1H-02

集合知に基づく現代日本文学研究のアプローチ

瀬山透矢† 加藤陸斗† Astremo Amilcare† 天野樹† 中山佳大† 安達由洋†
東洋大学総合情報学部総合情報学科†

1. はじめに

現代日本文学研究は、自己組織化マップやファジィクラスタ分析などのディジタル技術を一部に使っている研究もあるが、研究者あるいは評論家が主観に基づいて研究・分析する職人芸によるアナログ・アプローチがほとんどである。一方、自然言語処理および AI の分野では、日本語 Wikipedia 全ページを学習した単語意味分散表現や、多数の小説と各種レビュー文などを教師データとして収集した感情語辞書などのビッグデータを用いた集合知に基づくテキスト分析技術の研究が進んでいる。

本研究では、集合知に基づいて個人の主観に偏らない、すなわちバイアスのかからない文学研究のディジタル・アプローチを提案する。意味分散表現を用いた話題に基づくテキスト分析技術と、感情語辞書を用いた感情に基づくテキスト分析技術を用いて各作家と各作品の特徴付けを行い、それらの情報を利用して現代文学を研究するものである。すなわち、文学研究のデジタルトランスフォーメーション（DX）の提案である。そして、ディジタル・アプローチによる研究の第一歩として小説の感情分析に取り組んだ成果を報告する。

2. これまでの文学研究アプローチ

現代文学の研究は、研究者あるいは評論家が作品論、作家論、テクスト論などの切り口で主観に基づいた職人芸によるアナログ・アプローチがほとんどである。ディジタル技術を一部に使っている研究もあるが、Wikipedia 全ページ、SNS、あるいは Amazon や Google レビューなどのビッグデータから得た集合知に基づいて、個人の主観に偏らない、すなわちバイアスのかからない文学研究のディジタル・アプローチは提案されていない。

3. 集合知に基づくディジタル・アプローチ

日本語 Wikipedia 全文から学習した日本語 Wikipedia エンティティベクトル[1]や日本語 BERT 事前学習済みモデル[2,3]により求めた日本語文意味分散表現を用いて、話題による日本語テキストの分類・検索技術が研究されている[4]。また、多数の小説、Amazon や Google などの各種レビュー文などから収集した感情語データに基づく感情分析技術の研究も進められている[5,6]。これらの研究成果を踏まえて、日本語 Wikipedia やレビューサイトなどのビッグデータから得た集合知に基づく現代日本文学研究のディジタル・アプローチに取り組む。

4. EEAS による感情分析

現代日本文学研究のディジタル・アプローチの第一歩として、小説の感情分析研究について報告する。

4.1 分析対象

本研究では、村上春樹の小説 8 タイトルとエッセイ 6 タイトル、池井戸潤の小説 4 タイトル、東野圭吾の小説 4 タイトル、又吉直樹の小説 3 タイトルとエッセイ 2 タイトル、そして湊かなえの小説 5 タイトルの合計 32 タイトルを分析対象とした。

4.2 作家毎の感情語の出現頻度

感情分析には、7,234 語の語彙を持つ感情語辞書を用いた感情表現分析システム EEAS[5]を用いている。EEAS は非常に処理が高速で、小説一冊に対して数秒で 11 感情の分析をする。図 1 に作家毎の感情語密度を示す。感情語密度は小説中に出現する感情語の出現回数（頻度）を文章数で割ったものである。図 1 に示されるように、村上春樹は分析対象とした作品数が多いのも一因で感情語密度のばらつきが大きく、また小説とエッセイで感情語密度が異なる。東野圭吾の小説では感情語密度が 0.3～0.32 と低く一定であり、池井戸潤の小説も感情語密度のばらつきが小さい。又吉直樹は小説およびエッセイともに感情語密度が高い。湊かなえの小説では感情語密度が 0.41～0.53 と比較的高くばらつきが大きい。

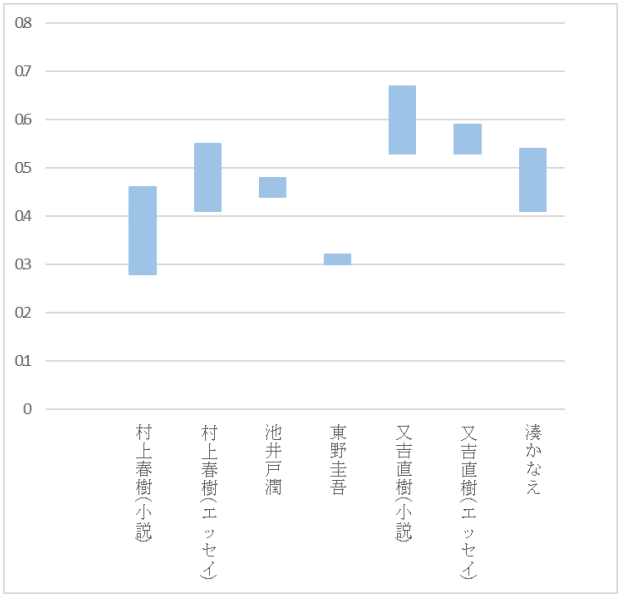


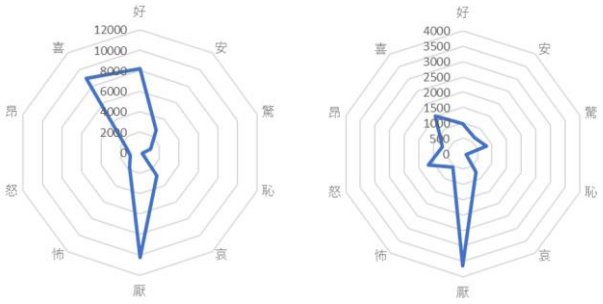
図 1. 作家毎の感情語密度

An approach to modern Japanese literature research based on collective knowledge
† Yukiya SEYAMA, Rikuto KATO, Amilcare ASTREMO, Miki AMANO, Keita NAKAYAMA, Yoshihiro ADACHI・Toyo University

