Web書評を対象としたカテゴリー分析と読み手が受けた印象や感情の自動抽出

佐々木若菜　関洋平　青野雅樹
豊橋技術科学大学　情報工学系
wakana@kde.ics.tut.ac.jp, {seki,aono}@ics.tut.ac.jp

1. はじめに

Web上には，Amazon.co.jp[1]やオンライン書店ビーケーワン[2]等といった書籍を扱うサイトから個人のBlogにまで，数多くの書評が存在する．書評は，対象とする書籍を読んだ人がどのような印象を得たのか, あるいはどのようなストーリーなのか等を知ることができる．

書評の情報は，書評を読んだ人物が，対象とする書籍を購入する際の判断材料として用いられる事がある．しかし，Web 上に存在する書評の数は多く，たまたま読んだある一つの書評に自分が欲しい情報が必ずしも記述されているとは限らない．そこで，対象とする書籍に関する書評の集合から, 書籍の判断材料として読み手が受けた印象を抽出する. その情報を書籍ごとに付与することで, その情報が類似した書籍の推薦を行うことを考えた．

本稿では，書籍のジャンルをミステリーに絞り，10 書籍に関する 87 の書評について分析を行い, 推薦に必要と思われる 5 つのカテゴリーを定義した．次に, 作品中の出来事や読み手の印象といった作品を表す特徴の自動抽出を行う枠組みを提案した．

2. 関連研究

レビュー等の評価情報からの評判･感性抽出に関する研究は, 様々なアプローチから数多く行われている[3][4].

原田ら[2]は，書評データ中の感性語を元に，各図書に感性項目の値を設定しておき, ユーザが予め用意してある感性項目をチェックすることで, 書籍の推薦を行うシステムを作成している．

本研究では, 書評データを用いるという点では原田らの研究と同じであるが, 作品全体に対して読者が受けた感性だけではなく, 作品中で起きた個別の事象に対して読者が持つ印象や感性を反映した推薦システムの実現を目指している．また，ユーザが感性の値を指定することなく, ユーザが好きな本を選択することで，その書籍に似た事象が作品中で起きるか，または似た雰囲気を持つ作品かといった様々な尺度での推薦を目指す.

[1] http://www.amazon.co.jp/
[2] http://www.bk1.co.jp/

3. カテゴリー分析

書評の中には様々な情報が含まれている．しかし，それらの情報全てが必ずしも推薦に必要な情報であるとは限らない. 本研究では，推薦に必要となる情報の定義付けを行った．

3.1 カテゴリーの設定

本研究では推薦に必要となる情報の性質を「作品の特徴を表す節や句」と仮定する．「作品の特徴を表す節や句」は，評価者（レビュアー）の体験談などと比べ，作品の特徴を得る上で重要度は高い．この仮定に基づき，ジャンルをミステリーに絞り，オンライン書店ビーケーワンから得た 10 書籍 87 の書評に関して，約 2 ヶ月に渡り分析を行った．分析は，グラウンデッド・セオリー[1]の技法を参考に行った．具体的な分析方法としては，①書評を一文ずつ分析し，各文が表す“現象”を明らかにする，②同じ“現象”に属するように見えるもの同士に分類する, という作業を繰り返し行った. その結果，作品の特徴を表す節や句を以下の 5 つのカテゴリーに分類した．

1. 事象解析
2. 作風
3. ストーリー
4. 評価者推薦
5. ジャンル分類キーワード

各カテゴリーはそれぞれ，

1. 書籍中で起こる事象の要素に対する属性･評価
2. 作品を通した抽象的な性質に対する属性･評価
3. 作品のストーリー
4. 評価者（レビュアー）による作家・作品の推薦
5. ジャンルに特有のキーワード

という性質を持つものとする．

本稿では, レビュアーが読むことで得られた印象や感情が書籍の推薦において重要と考え, これらを含む 1 と 2 に着目した. 以下では 1 と 2 のサブカテゴリーについて説明する．

3.2 事象解析

“事象解析”とは，書籍中で起きた具体的な事象に関する属性・評価を表すものである．しかし，書籍中で具体的に起きた事象にも様々な事柄がある.本稿では詳細な分析に基づき，以下の 8 つのサブカテゴリーを設定した.

