シズルワードを利用した土産レビュー文抽出の検討

池田 流弥　　安藤 一秋
香川大学大学院 工学研究科　香川大学 創造工学部
s18g454@stu.kagawa-u.ac.jp ando@eng.kagawa-u.ac.jp

1 はじめに

土産に関するアンケート [1] によると，旅行に行った際，9 割以上の人が土産を購入することや，土産を選ぶ際，現地に行かないと手に入らないものが重視されるということが報告されている．オンラインショップの普及により，多種多様な商品が手軽に購入できるようになり，現地でしか購入できない土産の需要が高まっている．土産情報を提供する各種の Web サービス [2] が存在するが，「現地でしか購入できない」という情報は提供されておらず，情報が提供される土産のほとんどはオンラインショップで購入できるものである．そこで，本研究では，現地でしか購入できない土産に関する情報を Web 上から自動で収集・整理し，ユーザに提示するシステムの構築を目的とする．なお，本研究では，菓子類の土産のみを対象とする．

システムを構築するためには，現地でしか購入できない土産に関する情報を収集する必要がある．これらの情報はブログ記事や Q&A サイトなどの Web 上に散在している．我々はこれまでの研究において，系列ラベリングにより，ブログ記事から土産の品名，販売店舗名を抽出する手法 [3] を提案した．本稿では，食べ物の美味しさなどを表現する「シズルワード [4]」を利用して，ブログ記事から土産の味覚や食感に関するレビューを抽出する手法を提案する．

2 問題設定

現地でしか購入できない土産情報を提供するシステムを構築するには，ブログ記事や Q&A サイト，SNS などの情報源から自動で土産情報を収集する仕組みが必要である．本稿では，土産情報の内，商品の味覚や食感についてのレビューに注目する．商品のレビューは土産物を購入する重要な情報の 1 つである．現地でしか購入できない土産のレビューはオンラインショップから収集できないため，土産について書かれた情報源から，商品のレビューに該当する箇所を自動抽出する手法が必要となる．

情報抽出の観点からレビューに注目した研究としては，レビュー文から商品に対する意見の観点を抽出する手法 [5] や，レビューと製品名や評価観点を入力することで，入力対象に対する意見の系列を抽出する手法を提案した研究 [6] がある．これらは，Aspect-Based Sentiment Analysis(ABSA) と呼ばれるタスクに取り組んだ研究例である．通常の感情極性推定は，文書全体の極性を推定するが，レビュー文書の場合，評価対象を複数の観点で評価する傾向があるため，文書全体に対する極性推定は適切ではない．ABSA は，このような問題に注目したタスクであり，文書から意見対象とそれに対する意見を抽出した後，意見の極性を推定し，対象と意見を紐づけることを目的としている．

レビュー抽出に関する研究の多くは，レビューを含むことが自明な文書からレビュー文やレビュー系列を抽出するものである．一方，本研究は，レビューの存在が自明ではない文書から製品のレビューを抽出する点で問題設定が異なる．

3 提案手法

本研究では，以下の手順でブログ記事から土産に関するレビューを抽出する手法を提案する．提案手法で処理する流れを以下に示す．

1. ブログ記事本文を文単位に分割
2. 文にレビューが含まれるかを判定して抽出
3. 抽出した文と対象となる土産名の紐づけ

3.1 文単位分割

ブログ記事は，文境界として使われる記号や記述法が多様であり，「。」や「！」などの記号を用いて文境界を推定できない場合が多い．ノイズを含む日本語文書の文境界を推定する研究としては，書き言葉に注目したもの [7] やマイクロブログに注目したもの [8] などがある．これらの研究では，系列ラベリングにより文境界を推定しており，9 割程度の性能が得られている．提案手法では，これらの手法を参考に系列ラベリングで文を分割する．

3.2 レビュー文の判定と抽出

ブログ記事本文の全てが土産に関するレビューではない．本文中には，土産名や土産を販売している店舗名のみで構成される文や土産情報と関係ない文が多く含まれている．そこで，レビューを含む文（レビュー文）を判定し，抽出する．

3.3 土産名とレビュー文の紐づけ

先行研究 [3] で提案した手法を用いて抽出した土産名と，先のステップで抽出したレビュー文を紐づける．多くの場合，土産名と土産情報は同じ文中に存在しないため，一文を対象とした関係抽出法を適用できない．Document Level の関係抽出手法や，レビューと製品名を入力し，入力対象に対する意見を抽出する Fan らの研究 [6] を参考に手法を今後検討する．

