『感情表現辞典』を用いたランダムフォレスト法による文章の感情分類

西村 淳† 久野 雅樹††
†電気通信大学 情報理工学部 総合情報学科
††電気通信大学大学院 情報理工学研究科 情報学専攻

n1310107@edu.cc.uec.ac.jp†, hisano@uec.ac.jp††

1 研究背景と目的

人類は古くから数え切れないほど多く物語を生み出し続け，1900 年代に構造主義の考えが広まるとともに物語を分析しようという試みが発展していった．コンピュータが導入される以前の分析は研究者の主観を排除することが難しく，また 1 回の研究で少数の物語しか分析できなかったため，限定的な量での分析しか行われることがなかった．しかし，コンピュータとインターネットの普及にともなって，物語をデータ化して大量に分析しようとする動きが活発になっていった．例えば，単語数のカウントと形態素解析がコンピュータで処理できるようになったことから，単語の共起回数を見る，活用形をまとめて見ることができるようになるなど，大量のデータを分析する方法が次々に開発されていった．

日本語の文書データ集としては国立国語研究所の現代日本語書き言葉均衡コーパスや，国立国会図書館の国会会議録検索システム等があるが，今回は『感情表現辞典 [1]』の例文をデータ化し，機械学習に活用する．『感情表現辞典』は中村明の著書で，平成 17 年 (2005 年) に東京堂出版から出版された第 11 版を実験に使用した．この著書では，感情は日本の六情から派生したと考えられる，喜，怒，哀，怖，恥，好，厭，昂，安，驚の 10 種類で分類されており，感情ごとに文学作品から抜き出した単語と，その単語を含んだ原文を例文として掲載している．感情表現に着目する数少ない資料の 1 つであるため，物語文の感情分析にはよく用いられている．

感情を分類しようという試みは東西問わず古くから行われてきた．代表的な分類は表 1 のとおりである。

表 1: 感情分類一覧

分類名	時期	分類数
日本の六情	中世	6
ダーウィンの基本的感情	1872	7
プルチックの基本分類	1980	8
エクマンの基本感情	1990 年代	18
『感情表現辞典』	1993	10
パロットの 3 層分類	2001	25

今回の研究では『感情表現辞典』以外の感情分類を選択した場合は，感情に当てはまると思われる文章を何処か別のコーパスから用意しなければならない．しかし,『感情表現辞典』を用いれば感情ごとですでに分けられた例文が用意されているため分類の信頼度が高いと考え，感情表現辞典の 10 分類を実験に用いることに決めた．

本研究では『感情表現辞典』の例文をデータ化して機械学習を行い，文書の感情を適確に分類させることを目的とする．物語の感情に着目したこれまでの研究では，感情の分類ごとにその感情をよく表していると考えられる単語を決めて辞典を作成し，その単語の単純な数やその単語と他の単語の共起状況を見ることで分析を行ってきた．しかしこの手法では，感情ごとの単語を集める段階で漏れや片寄りが発生する問題点がある．そこで，全文が感情に紐付いた『感情表現辞典』例文とすべてのデータに共起頻度を付与する機械学習を組み合わせることで，感情との関連性が高く，幅広い文章表現をカバーした学習で分析できると考えた．また，SVM（サポートベクターマシン）法など様々あるなか機械学習の手法としては，多クラス分析やビッグデータに柔軟に対応できることを理由に，ランダム

フォレスト法を採用した．

2　関連研究

本研究のように機械学習を用いて文書データに感情分析を適用する研究はいくつも行われているが，最も活発に研究されている文書データは Twitter である．佐藤ら [2] の研究では，Twitter での感情分類を様々な方法で行い，比較検討を行っている．この研究では，『感情表現辞典』に記載されている感情語を含むツイートをまとめ,『感情表現辞典』で用いられている 10 の感情分類の情報をツイートごとに付与した．その情報を用いて，6 種類のベクトル表現法と 3 種類の学習アルゴリズムで分類を行い，その結果を一覧にまとめている．

Twitter ではなく，文学作品に機械学習を適用している研究の 1 つに金ら [3] の研究がある．金らは青空文庫という著作権が切れた，または著作者が許可した文学作品をデータ化して公開しているサイトから文書データを収集し，文章の書き手の同定をランダムフォレスト法で行っている．結果の有効性を確認するために，SVM 法，学習ベクトル量子化法，バギング法，ブースティング法といった他の手法でも同じように分類したが，ランダムフォレスト法が最も正解率が高い結果となった．

