複数の属性に対する評価を含む宿泊施設レビューに対する多様な返信の自動生成

村越 裕太　　白井 清昭
北陸先端科学技術大学院大学 先端科学技術研究科
{yuta.murakoshi,kshirai}@jaist.ac.jp

概要

本研究は，宿泊施設に関する低評価レビューに対し，定型的で当り障りのない表現を用いずに，レビューワーが苦情を述べている全ての属性について言及した返信を生成することを目的とする．レビュー中の全ての属性に言及するために，レビューを文に分割し，それぞれの文に対して系列変換モデルで返信を生成した後，それらを統合する．系列変換モデルを学習する際には，属性に言及していない返信，定型的な表現が含まれる返信を訓練データからあらかじめ削除する．実験の結果，特に複数の属性を含むレビューに対し，提案手法によって自動生成された返信の品質が向上することを確認した．

1　はじめに

宿泊施設のオンライン予約サイトの中には，ユーザがレビューを書き，宿泊施設がそれに対して返信できるサイトも存在する．宿泊施設にとって，大量のレビューに対して返信することの負担は大きいが，低い評価のレビューに対して返信を返さないとユーザの不満が解消されず，宿泊施設の評判を落とすことにつながる．そのため，ユーザのレビューに対する返信を自動生成する技術が求められる．

本研究は，宿泊施設に関する低評価レビューに対し，それに対する適切な返信を生成することを目的とする．この際，以下の 2 つの点に留意する．ひとつは多様な返信の生成である．生成モデルは当たり障りのない一般的な表現を生成する傾向がある [1] が，「申し訳ありません」「ご迷惑をおかけしました」など謝罪を表す定型的な表現を生成するだけでは，ユーザは宿泊施設が事務的な対応をしているという印象を持ち，ユーザの不満が解消されない可能性がある．もう一つは，ユーザが宿泊施設に関する複数の属性に対して不満を表明しているとき，その全ての属性に言及することである．例えば，ユーザが「部屋の清掃」「フロントの対応」の 2 つの点について苦情を述べているのに，それらに対して言及しない，あるいはどちらか一方にしか言及しない場合，ユーザに不誠実な印象を与える．上記の目的を達成するため，ユーザレビューを入力，返信を出力とする系列変換モデルを学習するが，定型的な表現の抑制やレビュー中の属性に対する網羅的な言及を実現するための手法を探究する．

2　関連研究

アプリのレビューやホテルのレビューに対する返信を自動生成する先行研究について述べる．Gao らは，アプリケーションのユーザレビューを入力，それに対する返信を出力とする RNN による Encoder-Decoder モデルに，カテゴリー，レビューの長さ，ユーザー評価，感情スコアの 4 つの情報を attention 機構を用いて組み込む手法を提案した [2]．Kew と Volk は，レビューと返信の組からなるデータセットから返信生成モデルを学習すると，多くのレビューに対して用いられる汎用的な表現を含む返信が生成されやすいという問題に対処するため，訓練データにおける返信の汎用性のスコアを算出し，そのスコアが閾値以上の返信を訓練データから除外する手法を提案した [3]．日本語で書かれたレビューに対して返信を自動生成する研究として，伊草と鳥海は，宿泊予約サイトに書き込まれたレビューに対して適切な返信を自動生成するために，ユーザーによる評価値と返信の長さの情報を組み込んだ RNN による Encoder-Decoder モデルを提案した [4]．

先行研究では，ユーザレビューにおける複数の評価対象の属性に網羅的に言及することは留意されていない．本研究では，宿泊施設の低評価レビューについてはユーザの不満の全てに言及することが重要と考え，それを実現する方法を探究する．また，