4 フィルタリングと機械学習によるレビュー文の抽出

本稿では，ブログ記事からレビュー文を抽出するため，シズルワードによるフィルタリングと機械学習を用いたレビュー文の判定法を提案する．

4.1 シズルワードによるフィルタリング

シズルワードとは，食べ物の美味しさなどを表現する言葉のことで，味覚・嗅覚に関わる「味覚系」，触覚・聴覚に関わる「食感系」，感覚ではなく知識や認知に関わる「情報系」の 3 表現からなる [4]．表 1 に，シズルワードの例を示す．

本研究では，土産の味覚や食感に関するレビューの多くにはシズルワードが含まれていると仮定し，まずは，シズルワードが含まれている文のみを土産の味覚や食感に関するレビュー文の候補として抽出する．ブログ記事中の多くの文はレビューを含んでおらず，負例が多いデータ集合になる．そこで，データ集合から負例を減らすため，シズルワードでフィルタリングする手法を提案する．本稿では，シズルワードとして，[4] に掲載されている，味覚系表現 101 件，食感系表現 124 件，情報系表現 126 件を使用する．

また，シズルワードは食品のマーケティングにも利用されるため，表記が限定的である．ブログ記事中の文章は表記揺れが多いため，シズルワードとの単なるマッチングでは正例の取りこぼしが多くなる．例えば，「濃厚である」を含む文はレビュー文である可能性があるが「濃厚な」というシズルワードと完全一致しない．そこで，品詞･文法に注目して，形容詞の名詞化と活用系の追加，表記揺れへの対応として，ひらがな化と一部の漢字化の処理を施し，シズルワードを 2,467 件まで拡張し，フィルタリングに利用する．

4.2 機械学習によるレビュー文の判定

文中にシズルワードを含んでいるがレビュー文ではないものも存在する．そこで本研究では，レビューを含むか否かの二値分類問題として，レビュー文を抽出する手法を提案する．分類器には様々なモデルが存在するが，本稿では初期検討として，教師あり学習の代表的アルゴリズムである，ロジスティック回帰モデルを使用する．

表 1: シズルワードの例

味覚系	食感系	情報系
スイート	ジューシー	季節限定
リッチな	もちもち	本格的
うま味のある	もっちり	鮮度のよい
コクのある	サクサク	新鮮な
濃厚な	とろける	揚げたて

5 評価実験

5.1 実験データ

土産名をクエリとして，Yahoo!ブログの菓子･デザートカテゴリ内でヒットしたブログ記事からランダムに 380 エントリを選択する．そして，それらに含まれる 6,200 文に対して，人手で土産の味覚や食感に関するレビューを含むかどうかの観点で正負ラベルを付与する．本実験で利用する実験データは，821 件の正例文と 5,379 件の負例文で構成され，シズルワードを含まない正例とシズルワードを含む負例も含まれている．

5.2 シズルワードによるフィルタリングの評価

シズルワードによるフィルタリングの性能を評価する．実験データ 6,200 文に対してフィルタリングを適用し，正例文の候補と判定された文を用いて評価する．

評価尺度は正例に対する precision，recall，f1-measure とする．フィルタリングにより正例候補として抽出される文が少ない場合，レビューの取りこぼしが多くなるため，本実験では recall を重視する．

フィルタリングに用いるシズルワード集合として，以下の 4 種を検討する．

- [4] に掲載されているシズルワード 351 件（sizzle）
- sizzle から情報系シズルワードを除いた 226 件（sizzle - Info）
- 4.1 に示す手法で拡張したシズルワード 2,467 件（extension）
- extension から情報系シズルワードを除いたシズルワード 1,810 件（extension - Info）

情報系シズルワードは，菓子類や食品と共起する種類が少ないことを確認したため，情報系シズルワードを除いた集合によるフィルタリング性能も評価する．

表 2 に，各シズルワードによるフィルタリングの評価結果を示す．表 2 より，extension を用いてフィルタリングした場合の再現率が最も高いことが確認できる．ただし，適合率は最も低いことから，多くのノイズが混ざり込んでいることがわかる．このフィルタリングを用いた場合，全正例文 821 件中の 777 件と全負例文 5,379 件中の 1,638 件が正例文の候補と判定され，

