意見の重要度と客観的補足情報を考慮したレビュー要約

唯野 良介　　嶋田 和孝　　遠藤 勉
九州工業大学大学院 情報工学府情報科学専攻
{r_tadano, shimada, endo}@pluto.ai.kyutech.ac.jp

1 はじめに

近年，CGMの発展に伴い，個人が意見やレビューを投稿する機会が増えている．これらの主観的な意見（評価情報）は有益な情報と考えられる．そのため，レビューなどの評価情報を自動的に抽出・要約するシステムの必要性が高まっている．

従来のニュース記事などを対象とした文書要約では，単純に単語の重要度を基にした手法が一般的である．一方，レビュー要約では意見の極性の考慮や主要な意見の把握など様々な点にも着目する必要がある [7]．また，レビューに記述された意見は様々なアスペクト（カメラの “画質” や “操作性” など）に基づいており．それらのアスペクトを考慮したレビュー要約手法も提案されている [1][3]．しかし，レビュー要約において主観的情報が主に扱われる一方で，レビュー要約での客観的情報の扱いに関しては十分な議論がなされていない．

我々は現在，アスペクトを考慮したレビュー要約手法の構築に取り組んでいる [8]．本稿では手法の改良と新たに客観的情報とレビュー要約の統合を行い，その有効性を検証する．

本研究で扱うレビューは，あらかじめ複数の評価視点（アスペクト）が与えられており，レビューアにより評点（5段階評価など）が付与されているものとする．そして，自由にコメントが記述されている．レビュー要約は，レビュー集合から重要文を抽出し，それらを要約文とすることで生成する．重要度を決定する要素には単語の特徴度を表す $tfidf$ 値と，類似文クラスタリングによって得られた言及の多さを用いる．加えて，レビューの評点情報を用いることで，極性のバランスを考慮した要約文の選定を行う．客観的情報に関しては Wikipedia から情報を抽出し，補足情報としてレビュー要約へ導入する．これらを統合利用することで，より有効な要約の生成を目指す．

2 文抽出に用いる要素

重要文抽出に用いる要素として，レビューの評点情報，単語の特徴度，言及の多さの3つについて以下の節で述べる．

2.1 レビューの評点

要約を生成する場合，要約対象全体の意見をバランス良く反映させる必要がある．この点を考慮するための要素として，レビューに付与されている評点に着目する．レビューにはアスペクトごとにあらかじめ評点が与えられていることを利用し，レビュー内で出現した各評価文（何らかの対象について評価をしている記述）に評点を割り当てる．各評価文の評点は，評価文に付与されたアスペクトに対応した点数となる[*1]．作成した評価文と評点のペアは，最終的に要約を生成する段階で意見のバランスを考慮するために用いる．

2.2 単語の特徴度

要約対象において固有な単語を含む文ほど重要度が高いと考えられる．そこで，各単語の特徴度を $tfidf$ 法を用いて算出する．レビューに含まれるテキストを形態素解析[*2]し，要約対象レビュー群 T に対する各単語 w の $tfidf_w$ 値を次式で算出する．算出対象は，名詞（非自立語，接尾語，代名詞，数を除く）と形容詞のみとする．

$$tf_w = \frac{\log_2(\text{レビュー群 } T \text{ での単語 } w \text{ の頻度} + 1)}{\log_2(\text{レビュー群 } T \text{ での単語の異なり数})}$$

$$idf_w = \log_2\left(\frac{\text{対象ドメインの全レビュー数}}{\text{単語 } w \text{ を含むレビュー数}}\right) + 1$$

$tfidf_w$ 値は $tf_w \times idf_w$ となる．

次に，この値を用いて各評価文の $tfidf$ 値を求める．文 s の $tfidf_s$ 値は次のように定義する．

$$tfidf_s = \frac{\text{文 } s \text{ 中の単語の } tfidf_w \text{値の和}}{\text{文 } s \text{ 中で } tfidf_w \text{値をもつ単語数}}$$

