E-047

シズルワードの拡張による土産レビュー文抽出の検討
Extraction of Sentences with Reviews of Souvenirs Based on Expansion of Sizzle Words

在間 雅人† 安藤 一秋‡
Masato Zaima Kazuaki Ando

1. はじめに

ブログや SNS などユーザが発信するテキスト上には，現地でしか買えない土産に関する情報が散在している．新型コロナ禍の巣ごもり消費の拡大とともに，オンライン販売を実施する店舗がさらに増加しており，現地でしか買えない，手に入りにくいといった希少価値のある土産の選択・入手がより一層難しくなっている．当研究室では，ユーザが発信するテキストから食品系の土産情報の抽出・整理して，“現地でしか買えない”や“期間限定”といった希少価値を判定し，レビュー情報とともに土産情報を提示するシステム[1][2]の構築を進めている．本稿では，先行研究[1][2]で提案したブログ記事からレビュー文を抽出する手法の性能向上を目指し，シズルワードの拡張による改良手法について検討する．

2. 先行研究

池田らの研究[1]では，Yahoo!ブログ（2019 年サービス終了）を対象に，味に関するレビュー文を抽出する手法を提案している．ロジスティック回帰モデルとシズルワード[3]によるフィルタリングを用いた手法で抽出実験の精度評価をしており，正例に対して F 値が 71.64%，負例に対して F 値が 95.70%という結果が得られている．池田らの研究では，シズルワード辞書を用いて，レビューを含まない文をフィルタリングする手法を提案しているが，未知のシズルワードによる正例の抽出漏れを課題に挙げていた．

また，我々の先行研究[2]では，食べログ[4]を対象に味覚に関するレビュー文を抽出する手法を提案した．まず，Yahoo!ブログはサービスが終了していることから，食べログのレビューに注目して，データセットを新規に構築した．そして，池田らの課題であった未知のシズルワードに対応するため，味に関する表現を人手で収集することで拡張シズルワード辞書を作成し，池田らの素性に加えて，拡張シズルワードの出現回数を追加することで，レビュー文の抽出性能の改善法を検討した．また，分類器には，ロジスティック回帰と Support Vector Machine，Random Forest を採用して比較実験を実施した．実験の結果，Random Forest と拡張シズルワードの出現回数を追加したモデルで，正例に対して F 値が 78.83%，負例に対して F 値が 94.32%という結果が得られた．今後の課題として，拡張シズルワードにも含まれていない未知のシズルワードへの対応と，素材により味を伝えている表現の取り扱いを挙げた．

本稿では，固有表現抽出を用いて新たなシズルワードを自動抽出し，素性に利用することで，レビュー文抽出モデルの抽出性能に与える影響を検証する．

†香川大学 大学院 創発科学研究科，Graduate School of Science for Creative Emergence, Kagawa University
‡香川大学 創造工学部，Faculty of Engineering and Design, Kagawa University

3. データセット

本稿では，以下の手順で構築した新しいデータセットを用いてレビュー文の抽出性能を評価する．食べログの「洋菓子ジャンル」，「和菓子・甘味処ジャンル」に属する任意の 647 店舗に寄せられる口コミを収集し，
タグと句読点，日本語文境界判定器 Bunkai[5]を用いて，テキストを文単位に分割した．そして，商品に対して味や食感について言及しているレビュー文に人手でアノテーションしてデータセットを構築した．本データセットには，訓練データとして正例，負例各 1,317 件，未知データとして正例 327 件，負例 1,601 件のデータが含まれている．

4. シズルワードの自動抽出

本稿では，Conditional Random Fields (CRF)による系列ラベリングを用いて新しいシズルワードを自動抽出し，レビュー文抽出モデルの素性としての有用性を検証する．

