

擬似正解コーパスを用いたレストランレビューのコンテキスト分類

藤岡 寛子^{†,a} 山西良典^{†,b} 西原陽子^{†,c}

† 立命館大学 情報理工学部

a) is0374ix@ed.ritsumei.ac.jp b) ryama@media.ritsumei.ac.jp c) nisihara@fc.ritsumei.ac.jp

概要 レストラン検索では、各ユーザの利用目的が検索の重要な観点となる。利用目的は「デート」「宴会」「女子会」などがあり、利用目的に応じてレストラン選択の際に重要とする観点が異なる。レストランを選択する際の観点は、大きく「料理」と「雰囲気」の2種類に分類できると考えられる。これらの観点に基づき、レストランを検索する上でレストランのレビューは有用な情報源となりうる。しかし、レストランのレビューには、「料理」と「雰囲気」についての記述が文単位またレビュー単位で混在している。本研究では、単語分散表現を用いた文書クラスタリングにより、「料理」と「雰囲気」のコンテキストに関わるレビューを文単位またはレビュー単位で自動分類する。分類にあたり、レストランのレビュー自体にラベリングするのではなく、擬似正解データを仮定し、文またレビュー単位のコンテキストに基づく分類を行う。擬似正解データとしては、「料理」にはクックパッドのつくればを、「雰囲気」には楽天トラベルのホテルレビューをそれぞれ用いた。結果として、擬似正解データを用いてレビュー単位でコンテキストの分類を学習したモデルにより、レストランレビューのコンテキストに基づく分類が実現可能と示唆された。

キーワード 食メディア処理, レビュー分類, 単語分散表現

1 はじめに

「思っていたの違う...」。レストランを検索して、実際に行ってみると「思い描いていたレストランの雰囲気と違う」「求めていた味のメニューがない」と感じたことは、誰しも一度はある経験だろう。レストラン利用では、「どのような関係の人と」「どのような場面で」「どんな料理を」といった利用目的が重要となる。利用目的に応じて、ユーザが満足するレストランの特性は大きく異なる。例えば、上司や先輩とのビジネスの会話をする場合、宴会で利用される傾向が高い飲食店では騒がしく落ち着いた会話はできない。一方で、打ち上げなどを利用目的とする場合には、落ち着いたバーなどは適さない。

利用目的に合ったレストランを検索するためには、レビュー情報の利用が考えられる。レストラン検索サイトなどを用いる場合、スコアを参照するよりも、レビュー内容を確認した方がより正確性、緻密性の観点からも、レストランの特性に関する情報を得る上で有用と考えられる。しかしながら、レビューには様々な観点でレストランの評価が記述されており、一見してそれぞれのレストランの特性を判断して読み解くことは難しい。本稿では、利用目的に大きく影響を受けるレストランの特性として、料理と雰囲気の2種類の観点を考える。ここで、料理は調理方法や食材・味つけなど、雰囲気は店の照明や居心地のよさ、インテリアや接客などをそれぞれ示す。これら2種類のコンテキストが混在しているレストランのレビューを、料理と雰囲気に分類可能となれば容易にレストランの特性を把握可能となり、利用目的に応じた

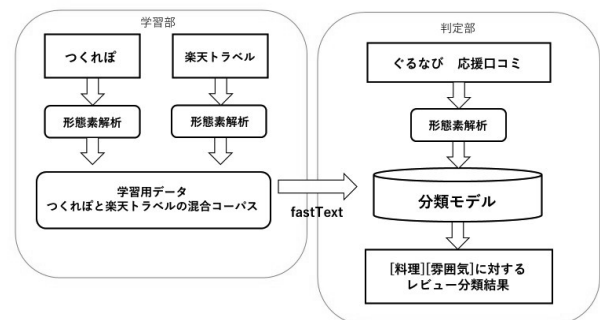


図1 提案手法の概要。

レストラン検索が実現できると考える。レビュー分類における各ラベルのデータ数不均衡の問題に対して、半教師ありアルゴリズム [1] や評価表現の組み込み [2] ではなく、擬似正解コーパスを用いることで解決を試みる。

2 提案手法

図1に、提案手法の概要を示す。提案手法は、大きく学習部と判定部から構成される。本研究では、単語分散表現技術を用いて、レストランレビュー中の各文または各レビューを高次元実数ベクトルで表現する。得られた単語分散表現をもとに、レストランレビューの各文、あるいは、各レビューを「料理」と「雰囲気」の2種類のコンテキストへ分類する。単語分散表現の獲得ではfastText¹ (Skip-Gram) [3] を用いる。fastTextは、大量のテキストをもとにニューラルネットワークを用いて、特に単語のsubwordに着目して、単語間のコンテキストを学習し、多次元のベクトルで表現することができる。

