7T-05

# 小説を対象としたジャンル推定手法の検討

三輪拓也† 松本忠博‡
岐阜大学大学院 自然科学技術研究科† 岐阜大学 工学部‡

## 1 はじめに

近年,電子書籍や Web 小説などで小説を入手しやすくなり,小説を使用した研究が多く行われるようになった.

文書分類は,与えられた文書を指定したクラスのいずれかに分類する問題であり,ニューラルネットワーク（以下 NN）などを使用した研究が行われている．しかし,ニュース記事等の情報伝達を目的とした文章を対象とした研究が多く,小説を対象とした研究は少ない.

本研究では小説を対象とした文書分類の精度向上を目的として,小説の特徴を考慮したジャンル推定手法を検討した.

## 2 分類対象としての小説の特徴

情報伝達を目的とした文章は,読書に読みやすく,伝わりやすくする必要があるため,客観的な事実や意見を一定の語彙の範囲で簡潔に述べたものが多い．だが,小説は表記上の制約が少なく,使用語彙や文章の長さの制限がなく,文体は自由であり,編集者等によって大きく編集されることは少ないと予想される．また,作品によっては複数のジャンル要素を含む場合もある.

筆者らは，小説の分類が新聞記事の分類より難しいことを実験により確認した[1].

## 3 関連研究

Worsham ら[2] と Pieters ら[3]は，それぞれ小説と映画の字幕を対象とした分類を行い,共に最良の学習モデルは対象とするデータに依存し,必ずしも最先端の複雑なモデルが最良の結果を出すとは限らないことを示した.

Worsham ら[2]は小説の多クラス分類を行ったが,本研究ではマルチラベル分類により,複数のジャンル要素を含む作品に対応させる.

Pieters ら[3]は映画の字幕を対象とし，複数のモデルで分類を行ったが,本研究では小説を対象とし，推定に使用する単語の選択方法を変更することで分類性能の向上を図る.

Examination of genre estimation method for novels
† Takuya Miwa, Graduate School of Natural Science and Technology, Gifu University
‡ Tadahiro Matsumoto, Faculty of Engineering, Gifu University

## 4 ジャンル推定手法の検討

### 4.1 前処理

内容語（名詞,動詞,副詞,形容詞,未知語）のみを使用する．動詞,形容詞は原形を抽出し，数字,カタカナ,アルファベットの文字種は統一する．小説は多くの種類の単語が使用されるが,作品全体で出現頻度が高い単語や低すぎる単語は分類においてあまり役に立たない．そこで,使用する全作品中,出現頻度が 70%以下かつ,出現頻度が高い上位 15000 単語を使用する．また,文章全体をデータとして使用すると単語数が多く,作品によって単語数の差があるため，作中の 6000 単語のみを使用し，6000 単語に満たない場合はパディングをする.

### 4.2 ベースライン手法

上述のように必ずしも複雑な NN モデルが最良というわけではない．本研究では NN において,簡単なモデルである多層パーセプトロン（MLP）を使用する．データとして使用する 6000 単語は,作品の冒頭から収集する.

### 4.3 n 分割法

小説はストーリー全体でその作品のジャンルを判断するが,ストーリー中にその作品のジャンル要素を感じることが難しいシーンが存在することがある．ベースライン手法では小説の冒頭から指定した単語数を使用するため,ストーリーの序盤のみでジャンルの推定を行うことになる.そこで,作品のストーリー全体でジャンル推定を行うため,作品全体の単語数を分割して,分割先から均等に単語を収集する．なお，分割は 4.1 節の前処理を行ったデータに対して行う.

### 4.4 追加データ

上述の 15000 語とは別に，ジャンル推定に役立ちそうな語のリストを作成し，小説の特徴情報として利用する.

(a) 場所・時間を表す単語

「SF」や「ファンタジー」のように,その作品の世界観によって判別できるジャンルが存在する．「ホラー」は恐怖感を与えるために夜の時間帯にストーリーが進みやすく,「推理」は事件

を解決するために時間を指す単語が多く使用されると考えられる．よって,場所と時間を指す単語は分類に役に立つのではないかと考えた．

場所と時間を表す単語は『日本語語彙大系』の意味属性体系を参考にして収集した．

(b) ジャンルごとの高頻度語

ベースラインの手法では,ある単語 a はジャンル A おける出現頻度が高くても,それ以外のジャンルにおける出現頻度が低いため,全体での出現頻度が低くなる．その結果,ジャンル A と分類するために役に立つ単語 a が,使用する単語内に存在しない場合がある．よって,各ジャンル出現頻度が 50%以上の単語を収集した．ただし,複数のジャンルで出現頻度が 70%以上の単語は除く．