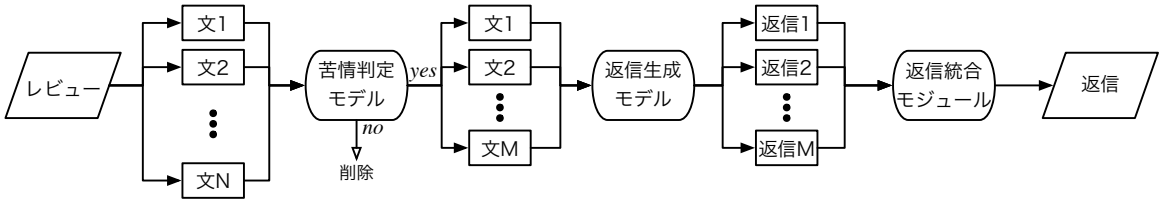

図 1　提案手法の概要

Kew と Volk の手法 [3] を参考に，当り障りのない表現の生成を抑制することにも取り組む．

3　提案手法

3.1　概要

提案手法の概要を図 1 に示す．まず，句点を文境界として，レビューを文に分割する．次に，苦情判定モデルを用いて，個々の文が苦情か否かを判定し，苦情と判定されなかった文を削除する (3.2 項)．苦情を含むレビュー文に対し，系列変換モデルを用いて宿泊施設の返信を生成する (3.3 項)．最後に，生成された返信文を統合して，最終的な返信を生成する (3.4 項)．

レビュー全体を入力として返信を生成すると，複数の苦情を含むレビューに対してはその全てが言及されない可能性がある．そのため，複数の属性は異なる文に現れると仮定し，それぞれの文から属性に対する言及を含む返信を生成する．これにより，レビュー内の複数の属性 (苦情) に対して網羅的に返信することを狙う．

返信生成モデルならびに苦情判定モデルの学習には楽天データセット [5] における楽天トラベルのデータ (以下，「楽天トラベルデータセット」と呼ぶ) を用いる．同データセットは宿泊施設に対するユーザレビューとそれに対する宿泊施設の返信を含む．また，ユーザレビューには「苦情」「感想・情報」などのラベルが付与されている．

3.2　苦情判定

本研究では苦情に対する返信を生成することを主たる目的としているため，レビューから苦情を含まない文をあらかじめ削除する．このため，文がユーザの苦情を含むか否かを判定する二値分類器を学習する．分類モデルとして Bidirectional Encoder Representations from Transformers(BERT)[6] を用いる．具体的には，東北大学が公開している BERT base Japanese [7] をファインチューニングする．訓練データとして，楽天トラベルデータセットにおける「苦情」ラベルが付与されたレビューを正例，それ以外のレビューを負例として用いる．正例と負例の数は同数とする．ただし，提案手法における苦情判定の対象はレビューではなく文であるため，楽天トラベルデータセットにおけるレビューのうち 1 つの文から構成されるレビューのみを訓練データとして用いる．

3.3　応答文生成モデル

苦情を含むと判定されたレビュー文に対し，それに対する宿泊施設の返信を生成する．このため，レビュー文を入力，宿泊施設の返信を出力とする系列変換モデルを学習する．系列変換モデルとして BART を利用し，日本語事前学習済みモデル [8] をファインチューニングする．ファインチューニングのための訓練データとして，楽天トラベルデータにおける「苦情」のラベルが付与されたレビューとそれに対する返信の組を用いる．ただし，1 節で述べた我々が望ましいと考える返信を生成するため，訓練データに対して 2 種類のフィルタリングを行う．

3.3.1　属性に言及しない応答のフィルタリング

ユーザがレビュー上で述べている宿泊施設の属性に対する返信を生成するため，属性に言及していない返信を訓練データから除外する．あらかじめ宿泊施設の属性を表す単語 (属性語) の集合 A を定義し，レビューと返信の両方に同じ属性語 a_i $(\in A)$ が出現していない組，ならびに属性語が 1 つも出現していない組を削除する．

属性語は宿泊施設のレビュー集合における特徴的な単語とする．具体的には，式 (1) によって各単語のスコアを算出し，その上位 500 件の単語を属性語集合 A とする．

$$S(w_i) = \mathrm{ave}_{r_j \in TOP_{1000}(w_i)} \text{TF-IDF}(w_i, r_j) \qquad (1)$$

ここで，R は苦情ラベルが付与されたレビューの集合，r_j はそのレビュー，TF-IDF(w_i, r_j) は R を全文書集合としたときの単語 w_i のレビュー r_j におけ

る TF-IDF，$TOP_{1000}(w_i)$ は TF-IDF 値の大きい上位 1000 件のレビューの集合であり，$S(w_i)$ はその 1000 件の TF-IDF の平均値と定義する．抽出された属性語の例を表 1 に示す．

表 1 宿泊施設の属性語の例

駐車 部屋 排水 予約 タバコ シャワー 臭い 風呂 対応 朝食 タオル 掃除 エアコン 温度 ポイント トイレ 髪の毛 バス 禁煙 喫煙 清掃 空調 ルーム プラン カード 換気 匂い カーテン 料理 冷蔵