この結果，特徴的な語に着目した重要度が得られる．

2.3 言及の多さ

レビューが複数ある場合，要約文を抽出する際に同じ内容の意見が重複し，冗長性を生む可能性がある．一方で，言及の多い意見は重要な意見ともいえる．そこで，言及の多さを特徴として抽出するために類似文のクラスタリングを行う．各クラスタを一つの話題とし，クラスタの大きさを言及の多さとする．

クラスタリング手法には，広く用いられている k-$means$ 法を採用する．分割クラスタ数 k は動的に決定する [9]．各評価文は，形態素解析で得られた形態素を

[*1] 本稿では要約形成部分に焦点を当てるため，評価文の特定や評価文とアスペクトの対応付けは完了済みと仮定する．
[*2] 形態素解析器には MeCab を用いた．
http://mecab.sourceforge.net/

素性とした特徴ベクトルで表す．素性とする対象は名詞（非自立語，接尾語，代名詞，数を除く），形容詞，動詞のみとする．各素性の値は，2.2 節で $tfidf$ 値が得られている語にはその値を与え，それ以外には評価文内での頻度を与える．また，文章中の話題格や主格，目的格などはその文章において主題性が高いとし，それらの語に対しては重み付けを行う．これらのアルゴリズムを用いてクラスタリングを行い，複数の類似文クラスタを形成する．また，クラスタの過分割に対応するため，クラスタリング結果の補正を行う [8].

3 要約器の構築

2 節で述べた 3 つの特徴要素を統合し，要約を生成する方法について述べる．

まず，各文の $tfidf_s$ 値と言及の多さ (クラスタリング結果) を組み合わせて，各クラスタの総合的な重要度を求める．あるクラスタ C の重要度を $Imp(C)$ としたとき，その重要度を次の式で定義する．

$$Imp(C) = Mean_{tfidf_s}(C) \times \log(|C| + 1)$$

$Mean_{tfidf_s}(C)$ はクラスタ C に属する文の $tfidf_s$ 値の平均値であり，$|C|$ は C に属する文の数である．

この重要度とレビューの評点を考慮し，最終的な要約文の決定を行う．要約文の決定手順を次に示す．

1. 各クラスタから代表文を抽出する．
2. 代表文が属する評点別に，代表文を分類する．
3. 各評点に属する代表文数の割合を考慮し，評点ごとに $Imp(C_i)$ の高い文から要約文として選択する．

今回は各クラスタの代表文として，各クラスタで中心に最も近い文（セントロイド）を利用する．

また，各クラスタ内には異なる極性（肯定/否定）の意見が混在しているため，どちらの極性から代表文を抽出するかが重要となる．しかし先行研究 [8] ではこの点を考慮しておらず，代表文がクラスタの極性の傾向を反映していないという問題があった．そこで，評価文に付与された評点からクラスタ内でどちらの意見が多いかを推定する．例えば評点が 5 段階評価の場合，1,2 点を否定的，4,5 点を肯定的とすることでクラスタの極性を判定する．意見が多い方の極性を持つ評価文から代表文を抽出する．

4 客観的情報の補足

客観的情報とは対象のスペックや売上などの事実情報を指す．また，対象について中立的な観点から述べている記述に関しても本手法では客観的情報として扱う．これらの情報をレビュー要約に組み込むことで，より効果的な情報提示が可能と考えられる．

本稿では情報の抽出源として Wikipedia を用いる．まず，要約対象をエントリとする Wikipedia ページ内の記述を抽出し，辞書を作成する．この際，表や箇条書きなどの記述構造がある場合はその構造を利用し，キーワードとその説明記述を関連付ける．補足情報の提示方法は，要約文内の単語ごとに提示する形式とする．要約文に含まれる各キーワードに対して辞書とマッチングを行い，キーワードと関連付けられた記述およびキーワードを含む記述を提示する．

5 実験

提案手法を実際のレビューに適応し，有効性を確かめた．実験データには，Web サイト[*3]から抽出したゲームレビューを用いた．このゲームレビューには，各レビュー記事ごとに 7 つのアスペクト（評価項目）が含まれている．アスペクトは「オリジナリティ (o)」,「グラフィックス (g)」,「音楽 (m)」,「熱中度 (a)」,「満足感 (s)」,「快適さ (c)」,「難易度 (d)」である．実験対象として，Nintendo DS ソフト「New スーパーマリオブラザーズ」のレビューを用いた．対象ゲームには 170 のレビュー記事が存在しており，本実験では任意に抽出した 25 レビュー（約 450 文）を要約対象とした[*4].