表 2: フィルタリングによる正例文候補の推定結果

	precision	recall	f1-measure
sizzle	60.02	66.01	62.87
sizzle - Info	**69.53**	61.14	**65.07**
extension	32.17	**94.64**	48.02
extension - Info	40.34	85.51	54.82

また，44 件の正例文と 3,741 件の負例文が正例文の候補から除外されたことになる．したがって，extension を用いた場合，正例をほとんどを減らすことなく，負例集合の約 7 割 (3,741 / 5,379) をフィルタリングのみで除外することが可能であるといえる．この結果から，土産の味覚や食感に関するレビューのほとんどはシズルワードまたは意味的にシズルワードと近い語を含んでいると考えられる．

5.3 機械学習によるレビュー文判定の評価

フィルタリングによって得られた正例候補からレビュー文を判定する手法の性能を評価する．本実験では，形態素解析に ipadic-neologd[2] を利用した MeCab[1] を利用する．分類器にはロジスティック回帰モデルを，素性は以下を用いる．

- SWEM[9] によって得られた文ベクトル
- 文に含まれる形態素数
- シズルワードが含まれているか否か

SWEM[9] は，文中の単語のベクトルを利用して文ベクトルを得る手法である．SWEM では，文ベクトルを構築する手法が 4 種類提案されている．本実験ではハイパーパラメータチューニングの結果を基に，hierarchical-pooling を用いることにした．単語ベクトルには，文中の名詞，動詞，形容詞，副詞の分散表現を用いる．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[3]を利用する．

評価指標は accuracy，precision，recall，f1-measure とし，10 分割交差検証により判定性能を評価する．交差検証の際，データセットを 8:1:1 の割合で学習，開発，テストデータに分割し，開発データを用いて，ハイパーパラメータチューニングを行う．本研究では，シズルワードによるフィルタリングを適用した後，機械学習によるレビュー文を判定することを想定している．そこで本実験では，交差検証毎に得られる開発･テストデータに対してシズルワードによるフィルタリングを適用し，正例候補として得られる文のみを開発･テストデータに利用する．学習データについては，フィルタリングを適用しない．

フィルタリング後の開発･テストデータを利用する実験のため，正例および負例に対する precision，recall を以下のように拡張して評価に利用する．

$$positive\_precision = \frac{TP}{TP+FP} \quad (1)$$

$$positive\_recall = \frac{TP}{TP+FN+FNF} \quad (2)$$

$$negative\_precision = \frac{TN+TNF}{TN+FN+TNF+FNF} \quad (3)$$

$$negative\_recall = \frac{TN+TNF}{TN+FP+TNF} \quad (4)$$

これらの式において，TP は正例で正解した数，TN は負例で正解した数，FN は正例で誤った数，FP は負例で誤った数，FPF はフィルタリングによってテストデータから除かれた正例の数，TNF はフィルタリングによってテストデータから除かれた負例の数である．フィルタリングによりテストデータから除かれた文は，負例と判定されたと見なすことができる．また，フィルタリングにより正例が除かれた場合は誤り，負例が除かれた場合は正解と見なすことができる．

表 3 にロジスティック回帰モデルによるレビュー文判定の結果を示す．baseline はテストデータにフィルタリングを適用しない場合の結果，その他は各シズルワード集合を用いたフィルタリングによりテストデータから得られた正例候補に対してレビュー文判定を行った結果である．表 3 の baseline とその他のデータセットを用いた場合の性能を比較すると，シズルワードによるフィルタリングを適用することで，負例に対する precision，recall の向上が確認できる．特に，recall については，baseline と sizzle-Info を比較すると，性能が約 8 ポイント向上していることがわかる．このことから，機械学習で負例と判定することが難しい文を，フィルタリングにより一部取り除くことができたと考えられる．また，表 3 の baseline と extension の正例および負例の f1-measure に注目すると，extension によるフィルタリングは，正例の判定性能を低下させることなく，特に負例に対する判定性能を向上できたといえる．ただし，baseline の正例に対する recall と比べ，extension の正例に対する recall は 6 ポイント以上の低下が見られる．このことから，シズルワードを拡張したとしても正例に含まれる味覚や食感を表す表現を網羅できていない部分があるといえる．

6 考察

負例文に対して，機械学習による分類では正例と誤判定されるが，フィルタリングでは取り除くことができる文例を以下に示す．

[1] http://taku910.github.io/mecab/
[2] https://github.com/neologd/mecab-ipadic-neologd/
[3] https://www.hottolink.co.jp/blog/20190304_101414/

