

## レビューを対象とした単一文書要約

小池 将郎<sup>†</sup>      笹野 遼平<sup>††</sup>      高村 大也<sup>††</sup>      奥村 学<sup>††</sup>  
 東京工業大学 総合理工学研究所<sup>†</sup>      東京工業大学 精密工学研究所<sup>††</sup>

mkoike@lr.pi.titech.ac.jp, {sasano, takamura, oku}@pi.titech.ac.jp

### 1 はじめに

近年、ブログやレビューサイトなどの登場により、Web上で商品や施設などに対して一般の人が気軽にレビューを投稿、閲覧できるようになった。しかし、レビューは大量に投稿されていることが多く、有益な情報を探し出すことが難しくなっている。そこで、情報を効率的に得ることを目的として、レビューを対象とした自動要約の研究が盛んに行われてきた。その多くは同一の対象に対する複数のレビューの集合から、多数の著者に共通する意見や、著者に特徴的な意見を抽出することを目的とした複数文書要約として問題を扱ったものである。

しかし、複数文書から生成された要約だけでは、実際に各意見がどのような文脈や条件で書かれたものかが分からず、関連するレビューに目を通すことになる場合も多い。しかし、その際は本文を全て読まなくてはならず、時間的コストがかかるなどの問題がある。したがって、各レビューを要約の対象とし、著者が最も主張したいことを捉えることも同様に重要なタスクであると考えられる。しかし、レビューは個人が商品に対して意見や感想を述べているという点で、新聞記事のような他の文書とは異なる性質を持っている。そのためレビューにおける著者の主張をより正確に捉えるためには、レビュー特有の情報を考慮し、どのような手がかりがレビューの要約に効果的であるかを明らかにする必要がある。

そこで本研究では、レビューの持つ特徴として、感情極性、商品ごとの評価の観点などの評判情報や、そのような情報の文間の遷移、及び、文とそのレビューのタイトルとの関係性などに着目し、各レビューを対象とした単一文書要約手法の開発を目指す。

### 2 関連研究

単一文書を対象とした評判要約に関する研究には、以下のものがある。Beinekeら[1]は、映画のレビューサイトにおいてレビューに付属する要約を正解とし、ナイーブベイズ分類器とロジスティック回帰を用いてレビュー中の各文を要約に含めるか否かを分類する問

題を扱い、文の位置と単語の種類が要約に有効な素性であることを示した。本研究ではほぼ同じ素性を後述する基本素性内に取り入れている。Maoら[7]は、文書内での感情の起伏を可視化し、予測した流れの極大値と極小値に当たる文を、最初と最後の文に加えて選択することで簡単な要約が可能であることを示した。Pangら[9]は、文が主観的かどうかを分類するタスクを扱い、主観的な文を全て選択したものが要約になると述べている。しかし、これらの研究では要約に対する一部の手がかりの有効性が示されるに留まっており、これだけでは不十分であると考えられる。また、Maoら、Pangらの研究は要約が研究の主目的ではない。

### 3 文のランキング学習のための素性

本研究では、単一文書に対する要約を、重要度に関する文のランキング問題として捉える。文の正しいランキングが得られれば、上位の文を選択することで短い要約文書を作成することや、重要度が高い文をハイライトして表示するなどのインターフェイスを構築することが可能である。ここでは、レビュー文のランキング学習に効果的な手がかりを調査する。対象文から得られる手がかりに加え、対象文とタイトル、対象文とレビュー全体、対象文と他の文との関係から得られる手がかりについても調査を行う。

#### 3.1 基本素性

ランキング学習に用いる文の素性として、まずは従来の単一文書要約に広く用いられており、レビューも共通して持つ手がかりの導入が考えられる。本研究では、単一文書要約の研究であるNishikawaら[8]が文の素性に用いたものを参考にし、以下を基本素性として作成した。

- 内容語<sup>1)</sup>の文書内頻度の合計
- 単語ユニグラム (基本形, 品詞, 表層形)
- 文長
- 文の位置 (最初, 最後から数えて何文目か, レビュー全体を5分割した際の位置)