この 25 レビューに対して，3 人のアノテータがそれぞれ評価文の抽出と評価文とアスペクトの対応付けを行った．加えて，人手要約として 25 レビューから 50 文を各アノテータが抽出した．

5.1 レビュー要約の評価

提案手法の有効性を確かめるために，以下の 2 つの手法のレビュー要約を比較した．

手法 1. $tfidf_s$ 値のみを用いた要約文抽出.
手法 2. 提案手法による要約文抽出.

ここで**手法 1** はベースラインであり，これは従来の重要度ベースによる要約手法である．

要約の例として，「熱中度」に対する**手法 2** の適応結果を表 1 に示す．代表文とその $Imp(C)$，評点に加えて，所属するクラスタとの関係性を調べるために，代表文が属するクラスタ内の評点平均値を共に示している．

代表文の評点と所属クラスタ内の評点平均値が近いことから，クラスタ内の評点平均値に近い評点を持つ代表文を抽出できたといえる．これは 3 節で新しく加えたクラスタ内の極性判定により，代表文がクラスタの極性の傾向を反映できたためと考えられる．

ただし問題点として，評点と意見の内容が一致しない例が存在した．例えば「隠し要素... あってよかった」の代表文は意見の内容と評点が矛盾する．これは要約文の抽出元であるレビューの評点が全体的に低い場合や，意見の一貫性がないことが原因であった．また，2.1 節で述べたレビューの評点を用いた評価文の評点推定の精度に関しては，0,1,2 点を否定的，4,5 点を肯定的として主観的に精度を調査したところ，およそ 8 割程度の正解率であった．より正解率を高めるには，評価表現辞書などの別知識の利用が考えられる．

[*3] http://ndsmk2.net/
[*4] *tf-idf* 値の算出には 170 レビュー全てを用いている．

表 1　**手法 2** による要約文抽出結果（熱中度）.

要約文	Imp(C)	評点	評点平均
ストレスは溜まるし，熱中させるものがないので 2 日で飽きました	2.64	0	0.00
とにかくクリア特典が全然無いのが一番最低でした	4.26	1	1.80
隠し要素　まあ，多すぎず，少なすぎず，あってよかった	3.45	2	2.00
隠しステージやクリアー後のオマケが無い	4.96	3	3.50
スターコインなどを集めるやりこみ要素が良かった	8.43	4	3.36
巨大キノコとか出てくるとものすごく笑える	3.03	5	5.00

表 2　自動要約と人手要約の間における ROUGE-1 スコア.

	a	c	d	m	g	o	s	*mean*
手法 1	**0.301**	**0.506**	0.357	**0.476**	0.095	**0.373**	0.351	**0.351**
手法 2	0.275	0.453	**0.430**	0.303	**0.205**	0.341	**0.354**	0.337

表 3　人手による要約文の一致度評価（数字は %）.

	a	c	d	m	g	o	s	mean
手法 1	14.3	**37.5**	21.4	25.0	33.3	27.3	20.0	25.5
手法 2	**28.6**	31.3	**50.0**	25.0	**66.7**	**36.4**	**25.0**	**37.6**

次に要約の質を評価するため, 定量的評価として提案手法による自動要約と人手要約の比較を行なった．**手法 1**，**手法 2** を用いて 50 文を要約文として抽出した．各アスペクトに対して何文抽出するかは，各アスペクトに対応付けられた評価文数の割合から推定した．要約の比較には ROUGE-N[2] を利用した．ROUGE は n グラムの一致度を用いて，対象文書の正解文書に対する再現率を表す.