3.3.2 定型文のフィルタリング

多様な返信を生成するために，すなわち紋切り型の返信が生成されるのを抑制するために，訓練データにおける返信から定型文を除外する．文献 [3] に倣い，宿泊施設の返信を文に分割し，それぞれの文 s_i に対して定型度スコア $C(s_i)$ を算出し，その上位 30%の文を定型文として削除する．この処理の後の訓練データは，レビューと，それに対する元の返信から定型文を除いたテキストの組となる．

定型度スコア $C(s_i)$ は式 (2) のように定義する．

$$C(s_i) = \text{ave}_{tg_{ij} \in s_i} fre(tg_{ij}) \tag{2}$$

ここで，tg_{ij} は文 s_i に出現する単語 tri-gram，fre は訓練データにおけるその出現頻度であり，定型度スコアはその平均と定義する．

3.4 返信文の統合

分割したそれぞれのレビュー文から返信生成モデルによって生成された返信文をマージし，最終的な返信を得る．返信文の順序は，生成元のレビュー文のレビューにおける出現順序と同じとする．ただし，返信文は独立に生成しているため，類似した文や表現が重複して返信に含まれる可能性がある．重複する表現を除外するため，2 つの返信文間の距離を正規化された編集距離 [9](編集距離を 2 文の長さの和で割った値) で測り，それが 0.1 以下の場合，元のレビューにおける出現順序が後である返信文を残し，もう一方の返信文を除外する．ただし，属性語を含む返信は常に除外しないものとする．

4 評価

4.1 苦情判定モデルの評価

楽天トラベルデータから，苦情ラベルが付与されているレビューとそうでないレビューを同じ数だけ取得し，訓練データ (約 32,000 件) とテストデータ (約 8,000 件) を作成した．これらのデータを用いて苦情判定モデルを評価したところ，苦情クラス検出の F 値は 0.890，苦情か否かの二値分類の正解率は 0.887 となり，十分に高いことを確認した．

4.2 返信生成モデルの評価

4.2.1 実験データ・実験条件

返信生成モデルの学習ならびに評価に用いたデータの統計を表 2 に示す．楽天トラベルデータにおけるレビューと返信の組のうち，90%を訓練データ，5%を開発データ，5%をテストデータとした．開発データは研究の初期段階で訓練データのフィルタリング手法を検討するために，テストデータは自動評価 (結果は付録 B で報告する) のために用いた．3.3.1 と 3.3.2 で述べたフィルタリング処理により，訓練データのデータ数は約 29%減少した．

表 2 返信生成モデルの実験データ

訓練データ	147,749
訓練データ (フィルタリング後)	105,241
開発データ	8,209
テストデータ	8,209

返信生成モデルに使用した BART をファインチューニングする際のハイパーパラメタとして，最大エポック数は 5，学習率は $3e^{-5}$，ドロップアウト率は $p = 0.3$ と設定した．

4.2.2 返信生成モデルの評価

提案手法によって生成された返信生成モデルを人手により評価する．ここでは以下の 5 つの手法を比較する．また，返信生成タスクの上限としてデータセットの返信も評価する．

BASELINE ベースライン．訓練データに対するフィルタリングは行わない．図 1 に示したようにレビューを文に分割しそれぞれの文から生成された返信を統合する処理は行う．

PRO-A-S 訓練データに対して属性に言及しない応答のフィルタリング (3.3.1) を行う．文分割と返信文の統合処理も行う．

PRO-C-S 訓練データに対して定型文のフィルタリング (3.3.2) を行う．文分割と返信文の統合処理も行う．

PRO-AC レビューを文に分割してから返信を生成するのではなく，レビュー全体を入力として返信を生成する．訓練データに対する上記 2 つのフィルタリング処理も行う．

PRO-AC-S　訓練データに対して 2 つのフィルタリング処理を行い，文分割と返信文の統合処理も行う．

GOLD　データセットにおいて宿泊施設が実際に書いた返信

表 2 に示したテストデータからランダムに 50 件のレビューを選択し，上記の 5 つの手法で生成された返信ならびに GOLD を人手で評価する．評価者は著者 2 名を含む日本語母語話者 7 名である．評価項目を以下に述べる．

