6V-08

旅館レビューのカテゴリ分類のための重文自動分割

髙師 遥* 阿部 遥佳* 鈴木 信太郎* 大和矢 悠仁* 延澤 志保*
*東京都市大学知識工学部 *東京都市大学情報工学部

1 研究背景

楽天トラベル[1][1] などの旅館予約サイトでは，利用者がレーティング評価やクチコミと呼ばれる自由レビュー文を投稿するシステムがある．旅行サイト TripAdvisor は 2019 年にクチコミを参考にするという人が 7 割以上であるというデータを発表している[2]．しかし，総合評価などのレーティングの数値を見るだけでは具体的な情報が手に入らないこと，投稿されたレビュー文では何について書かれているか内容がまとまっていないこと，レビュー数は膨大でありすべてを参照することが困難であることを課題として，レビューをより活用するためにさまざまな研究が行われている．安倍らは，同じ項目の評価でもユーザによってニュアンスが違うことに着目をして，ユーザが重視している観点の評価が高いホテルを推薦するために，詳細化された評価項目にスコア付けし推薦する手法を提案し，評価している内容はユーザによって違い，膨大なレビューを読まなければ本当に高いのか判断が難しいことが課題とした [2]．伊草らは，宿泊レビューの注目すべき点を可視化し，適切な返信例を作成することを目的とし，レビューについてユーザ評価項目ごとの分類と極性判定を行い，可視化モデルの構築を行っていた [3]．

2 レビュー文中の重文の分割

2.1 提案手法の概要

本研究は宿泊施設ごとのレビュー文をカテゴリごとに箇条書きで出力することでレビューの評判情報を簡潔にまとめることを目的とする．

レビュー文では「景色も綺麗で、干物や舟盛りもとても美味しくて本当に満足です。」のように 1 文の中に立地や食事など複数のカテゴリの内容が記述されている場合があり，レビュー文のカテゴリ分類にはこういったレビュー文の分割処理が必要となる [4, 5]．そこで本研究では，レビュー文の構造に着目し，係り受け関係を基に重文の自動分割を実現する手法を提案する．

2.2 文の分割

レビューには複数のカテゴリの内容を含む文が存在するため，レビュー文をカテゴリごとに分けるにはレビュー文の分割が必要となる．そのため本研究では，係り受けを考慮し，重文の分割を行う．

本研究の重文分割の目的はカテゴリ分類であるため，カテゴリを示す語句を含まない文は不要である．しかし重文分割の時点ではカテゴリ分類に有効な各カテゴリの特徴語は明確でないため，重文分割の段階では，特徴語の候補としての重要語を想定し，これを含む文をすべて出力することを目的とする．カテゴリ特徴語にはならない語を重要語として抽出された文は，カテゴリ分類の際にどのカテゴリにも当てはまらず，結果的に除去されることになる．また，複数の重要語を含む文については，重文分割の段階ではどの重要語がカテゴリ特徴語かわからないため，ここでは複数の文候補を出力し，カテゴリ分類の時点で冗長な文を除去する形で絞り込む．本研究では重要語として固有名詞，一般名詞，サ変接続名詞の 3 種類の名詞を用いる．

レビュー文分割の提案手法を図 1 のフローチャート [4] で説明する．文単位で解析を行うため，レビュー群を句点で区切り，これを単位文とする．それぞれの単位文に対して係り受け解析を行い，重要語の係り受け先を文末までたどって 1 文として出力する．これにより，「景色も綺麗で、干物や舟盛りもとても美味しくて本当に満足です。」というレビュー文はそれぞれ「景色」を重要語とする「景色も綺麗で満足です。」と「干物」や「舟盛り」を対象とする「干物や舟盛りもとても美味しくて満足です。」の 2 文に分かれ，それぞれ立地カテゴリと食事カテゴリとに適切に分けることが可能になる．半面，「干物や舟盛り」のような並列関係に対しては，「干物」を重要語としてたどることで得られる「干物や舟盛りも～満足です。」と「舟盛り」を重要語としてこれ以降をたどることで得られる「舟盛りも～満足です。」の 2 文がそれぞれ出力されてしまう問題が残る．

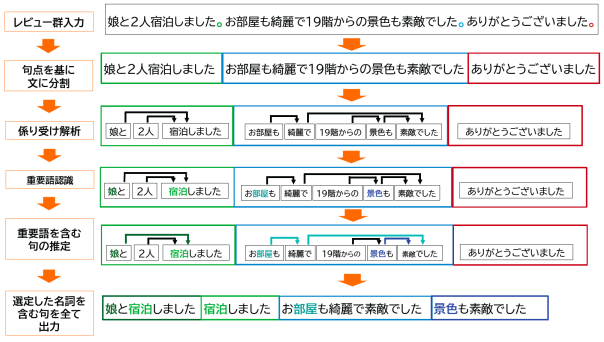

図 1: レビュー文分割の処理の流れ [4]

Compound Sentence Segmentation for Categorization of Accommodation Review Texts.
Haruka Takashi*, Haruka Abe*, Shintaro Suzuki*, Yuji Yamatoya*, and Shiho Hoshi Nobesawa*.
* Faculty of Knowledge Engineering, Tokyo City University
* Faculty of Information Technology, Tokyo City University

[1]楽天トラベル，https://travel.rakuten.co.jp/.
[2]トリップアドバイザー株式会社，「トリップアドバイザー、口コミの影響力に関する調査結果を発表」，https://www.tripadvisor.jp/blog/wp-content/uploads/2019/07/190725_TripAdvisorPressRelease.pdf.

提案手法では重要語を持たない句は切り捨てられるため，宿泊施設のレビューによく見られる「また来ます」や「ありがとうございました」などカテゴリに属さない文の削除も文分割と同時に行うことが可能である (図 1).

