1C-05

Sentence-BERT を用いた文学作品の話題分析

天野樹† 圓谷顯信† 上原稔† 安達由洋‡
東洋大学総合情報学部総合情報学科† 東洋大学工業技術研究所‡

1. はじめに

現代文学研究のアプローチ方法は、研究者や評論家の主観に基づいたアナログアプローチと、デジタル情報処理技術を駆使した主観に偏らないデジタルアプローチに大別される[1]。デジタルアプローチによる研究として、感情語辞書と自己組織化マップを用いた小説の感情分析研究[2]、ファジークラスタ分析を用いた研究[3]が報告されている。また近年では、機械学習技術やビッグデータを利用した研究も報告されている。例えば、日本語 Wikipedia で事前学習した BERT を用いた感情分析研究[4]や、大量のレビュー文や小説などから収集した感情語辞書を用いた感情表現分析システムによる感情分析研究[1]が報告されている。しかし、文学作品の中で記述されている話題に対するデジタルアプローチ研究はほとんど見当たらない。

本研究では、機械学習技術を用いて話題から各文学作品を特徴付けし、その情報を利用して作家や作品の個性や傾向などを分析する[5]。文学作品を構成する各文の話題特徴表現には、Japanese Sentence-BERT (JSBERT)[6]により生成された分散表現を用いる。

2. 話題に基づく分析手法

本研究の話題に基づく文学作品分析は、次の手順に基づいて行う[5,6]：

手順１．JSBERT を用いて作品を構成する各文の分散表現を生成する。

手順２．作品を構成する文全体に対する分散表現集合をクラスタリングする。

手順３．各クラスタに適切な話題単語リストをラベル（ラベル語リスト）として自動的に付加する。

なお、JSBERT は NICT BERT 日本語 Pre-trained モデル[7]を事前学習モデルとして使用し、JSNLI コーパス[8]でファインチューニングして構成した。クラスタリングは Ward 法を用いた。文献[5]では、12 個の話題で教師ラベルを付けた 2,385 文のテストコーパスを作成して上記の手順で話題分類した結果、F1-score が約 0.896 の高精度を得ている。

各クラスタのラベル語リストには、クラスタに出現する単語をラベル語評価式で得た評価値で降順にソートして、上位 10 語を選択した。ラベル語評価式は次式で定義する。

$$\alpha \times TFIDF + (1-\alpha) \times Cosim \quad (0 \leqq \alpha \leqq 1)$$

ここで、*TFIDF* は cluster-based TFIDF であり、*Cosim* は各単語の分散表現とクラスタセントロイドとのコサイン類似度である。

Topic-Based Analysis of Representative Literary Works Using Japanese Sentence-BERT
† Miki AMANO, Kenshin TSUMURAYA, Minoru UEHARA ・Toyo University
‡ Yoshihiro ADACHI・Toyo University

3. 文学作品の話題に基づく分析

3.1 分析対象作品

本研究の話題分析では、現代文学作品として村上春樹の作品 14 タイトル（小説 8、エッセイ 6）、池井戸潤の小説 3 タイトル、東野圭吾の小説 4 タイトル、又吉直樹の作品 4 タイトル（小説 3、エッセイ 1）、湊かなえの小説 5 タイトルの 5 作家の合計 30 タイトルを分析対象とした。

3.2 クラスタ数とα値の検討

2 節の分析手法ではクラスタ数とα値の選択が重要である。まず、クラスタ数 10,15,20,30 でクラスタリング実験を行った。クラスタ数が少ないと 1 クラスタ内に幾つかの話題が含まれ、クラスタ数が多過ぎると同じ話題のクラスタが幾つも生成される。実験の結果、本研究で対象とする文学作品ではクラスタ数 20 が適切であると判断した。ラベル語リスト選択は、α= 0, 0.25, 0.5, 0.75, 1 で実験を行った。例として、村上春樹の「世界の終りとハードボイルド・ワンダーランド」のあるクラスタのラベル語リストを表１に、又吉直樹の「火花」のあるクラスタのラベル語リストを表 2 に示す。表 1 では、α = 0.5 のとき、ラベル語リストで“食事”と“料理”、“食べ物”が 2 位、3 位、4 位となっており、わかりやすい単語が揃っている。また、表 2 のような“私”や “僕”など一人称を表す単語が集まったクラスタの場合にはセントロイドの割合が高い、すなわちαが 0.5 以下のとき適切なラベル語リストが選択される。以上の実験より、ほとんどすべての作品のクラスタで、α= 0.5 のとき適切にラベル語リストを付与できることを確認した。