1a. キャラクター：登場人物の特性を表す
1b. 会話：作品中での会話の特性を表す
1c. 事件：作品中で起きた事件の特性を表す
1d. 謎：作品中で提示された未解決の問題の特性を表す
1e. 解決：作品中の事件解決の特性を表す
1f. 一般：一般的な特性を表す
1g. 展開：作品の展開を表す
1h. 設定：作品の設定を表す

具体的な例を表 1 に示す.

表 1“事象解析”のサブカテゴリーの具体例

サブカテゴリー		具体例
1a	キャラクター	悪人というのは滅多に登場しない
1b	会話	会話が説明口調
1c	事件	いかにも理系な事件ばかり
1d	謎	日常の謎
1e	解決	科学の力でアッサリ綺麗に解決
1f	一般	美味しそうな描写
1g	展開	実際見てきたように、体験してきたように描写されるストーリー展開
1h	設定	設定に無理があった

“事象解析”は書籍中で起きた具体的な出来事を表すため,対象とした書籍のジャンルに強く関わっている．今回ミステリーを対象とした分析では 1c～1e はミステリー特有のものであることがわかる.

“事象解析”は，具体的な事象を表す語の長さが比較的短く,その語の表現のゆれが少ないことから,“事象解析”は自動抽出が行いやすい．また，自動抽出の結果は,作品中で具体的に起きた事象で読み手の印象に残ったものを抽出している．ゆえに，具体的に起きた事象の傾向における推薦が可能となる.

3.3 作風

“作風”とは，作品が持つ抽象的な性質に対する属性・評価を表したものである．抽象的な性質とは，作品の文体や文章の構成力,作品そのものが持つ雰囲気等といったものを表す．すなわち，作品を表す特徴の一部ではあるが,作品中で起きた具体的な現象を表すものではないとする.抽象的な性質にも様々な種類がある．そのため，“作風”に関しても“事象解析”と同様にして分析を行い，5 つのサブカテゴリーを設定した.

2a. 文体：書き方,文章を表す
2b. 視点：物語を進めていく上での視点を表す
2c. 雰囲気：作品における雰囲気を表す
2d. 文章構成：文章の構成力を表す
2e. 読後感：読後に感じたものを表す
-A 感性：読後の感性を表す
-B 筆力：物語に惹き込まれたかを表す

具体的な例を表 2 に示す.

表 2“作風”のサブカテゴリーの具体例

サブカテゴリー		具体例
2a	文体	軽い文章
2b	視点	本の中に本があるという「作中作」の方式が取られている
2c	雰囲気	和気藹々とした雰囲気
2d	文章構成	構造上ネタバレに弱い
2e-A	読後感-感性	心にしみるものでした
2e-B	読後感-筆力	ぐいぐいとひきこまれます

“作風”は作品が持つ性質を現すため，ジャンルなどには強く左右されない.

“作風”が持つ抽象的な性質は，その性質を現す語の表現のゆれが多い．そのため，“作風”を表す語の特定が困難であり,キーワードを手掛かりとした自動抽出が難しい.自動抽出方法の提案は今後の課題である.

4. カテゴリーの自動抽出

本稿では，これまでに取り上げた“事象解析”・“作風”のうち，ジャンルに特徴的な言葉が多数出現する“事象解析”について専門用語辞書を用いた自動抽出を試みた.

4.1 抽出する要素とその組

3 章で設定した“事象解析”のデータに対して，抽出するキーとなる要素について分析を行い，“事象解析”は，“対象”，“属性”，“出現”，“評価”，“対象（属性)”，の 5 つの要素から構成されるものとした.