3 実験結果

分割の精度を測るため，楽天トラベル 204 レビューを対象として重文分割を行い，人手で作成した正解データと比較を行った．この 204 レビューから作成された正解データは 90 文であった．重文分割の再現率を表 1 に示す．重文分割の結果，90 文の正解データに対応する文は 160 文出力された．これは，例えば「フロントのスタッフの対応がひどい」という正解文が分割されて「フロントのスタッフがひどい」と「スタッフの対応がひどい」の 2 文になって出力されるような場合があり，結果的に出力文数が増えたものである．表 1 に示した成功の 81 文

表 1: 重文自動分割結果 (再現率)

分割レベル		文数	
成功	正解 (過不足なし)	40	
	意図保存 (過不足あり)	41	81
失敗	出力なし (主語述語あり)	27	
	出力なし (主語なし)	37	
	出力なし (述語なし)	6	
	出力なし (主語述語なし)	9	79
	計		160

は，正解 90 文のうち 54 文に対応しており，正解の 60% を正しく取得することができたと言える．ここでの過不足は主に副詞などの消失で，意図を損なうものではない．過不足の発生は主に係り受け解析結果によるものである (図 2)[4]．提案手法は重要語から係り受けをたどるため，

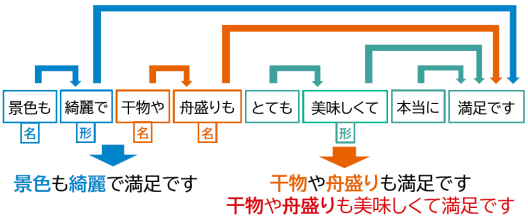

図 2: 重要文分割失敗例 [4]

重要語から係らない語句は削除される．抽出に失敗した文の中で，「6 時半のバスに乗ると朝食が食べれない」などの文の「6 時半」の部分が抜けてしまうことや，「この値段」などの連体詞を使用した文も「この」や「あの」などは出力されないため，文によっては意図が不明になる．

提案手法では，重要語の係り受けの有無を抽出の判断基準としている．そのため表 1 の失敗事例のうち「主語述語あり」の 27 文以外は提案手法では抽出できない文である．これは「また来ます」や「ありがとうございました」といった名詞が含まれない文をカテゴリ分類に不要であるとして消去するためだが，これにより「美味しいです」などの主語は無いが食事の項目に分類されるであろう文も削除対象となってしまった．提案手法では重要語として固有名詞，一般名詞，サ変接続名詞を設定したが，90 文中 36 文の抽出に失敗したことから，重要語についてはさらに検討が必要である．

次に，重文分割の適合率を表 2 に示す．表 2 に示すと

表 2: 重文自動分割結果 (適合率)

分割レベル		文数	
適切	正解	40	
	意図保存 (過不足あり)	41	81
不適	不十分	82	
	意味不明	129	211
	カテゴリ外	106	106
	計		398

おり，出力すべき文は 90 文だったにも関わらず，実際の出力は 398 文に上る，これは，重要語となり得る語句を含む文を全て出力するアルゴリズム (2.2 節) で同じ文に対して複数の重要語による抽出を許しているためである．

出力された文に過不足が多くなる原因として，レビュー文の大半が口語で書かれているために係り受け解析が正しく行われていないことがあげられる．レビューは形式が決まっていないため，句点のない文や箇条書き，顔文字の利用も多く見られる．出力された 398 文中 4 文に箇条書き記号による単位文の分割の誤りが見られた．

4 まとめ

トップページで表示される評価の数字だけでは具体的な情報が得られないこと，レビュー文ではどこに何が書いてあるのかわかりにくいこと，レビュー数が膨大でありすべて読むことが困難であることを課題として，本研究ではレビュー文を句に分割し評価視点の各カテゴリごとに箇条書きすることを目的に文の分割を行った．

名詞を認識し係り受けを文末までたどることでカテゴリ分類に不要な文の削除を行えた．しかし，レビューの大半が口語であるため，係り受けが正しく行えず過不足が多く見られたが，結果として精度は約 60 %が正しく分割されている文として出力された．

形式が決まっていないレビューだからこそ，句点が無い文，箇条書きで用いられる記号，顔文字などはレビュー解析をする上で今後の課題になると考えられる．

謝辞

本研究では，国立情報学研究所 IDR データセット提供サービスを通して楽天グループ株式会社様よりご提供いただいた「楽天データセット (https://rit.rakuten.com/data_release/)」を利用しました．心より感謝いたします．

参考文献

[1] 楽天グループ株式会社, 楽天データセット, 国立情報学研究所情報学研究データリポジトリ, https://doi.org/10.32130/idr.2.0, 2014.
[2] 安部 克, 中島 伸介,“レビュー自動スコアリング方式に基づくホテル推薦システム,” データ工学と情報マネジメントに関するフォーラム論文集, P1-28, p.093, 2020.
[3] 伊草 久峻, 鳥海 不二夫,“宿泊予約サイトにおけるレビュー自動分類,” 2020 年度人工知能学会全国大会, pp.1–4, 2020.
[4] 髙師 遥, 阿部 遥佳, 鈴木 信太郎, 大和矢 悠仁, 延澤 志保, “レビュー文のカテゴリ分類のための重文自動分割,” NII-IDR ユーザフォーラム, no.P13, 2021.
[5] 髙師 遥, 阿部 遥佳, 蒲生 奏衣, 高貴 達之, 延澤 志保, “旅館予約サイトを対象としたレビュー文の自動カテゴリ分類,” NLP 若手の会第 16 回シンポジウム (YANS2021), no.P1-17, 2021.