## 5 評価実験と考察

### 5.1 使用データ

小説のデータは「小説を読もう[1]」から推理,ホラー，歴史，SF，恋愛，ファンタジーの 6 つの要素いずれかを含み,30000 文字以上ある合計 23178 作品を収集した．また,「小説を読もう」におけるジャンルとキーワードを参考に，各作品に対し 6 種類のラベルを付けた．

### 5.2 実験方法

形態素解析には MeCab と拡張辞書 mecab-ipadic-NEologd を使用した．NN モデルは全結合層 3 層とし，1 層目のドロップアウト率を 0.5 とした．また，最初の全結合層の前段に Embedding 層（300 次元）を追加した．閾値は 0.5 とする．n 分割法の n は 4 とした． ベースライン，n 分割法は単語に番号を振り分けた値，追加データは TF-IDF を特徴量とする．場所，時間データは抽出した単語から，全作品中の出現頻度が 0.1%以下の単語を削除した．場所を表す単語は 3694 単語，時間を表す単語は 214 単語，高頻度の単語は 2209 単語を使用した．なお，追加データは，使用するデータごとに分類器を作成し，各分類器の出力結果に適切なバイアスを付け，足し合わせた値を出力とする．

### 5.3 評価方法

マルチラベル分類を 5 分割のクロスバリデーションで行い，正解率，再現率，F 値で比較する．なお，本研究では正解率をデータ単位で算出し，対応するラベルがすべて正しければ正解，1 つでも間違えれば誤りと判定する．また，再現率と F 値は各ジャンルの値を平均した値とする．

### 5.4 実験結果・考察

実験の結果を表 1 に示す． n 分割法，場所・高頻度語データを使用した手法では，正解率は 0.19～2.47%，再現率は 0.43～1.68%，F 値は 0.92～4.22%向上した．これらの手法では，ジャンルごとの F 値のほとんどが向上していたことから，小説に対して有効と考えられる．一方，時間データは「恋愛」以外のラベルの分類性能が低下した．また，追加データを多入力モデルで行った結果，正解率は 51.84%，再現率は 79.93%，F 値は 69.75%が最良の値だったことから，F 値は低下するが，分類器を複数組合わせる手法は正解率を向上させるのに有効であった．

表 1 実験結果

| 手法 | 正解率(%) | 再現率(%) | F 値(%) |
|---|---|---|---|
| B | 51.04 | 86.01 | 63.68 |
| S | 51.23 | 85.53 | 64.73 |
| B+P | 51.80 | 87.05 | 64.60 |
| B+T | 50.97 | 86.67 | 63.02 |
| B+F | 53.12 | 86.44 | 67.64 |
| S+P+F | 53.51 | 87.16 | 67.90 |
| S+P+T+F | 53.33 | 87.69 | 67.41 |

表中の記号は，B：ベースライン，S： n 分割法，P： 場所データ，T ：時間データ，F ：高頻度語データを表す．重み付けは次の通り．
B：P＝1：1，B：T＝8：2，B：F＝1：1，
S：P：F＝4：2：4，S：P：T：F＝4：1：1：4

## 6 まとめ

本研究では小説のジャンルを対象としたマルチラベル分類において，小説の特徴を考慮した手法を提案して性能向上を図り，実験によりその有効性を検証した．

今後は，Worsham ら[2]のように NN より簡単な構造である分類モデルで行ってみる必要がある．また，本研究で使用したデータはラベルごとにデータ数に差があったことも原因の 1 つでもあるが，ラベルごとの分類性能に差があったため，小説の各ジャンル専用のモデルを作成することで，性能の向上が図れるのではないかと考える．

### 参考文献

[1] 三輪拓也，松本忠博，小説を対象とした文書分類手法の検討，言語処理学会第 25 回年次大会発表論文集，pp. 594-596，2019

[2] Joseph Michael Worsham，Towards Literary Genre Identification：Applied Neural Networks For Large Text Classification，University of Colorado Colorado Springs，2018

[3] Mathijs Pieters, Marco Wiering，Comparison of Machine Learning Techniques for Multi-label Genre Classification，DOI：10.1007/978-3-319-76892-2_11，2018

[1] https://yomou.syosetu.com/