<sup>1)</sup>名詞, 動詞, 形容詞

表 1: 提案素性一覧

番号	素性名	素性の説明
1	P, A <sub>1,2</sub>	極性, 属性 *
2	{P, A <sub>1,2</sub> }prob	極性, 属性の各値の確率
3	{P, A <sub>1,2</sub> }probMax	極性, 属性の各値の確率における最大値 *
4	PA <sub>1,2</sub> probMax	極性と属性の各値の確率における最大値の積
5	Twords	タイトルの内容語の有無
6	TwordsTPprobMax	タイトルの内容語を含むとき, タイトルの極性の各値の確率における最大値
7	TwordsCov	タイトルと内容語の一致度 (文の内容語の数で正規化)
8	TwordsCovTPprobMax	上記の一致度とタイトルの極性の各値の確率の最大値との積
9	Tmatch{P, A <sub>1,2</sub> }	タイトルとの極性, 属性の一致
10	Tmatch{P, A <sub>1,2</sub> }T{P, A <sub>1,2</sub> }probMax	タイトルと極性, 属性が一致するとき, タイトルの極性, 属性の各値の確率における最大値
11	Tcos{P, A <sub>1,2</sub> }	タイトルとの極性, 属性の各値の確率の余弦類似度 *
12	DcosW	レビュー全体と内容語のみで構成された Bag of Words の余弦類似度
13	DmatchP	レビュー全体との極性の一致
14	DmatchPTPprobMax	レビュー全体と極性が一致するとき, レビュー全体の極性の各値の確率における最大値
15	Dcos{P, A <sub>1,2</sub> }	レビュー全体との極性, 属性の各値の確率の余弦類似度 *
16	PprobMaxDiff	極性の各値の確率における最大値が, 直前 (直後) の文より大きい
17	PprobMaxDiff2	極性の各値の確率における最大値が, 直前と直後の両方の文より大きい
18	Pratio	同じ極性を持つ文のレビュー内で占める割合
19	rank, rank1, rank3	* が付く素性の確率又は数値の大きさのレビュー内での順位 (順位そのまま, 1 位か, 3 位以内)
20	RSTrel	DEP-DT における係り先の文との関係
21	RSTrelDiff	上記の関係が直前の文と異なるか
22	RSTdepth, RSTdepth3, 10	DEP-DT における根からの深さ, 及び全体の深さに対する割合 (3, 10 分割)
23	RSTdepthN	DEP-DT における直近の核となる文からの深さ
24	RSTchild, RSTdes	DEP-DT における子ノード, 子孫ノードの数
25	RSTparentDiff	DEP-DT における親ノードが, 直前の文と異なるか

### 3.2 提案素性

基本素性に加えて, レビューの要約に際し, 有効であると考えられる, 大きく分けて 4 種類の提案素性を導入した. 提案素性の一覧を表 1 に示す<sup>2)</sup>.

1 つ目は, 文の持つ評判情報に関する素性であり, 表 1 では番号 1 から 4 ままでが対応する. レビューは商品に対する意見や感想を述べた文書であるため, 感情極性 (極性) や, 商品に対する評価の観点 (属性)<sup>3)</sup> といった評判情報が手がかりとして利用できる. 本研究では, 素性の作成にあたり, 極性はポジティブかネガティブかの 2 値分類, 属性は「書き出し部」「各属性に関する記述」「総評」の 3 値分類と, どの属性か (ドメイン<sup>4)</sup> ごとに種類は異なる.) についての分類を行った. 各文に対してそれぞれの確率を推定し, その結果を用いて素性を作成した. ここで, 素性名において, 極性は P, 属性は上記の分類の種類に基づいて前者を A<sub>1</sub>, 後者を A<sub>2</sub> と表現している.

2 つ目は, 文と, レビューに付属するタイトルとの関係を表す素性である. 投稿されたレビューにタイトルが付与されていた場合に, その部分に著者の主張が含まれていることが多いという考えから, 単語の一致度や前述の評判情報などを用いてタイトルとの関連度を表現する素性を作成した. 番号は 5 から 11 ままでが対応している.

3 つ目は, 文とレビュー全体との関係を表す素性である. これは, 著者が強く主張したいことは, レビューにおいて複数回, 言及されるという仮定に基づいてい

<sup>2)</sup>表 1 の素性 7, 11 は従来の研究でも頻繁に用いられる基本素性であるが, 有効性を再度確認するために本研究では提案素性のリストに加えた.

<sup>3)</sup>商品「ノートパソコン」に対する「処理速度」や「デザイン」といった商品を実際の観点から評価する上での着眼点のこと.

<sup>4)</sup>商品の属するカテゴリ

る. 対応する番号は 12 から 15 までである.