表 1.「世界の終りとハードボイルド・ワンダーランド」のあるクラスタのラベル語リスト

セントロイド	TF-IDF	α=0.25	α=0.5	α=0.75
味	ビール	グラス	グラス	グラス
気持	ウィスキー	食べ物	食事	料理
飲み物	サンドウィッチ	味	料理	食事
食べ物	コーヒー	食事	食べ物	我々
ひとくち	グラス	ひとくち	ひとくち	ミルク
すずき	彼女	気持	我々	食べ物
ブランディー	ワイン	飲み物	ミルク	ひとくち
ト	テーブル	料理	味	味
スフレ	パン	我々	気持	気持
食	料理	すずき	飲み物	飲み物

3.3 各作家と作品の話題による特徴

本節では、クラスタ数 20、α= 0.5 の条件の下で、文学作品の文の分散表現をクラスタリングしてラベル語リストを付加した実験結果をまとめる。まず、湊かなえ作の「N のために」に対する分析で生成されたクラスタとラベル語リストを表 3 に示す。

表 2.「火花」のあるクラスタのラベル語リスト

セントロイド	TF-IDF	α=0.25	α=0.5	α=0.75
僕	車両	僕	僕	自分
奴	世間	自分	自分	相方
人	両手	奴	相方	僕
自分	風景	相方	奴	ざる
違和感	乗客	人	人	奴
相方	自分	違和感	違和感	人
彼	習慣	彼	ざる	違和感
等	ホーム	等	彼	彼
動揺	相方	動揺	等	等
真樹	嘔吐	真樹	動揺	動揺

表 3.「N のために」のクラスタとラベル語リスト

クラスタ名	ラベル語
c1	わたし，あなた，究極，きみ，自分，行為，言葉，関係，愛，証拠
c2	部屋，ドア，廊下，リビング，ぼく，隣室，西崎，巻，書斎，野口
c3	西崎，貝殻，珊瑚，安藤，人，島，奈央子，場所，経歴，話
c4	料理，食事，西崎，タッパー，テーブル，給仕，総菜，いるか，成瀬，安藤
c5	サービス，出張，会社，受付，入社，安藤，仕事，就職，人，ちよ
c6	安藤，西崎，成瀬，奈央子，相手，相談，たら，言葉，ヤツ，人
c7	自分，場所，世界，安藤，野望，人間，西崎，言葉，才能，気持ち
c8	母親，父親，ママ，子ども，洋介，わたし，人，行為，父さん，両親
c9	彼女，女，言葉，ぼく，棋譜，人間，ヒステリック，場所，気，人
c10	アパート，マンション，ビル，部屋，住人，みどり，安藤，ホテル，介護，人
c11	学校，先生，授業，クラス，高校，担任，奨学，同窓，進学，成瀬
c12	奈央子，野口，暴力，自分，人間，わたし，西崎，からだ，行為，気持ち
c13	奈央子，ドレッサー，肉じゃが，マーくん，おくに，ヤッ，インターフォン，ゴメン，出だし，簀子
c14	あいつ，自分，ぼく，ヤツ，俺，努力，状況，おまえ，僕，そいつ
c15	文学，原稿，小説，作品，感想，昇華，西崎，バード，現実，世界
c16	花屋，誕生日，広田，かな，おまえ，住所，野口，ドレッサー，成瀬，西崎
c17	放火，火事，灼熱，炎，火，明かり，ボッ，出来事，空想，建物
c18	杉，いるか，ノゾミ，足場，一覧，猶予，けり，おまえ，模様，安藤
c19	野口，奈央子，安藤，西崎，エゴイスト，わたし，人，善，土産話，貴弘
c20	将棋，対局，安藤，野口，ブレーン，棋譜，相手，戦法，手段，わたし

表 3 からこの小説には“部屋”や“アパート”などの場所（c2,c10）、“食事”（c4）、“母親”や“父親”など家族（c8）、“学校”（c11）、“小説”（c15），“火事”（c17）、“将棋”（c20）に関する話題が含まれていることがわかる。この小説は「ミステリー」や「恋愛小説」のジャンルに分類されているが、話題分析によってさらに細かい内容を検出することが可能となる。