機械学習を用いていないが，文学作品を感性語を使って分類する研究に原田 [?] の研究がある．原田は図書 100 冊を選出し，形容詞，形容動詞，名詞を抜き出し，人手によって 15 軸の感性概念を構築した．実験の評価として，25 冊の図書を対象にこの感性概念を本文中の感性語数と結びつけて数値化すると，第 1 候補と判定 (最も感性の振れ幅が大きい軸と) されたものが，人手での判定結果と 10 冊（0.40）で一致する結果を得た．

3　感情分類手法

3.1　文書データの収集

本研究で用いる文書データは『感情表現辞典』より収集した.『感情表現辞典』は構造として，全体が 10 種類の感情に分かれており，その中に文学作品から抜き出した見出し語とその語を含む例文（文学作品より抜粋）がセットになる形で，多数収録されている．以下に文書データの一部を例示する．

1, 喜
（例文）体内に充実感がみなぎる
（見出し）体内には光のように峻厳な充実感がみなぎっていた〔檀=花筐〕
2, 怒向っ腹を立てる
汽車が大阪に着くと、向っ腹を立てたあげくボーイと喧嘩して、ぐわんと一つ殴ってやってから汽車から飛び降りた〔阿部知=冬の〕
3, 哀
悲しくて眼を赤くする
お嫁に行くというのにやっぱりこの家を出るのが悲しいんですって、まあごらんなさい、こんなに眼々をあかくして、ほほほ〔宇野千=色ざ〕
4, 怖
恐ろしいと思う
死後の静寂に親しみを持つにしろ、死に到達するまでのああいう動騒は恐ろしいと思った
5, 恥
てれくさいこと夥しい
近所の乗客まで振り返って見たから、照れ臭いこと夥しい〔小沼=風光〕
6, 好
慕わしく感じる
離れの二階でこうして二人きりに倚り添っている安子さんのことを、不意に慕わしく感じ始めた。〔福永=廃市〕
7, 厭
激しい嫌悪
通りすがりの男に、躯をあたえている玉枝への激しい嫌悪となって喜助を苦しめた。〔水上勉=千羽〕
8, 昂
興奮して眠れない
その気持の張りと柳吉が帰ってきた喜びとで、その夜興奮して眠れず、眼をピカピカ光らせて低い天井を睨んでいた〔織田=夫婦〕
9, 安
くつろいだ気持ち
誰もいないきやすさに、くつろいだ気持ちで、押入れの汚れものを探してみる。〔林芙=放浪〕
10, 驚
心臓が止まる
「もし、もし、君！」私を呼びかける太い声がした。心臓が止った。〔井伏=夜ふ〕

『感情表現辞典』は書籍を電子データ化した後に例文のみを取り出し，例文をテキストファイル形式にまとめた．感情ごとの文字数は表 2 のようになった．

表 2: 感情ごとの文字数

感情	文字数
喜	48439
怒	25779
悲	35144
怖	19456
恥	8901
好	7191
厭	39609
昂	24158
安	3156
驚	19983

3.2 機械学習

『感情表現辞典』で作成した感情ごとの文書データを機械学習で分類するため，本研究ではプログラミング言語 Python を利用した．機械学習で用いられる言語は Python, R, Java, C++と様々であるが，メソッド，ライブラリが充実してプログラミングが簡便であること，メモリ上限が大きく取れることを考慮し，Python を本実験では採用した．

ランダムフォレスト法を利用して文章データを分類するためには，分類する前に文書データをベクトル化する必要がある．本研究では自然言語処理のために用意された Python ライブラリである Gensim を利用し，Bag of words(BoW) による出現回数を基にしたベクトル生成を行った．BoW モデルはドキュメントの集合全体から単語 1 つ 1 つに ID を割り振り，文書内で出現した単語をカウントしてベクトルを生成するという，非常にシンプルな考えで作られている．

ランダムフォレスト法は決定木分類器を組み合わせて結果を出すアンサンブル学習の 1 つである．決定木学習はあるデータ群に質問を繰り返し行うことによって，データを分ける学習方法を指す．このデータを分ける質問を設定する時に考えるのが情報利得という数値である．情報利得はカイル･バックラー情報量とも呼ばれ，2 つの情報の情報量の差分の期待値で表される．決定木学習ではこの情報量を高く保つことで，決定木の葉が不純度の低いが特徴によってまとまったデータになるのである．ただし，葉が完全に純粋になるまで分割すると，質問の多い，非常に深い決定木になってしまうため，過学習に陥ってしまう．そのため，分析では決定木の最大の深さを決めて行うのが基本になっている．

アンサンブル学習とは，いくつかの分類器を 1 つに組み合わせることで強い分類器を作る学習方法である．分類器 1 つ 1 つに分類の結果を出させ，分類器の出した答えを多数決などの手法で総合判断して 1 つの結果を出している．