最後は文同士の関係性を表す素性であり, これは, 各文の持つ情報がそのレビュー内で特徴的かどうかを表す手がかりとなる. まず, 前述の評判情報を同じレビュー中の他の文と比較した結果を用いて番号 16 から 19 までに対応する素性を作成した. 次に, 修辞構造理論 (Rhetorical Structure Theory: RST)[6] に基づく談話構造解析の結果を素性の作成に利用した. 談話構造解析では, 特定の文との関係を推定するだけでなく, 文書全体を依存関係に基づく木構造 (DEP-DT[3]) として表現することが可能であり, より多くの手がかりを得ることが期待できる. 作成した素性は, 番号 20 から 25 ままでが対応する.

## 4 評価実験

### 4.1 データセットの作成

実験に用いるデータセットをクラウドソーシングサービス Lancers<sup>5)</sup> を利用して作成した. アノテーションは文抽出による要約手法を用いるため, 各レビューにおいて著者の主張の強い順に 3 文選択してもらう形式で行った. この際, データの信頼性を保つため, 選択してはならない文を追加した設問を 5 問用意し, その正解率が高い回答者の回答だけを利用した.

データの作成に用いたレビューとして, Amazon<sup>6)</sup> における 5 つの異なるドメイン (「ノートパソコン」「デジタルカメラ」「掃除機」「携帯音楽プレーヤー」「スマートフォン」) のレビューを「ノートパソコン」のみ 480 件, 残りの 4 ドメインについては 500 件のレ

<sup>5)</sup><http://www.lancers.jp/>

<sup>6)</sup><http://www.amazon.co.jp/>

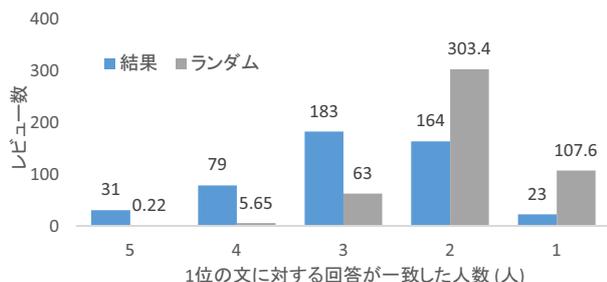


図 1: 5 人の回答の 1 位の最大一致数によるレビューの分類  
ビューを使用した。これらのレビューからは、前処理によって文字数 (200 文字以上 1000 文字以下) や文数 (5 文以上 20 文以下) などに関する条件を満たさないものは取り除かれている。

ノートパソコンドメインの各レビューに対して、5 人中最も多くの回答者が 1 位に選択した文の一致した回答者の人数の分布を図 1 に示す。縦軸はレビュー数を表し、横軸は回答者の最大一致数を表している。3 人以上が一致したレビューの割合は 61%であったが、ランダムに選択した期待値よりはるかに高い一致率であることが確認できる。

## 4.2 実験設定

作成したデータセットを用いて、レビュー中の各文を著者の主張が強い順にランキングする実験を行った。学習器はランキング SVM を使い、実装は LIBLINEAR (Large-scale linear rankSVM<sup>7)</sup>) を利用した。各文の正解のスコアはアノテーションされた順位に応じて、1 位に 4 点、2 位に 2 点、3 位に 1 点を付与し、それを 5 人分合計したものを利用した。実験は用いるデータの種類に基づいて、以下の 3 つの設定で行った。

- (i) 各ドメイン: 訓練, パラメータ調整, 評価に同一ドメインのレビューを用いる
- (ii) 全ドメイン: 訓練, パラメータ調整, 評価に全 5 ドメインのレビューを全て混合したものを用いる
- (iii) クロスドメイン: 訓練, パラメータ調整は 4 ドメインを混合したもの, 評価は残り 1 ドメインのレビューを用いる

実験 (ii) では、各ドメインから 100 ずつランダムに抽出したレビューを用いる場合と、各ドメインのレビュー全てを用いる場合の 2 種類の実験を行った。また、実験 (iii) で訓練に用いるレビューは各ドメインからランダムに 110 ずつ抽出したものを利用した。評価指標は、スコアが最大の 1 文を正解とした場合の recall@1, スコアが高い順に上位 3 文を正解とした場合の recall@3, Mean Average Precision (MAP) @3, そして、上位を重視する評価指標である Mean normalized Discounted Cumulative Gain (nDCG) @N (N: レビューの文数) [4] の 4 種類を用いた。Mean nDCG@N