流暢性　返信が自然な日本語であるかを 5 段階で評価する．

非冗長性　返信に同じような表現が繰り返されていないかを 5 段階で評価する．表現の繰り返しが多いほど低い評点を与える．

総合評価　苦情を書いたレビュワーの立場から見て，宿泊施設からの返信として適切であるかどうかを 5 段階で評価する．

属性への言及　レビューでユーザが苦情を述べている属性のそれぞれについて，返信でそれについて触れているか否かを判定する．属性はあらかじめ被験者以外の人が抽出しておく．

実験結果を表 3 に示す．評価値は 7 名の被験者による評点の平均である．属性言及率とは，評価対象の 50 件のレビューに出現する属性のうち，返信内で言及されたものの割合である．アスタリスク (*) は t 検定によって BASELINE との有意差 ($p < 0.01$) が確認できたことを示す．

表 3　返信生成モデルの人手評価の結果

	流暢性	非冗長性	総合評価	属性言及率
BASELINE	4.58	4.44	2.53	0.338
PRO-A-S	4.34*	4.16*	2.72*	0.581*
PRO-C-S	4.63	4.50	2.80*	0.415*
PRO-AC	4.66	4.71*	2.98*	0.450*
PRO-AC-S	4.48	4.17*	2.77*	0.538*
GOLD	4.63	4.84	3.99	0.652

BASELINE と比べて，属性に言及しない応答のフィルタリングを行う提案手法 (PRO-A-S，PRO-AC-S) では属性言及率が高い．ユーザが苦情を述べている属性に対して何らかの返信をするという本研究の目的がある程度達成できている．一方，流暢性と非冗長性のスコアは BASELINE と比べて低くなっている．属性に対する言及が増えたことにより，同じような表現の繰り返しが増え，流暢性も損われたと考えられる．提案手法では，個々のレビュー文に対する返信文を統合する際に類似した返信文を除外する処理をしているが，属性を含む文は除外しないことにしているため，似ている表現を完全に排除できていない．属性言及率と流暢性・非冗長性はトレードオフの関係にあると言える．

定型文のフィルタリングに着目すると，PRO-C-S は BASELINE と比べて非冗長性が改善され，流暢性や総合評価も高い．紋切り型の表現の生成が抑制されていることが確認できた．

PRO-AC と PRO-AC-S を比較すると，属性言及率は PRO-AC-S の方が高いが，流暢性・非冗長性の指標は PRO-AC の方が高い．文ごとに返信を生成し，最終的にそれを統合する方法では，レビュー中の属性に対して漏れなく言及することができるが，レビュー全体を一括して処理する手法と比べて文の自然さが損われたり類似表現の繰り返しが生じているためである．総合評価では PRO-AC の方が高いことから，文ごとに返信を生成する方式の有効性は認められない．ここで，属性が 2 つ以上存在する 22 件のレビューのみを対象にした評価結果を表 4 に示す．PRO-AC の属性言及率は BASELINE よりも悪く，総合評価も PRO-AC-S と比べて低い．属性が 1 つしかないレビューに対しては，文ごとに返信を生成しても属性言及率を上げる効果は少ないが，複数の属性を含むレビューについては，レビューを文に分割してから返信を生成する提案手法が有効に働くと言える．

表 4　複数の属性を含むレビューに対する返信生成モデルの人手評価の結果

	流暢性	非冗長性	総合評価	属性言及率
BASELINE	4.54	4.27	2.48	0.296
PRO-A-S	4.27*	3.99*	2.58	0.473*
PRO-C-S	4.54	4.30	2.81*	0.329*
PRO-AC	4.60	4.73*	2.61	0.232
PRO-AC-S	4.50	3.97*	2.73*	0.466*
GOLD	4.56	4.82	3.94	0.596

提案手法による返信の生成例を付録 A に示す．

5　おわりに

本研究では，複数の属性に対して苦情を述べたレビューに対する宿泊施設の返信を自動生成する手法を提案した．今後の課題として，返信文を統合する処理の改善が挙げられる．例えば，抽象型要約生成モデルによって複数の返信文から要約を作成することにより，より自然で同じような表現の繰り返しが少ない返信を生成する方法が考えられる．

参考文献

[1] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. **International Conference on Learning Representations**, pp. 1275–1312, 2020.

[2] Cuiyun Gao, Jichuan Zeng, Xin Xia, David Lo, Michael R. Lyu, and Irwin King. Automating app review response generation. In **Proceedings of the 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)**, pp. 163–175, 2019.

[3] Tannon Kew and Martin Volk. Improving specificity in review response generation with data-driven data filtering. In **Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)**, pp. 121–133, 2019.