それぞれの要素は，

・ “対象”とは，起きた現象に関する話題の中心

となる語（例：登場人物,会話,事件）
- “属性”とは，“対象”の属性となる語（例：個性,テンポ,要素）
- “出現”とは，作品中で現象が起きたことを表す語（例：出る,ある,登場する）
- “評価”とは，“対象”に対する評価を表す語（例：やさしい,ありきたり,日常の）
- “対象（属性）”とは，対象となりかつ意味に属性が含まれている語（例：悪人,殺人,ミステリー）

とする．抽出は，これらの要素の組み合わせで行う．

これらの要素の組を抽出する条件としては，分析の結果から，以下の条件を満たす組が“事象解析”として意味を成すと判断した．

- “対象”または“対象（属性）”が抽出する組の中に含まれている．
- “出現”または“評価”が抽出する組の中に含まれている．
- “対象”と“出現”だけの組は抽出しない．
- “属性”は抽出される組に含まれていても含まれていなくてもどちらでもよい．

以上の条件を満たす組を抽出する．

4.2 抽出方法

抽出は“対象”を中心とした係り受け関係を用いて行う．係り受け解析には，CaboCha[3]を用いた．

抽出システムへの入力は一文ずつとする．抽出手順は以下の通りである．

i) 入力した一文内に，“対象”・“出現”・“評価”・“属性”・“対象（属性）”辞書に存在する単語があるか調べる．
ii) a と b で得た候補について，“出現”・“評価”・“属性”の辞書に登録されている単語が存在するかどうか調べる．
   a）“対象”または“対象（属性）”を中心とする前後の係り受け関係を持つ単語を抽出要素の候補とする．
   b）“対象”または“対象（属性）”からの係り受け先について，係り受け先に係っている別の要素を抽出要素の候補とする．
iii) ii a と ii b で得た要素の組が 4.1 節で述べた条件を満たした場合，その組を抽出する．

今回は，ii a，b で得る関係を調べる距離を変化させて実験を行った．具体的に，ii a で得る関係の距離を 1～5 で，ii b で得る関係の距離を 0～5 に変化させて実験を行った．

5. 実験

4.2 節で述べた方法をプログラムで実装し，自動抽出実験を行った．そして，あらかじめ作成しておいた正解セットと抽出データを比較することで，本自動抽出プログラムの再現率，精度，F-値をサブカテゴリーごとに求めた．

5.1 辞書の作成

本実験では，自動抽出は辞書を用いて行う．

辞書の作成は，カテゴリーの設定を行う際に用いたデータと同じ，10 書籍 87 の書評を用いて行った．具体的には，それらの書評について分析を行い，4.1 節で述べた要素を取得し，各要素に対応した辞書を人手によって作成した．辞書は“対象”，“出現”，“評価”，“属性”辞書の 4 つを作成した．“対象（属性）”は，あらかじめ属性が付与している語ということ以外は，“対象”と同様に，抽出のキーとなる語である．そのため，一つの“対象”辞書としてまとめ，抽出に用いた．ただし，4.1 節で述べた条件を適用するために，辞書の中で“対象”と“対象（属性）”が区別できるように登録した．また，辞書はサブカテゴリーごと要素が区別できるように登録を行った．

5.2 実験方法

実験では，入力データとして，カテゴリーの設定に用いた 10 書籍 87 の書評を用いた．それらの書評に対する正解セットを作成し，抽出結果と正解セットを比較することで再現率・精度・F-値を求めることで，どれだけ係り受け関係に基づく抽出が有効であるかを確認した．

実験では，4.2 節で述べた ii a，b の距離を変化させ，各場合についてのサブカテゴリーごとの再現率・精度・F-値を求めた．また，各サブカテゴリーの F-値の平均が最も高くなる ii a，b の距離を求めた．