<sup>7)</sup><https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/>

の計算に用いる文のスコアは、アノテーションによる正解スコアの 5 人の平均値を使用した。各実験では 10 分割交差検定を行い正則化パラメータ ( $C$ ) の調整、及び、素性選択を行った。具体的には、実験 (i), (ii) では 10 分割した内 7 つを訓練データとして、1 つをパラメータ調整用、1 つを素性選択用の開発データとして利用した。また実験 (iii) では、訓練データ内で 10 分割交差検定を行ってパラメータを調整した後、訓練に用いたドメインのレビューを各ドメイン 125 ずつ用意し、素性選択用の開発データとして利用した。パラメータ調整に用いる評価指標には、Mean nDCG@N を利用した。実験は基本素性に対して提案素性を 1 つずつ加えて評価する形式で行った上で、開発データで有効に働いた素性の内、評価値が高かった上位 10 個を使用した実験を行った (base+top10)<sup>8)</sup>。また、導入した素性の内、余弦類似度などの実数値は全てバイナリ値に置き換えた。

一方、素性の作成に使用する文の極性と属性の推定にはロジスティック回帰を用い、こちらも実装は LIBLINEAR を利用した。訓練データには、極性に関しては Kaji ら [5] による ACP コーパス<sup>9)</sup>、属性に関しては価格.com<sup>10)</sup> に投稿されたレビュー (ドメインに対応した属性ごとに記述されたもの<sup>11)</sup>) を利用し、それぞれ 10 分割交差検定によって正則化パラメータ ( $C$ ) の調整を行った。また、素性には単語ユニグラム (基本形) を利用した。なお、談話構造解析には、日本語談話依存関係 [10] を訓練データとして、Heilman らの手法 [2]<sup>12)</sup> を参考にして開発された談話構造解析器を利用した。

## 5 結果

実験 (i), (iii) の結果を表 2 に、実験 (ii) の結果を表 3 に示す。ここで、表中の太字は基本素性のみ結果を上回っていることを示し、<sup>†</sup> は並べ替え検定 (permutation test) によって統計的に有意であることを示す<sup>14)</sup>。これらの結果から、ほぼ全ての実験で提案素性が有効に働いていることがわかる。特に、全ドメイン混合で全データを使用した場合に最も良い結果を示しており、全ての評価指標で基本素性のみの場合を上回っている。また、表 3 より、素性設計やドメインを考慮するより

<sup>8)</sup> 全ての素性を使用した場合、いずれの実験においても基本素性のみより低い性能となったため、素性選択を行った。

<sup>9)</sup><http://www.tkl.iis.u-tokyo.ac.jp/~kaji/acp/>

<sup>10)</sup><http://kakaku.com/>

<sup>11)</sup> 価格.com ではレビューの投稿時にこのような属性ごとに記述するフォーマットが選択できる。例えば、ノートパソコンドメインでは「デザイン」「処理速度」「グラフィック性能」「拡張性」「使いやすさ」「携帯性」「バッテリー」「液晶」の 8 種類の属性が存在する。

<sup>12)</sup><https://github.com/mheilman/discourse-parsing>

<sup>13)</sup> 各ドメイン、デジタルカメラのみ最大有効素性数に基づき top6

<sup>14)</sup> 有意水準は 0.05 とした。

表 2: 各ドメイン (In), クロスドメイン (Cross) の実験結果 (Mean nDCG@N)

	ノートパソコン		デジタルカメラ		掃除機		音楽プレイヤー		スマートフォン	
	In	Cross	In	Cross	In	Cross	In	Cross	In	Cross
基本素性 (base)	0.7803	0.7804	0.7788	0.7772	0.7826	0.7674	0.7695	0.7695	0.7873	0.7814
base+top10 <sup>13)</sup>	<b>0.7940</b> <sup>†</sup>	<b>0.7896</b> <sup>†</sup>	<b>0.7814</b>	<b>0.7804</b>	<b>0.7830</b>	<b>0.7729</b>	<b>0.7761</b> <sup>†</sup>	<b>0.7806</b> <sup>†</sup>	0.7822	<b>0.7842</b>

表 3: 全ドメインの実験結果

	各ドメインランダムに 100 ずつ抽出				全て使用			
	recall@1	recall@3	MAP@3	MnDCG@N	recall@1	recall@3	MAP@3	MnDCG@N
基本素性 (base)	0.3840	0.5886	0.7073	0.7844	0.3935	0.6299	0.7400	0.8053
base+top10	0.3760	<b>0.5993</b>	<b>0.7171</b>	<b>0.7868</b>	<b>0.4048</b>	<b>0.6325</b>	<b>0.7461</b>	<b>0.8114</b> <sup>†</sup>

表 4: 各ドメインの実験で有効に働いた素性 (上位 10 個)

