感情極性に基づく小説の俯瞰分析への取り組み

吉田　知世　(指導教員：小林 一郎)

1　はじめに

近年，電子化されたテキストが膨大に存在し，これらのテキスト情報を有効に利用する方法として，テキストマイニング技術の研究が盛んに行われている．また，その一つとして感情抽出の研究も進められている．感情抽出の対象となるテキストとして，レビュー記事やアンケートを対象とした研究が多く行われているが，本研究では読書対象となる機会の多い小説に対して感情抽出を試みる．一つの小説内での感情遷移を分析し，複数の小説における類似性を感情の観点から見ることで，小説全体の雰囲気や感情の流れを俯瞰しながら，本の選択ができるようにすることを目的とする．

2　感情表現の抽出

感情表現抽出手法には，収集した感情表現から辞書を作成して用いることで感情表現を抽出する方法などがある．高村ら [1] は語彙ネットワークを用いて単語の感情極性値を自動計算し，単語と感情極性値の対応表を作成した．本研究においても，語彙に対して感情の種類と正負の極性値が付与された辞書を用いて感情表現を抽出する．

2.1　感情表現抽出の流れ

本研究ではインターネット電子図書館である青空文庫 [2] に児童文学として収録されている小説から 88 作品を用いた．図 1 に感情表現抽出の流れを示す．小説を対象テキストとして形態素解析を行い，解析結果に対して，感情表現辞典と感情極性対応表を融合した感情語辞書を用いて感情表現を抽出し，感情値の算出を行う．抽出されたデータを用いて，一文書のデータに対しては，感情の変遷を分析し，複数文書のデータに対しては，感情の類似度の観点から文書の類似判定を行う．さらに，一文書全体に対して KeyGraph[3] を用いて感情表現と内容の関係について調べる．

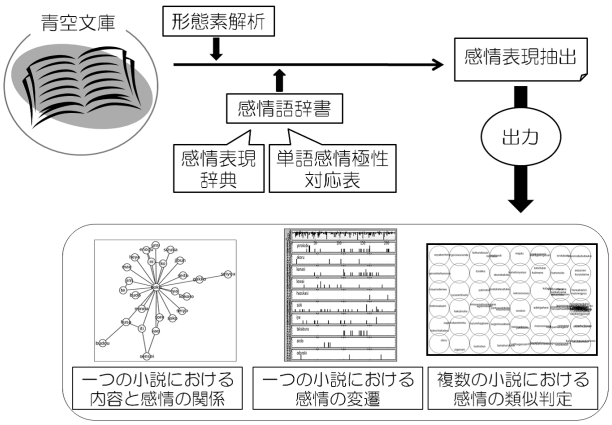

図 1: 感情表現抽出の流れ

2.2　感情語辞書

本研究では “感情表現辞典 [4]”から得られる感情表現の種類ごとのスコアと “単語感情極性対応表 [1]”から得られる感情極性値をまとめ，感情語辞書を構築した．感情表現辞典は近現代作家の作品から感情表現を収録したものであり，感情表現の用例や語句が 10 種類の感情 (喜，怒，哀，怖，恥，好，厭，昂，安，驚) に分類され収録されている．単語感情極性対応表では，感情極性値として－1 から＋1 の実数値を割り当てており，単語の感情極性が positive であるほど＋1 に近く，negative であるほど－1 に近い値が与えられている．感情極性の対象となる語には岩波国語辞書 (岩波書店) の単語 55,126 語が用いられている．

3　実験

感情値の算出が行われた文書に対し，感情に基づく分析を行う．一文書に対する分析では，対象文書の行が進むごとに話が進むと捉え，感情の変遷をグラフに表す．また，内容と感情の関係をグラフ表示する．複数文書に対する分析では，自己組織化マップを用いて感情が類似する文書の判定を行う．