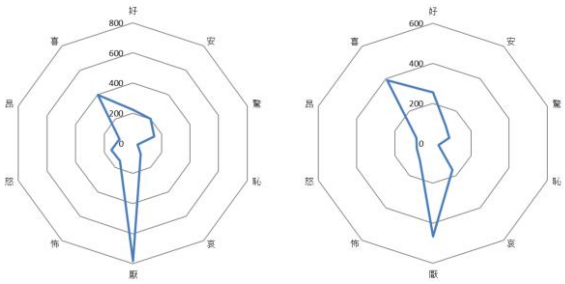
4.3 作家毎の感情レーダーチャート

作家毎の全ての作品に出現する感情語の‘望’感情を除いた 10 感情カテゴリ別に可視化した感情レーダーチャートを求めた。感情レーダーチャート[5]とは、正多角形の軸上に各感情カテゴリを配置して小説全体での各感情の出現頻度の特徴を一目で把握できるようにした図である。positive 感情を上側に、negative 感情を下側に、また‘好’と‘厭’など互いに対になる感情は反対側の位置になるように配置している。

図 2 に村上春樹と池井戸潤の全作品に出現する感情語の感情レーダーチャートを示す。図 2 が示すように、感情語の出現には作家毎に作品に共通する大きな特徴がある。例えば、村上春樹の小説ではほとんどすべてで図 2(a)のような感情分布をしており、また池井戸潤の小説はほとんどすべてで図 2(b)のような感情分布をしている。また、図 2(c)に東野圭吾の「マスカレード・ホテル」に出現する感情語の感情レーダーチャート、図 2(d)に湊かなえの「告白」に出現する感情語の感情レーダーチャートを示す。これらの図が示すように、同じジャンル“ミステリ”に属する小説でも作家によって感情語の出現分布が異なっている。



(a)村上春樹の全小説　　(b)池井戸潤の全小説



(c)「マスカレード・ホテル」　　(d)「告白」

図 2. 感情レーダーチャート

4.4 感情系列図

感情系列図は横軸に時間あるいはテキスト中の各文の出現順序をとり、縦軸に各感情カテゴリに属する感情語の出現頻度を表示した図である。各感情系列は、上から順に[喜，好，安，驚，昂，望，恥，哀，怒，怖，厭]の 11 感情カテゴリに対応しており、上側の系列が positive 感情に下側の系列が negative 感情を表している。感情系列図により、小説の中での感情の変化を可視化することができる。図 3 に、村上春樹の 1Q84 分冊 3 の感情系列図を示す。この図が示すように、小説の中盤あたりで‘昂’と‘哀’感情語の出現頻度が高くなっている。この部分は「親しい友人の死による主人公の感情の大きな揺れ」が起こった部分であり、話の重要なポイントである。

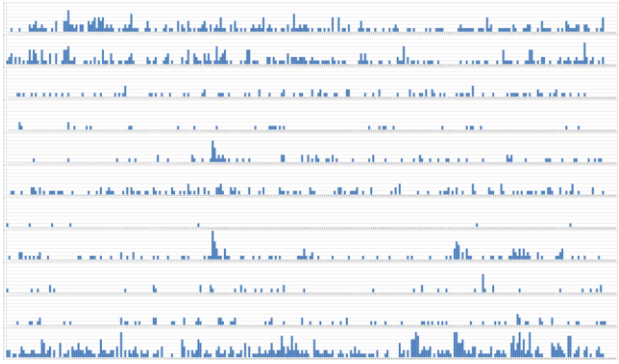


図 3. 1Q84 分冊 3 の感情系列図

5. まとめ

現代日本文学研究の集合知に基づくディジタル・アプローチを提案し、その第 1 歩として感情分析技術による現代作家の小説 32 タイトルの分析研究を行った。その結果、小説の感情分析に基づく特徴付けにより各作家の分類が可能であることを検証した。

今後は、日本語 Wikipedia エンティティベクトル[1]や日本語 BERT 事前学習済みモデル[2,3]により求めた日本語文の分散表現を用いて、話題による現代日本文学の分析にも取り組む。小説の中に現れる種々の話題の頻度により各作家の特徴付けをして分析できることが期待される。

参考文献

[1] 東北大学乾・鈴木研究室，日本語 Wikipedia エンティティベクトル，http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/ (2022/01/04 アクセス).
[2] 情報通信研究機構データ駆動知能システム研究センター，“NICT BERT 日本語 Pretrained モデル ”，https://alaginrc.nict.go.jp/nict-bert/index.html (2022/01/04 アクセス).
[3] 東北大学乾・鈴木研究室 “Pretrained Japanese BERT models”，https://github.com/cl-tohoku/bert-japanese (2022/01/04 アクセス).
[4] Yoshihiro Adachi and Takanori Negishi, “Development and evaluation of a real-time analysis method for free-description questionnaire responses”, IEEE ICCSE2020 (2020).
[5] 安達由洋，近藤友啓，小林孝充，惠谷菜央，石井解人，「感情語辞書を用いた日本語文の感情分析」，可視化情報 Vol.41 No.161 (2021).
[6] 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 年 3 月 発表予定).