順位	ノートパソコン	デジタルカメラ	掃除機	音楽プレイヤー	スマートフォン
1	A <sub>1</sub> prob	A <sub>1</sub> prob	DMatchP	A <sub>1</sub> prob	A <sub>1</sub> rank
2	TcosA <sub>1</sub> rank	RSTdepth3	A <sub>1</sub> rank3	DcosA <sub>1</sub>	DcosA <sub>2</sub>
3	RSTchild	A <sub>1</sub> probMaxrank1	DcosA <sub>1</sub> rank3	A <sub>2</sub> probMax	A <sub>1</sub> rank1
4	PprobMaxDiff2	A <sub>1</sub>		A <sub>1</sub>	Prank3
5	PprobMaxDiff	RSTchild		A <sub>2</sub> probMaxrank1	DcosPrank
6	TcosA <sub>1</sub>	DcosA <sub>2</sub>		PprobMaxPA <sub>1</sub>	A <sub>1</sub> prob
7	PprobMaxrank3	A <sub>2</sub> probMaxrank1		PprobMaxDiff	RSTdepthN
8	TwordsCov	DcosA <sub>1</sub>		Pratio	TwordsTPprobMax
9	Prank3	RSTparentDiff		DcosA <sub>1</sub> rank1	TcosA <sub>1</sub> rank3
10	A <sub>1</sub> probMaxrank3	A <sub>2</sub> rank1		RSTdepth10	RSTparentDiff

もデータを増やすことが重要であることがわかる。このことは、表 2 に示すように、同ドメインで訓練した場合と、クロスドメインで訓練した場合で、結果に大きな差がないことから確認できる。

表 4 に、各実験において有効に働いた素性上位 10 個を載せる。ここでは、特定のドメインにのみ有効に働く素性と多くのドメインで有効に働く素性があることがわかる。特に属性 A<sub>1</sub> (「書き出し部」「各属性に関する記述」「総評」) に関する素性は、全体的に強く効いていることがわかる。

問題点として、評価値が全体的に高くないことが挙げられる。これは、レビューの要約において本質的な課題を残している可能性を示唆している。要因の 1 つとしては、素性作成に用いた極性等の分類器の精度の問題が考えられる。本稿における実験では、比較的少ない種類の素性を用い、分類対象のデータとは異なる種類のデータで訓練した分類器を使用した。想定以上に低い分類精度であった可能性が考えられる。また、各レビューの内容のばらつきに対して、データ量が小さかったことも要因の 1 つとして考えられる。

## 6 まとめと今後の課題

本研究では、レビューを対象とした単一文書要約を目的とし、レビュー中の文のランキングを行った。実験の結果、基本素性にレビュー特有の手がかりを用いた素性を新たに加えることで、ランキング精度が向上することを示した。今後は、ランキング精度の向上を図ると同時に、提案したランキングモデルが実際にユーザの役に立つかを検証するために、出力した要約に対して人手による評価を行いたいと考えている。

謝辞 本研究は JSPS 科研費 26280080 の助成を受けた。

## 参考文献

- [1] P. Beineke, T. Hastie, C. Manning, and S. Vaithyanathan. Exploring sentiment summarization. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 2004.
- [2] M. Heilman and K. Sagae. Fast rhetorical structure theory discourse parsing. *arXiv preprint arXiv:1505.02425*, 2015.
- [3] T. Hirao, Y. Yoshida, M. Nishino, N. Yasuda, and M. Nagata. Single-document summarization as a tree knapsack problem. In *EMNLP*, pp. 1515–1520, 2013.
- [4] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM (TOIS)*, Vol. 20, No. 4, pp. 422–446, 2002.
- [5] N. Kaji and M. Kitsuregawa. Automatic construction of polarity-tagged corpus from html documents. In *COLING*, pp. 452–459, 2006.
- [6] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, Vol. 8, No. 3, pp. 243–281, 1988.
- [7] Y. Mao and G. Lebanon. Isotonic conditional random fields and local sentiment flow. In *NIPS*, pp. 961–968, 2006.
- [8] H. Nishikawa, K. Arita, K. Tanaka, T. Hirao, T. Makino, and Y. Matsuo. Learning to generate coherent summary with discriminative hidden semi-markov model. In *COLING*, pp. 1648–1659, 2014.
- [9] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL*, pp. 271–278, 2004.
- [10] 横山憲司, 難波英嗣, 奥村学. Support vector machine を用いた談話構造解析. 情報処理学会自然言語処理研究会報告, NL-155, pp. 193–200, 2003.