4.1 自動抽出用のデータセット

レビュー文抽出用データセットの訓練データである 2,634 文を用いて，シズルワードの自動抽出用データセットを構築する．ipadic-neologd を利用した MeCab を用いて各文を形態素解析し，「滑らか」や「濃厚」といった味に関する表現に人手でタグを付与した．タグ形式には，BIO を採用した．タグ付け例を表 1 に示す．表中の B-SZL タグが開始位置を示し，後続する I-SZL タグを含めた系列が味覚に関する表現である．O タグはその他を意味する．最終的に，2,020 件の味覚表現を含むデータセットを構築した．

表 1．BIO 方式に基づく味に関する表現へのタグ付け例

甘	さ	控え目	で	最高	の	味わい
B-SZL	I-SZL	I-SZL	O	O	O	O

4.2 シズルワードの抽出実験

CRF の実装に CRFsuite[6]を用いる．素性は表記と品詞を，ハイパーパラメータはデフォルト値を用いた．評価指標を適合率，再現率，F 値として 10 分割交差検証で評価する．

実験結果を表 2 に示す．実験の結果，訓練データに含まれる既知の表現に対して，F 値が 72.12%，未知の表現に対して F 値が 12.41%であることがわかった．この結果から，未知の味覚表現に対する抽出性能の向上が必要であることがわかった．しかし，既知の表現に対する適合率が 80.79%であることから，ノイズが少ないモデルが構築できていると仮定し，本稿では，このモデルで抽出した結果をレビュー文抽出の素性に利用して，性能を評価する．

表 2．シズルワード抽出実験の結果

	適合率	再現率	F 値
既知	80.79	65.24	72.12
未知	21.85	8.70	12.41
全体	71.03	49.13	58.04

5. レビュー文抽出の実験

5.1 素性

本実験では以下の 5 素性を利用する．素性①②③④は先行研究[1][2]で提案された素性であり，素性⑤が本稿で提案する素性である．

① SWEM の average-pooling で得られた文ベクトル

株式会社ホットリンクが提供する学習済みモデル[7]を基に名詞，動詞，形容詞，副詞の 200 次元単語ベクトルを利用して，SWEM の average-pooling により 200 次元の文ベクトルを計算する．なお，先行研究[1][2]では，hierarchical-pooling を用いていたが，予備実験により average-pooling を採用することにした．

② 文に含まれる形態素数

③ シズルワードの文内の出現有無

シズルワード[3]のうち，味覚系・食感系のシズルワード 226 個を使用し，出現するか否かを素性とする．

④ 拡張シズルワードの文内での出現回数 or 出現有無

訓練データ内に存在するシズルワードと素材により味を伝えている表現のうち，素性③のシズルワードと重複するものを削除して得られる拡張シズルワード 1,360 個を基に，文内での出現回数または出現有無を素性に利用する．

⑤ B-SZL タグの出現回数 or 出現有無

CRF の抽出性能に課題が残ることから，CRF を用いたシズルワードの自動抽出モデルの出力結果のうち，シズルワードの開始を意味する B-SZL タグの各文内で出現回数または出現有無を素性に利用する．

5.2 評価実験

本実験では，レビュー文抽出における新素性⑤の有用性を検証する．素性④と素性⑤については，出現回数と出現有無を素性とした場合の両方について有用性を検証する．

分類器には，先行研究[2]の実験において，最も抽出性能が高かった Random Forest (RF)を用いる．利用する素性（モデル）を以下に示す．なお，モデル A は先行研究[1]で使われた素性，モデル B は先行研究[2]で使われた素性である．

A) RF+①②③素性
B) RF+①②③④素性（④：出現回数）
C) RF+①②③④素性（④：出現有無）
D) RF+①②③⑤素性（⑤：出現回数）
E) RF+①②③⑤素性（⑤：出現有無）
F) RF+①②③④⑤素性（④：出現回数，⑤：出現回数）

訓練データ対して，グリッドサーチにより各モデルのパラメータの最良値を決定し，未知データに対する適合率，再現率，F 値を評価指標として各モデルの性能を評価する．実験の結果を表 3 に示す．表 3 の太字は最良値を示す．

