楽曲の類似度計算、分類手法等は 2 章と同様だが、データラベルは作詞家名とし、対比される作詞家による歌詞約 400 曲を追加で用いた。さらに TF-IDF[10]でも同様の実験を行なった。

結果は、TF-IDF のみが二者を完全に区別した。このことから、二者は語句選択や使用頻度が特徴的であり、その傾向が明確に違うことがわかる。一方で、他の言語モデルでは両者が混在していることから、ベクトル化した際に類似度が大きくなる語句に、こうした特徴が表れると推測できる。なお、2 章の実験を TF-IDF で行なったところ、最適化条件を全く満たさなかった。

4.　おわりに

本稿では、ベクトル表現を用いた言語モデルを J-POP 歌詞の文体分析に適用することを目指し、複数言語モデルの比較を行なった。特定作詞家による楽曲分類では、Word2Vec や fastText が特徴的な傾向を示した。作詞家間の対比では、TF-IDF が極めて明確な区別を実現した。さらに、ベクトル表現を用いた言語モデルの結果と比較することで、各作詞家の語句選択における特徴を掴む手がかりを得た。言語モデルの学習最適化、使い分け基準、結果の解釈などに課題があり、今後はこれらの問題に取り組む予定である。

謝辞

本研究は JSPS 科研費 16KK0008 の助成を受けたものである。

参考文献

[1] 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.

[2] 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.

[3] 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.

[4] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient Estimation of Word Representations in Vector Space”, Proc. 1st Int’l Conf. on Learning Representations, 2013.

[5] P. Bojanowski, E. Grave, A. Houlin, T. Mikolov, “Enriching Word Vectors with Subword Information”, Trans. of Assoc. for Comp. Linguistics, Vol. 5, pp. 135 – 146, 2017.

[6] Q. Le, T. Mikolov, “Distributed Representations of Sentences and Documents”, Proc. The 31st Int’l Conf. on Machine Learning, Vol.32, pp. II-1188 – 1196, 2014.

[7] D. M. Blei, A. Y. Ng, M. I. Jordan, “Latent Dirichlet Allocation”, J. of Machine Learning Research, Vol. 3, 2003.

[8] 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.

[9] L. McInnes, J. Healy, J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”, ArXiv e-prints 1802.03426, 2018, 2020 年 1 月 10 日取得.

[10] G. Salton, M. J. McGill, “Introduction to Modern Information Retrieval”, McGraw-Hill Inc., USA, 1983.