4P-04

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

在間 雅人† 安藤 一秋†

†香川大学

1. はじめに

オンランショップの普及により，旅行に行かずに全国各地の様々な土産を購入できるようになった．また，新型コロナウイルス感染症の影響により，オンライン販売を開始した店舗も存在している．一方で，旅行時に現地でしか購入できない土産を求める声[1]は多いということが，アンケート調査で報告されている．また，Web 上には土産情報を提供するサービスがいくつか存在するが，現地でしか購入できないといった情報は，ほとんど提供されていない．

そこで当研究室究では，現地でしか買えない食品系の土産に関する情報（品名や店舗名，レビューなど）を Web 上から収集し，ユーザに提示するシステムの構築を進めている．本稿では，先行研究[2]で取り組まれていたレビュー文抽出手法の改善法について検討する．

2. 関連研究

先行研究[2]では，Yahoo!ブログ（2019 年サービス終了）を対象に，おいしいを感じる言葉「シズルワード[3]」を利用して，レビューを含まない文のフィルタリング手法と，フィルタリングにより得られたレビュー文候補からロジスティック回帰モデルを用いてレビュー文を抽出する手法を提案している．正例 777 文と負例 1,638 文を用いて，10 分割交差検証でレビュー文抽出手法の性能を評価した結果，正例に対して F 値 71.64，負例に対して F 値 95.70 という結果が得られている．この結果から，正例に対する抽出性能の向上が課題として挙げられている．

先行研究では，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しているため，未知語の影響により正例が漏れてしまう問題がある．また，ロジスティック回帰モデルのみでの実験であるため，他の分類器を用いることで性能を改善できる可能性がある．

そこで，本稿では，フィルタリングを用いずに，素性の追加と分類器の変更により，正例に対する抽出性能の向上を目指す．

3. データセットの構築

先行研究[2]では，Yahoo!ブログを利用していたが，現在はサービスを終了しているため，土産に関する新しい情報を得ることができない．そこで，本稿では，食べログ[4]の口コミに注目する．食べログは，ユーザの口コミとランキング情報でおいしい店を探せるサイトであるが，事前調査により，口コミの中には現地でしか買えない食品系の土産に関する情報が存在していることを確認している．

レビュー文の抽出性能を評価するデータセットを新規に構築するため，まず，食べログの「洋菓子ジャンル」，「和菓子・甘味処ジャンル」に属する店舗から任意の 647 店舗に寄せられる口コミを 1 件ずつ収集した．次に，口コミ文章に対して，
タグと「，」「、」，日本語文境界判定器 Bunkai[5]を利用して文単位に分割した．そして，各店舗が取り扱う商品に対して味覚・食感について言及しているレビュー文を人手でアノテーションした．最終的に，1,644 件の正例（レビュー文）と，7,986 件の負例（レビュー文以外）が得られた．これらのうち，各 1,310 件をデータセットに利用する．

4. 提案手法

4.1 素性

本実験では，以下の 4 つの素性を利用する．素性①，②，③は先行研究[1]で提案された素性であり，素性④は，本稿で提案する素性である．

① SWEM の hierarchical-pooling で得られた文ベクトル

単語ベクトルとして，名詞，動詞，形容詞，副詞の分散表現を使用する．単語の分散表現には，株式会社ホットリンクが提供する学習済みモデル[6]を利用する．

② 文に含まれる形態素の数

③ シズルワードの文内での出現有無

シズルワードには味覚系・食感系・情報系がある．先行研究と同様，情報系を除いた味覚系・食感系のシズルワード 226 個を使用し，出現するか否かの 0 or 1 を素性とする．

④ 拡張シズルワードの文内での出現個数

素性③のシズルワード集合には含まれていないが，食品の味覚や食感などを伝える表現が存在する．そこで，「嫌味のない」や「ムニュ～」などの味覚や食感に関する表現，「イチゴ」や「チョコ」など素材により味を伝えている表現 1,360 個を新たに収集した．これらと 226 個のシズルワードの集合から重複した表現を削除した 1,553 個を拡張シズルワードとして利用する．

Extraction of Sentences with Reviews of Souvenirs for Construction of a Souvenir Presentation System