各アスペクトでの ROUGE-1 のスコアとその平均値「mean」を表 2 に示す．ROUGE-1 の結果では，提案手法である**手法 2** よりも**手法 1** の平均スコアが高い結果となった．しかし，アスペクトによっては**手法 2** のスコアが高い場合も存在した．また，ROUGE は単語の重なりを測る手法であるため，表記は異なるが内容的には一致するというケースは扱うことができない．そのため，要約の冗長性に関しては正しく評価することが難しい.

そこで新たに，2 人の被験者に要約の内容的な一致度を評価してもらった．被験者は自動要約文と正解要約文の 2 文を提示され，2 文が内容的に一致するかどうかを判断する．ただし，一つの正解要約文に複数の自動要約文が重複して一致する場合は冗長とし，1 度目の一致以外は一致度の計算に含めない.

人手による評価結果（再現率）を表 3 に示す．表 2 とは異なり，表 3 では**手法 2** のスコアが**手法 1** よりも全体的に高い結果となった．これは手法 1 では冗長性が多く含まれており，一つの正解要約文に対して複数の自動要約文が一致したのが原因であった．この結果から提案手法は，人手要約との内容的な一致度の向上に有効であるといえる.

5.2　補足情報の評価

生成したレビュー要約に対して補足情報の統合を行なった．実験をするにあたり，要約提示用の GUI ツールを試作した．図 1 に GUI ツールと補足情報の例を示す．補足情報の有効な提示方法を調査するため，複数の提示方法を用意した．被験者は 4 名である．評価項目とそれぞれの評価結果を次に示す.

補足対象　補足情報を提示する場合，どの単語をその処理の対象とするかが問題となる．そこで，情報を補足する対象を文章内の全ての名詞にした場合と $tfidf_s$ 値の高い単語のみにした場合の 2 種類を用意し，比較した.
調査の結果，$tfidf_s$ 値の高い単語のみを対象とした方が効果的であることがわかった．これは，全名詞を対象にするとノイズとなるキーワードが増えてしまうことが原因であった.

提示順　補足情報が複数ある場合，情報の提示順が理解度に影響すると考えられる．そこで補足情報の提示順として，Wikipedia ページ内での出現位置順, 文章の $tfidf_s$ 値順, キーワードの主題性順[*5]の 3 種類を比較した.
評価の結果，3 種類のうちキーワードの主題性順の評価が最も高かった．これは，キーワードの主題性が補足情報の有用度の評価に有効であることを示している.

提示数　補足情報として必要な提示数を調査するため，最大提示数として，3 文，5 文，全文の 3 種類を用意した.
その結果，最も評価が高い提示数は 5 文であった．3 文では情報量が少なく，全文では数が多いため見辛いことや重要でない情報も提示されることが原因であった.

有効性　客観的情報の補足の有効性を検証するため, 情報の理解促進としての効果を 1–5 点で評価してもらった.

[*5] 2.3 節での素性の重み付けと同様に，構文解析によって補足文章におけるキーワードの主題性の高さを推定した.

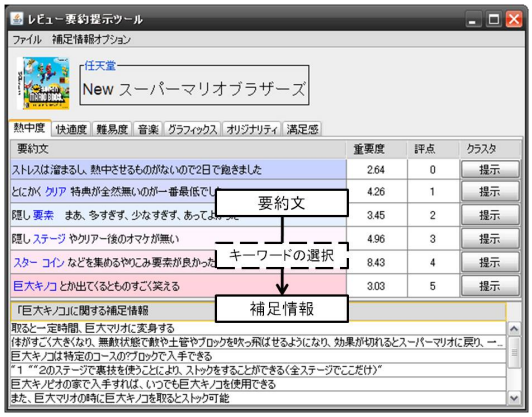

図 1　GUI ツールの外観と補足情報の例.