ランダムフォレスト法の具体的な手法を説明する．まず，元の文章データからある大きさのランダムな標本を作成する．その標本を用いて決定木分類を行って結果を出す．この標本作成と決定木分類を何度か行なって，決定木分類の結果の多数決で最終的な結果を表示する．ここまでが具体的なランダムフォレストの動きである．

本実験の学習は，教師あり学習という方法を採用している．データに分類の正解をつけて行う教師あり学習は，正解率が最も高くなるようなモデル構築をして，そのモデルを評価できる．そのため，高い評価のモデルを構成できれば，未知のデータでも適確な分析が行えると考えられる．本実験では，分類された各感情ごとに 6 割をモデル構成用の学習データとし，残りの 4 割のデータをモデルの評価を行うテストデータとした．

4 結果・考察

4.1 分類の結果

データ化された『感情表現辞典』の分類を基に，ランダムフォレストで分類した正答率は表 3 のようになった．自立語（動詞，名詞，形容詞，副詞，接続詞，連体詞）をすべて含んだ状態と品詞それぞれの状態で分類した場合の 5 種類に分けて行った．

表 3: 品詞ごとの分類正答率

品詞	正答率
自立語全て	0.569
動詞のみ	0.574
名詞のみ	0.534
形容詞のみ	0.552
副詞のみ	0.538
接続詞のみ	0.536
連体詞のみ	0.547

4.2 考察

実験結果の正答率は自立語全てで 5 割 7 分ほどという結果になったが，ツイートの感情分類を試みた佐藤らの研究は同じベクトル化の方法 (BoW) で平均 2 割

5 分ほどに留まったため，比較すると良好である．差が生まれた要因として，佐藤らの研究では『感情表現辞典』の見出し語を集め，見出し語と適合する 2400 ツイートを抜き出して学習，分類を行っているが，本実験では 4076 例文を分析に利用したため，分類に使用できる情報量に差があったことが考えられる．一方，金らの書き手の同定実験の分類は平均 9 割 5 分に正答率が達していた．この差は，書き手の同定は 10 人の 200 編の小説，11 人の 110 編の作文，6 人のワープロまたは手書きの日記 60 編をデータとして使ったのに対し，本実験では『感情表現辞典』で取り上げられた文書しか用いることができなかったため，分類に使えるデータ量が少なかったことから生じたと察せられる．

品詞ごとに結果を見てみると，動詞が最も高く，次に高いのは形容詞，最も低いのは接続詞という結果になった．感情を見るという観点から考えると，人物の行動を表す動詞や，感覚を表す形容詞が高くなったのは理にかなっていると考えられる．しかし，最も差が大きな動詞のみの分類と接続詞のみの分類でも，そこまで差がつかなかったのは意外であった．文学作品では，接続詞，副詞の用法が感情によって変わる可能性があると思われる．

5　今後の展望

今後の展望として，分類精度の向上をめざしたい．

分類精度を改善する方法として最初に考えられるのは，機械学習の方法をランダムフォレスト以外にとすることである．今回は金らの研究結果を参考にしたためランダムフォレスト法を用いたが，本実験でも SVM，K 近傍法といった別の手法でも実験を行い，分類の正答率を比較検討すべきである．

次に，文書データのベクトル化する方法も見直せると考えている．本実験では BoW をもちいて出現回数のみでベクトルを構成しているが，佐藤らの研究で書かれているように，ベクトル化するためのメソッドも様々用意されているため，これらの手法も試してみる必要があると考える．候補としてあげられるものとして，共起頻度に重要度を足した TF-IDF 法，ニューラルネットワークモデルを導入した Word2Vec，Word2Vec をさらに文章レベルでベクトル化する Doc2Vec がある．

また，今回の実験では用いることのできなかったグリッドサーチによるパラメータのチューニングを行いたいと考えている．グリッドサーチは最初に変数のリストを作成して，そのリストの変数の全ての組み合わせで正答率を算出し，最も正答率が高い組み合わせを探索するというものである．本実験では設定できる 17 の変数を全てデフォルトで決められている値で行ったが，すべての変数の値を検討することで，さらなる精度向上が見込まれる．

参考文献

[1] 中村 明, 『感情表現辞典』, 東京堂出版, 1993.

[2] 佐藤 一輝，尾崎 知伸, 単語埋め込み技術の違いによる日本語ツイート感情分類精度の比較実験, 第 31 回人工知能学会, 2017.

[3] 金 明哲，村上 征勝, ランダムフォレスト法による文章の書き手の同定, 統計数理, 第 55 巻第 2 号, 255-268, 2007.

[4] 原田隆史, 書評中の感性キーワードを用いた小説の分類, 情報知識学会誌, Vol.15 No.2, 57-62, 2005.