[4] 伊草久峻, 鳥海不二夫. レビュー特性を用いたレビュー返信の自動生成. 人工知能学会全国大会 (第 35 回) 論文集, pp. 2F3–GS–10g–01, 2021.

[5] 情報学研究データリポジトリ 楽天データセット, (2023-12 閲覧). https://www.nii.ac.jp/dsc/idr/rakuten/.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4171–4186, 2019.

[7] BERT base Japanese (IPA dictionary, whole word masking enabled) – Hugging Face, (2023-12 閲覧). https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking/.

[8] BART 日本語 Pretrained モデル – LANGUAGE MEDIA PROCESSING LAB, (2023-12 閲覧). https://nlp.ist.i.kyoto-u.ac.jp/?日本語 Pretrained モデル.

[9] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. **In Soviet physics doklady**, Vol. 10, pp. 707–710, 1966.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[11] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 110–119, 2016.

表 5　生成された返信の例

レビュー	返信		
	BASELINE	PRO-AC-S	GOLD
まず最初に、全室禁煙なら大々的に謳って欲しい。 サイトのどこかには記載があるだろうけど、ざっくり読む分には見つけきらない。 係りの方は雰囲気の良い方で、満足です。 客室のデザインは良いが、お風呂がない。 貸切風呂（一番左奥）は温度調整できず真夏に熱い風呂と罰ゲームでした。 シャワーが弱すぎる。最終的には湯船の熱いお湯を汲み利用しました。 大浴場は普通で良かった。 食事は素材の味を生かし、的な内容でした。悪く言えばそんなに手をかけていない。 っと思いきや、唯一一品だけ、大根のレモン煮（？）抜群に美味かった。 これだけは、感動しました。	貸切風呂の温度につきましては、ご迷惑をお掛け致しまして申し訳ございませんでした。お客様に快適にお過ごし頂けるよう、スタッフ一同努力して参ります。この度は当ホテルをご利用頂きまして誠にありがとうございます。	禁煙ルームのご案内につきましては、ご不便をお掛け致しました事、心よりお詫び申し上げます。ご指摘いただきましたサイトの件でございますが、お客様にご迷惑をお掛け致しましたことを深くお詫び申し上げます。貸切風呂の温度調整につきましては、お客様にご迷惑をお掛け致しました事を心よりお詫び申し上げます。シャワーの件では大変申し訳ございませんでした。今後このような事がないよう、スタッフ一同精進して参ります。	湯の宿　入船でございます。 先日はご宿泊頂きまして、誠にありがとうございました。 ご指摘頂きました箇所に関しましては、出来る箇所から順次、改善してまいりたいと思います。 今後とも、入船を宜しくお願い致します。

A　生成された返信の例

あるレビューに対する BASELINE，PRO-AC-S，GOLD の返信を表 5 に示す．このレビューでユーザが不満を表しているのは「禁煙室の案内」「風呂の温度」「シャワーの強さ」の 3 点である．BASELINE では「風呂の温度」に対してしか謝罪していないのに対し，PRO-AC-S では 3 つの属性全てに言及して謝罪している．

B　返信生成モデルの自動評価

訓練データのフィルタリングの効果を自動的に評価する．自動評価尺度として BLEU[10] と DISTINCT[11] を用いる．BLEU は，データセットにおける宿泊施設の返信を正解とし，自動生成された返信が正解とどれだけ似ているかを評価する．DISTINCT は，評価データに対して生成された返信テキストの多様性を評価する．いずれも単語 3-gram を基にした指標 (BLEU-3 と DISTINCT-3) を用いる．

実験結果を表 6 に示す．ここでは文毎に返信を生成して統合するのではなく，レビュー全体をモデルの入力としている．フィルタリング処理をした方が DISTINCT が高くなっていることから，定型文のフィルタリングにより，ありきたりな文の生成が抑制され，様々な表現の文が生成されるようになったと言える．一方，BLEU はフィルタリング処理をしない方が高い．これは，フィルタリング処理をしないデータから学習されたモデルでは定型文が生成されることが多いが，評価データにおける正解の返信にも定型文が多く，両者で単語 n-gram が一致することが多いためと考えられる．

表 6　返信生成モデルの自動評価結果

手法	BLEU-3	DISTINCT-3
フィルタリングなし	0.1576	0.0147
フィルタリングあり	0.0938	0.0256