同様の分析を分析対象 30 作品に行い、作家ごとに共通する話題を調べた。表 4 に村上春樹の小説 8 タイトル中の類似する話題クラスタのラベル語上位 5 語を示す。表中の番号は次の作品タイトルに対応する：

①1Q84
②ノルウェイの森
③海辺のカフカ
④騎士団長殺し(第 1 部)
⑤騎士団長殺し(第 2 部)
⑥色彩を持たない多崎つくると、彼の巡礼の年
⑦世界の終りとハードボイルド・ワンダーランド
⑧風の歌を聴け

表 4 より、村上春樹の小説の上記 8 タイトル全てに“彼女”や“女性”といった話題が含まれていることがわかる。村上春樹作品では、エッセイ 5/6 タイトル（6 タイトル中 5 タイトルを表す）で“小説”、小説とエッセイの 11/14 タイトルで“音楽”に関するラベル語リストを持つクラスタが生成された。

表 4. 村上春樹作品の類似クラスタのラベル語リスト

①	②	③	④	⑤	⑥	⑦	⑧
彼女	彼女	彼女	彼女	彼女	彼女	彼女	彼女
自分	あなた	女性	女性	女性	女性	あなた	女
意味	つた	少女	行為	自分	人間	女	溜息
出来事	気持	あなた	関係	女	自分	私	口調
人間	人	女	あなた	気持ち	出来事	自分	伝票

他の作家では、池井戸潤の小説 3/3 タイトルで“会社”や“銀行”、東野圭吾の小説 3/4 タイトルで“事件”、又吉直樹の作品 3/4 タイトルで“芸人”、湊かなえの小説 5/5 タイトルで“学校”、4/5 タイトルで“事件”や“母親”に関するラベル語リストが付加されたクラスタが生成された。作家全体で類似したクラスタとして、ラベル語リストの先頭に“部屋”がくるクラスタが、村上春樹 6/14、東野圭吾 3/4、又吉直樹 4/4、湊かなえ 5/5 タイトルに生成された。これは小説の中で部屋が舞台となるシーンが多いことが考えられる。

4. まとめ

JSBERT で生成した分散表現をクラスタリングし、生成された各クラスタにラベル語リストを自動付加する技法を用いて、現代文学の 5 作家 30 タイトルの作品を話題に基づき分析した。その結果、各作家の小説やエッセイについて共通する話題や登場頻度の高い話題を発見でき、作家ごとの個性や傾向を明らかにすることが可能であることを検証した。

今後の課題として、適切なクラスタ数とα値の自動設定、類似クラスタの自動検出、ラベル語リストの精度評価などに関する研究が挙げられる。また、本稿で提案したデジタルアプローチによる日本文学分析結果とアナログアプローチによる分析結果の比較検討、文学の専門家による本手法の評価も重要である。

参考文献

[1] 瀬山透矢，加藤陸斗，Astremo Amilcare，天野樹，中山佳大，安達由洋，“集合知に基づく現代日本文学研究のアプローチ”，情報処理学会第 84 回全国大会（2022）
[2] 吉田知世，小林一郎，“感情表現に基づく小説の俯瞰分析への取り組み”，DEIM Forum A8-6（2011）
[3] 菊地泰史，加藤千恵子，前城裕紀，“村上春樹文学に対するファジィクラスター分析によるテキスト解析”，可視化情報学会誌，31（Suppl.1），141-146（2011）
[4] 圓谷顯信，高橋宏和，安達由洋，“BERT による日本語文の感情分析と話題分析”，情報処理学会第 84 回全国大会（2022）
[5] M. Amano, K. Tsumuraya, M. Uehara, Y. Adachi, “An Analysis of Representative Works of Japanese Literature Based on Emotions and Topics” AINA2023 (2023, accepted)
[6] K. Tsumuraya, M. Amano, M. Uehara, Y. Adachi: Topic-Based Clustering of Japanese Sentences Using Sentence-BERT. CANDAR2022 (2022)
[7] 情報通信研究機構，“NICT BERT 日本語 Pre-trained モデル ”， https://alaginrc.nict.go.jp/nict-bert/index.html (2023/1/11 アクセス)
[8] 吉越卓見，河原大輔，黒橋禎夫，“機械翻訳を用いた自然言語推論データセットの多言語化”，情報処理学会研究報告，Vol.2020-NL-244 No.6（2020）