Masato Zaima†, Kazuaki Ando†

† Faculty of Engineering and Design, Kagawa University

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

表 2.　分類実験の結果

	正例			負例		
	適合率	再現率	F 値	適合率	再現率	F 値
LR+①②③素性	0.5869	0.8746	0.7025	0.9692	0.8653	0.9143
SVM+①②③素性	0.6161	0.9145	0.7362	0.9790	0.8753	0.9242
RF+①②③素性	0.6229	0.8803	0.7296	0.9712	0.8834	0.9252
LR+①②③④素性	0.6007	0.8831	0.7151	0.9715	0.8715	0.9188
SVM+①②③④素性	0.6422	0.9259	0.7584	0.9820	0.8871	0.9321
RF+①②③④素性	**0.6848**	**0.9287**	**0.7883**	**0.9830**	**0.9064**	**0.9432**

4.2　分類器

分類器には，先行研究で用いたロジスティック回帰モデル（LR）と Support Vector Machine（SVM），Random Forest（RF）を用いる．

5. 評価実験

5.1　実験設定

先行研究で提案された手法と提案手法の性能を比較するため，3 つの分類器に素性①，②，③と，3 素性に素性④を追加したモデルを使用する．まず，データセットを対象に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する．その後，新たに用意したテストデータ（正例：334 件，負例：1,532 件）に対する適合率，再現率，F 値を評価指標に利用して，各モデルの性能を評価する．なお，形態素解析器には，ipadic-neologd を利用した MeCab を使用する．

5.2　実験結果

実験結果を表 2 に示す．表 2 より，先行研究（LR＋①②③素性）の正例に対する F 値と比較して，SVM＋①②③素性で約 3.4 ポイント，RF＋①②③素性で約 2.7 ポイント向上していることを確認した．

さらに，提案素性④を追加した場合，正例に対する F 値は，素性を追加していないものに対して，LR で約 1.2 ポイント，SVM で約 2.2 ポイント，RF で約 4.9 ポイント向上していることを確認した．

先行研究（LR＋①②③素性）と最も F 値の高い最良手法（RF+①②③④素性）を比較すると，正例に対する F 値は約 8.6 ポイント，負例に対する F 値は約 2.9 ポイント向上していることを確認した．

これらの結果から，分類器として RF と SVM を用いることで分類性能が向上し，また，提案素性を追加することで，さらに分類性能が向上することを確認した．

5.3　エラー分析

最良手法の分類結果に対して，エラーを分析する．まず，正例を負例と予測した例として「皮がボサった感じ」という文が存在した．「ボサった」という表現は，拡張シズルワードにも含まれていない表現であるため，未知のシズルワード，特に美味しくない表現を含むレビュー文に対して，分類性能が下がってしまうことを確認した．この問題に対しては，美味しくない表現を収集・利用することで改善できると考える．

次に負例を正例と予測した例として，「生チョコを食べたことがあります」や「次はカシューナッツだ」のような素材を含むがレビューとしては成立していない文の存在を確認した．素材により味を伝えている表現の取扱いについて検討が必要であると考える．

6. まとめ

本稿では，食ベログに寄せられる口コミからレビュー文を抽出する手法について検討した．実験により，提案手法を用いることで，先行研究で提案された手法より，正例に対する F 値が約 8.6 ポイント向上できることを確認した．

今後の課題として，未知のシズルワード，特に美味しくない表現の抽出手法や，素材により味を伝えている表現の取扱いについて検討する．そして，商品名とレビュー文の紐付け手法と希少価値を示す表現を抽出する手法を検討し，土産情報のデータベースの構築および土産提示システムを実装する．

参考文献

[1]アサヒグループホールディング．
ハピ研：毎週アンケート第 641 回．
https://www.asahigroup-holdings.com/company/research/hapiken/maian/201707/00641/
[2] 池田流弥，安藤一秋，“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.
[3] おいしいを感じる言葉 Sizzle Word 2021，http://bmft.co.jp/publication/reports/sizzleword2021/
[4] 食ベログ，https://tabelog.com/
[5] Yuta Hayashibe and Kensuke Mitsuzawa, “Sentence Boundary Detection on Line Breaks in Japanese”, Proc. of the 6th Workshop on Noisy User-generated Text (W-NUT2020), pp.71-75. 2020.
[6] 原敬介，内海彰，“日本語類似度・関連度データセットの作成”，言語処理学会第 24 回年次大会発表論文集，pp.1011–1014，2018.

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.