7L-04

ブックレビューからの閲覧者の視点に沿ったレビュー文の抽出方法

寺尾　建登†　　亀井　清華‡　　藤田　聡‡

広島大学工学部†　　広島大学大学院工学研究科‡

1 はじめに

近年，商品の購入を検討する際にレビュー情報を参考にする人が多くなっている．それに伴って，インターネット上におけるレビュー情報量は年々増加している．また，レビューは購入者によって自由に記述されるため，様々な視点で書かれている．本に関するレビューにおいても，「著者情報」「お勧め対象」「読みやすさ」「読後感」「熱中度」 等の視点が考えられる．

そこで本研究では，これらの視点が混在したブックレビューの中から，閲覧者が本の購入を検討する際に重視する情報を含む文を抽出し，ハイライトして提示することを考える．具体的には，システムは上記の５つの視点毎に「タグ」を用意し，レビュー文毎にタグ付けをする．閲覧者が自身にとって重要であると考えるタグを選択すると，システムはそのタグのついた文をハイライトして表示する．これにより，閲覧者はすべての文を読まなくても済み，その負担と時間を軽減することができる．

2 関連研究

レビューの分類に関する研究はいくつか存在する．例えば桑田ら[1]は自由形式のレビューを評価視点毎に分類する手法を提案した．彼らの手法では，予め評価視点毎に記述されているレビューを元に教師あり学習を行う．しかし，そのような視点毎に記述されたレビューが十分に得られない場合には適用できない．また，岩井ら[2]は本のレビュー文がその本の「あらすじ」であるかどうかを分類判定する手法を提案した．

本研究では「あらすじ」ではなく，「著者情報」「お勧め対象」「読みやすさ」「読後感」「熱中度」の５つのタグを考える．前者の３つのタグは使われやすいキーワードが比較的明らかであるが，後者の２つのタグは投稿者の感情的な表現が多く，多様な表現がなされているため，抽出するのは容易ではないと予想される．

3　提案手法

本研究では，前述の５つのタグに該当するレビュー文を重要文と呼ぶことにする．ここでは各タグの重要文を抽出する手法を提案する．

提案手法では，複数のタグに該当するレビュー文を考慮し，「、，。．！？　（スペース）」の全角半角を含む 14 個の区切り文字でレビュー文を区切る．そして，区切られた部分文毎にタグ付けし，出力の際には元の文に戻してハイライトする．以下に，部分文のタグ付け方法として２つの手法を提案する．

手法１ではキーワードを含むものにタグ付けをする．まず，タグ名から連想される単語を代表語と呼ぶことにする．実際のレビュー文には，同義語や表記ゆれの単語が存在しているため，word2vec [3]の単語の類推を利用し，代表語とその同義語や表記ゆれの単語をキーワードとすることにする．word2vec はMikolov らによって提案された「似た意味の単語は似た文脈に出現しやすい」という分布仮説をもとに，ニューラルネットワークを用いて単語の分散意味表現を獲得する手法である．各タグの代表語の選択は，3 名の被験者による多数決で選出した．また，それらの代表語からの単語類似度が小さいものを順に３名の被験者に提示し，多数決でキーワードとして適切であると判断したものを追加した．また，「著者情報」タグのキーワードとしては，データベースからその作品の著者名を追加した．

手法２ではベクトル空間を利用する．各部分文のベクトルを作る為に word2vec の単語ベクトルと，その部分文が含まれるレビュー文章の各単語の TF-IDF 値を使用する．

TF-IDF 値とは文章中に含まれる各単語の重要度を評価する手法の１つであり，その単語の出現頻度 TF と逆文章頻度 IDF の積で表される．$n_{i,j}$をレビュー文章d_jにおける単語t_iの出現回数，$\sum_k n_{k,j}$をレビュー文章d_jにおけるすべての単語の出現回数の和とすると，$TF_{i,j} = n_{i,j} / \sum_k n_{k,j}$となる．|D|を全レビュー文章数，$|\{t_i \in d : d\}|$を単語$t_i$を含む文章数とすると，$IDF_i = \log(|\mathrm{D}| / |\{t_i \in d : d\}|)$となる．

v_{t_i}を単語t_iの単語ベクトルとすると，レビュー文章d_jのl番目の部分文$d_{(j,l)}$のベクトル$v_{d_{(j,l)}}$は以下の式で表すことにする．

$$v_{d_{(j,l)}} = \frac{\sum_{k \in d_{(j,l)}} (v_{t_k} * TF_{k,j} * IDF_k)}{k}$$

そして，各タグの代表文と各部分文とのベクト

Extraction of review sentences matching the user's view point from a collection of book reviews
† Kento TERAO, Hiroshima University
‡ Sayaka KAMEI, Hiroshima University
‡ Satoshi FUJITA, Hiroshima University