3.1　一つの小説における感情の変遷

今回，分析対象とした文書は小説「一房の葡萄 (有島武郎著)」である．図 2 に示すグラフが一房の葡萄の感情変遷を表したグラフである．感情極性値のグラフは正の極性を持つものと負の極性を持つものを分けて表示しており，その他の感情の種類のグラフは出現頻度を表している．

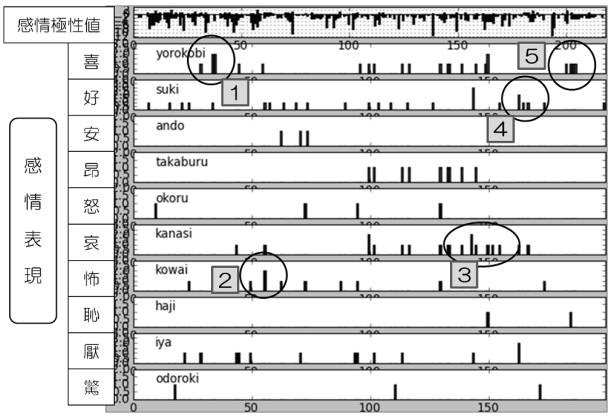

図 2: 「一房の葡萄」の感情変遷に対する解析結果

1〜5 の番号が振られた部分を話の流れに沿って順に見ていくと，本文中では 1 お弁当の時間で嬉しそうな級友の様子 (喜)，2 絵具を盗んだと級友に問い詰められる (怖)，3 先生に怒られそうで泣いている (哀)，4 級友と喧嘩をした主人公が好きな先生に励まされる (好)，5 険悪になっていた級友と仲直りする (喜) のようになっている．

3.1.1　考察

グラフで同じ項目の感情表現が集まっている箇所では，本文中でも感情を喚起する何らかの事象が多く起こっていることが分かった．グラフを通して感情表現

が表れている箇所が明確になり，感情の変遷を分かりやすく提示できた．しかし，抽出された感情表現が誰の感情であるかを判定できないといった問題や，感情表現辞典の感情の種類だけを見ても“驚”や“昂”など分類しにくい語句が存在するという問題があるため，感情表現の種類とともに，感情極性値を利用して表現のされ方をより詳細に見ていく必要があると考えられる．

3.2 一つの小説の内容と感情の関係

KeyGraph は大澤ら [3] によって提案された共起グラフの分割・統合操作によってキーワードを抽出する手法である．本研究では，KeyGraph のアルゴリズムのうち土台にあたる部分を形成することで，文書の内容を表すグラフを作成し，さらにそのグラフに感情語のノードを加えリンクさせることで，文書全体の内容と感情表現をひと目で見られるようにする．

文書 D からキーワードにならない語 (助詞，助動詞などの品詞をもつ語や，非自立語) を取り除き，それ以外の語集合を D_{terms} とする．D_{terms} 中の語を出現回数でソートし，上位 M 語 ($M = min(20, \mid D_{terms} \mid)$) を土台語として，以下の式で定義される語対の共起度 $co = (w_i, w_j)$ によって示される土台語の共起関係を図 3 に示す (但し，$\mid w \mid_s$ は文 s における語 w の出現回数).

$$co(w_i, w_j) = \sum_{s \in D} |w_i|_s * |w_j|_s$$

土台語と 3.1 節で取り出した感情表現の共起度を計算し，土台語ごとに共起する感情表現を表したものを図 4 に，小説に現れる感情表現と共起する土台語を表したものを図 5 にそれぞれ示す.

