4V-05

小説の内容を表すキーワードの抽出に関する研究

若井 龍弥 *
電気通信大学 *
情報理工学域

大久保誠也 †
静岡県立大学 †
経営情報学部

若月光夫 ‡
電気通信大学 ‡
大学院情報理工学研究科

西野哲朗 §
電気通信大学 §
大学院情報理工学研究科

1 はじめに

小説のジャンルに関する情報として, キーワードが存在する. キーワードは小説の属性や要素, テーマを表す単語であり, 小説家になろう [1] などの小説投稿サイトでは作者が自身の作品に対し自由に付与することが可能となっている. これにより, 読者はより細かく作品の情報を知ることができ, キーワードによって読みたい作品を探しやすくなっている.

しかし, キーワードは作者が作品に対して予め付与していなければならないという問題もある. 書籍で販売されている小説や, 図書館に架蔵されている小説にはキーワードの情報が付与されていない場合が多く, そのような小説にはどのような要素が含まれているかを判断することは難しくなっている.

この問題を解決する手段として, 小説テキストからキーワードを自動抽出する方法が挙げられる. 馬場ら [2] は小説に登場する人物の名前や特徴などを抽出する手法を提案した. しかし, この研究において, 抽出したキーワードがその小説の内容に関わる重要単語であるかどうかは考慮されておらず, 小説に対する抽出キーワードの重要性に関する研究もあまり行われていない.

2 研究目的

本研究では, より小説の内容に沿ったキーワードを抽出することを目的とし, 従来のキーワード抽出手法として広く知られている TF-IDF 法を用いた新たな抽出手法を提案する.

Research on the extraction of keywords representing the contents of novels

*Ryuya Wakai, School of Informatics and Engineering, The University of Electro-Communications
†Seiya Okubo, School of Management and Information, University of Shizuoka
‡Mitsuo Wakatsuki, Graduate School of Informatics and Engineering, The University of Electro-Communications
§Tetsuro Nishino, Graduate School of Informatics and Engineering, The University of Electro-Communications

3 前提知識

3.1 TF-IDF 法

TF-IDF 法 [3] は, 各文書中に含まれる各形態素が, その文書内でどれほど重要なのかを表す手法である. 出現頻度が高く, 他の文書ではあまり出現しない単語ほどスコアは高くなる. TF-IDF 法において, 単語のスコア w は以下の式で定義される.

$$w = tf(t,d) \times idf(t) \tag{1}$$

$tf(t,d)$ はある単語 t のテキスト d における出現頻度であり, 以下の式で表される.

$$tf(t,d) = \frac{\text{単語 } t \text{ の出現回数}}{\text{テキスト } d \text{ 内の全単語の出現回数}} \tag{2}$$

また, $idf(t)$ はある単語 t が出現する文書数の逆数であり, 以下の式で表される.

$$idf(t) = \log \frac{\text{テキストデータの総数}}{t \text{ を含むテキストの数}} + 1 \tag{3}$$

3.2 Word2Vec

Word2Vec[4] とは, 文章中の単語を数値ベクトルに変換してその意味を把握する自然言語処理の手法である. 2013 年に Google の研究者 Tomas らによって提案された手法で, 同じ文脈に現れる単語は類似した意味を持つという分布仮説を元にしている.

4 提案手法

本研究において利用する小説は, 小説家になろうに投稿されている小説を対象とした. 本研究において提案する手法は, 大きく分けて二つの異なる手法から得られる値を用いる.

抽出対象とするテキストを形態素解析し, 得られた単

語に対して TF-IDF 法によって重みづけを行う. なお, 対象とする単語は名詞に限定し, 代名詞や数詞などの単語は取り除いた. また, 抽出対象テキストは小説のあらすじである. これによって単語 t_i には重み w_i が付加されることになる.

抽出するキーワードは, 作品の内容を表している単語, なおかつそれらによって作品の概要をある程度把握できるような単語が望ましい. これは, 通常小説投稿サイトに投稿されている小説に付与されているキーワードは, これら二つの特徴を持つような単語が多いと考えられるからである. よって, 既に小説家になろうにて一般的によく用いられているキーワードと類似するような単語を抽出するために, Word2Vec で作成した学習モデルを用いてそれらのキーワード k_j と単語 t_i との類似度 $sim(t_i, k_j)$ を計算する. ここで, $k_j \in \boldsymbol{K}$ であり, $\boldsymbol{K}$ は小説家になろうで使用されているキーワード全体の集合である.

これらを踏まえて, 単語 t_i のスコアとして, 以下の式から算出される値を新たに用いる.

$$score(t_i) = w_i \times \max_{k_j \in \boldsymbol{K}} sim(t_i, k_j) \tag{4}$$

これにより, TF-IDF 法によって得られた単語の重みと, キーワードとの類似度を同時に考慮することが出来る. すなわち, そのテキスト内で重要な単語ほど *score* の値は大きくなり, キーワードとの類似度が高いほど同じように *score* の値が大きくなる. こうして得られた *score* を降順に並べた際の, 先頭から 10 単語をこの手法によって抽出されたキーワードとする.

5 実験

5.1 実験目的

提案手法によって抽出されたキーワードがその作品のキーワードとして相応しいかどうかを, 被験者実験によって明らかにする.

5.2 実験概要

小説のタイトルに加え, 既存のキーワード抽出手法として広く用いられている TF-IDF 法 (以下, 従来手法) による重みづけを用いて抽出されたキーワード 10 個を提示する. その後, 本研究の提案手法によって抽出されたキーワード 10 個を被験者に提示する. 両者を提示した後, その小説のあらすじを提示し, あらすじに対してそれぞれのキーワードがどれくらい相応しいと感じたかを 5 段階評価で回答してもらう. 回答は Google フォームによって収集した.

5.3 実験結果

四つの作品に対する回答を以下のようにまとめた. なお, 表中の値は, 5 段階評価の回答それぞれにスコアを設けた際の平均値となっており, 「20 %以下」が 1, 「21 %〜40 %」が 2, 「41 %〜60 %」が 3, 「61 %〜80 %」が 4, 「81 %以上」が 5 となっている.

表 1: 各抽出手法の評価

	作品 1	作品 2	作品 3	作品 4
従来手法	3.00	3.13	2.38	2.00
提案手法	3.00	4.38	3.25	2.75

5.4 考察

表 1 より, 作品 1 を除いた 3 作品で TF-IDF 法のみで抽出した従来手法よりも提案手法の方が被験者による評価が高いことがわかった. これは, 作品 1 の抽出対象テキストおよび被験者に提示した小説のあらすじが他の 3 作品よりも短いために, 被験者があらすじに対する適切な単語が分かりにくいことが原因の一つであると考えられる.

6 おわりに

より小説の内容に沿ったキーワードの抽出を目的とし, TF-IDF 法と Word2Vec を用いて新たな提案手法を提案した. 従来手法と提案手法を比較した被験者実験では, 表 1 より提案手法が有効であることがわかった.

参考文献

[1] 小説家になろう https://syosetu.com/ 閲覧日 2023/1/12

[2] 馬場こづえ; 藤井敦; 石川徹也. 小説テキストを対象としたジャンル推定と人物抽出. 言語処理学会第 11 回年次大会発表論文集, 2005, 157-160.

[3] 奥村 学. 自然言語処理の基礎. コロナ社. 2010.

[4] Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations." Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies. 2013.