ル距離が閾値以下の部分文にタグ付けをする．代表文としては，単語数が 3 以下の各部分文を 3 名の被験者に提示し，多数決で各タグの代表文にふさわしいと判断したものから 3 文を厳選した．

4　実験

システムが適切に重要文を抽出出来ているか評価実験を行った．

まず，楽天のレビューデータセットの商品ジャンル「小説・エッセイ」に分類されている 19151 件のレビューを word2vec の学習に使用した．word2vec の実装は Python の gensim モジュールを用いて Skip-gram で上記データセットを学習したモデルを作成した．単語ベクトルの次元数は 200，最小単語出現数は 1，学習反復回数は 20，それ以外のパラメータはデフォルト値を使用した．

次に，人手でタグ付けを行い，正解ラベルのついたレビューデータセットを作成した．ここでは，特にレビュー件数の多かった「流星の絆」「永遠の 0」「八日目の蝉」の 3 作品についての計 318 件のレビューのみを用いた．各タグを 3 人に担当してもらい，少なくとも 2 人がタグ付けしたら正解ラベルとして採用することとした．この正解ラベルをつけたレビュー集合を使用して各手法とそれらを組み合わせた手法 1+2 の適合率，再現率，F 値を計算した．

表 1 に各タグの代表文を示す．表 2 に各タグのキーワードを示す．これらのキーワードのうち太字が代表語である．表 3 に各タグの閾値を示す．また，表 4 に適合率，再現率， F 値を示す．

手法 1 ではすべてのタグにおいて適合率は良い結果が得られた．しかし再現率が低いタグが多く，キーワードだけでは網羅できなかった文が多いことが分かる．F 値を見ると，「著者情報」「お勧め対象」といったキーワードが比較的はっきりしているものについては，予想通り，最も良い結果となっている．手法 2 では「著者情報」「お勧め対象」の精度が悪く，それ以外のタグでは再現率が手法 1 より良くなっている．ベクトルを利用することで様々な表現に対応しやすいことが分かった．

5　おわりに

本研究では，レビュー文の中から閲覧者の視点に沿った箇所を抽出する手法を提案した．今後の課題としては，代表文や代表語の自動選択手法の検討を考えている．また，実際にシステムを使用することによる効果について評価することを考えている．

表 1　代表文

タグ名	代表文
著者情報	相変わらずの筆力。／すごい作家だ。／さすが売れっ子作家！
お勧め対象	お勧めです。／必読です。／一読あれ！
読みやすさ	読みやすい。／スラスラ読めます。／わかりやすい！
読後感	面白かった。／超感動！／読後感最高！
熱中度	一気に読めます。／はまります！／一気に読破！

表 2　キーワード

タグ名	キーワード
著者情報	**著者**，**作者**，**筆者**，作家
お勧め対象	**おすすめ**，オススメ，お勧め，**年代**
読みやすさ	**読みやすい**，**スラスラ**，ササッと，**簡単**，すらすらー，ささっと，ススっと
読後感	**感動**，**残念**，**面白い**，おもしろい，バットエンディング，総合的，すらすらー，むずい
熱中度	**あっという間**，**夢中**，**入り込む**，あっという間に，アッという間に，いっぺんに，更ける，睡眠不足，のめり込む，引き込む

表 3　閾値

	著者情報	お勧め対象	読みやすさ	読後感	熱中度
閾値	0.210	0.322	0.390	0.180	0.520

表 4　実験結果

	手法	著者情報	お勧め対象	読みやすさ	読後感	熱中度
適合率	1	**0.672**	**0.714**	**0.750**	**0.534**	**0.827**
	2	0.163	0.197	0.293	0.230	0.297
	1+2	0.429	0.383	0.291	0.256	0.341
再現率	1	0.652	0.400	0.102	0.172	0.289
	2	0.121	0.160	**0.523**	0.335	0.497
	1+2	**0.682**	**0.480**	**0.523**	**0.417**	**0.624**
F 値	1	**0.662**	**0.513**	0.180	0.261	0.289
	2	0.139	0.176	**0.376**	0.273	0.372
	1+2	0.526	0.426	0.374	**0.317**	**0.441**

謝辞

本研究では，楽天株式会社が提供し，国立情報学研究所が配布している「楽天公開データ」を利用させていただきました．ここに記して謝意を表します．

参考文献

[1]桑田大徳ら，“自由形式で記述されたレビューの評価視点別自動分類の提案”，情報処理学会第 73 回全国大会，2011

[2]岩井秀成ら，“レビュー文を対象としたあらすじ分類手法の提案とあらすじ非表示システムの開発”，情報処理学会インタラクション，2013

[3]T.Mikolov， et.al.,”Efficient Estimation of Word Representations in Vector Space.” In ICLRWorkshop. 2013