その結果，4 人の平均で 4 点となった．補足情報の量と提示方法を工夫することで，情報理解の手助けになる可能性は高いと考えられる．

また本実験では対象ドメインがゲームであったため，客観的情報としてはキーワードの詳細な説明などの情報が主であった．一方で，事実情報については十分な量を得ることができなかった．対象が他のドメイン（例えば PC や携帯電話などの多彩なスペック情報を持つドメイン）であれば多くの事実情報を取得可能であり，本手法の有効性もより高まると考えられる．そのため，今後は他のドメインを対象とした実験も行う必要がある．

6　関連研究

レビュー要約を対象とした先行研究として，Blair-Goldensohn ら [1] は WordNet やレビューの評点，最大エントロピー法を用いて文の極性値を算出し，極性値の高い評価文を要約文としている．この手法の利点として，各文の極性を高い精度で推定できる点が挙げられる．しかし，極性値の高い評価文が要約文として必ずしも重要であるとは限らない．

レビュー要約を最適化問題の一つとして解く手法も研究されている．高村ら [10] はレビュー要約を施設配置問題を用いてモデル化している．Nishikawa ら [5] は整数線形計画法を用いて，文抽出と可読性の考慮を同時に行っている．これらの手法は有効であるが，客観的情報の扱いに関しては触れられていない．

一方，Lu ら [3] は PLSA にアスペクトの概念を導入し，専門家によるレビューと Web 上に散在する意見の統合を行っている．統合する意見は，Web 上に存在する意見の数から話題の大きさを測ることで，重要な意見を選定している．Lu らも情報の補足を行っているが，彼らの目的は 1 つのレビューを基に，それに対する類似文や補助文を付加させることで情報量を増やすことである．そのため，多くの情報から重要な部分を選定する我々のタスクとは厳密には異なる．

本稿では評価文の抽出方法や，評価文が属するアスペクトの特定方法については議論しなかった．しかし，これらは本提案手法において重要なタスクである．Pang ら [6] は主観的な文章の抽出を最小カット問題として解くことで行い，それらが意見要約に利用可能と述べている．波多野ら [11] は機械学習を用いて評価文が属するアスペクトの特定を行っている．また，Lu ら [4] はアスペクトを対象文書から動的に抽出する方法について述べている．将来的にはこれらの手法を我々の手法に導入する必要がある．

7　おわりに

本稿では，アスペクトと客観的情報を考慮したレビュー要約手法を提案した．レビュー要約の生成では，レビューの評点，単語の特徴度，言及の多さに着目することで要約文の抽出を行った．加えて，生成したレビュー要約に対して Wikipedia から抽出した客観的情報を補足した．提案手法によるレビュー要約と人手要約を比較した結果，文の重要度と冗長性を考慮した要約の生成が確認できた．また，客観的情報の補足について評価をした結果，情報の理解促進としての有効性が確認できた．

今後はクラスタリングや評点情報の扱い方を改良することでレビュー要約の質を上げると共に，客観的情報との有効な統合方法について考察していきたい．

参考文献

[1] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. A. Reis, and J. Reynar. Building a sentiment summarizer for local service reviews. In *WWW 2008: NLPIX Workshop*, 2008.

[2] C. Lin. Looking for a few good metrics: Automatic summarization evaluation – how many samples are enough? In *NTCIR Workshop 4*, 2004.

[3] Y. Lu, C. Zhai, and N. Sundaresan. Opinion integration through semi-supervised topic. In *WWW 2008*, pp. 121–130, 2008.

[4] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *WWW 2009*, pp. 131–140, 2009.

[5] H. Nishikawa, T. Hasegawa, Y. Matsuo, and G. Kikui. Opinion summarization with integer linear programming formulation for sentence extraction and ordering. In *Coling 2010*, 2010.

[6] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL*, pp. 271–278, 2004.

[7] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, Vol. 2 (1-2), pp. 1–135, 2008.

[8] R. Tadano, K. Shimada, and T. Endo. Multi-aspects review summarization based on identification of important opinions and their similarity. In *PACLIC24*, pp. 685–692, 2010.

[9] 関, 嶋田, 遠藤. 表の属性と属性値の関係を利用した類義語抽出. 電子情報通信学会論文誌, Vol. J89-D, , 2006.

[10] 高村, 奥村. 施設配置問題による文書要約のモデル化. 人工知能学会論文誌, 25 巻, 1 号 A, 2010.

[11] 波多野, 嶋田, 遠藤. クラスタリングを利用した評価文のアスペクト推定. In *FIT 2010*, 2010.