表 3: ロジスティック回帰モデルによる分類結果

dataset	positive			negative			accuracy
	precision	recall	f1-measure	precision	recall	f1-measure	
baseline	55.20	**85.32**	66.96	97.54	89.53	93.36	88.96
sizzle	73.96	60.90	66.80	94.19	96.73	95.44	91.99
sizzle-Info	**76.43**	56.88	65.22	93.67	**97.32**	95.46	91.67
extension	69.16	74.30	**71.64**	**96.03**	94.95	95.49	92.21
extension-Info	73.76	67.11	70.28	95.05	96.36	**95.70**	**92.48**

- 和歌山の季節の物を食べられることに感謝です。
- 種が入ってるので、喉に詰めないように気をつけて食べて下さいね
- 広島県宮島のお土産と言えばやっぱり“もみじ饅頭”ですよね

これらの文は「食べる」,「土産」,「饅頭」など，レビュー文にも含まれやすい手がかり語を含んでいるが，土産の味覚や食感に関するレビューは含んでいない．フィルタリングを適用することで，機械学習で判定の難しい文も一部取り除くことができている．

次に，フィルタリングにより誤って取り除かれる正例を示す．

- カステラのザラメのじゃりじゃり感に似てはいるけど
- パサパサした感じも無くってこれならイッキ食いしちゃうなぁ
- 『中にあんことクリーム』が入っていてもう … めちゃうま

シズルワードには「モチモチ」や「フワフワ」など食感に関する擬音語が多数含まれているが,「じゃりじゃり」や「パサパサ」など，ネガティブなレビューで使われやすいものはほとんど含まれていない．これらの表現に対応するために，シズルワードを拡張する必要がある．

また，上記 3 文目の「うま」や例文にはないが「甘」など，短い表現を扱えないことが原因となるフィルタリング誤りがみられた．仮にこれらの表現をシズルワード集合に人手で加えた場合でも，単純な文字列マッチングではノイズとして働く場合が多いと考えられる．そこで，表記の出現有無でフィルタリングするのではなく，正規表現によるマッチングや形態素解析結果に対するマッチングなどの対応が必要である．

最後に，味覚や食感以外の土産のレビューについて考察する．味覚や食感以外の観点に基づくレビューも実験データ 6,200 文中 269 件（正例中の約 25%）存在している．それらのレビュー文には「可愛い」や「綺麗」などの土産の見た目の評価や「コスパ」や「高い」などの金額に関する評価などが散見された．今後，抽出法について検討する必要がある．

7 おわりに

本稿では，シズルワードを用いて，土産の味覚や食感に関するレビュー文の抽出法を提案した．評価実験により，シズルワードによるフィルタリングを適用することにより，機械学習で正例と誤判定する文を一部取り除ける可能性を確認した．提案手法は「食べ物」の美味しさを表現するシズルワードを用いているため，土産以外にも菓子や料理に関するレビューの抽出に利用できる可能性がある．

今後は，正例に対する判定性能の向上とレビュー文と土産名を紐づける方法を検討する．現状では，学習データ中の正例の割合が低いため，疑似的に正例を拡張できれば正例に対する判定性能が向上すると考えられる．また，土産の味覚や食感だけでなく，見た目や金額などのレビューを抽出する手法を検討する．その後，現地でしか購入できない土産を判定する手法を検討し，土産情報データベースの構築を目指す．

参考文献

[1] アサヒグループホールディングス. 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] OMIYA! https://omiyadata.com/jp/.

[3] 池田流弥, 安藤一秋. 固有表現抽出によるブログテキストからの品名・店名抽出. 情報処理学会, 自然言語処理研究会. NL243-5, 8pages, 2019.

[4] おいしいを感じる言葉 Sizzle Word 2019. http://bmft.co.jp/publication/reports/sizzleword2019/.

[5] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the ACL*, 2018.

[6] Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the NAACL*, pp. 2509–2518, 2019.

[7] 福岡健太, 松本裕治. Support vector machine を用いた日本語書き言葉の文境界推定. 言語処理学会第 11 回年次大会発表論文集, pp. 1221–1224, 2005.

[8] 難波悟史, 門内健太, 但馬康宏, 菊井玄一郎. マイクロブログに対する文境界推定および係り受け解析. 言語処理学会第 21 回年次大会発表論文集, pp. 107–110, 2015.

[9] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of the ACL*, pp. 440–450, 2018.