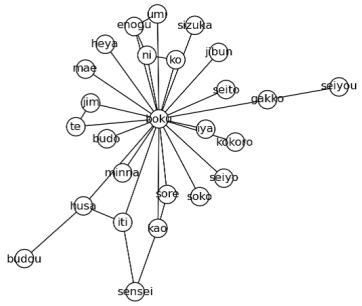

図 3: 土台語の共起関係のグラフ

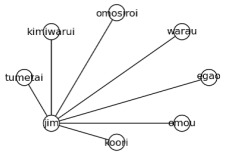

図 4: 土台語に関連する感情語

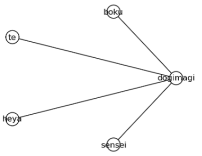

図 5: 感情語に関連する土台語

3.2.1 考察

図 3 から，主人公「僕 (boku)」を中心として内容が展開されていることが分かる．図 4 では土台語“ジム (人名)”と関連する感情表現として { 冷たい，気味悪い，面白い，笑う，笑顔，思う，氷 } が表れた．ジムと喧嘩をするが，仲直りをするため，負の表現だけでなく，{ 笑顔，笑う } という表現がある．図 5 では感情語として“どぎまぎ”と関連する土台語 { 僕，手，部屋，先生 } が表れた．これらを合わせることで，感情語と内容を考慮しながら，小説の選択が可能になる．

3.3 複数の小説における感情の類似判定

自己組織化マップ (SOM)[5] を用いて複数の小説の類似判定を行う．通常，複数の文書の類似判定を行う場合は文書全体に対する特徴量を一括したデータを用いる．しかし，本研究では感情変遷に関する情報も含めて類似度の判定を行うことを考え，テキストの全体を 10 分割し，分割されたブロックごとに 12 属性から成る特徴量 (正の感情極性値，負の感情極性値，喜，…，驚) を算出して一つの文書の特徴量を設定している．

青空文庫の児童書のうち，88 作品を対象にデータを作成し，SOM を利用して類似度の判定を行った．その結果を図 6 に示す.

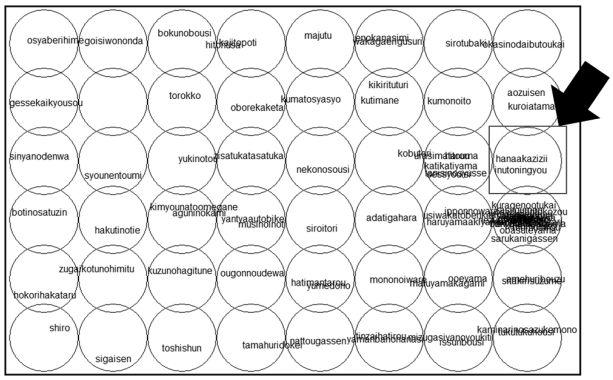

図 6: 文書の類似判定の結果

3.3.1 考察

マッピングの結果，類似していると判断された 2 作品 (図 6 中，矢印参照) を 3.1 節と同様なグラフで比較すると，“喜”，“昂”，“哀”，“厭”の 4 つの同じ項目で感情表現が表れていた．また，“喜”，“昂”，“哀”という 3 つの同じ項目においてはほぼ同時に感情表現が表れており，これらの理由により，この 2 つの小説が類似していると判断されたと考えられる．

4 おわりに

本研究では感情表現辞典と感情極性対応表を融合した感情語辞書を用いて，文書の感情表現を抽出し，一つの小説内における感情の変遷の分析と，複数の小説における感情の類似判定を行った．また，感情語と共起する語を調べることで，文の内容を考慮した感情分析を試みた．今後の課題として，内容分析のにおいてキーワード抽出の方法をさらに検討し，感情語との関連を再考するつもりである．

参考文献

[1] 高村大也，乾孝司，奥村学 “スピンモデルによる単語の感情極性抽出”，情報処理学会論文誌ジャーナル,Vol47 No.02 pp.627-637, 2006.

[2] 青空文庫，http://www.aozora.gr.jp

[3] 大澤幸生, Nels E.BENSON, 谷内田正彦 “KeyGraph：語の共起グラフの分割・統合によるキーワード抽出”, 電子情報通信学会論文誌,Vol.J82-D-I No.2 pp.391-400, 1999.

[4] 中村明「感情表現辞典」，東京堂出版，1993.

[5] T.Kohonen, “Self-organized formation of topologically correct feature maps.”, Biological Cybarnetics, 43: pp.59-69, 1982.