表 3．レビュー文抽出実験の結果

	正例			負例		
	適合率	再現率	F 値	適合率	再現率	F 値
A	61.96	88.68	72.95	97.46	88.88	92.97
B	68.15	93.57	78.86	98.58	91.06	94.67
C	68.08	92.66	78.49	98.38	**91.13**	94.61
D	62.97	90.51	74.27	97.87	89.13	93.29
E	62.63	90.21	73.93	97.80	89.00	93.19
F	**68.28**	**94.80**	**79.38**	**98.84**	91.00	**94.76**

実験の結果，自動抽出したシズルワードと拡張シズルワードを素性に利用したモデル F が正例，負例ともに F 値が最も高くなることがわかった．先行研究[1]の素性に基づくモデル A と自動抽出したシズルワードを素性に用いるモデル D を比較すると，モデル D のほうが正例の F 値に対して約 1.32 ポイント，負例の F 値に対して約 0.32 ポイント高くなることがわかった．また，先行研究[2]の素性に基づくモデル B とモデル D を比較すると，モデル B のほうが，正例に対する F 値が約 4.59 ポイント，負例の F 値が約 1.38 ポイント高くなることがわかった．このことから，自動抽出したシズルワードを単独で利用した場合は，人手で作成した拡張シズルワードに基づくモデルより抽出性能は劣るといえる．しかし，両シズルワードを利用したモデル F の性能が最良であることから，自動抽出におけるノイズの問題は残るが，性質の異なるシズルワードが得られたといえる．最後にモデル B とモデル C，モデル D とモデル E を比較した結果，いずれのモデルにおいても出現回数を素性としたほうが，正例，負例ともに F 値が高くなることがわかった．

5.3 考察

先行研究[2]の素性に基づくモデル B と最良性能を得たモデル F で抽出された文を比較する．固有表現抽出の結果を利用したモデル F は，「**グレープフルーツ**をいただきました」や「**バニラ**でよろしいですか？」のようなレビューではない文を一部取り除けていることがわかった．しかし，「**黒糖**ですけど」や「**いちご**にします」など，レビューではない文を誤判定している事例が残っていることも確認できた．また，「**シュー生地**で挟んだもの」や「**能登の粗塩**を組み合わせた」のような未知の味覚表現を含むレビュー文が抽出できていないことがわかった．これは，未知の味覚表現に対する固有表現抽出の性能が低いことが原因であると考える．今後は，シズルワードの自動抽出モデルの性能向上により，レビュー文抽出の性能向上を目指す．

6. まとめ

本稿では，土産に関するレビュー文の抽出性能の向上を目的に，CRF で自動抽出したシズルワードの素性としての有用性について検証した．実験の結果，先行研究[2]の素性に提案素性を追加することで，抽出性能がわずかに向上することを確認したが，人手で作成した辞書を用いたモデルと比較すると抽出性能は低い結果となった．今後は，シズルワードの自動抽出モデルの性能向上とともにレビュー文の抽出性能向上を目指す．

参考文献

[1] 池田他，“シズルワードを利用した土産 レビュー文抽出の検討”，言語処理学会第 26 回年次大会発表論文集，pp.1431-1434，2020.
[2] 在間他，“土産提示システム構築に向けた土産レビュー文抽出の検討”，情報処理学会第 84 回全国大会講演論文集，pp.1-573-1-574，2022.
[3] おいしいを感じる言葉 Sizzle Word 2021, http://bmft.co.jp/publication/reports/sizzleword2021/
[4] 食べログ，https://tabelog.com/
[5] Yuta Hayashibe, et al., “Sentence Boundary Detection on Line Breaks in Japanese”, Proc. of W-NUT2020, pp.71-75. 2020.
[6] CRFsuite，https://www.chokkan.org/software/crfsuite/
[7] 松野他，“日本語大規模 SNS+Web コーパスによる単語分散表現のモデル構築”，2019 年度人工知能学会全国大会論文集，pp.1-3，2019.