5.3 結果

5.2 節で述べた，各サブカテゴリーの F-値の平均が最も高くなった距離の組み合わせは，ii a の距離が 3，ii b の距離が 5 の時であった．

その距離のときの各サブカテゴリーの再現率，精度，F-値の値を表 3 に示す．

再現率，精度，F-値は（1）－（3）式で求めた．

今回行った実験と表 3 の結果から，以下のことが明らかとなった．

[3] http://chasen.org/~taku/software/cabocha/

$$再現率 = \frac{正解セット \cap 抽出データ数}{正解セット数} \quad (1)$$

$$精度 = \frac{正解セット \cap 抽出データ数}{抽出データ数} \quad (2)$$

$$F-値 = \frac{2 \times 再現率 \times 精度}{再現率 + 精度} \quad (3)$$

表 3 各サブカテゴリーの再現率，精度，F-値

サブカテゴリー	再現率	精度	F-値
キャラクター	0.681	0.416	0.516
会話	0.556	0.455	0.500
事件	0.706	0.818	0.758
謎	0.784	0.800	0.792
解決	0.714	0.690	0.702
一般	0.556	0.625	0.588
展開	0.477	0.320	0.383
設定	0.476	0.769	0.588

- サブカテゴリーのうち，“事件”，“謎”，“解決”の F-値は 0.7 を超える結果となった. この原因としては,“対象”とする語が限定されており，その語に対する表現の揺れが比較的小さかったためと考えられる.
- 一方，“展開”，“会話”，“キャラクター”は，精度が良くない結果となった. この原因として，表現の揺れが大きいことが考えられる.
- 再現率が良くなかった, すなわち検出漏れの原因は，抽出のキーとなる語の，各辞書への登録に不備があったためと考えられる.

5.4 考察

今回の実験では, 書評中からサブカテゴリーごとに作品の特徴を表す情報を抽出することで, 特徴ごとに読み手が受けた印象を確認することが可能となった.

－失敗の分析－

今回の抽出手法では, サブカテゴリーごとに要素の組の抽出を行った. 抽出する要素の組の条件は 4.1 節で指定したが，その条件を満たしていても，“事象解析”としては意味を成さない組み合わせが抽出される場合があることが分かった.

その一つとして，“出現”要素については，係り受けの距離が長くなる場合に，“事象解析”として適切でない組を抽出する場合がある.

例えば，＜少年 :“対象”，中性的“評価”，魅力“属性”，ある“出現”＞ならば，作品内に出てきた登場人物の具体的な特徴を表すことから“事象解析”となる．しかし，＜ジャンル :“対象（属性)”，ある :“出現”＞では，抽出元の文は，「ジャンル分けする必要がある」であった．この文の中で，“ある”と組み合わされるべき要素は，“必要”であり，“ジャンル”ではない．また，現在のシステムでは，抽出語の品詞を区別していないため，“である”のような助動詞も抽出していた．今回の実験の誤抽出の 8 割ほどはこのパターンであり，今後改善する予定である.

6. まとめ

今回の実験では，書籍推薦を最終目標として，作品の特徴を，書評から自動的に抽出することを試みた. まず，書評中の情報を詳しく分析し，カテゴリー分類を行った. 次に, 書籍の推薦を行うのに重要となる“事象解析”の自動抽出を実現した．そして，作品の特徴が, その内容により，よく使われる表記が異なる点に着目し, サブカテゴリーごとの抽出の難しさの傾向などを明らかにした.

謝辞

本研究の書評データは，株式会社図書館流通センター　ビーケーワン事業部の御好意により研究目的で使用させて頂いたものです. この場をお借りして御礼を申し上げます．また，本研究の一部は，科研費基盤（C）（課題番号 16500057），若手研究（B）（課題番号 18700241）の支援を受けて遂行したものです.

参考文献

[1] Anselm Strauss，Juliet Corbin，質的研究の基礎 グラウンデッド・セオリーの技法と手順，南裕子（監訳），操華子，森岡崇，志自岐康子，竹崎久美子（訳），（株）医学書院，東京，1999.

[2] 原田隆史，“最適解の付与に基づく重み付けの自動変更”, 情報知識学会誌, Vol.16, No.2, pp.19-22, May. 2006.

[3] 小林のぞみ，乾健太郎，松本裕治，“意見情報の抽出／構造化のタスク仕様に関する考察”，情報処理学会研究報告, NL171-18, pp.111-118 , Jan. 2006.

[4] 乾孝司，奥村学，“テキストを対象とした評価情報の分析に関する研究動向”，自然言語処理，Vol.13，No.3，pp.201-241，July. 2006.