¹<https://github.com/facebookresearch/fastText>

fastText が、ランダムフォレストや doc2vec と比較し優れた分類結果を持つことが既存研究 [4] で示されている。

ここで、fastText の学習データには NII が配布するクックパッドデータセットのつくれば²（以下、C データ）と楽天データセットの楽天トラベルのレビューデータ³（以下、R データ）を統合したデータを、レストランレビューのコンテキスト分類のための擬似正解コーパスとして用いる。ここで、C データと R データは、それぞれレストランレビューの「料理」についての記述の擬似正解データと「雰囲気」についての記述の擬似正解データに対応づく。C データを「料理」についての擬似正解データとして用いる理由は、つくればには料理についての具体的な記述（つまり、料理名や調理方法、使われている食材、料理の味の感想）を含むと考えられるためである。一方で、R データを「雰囲気」についての擬似正解データとして用いる理由は、「利用目的」「アクセス」「店員の態度」「内装」などの施設に関連する意見は、レストランの「雰囲気」の評価においても共通すると考えたためである。

2.1 学習部

図 1 の左側に相当する学習部では、C データと R データをレストランレビューを「料理」と「雰囲気」の 2 種類のコンテキストに分類するための擬似正解データとする混合コーパスを用意する。C データと R データはそれぞれ各文と各レビューに C と R のラベルを付与し、C データと R データの分類問題を解く過程で得られるニューラルネットワークの中間層を単語分散表現モデルとして扱う [5]。分類目的のコンテキストに類似した内容を含むコーパスをそれぞれ用意して擬似正解データとして用いることで、分類対象とするレストランレビューに対して「料理」と「雰囲気」のラベル付けを行うことなく、目的とする分類モデルを構築する。具体的な手順は以下の通りとなる。

1. C データと R データのそれぞれのレビュー、または、文に C または R のラベルを付与する。
2. 各文、または、各レビューを形態素解析する。形態素解析器には MeCab [6]、辞書には NEologd⁴ を用いる。
3. 上記で得られたコーパスを学習データとして fastText でモデルを構築する。モデル構築時のパラメータ設定については、epoch=200, dim = 300, thread = 2 とし、他はデフォルト値を用いた。

2.2 判定部

図 1 の右側に相当する判定部では、学習部で得られた分類モデルに対してレストランレビューを入力とすることで、レストランレビューの各レビュー、または、各文を「料理」「雰囲気」の 2 種類のコンテキストへと分類する。また、入力とするレストランレビューにはぐるなびの応援口コミ API⁵を用いた。具体的な手順としては、以下のように判定処理が行われる。

1. 入力する各文と各レビューを形態素解析する。学習部と同様に、形態素解析器には MeCab、辞書には NEologd を用いる。
2. 入力された各文と各レビューについて、「料理」と「雰囲気」それぞれのコンテキストである確率を softmax 関数を通して算出する。
3. 確率がより高いクラスを推定結果とする。

3 コンテキストの分類実験

本稿では、擬似正解コーパスにおけるコンテキスト分類精度に関するクローズドな検証と、レストランレビューに対する「料理」と「雰囲気」のコンテキスト分類のオープンな分類精度についての考察を行う。

3.1 擬似正解コーパスにおける分類精度に関する検証

擬似正解コーパス内での検証では、テストデータにも C または R のラベルが付与されているためそれぞれの正解ラベルの入力文に対しての分類精度について考察する。表 1 と表 2 に、C データと R データそれぞれのテスト入力に対する推定結果を示す。学習データとしては C データと R データをそれぞれ 10 万件ずつの文を使った。なお、テストデータは、C データと R データのうち学習には用いていないデータをそれぞれ 100 件の文を用いた。

テスト結果として、C データと R データにおける正解率はともに約 98 % となった。どちらの表からも、本来の正解ラベルを高い確率で推定しており、擬似正解コーパスに含まれる文については正しくコンテキスト分類が可能であることが確認された。

ここで、それぞれの出力結果に対しての考察を行う。まず、表 1 について、文 ID9, 10 のみ「雰囲気」である確率が高くなった。文 ID10 に関しては、ホテルレビューでも頻繁に出現する「感動」という単語が含まれているため、「雰囲気」に分類されたと考えられる。次に、表 2 について、文 ID9, 10 のみ「料理」である確率が高くなった。文 ID10 に関しては、「ごはん」「うまい」とい

²<https://www.nii.ac.jp/dsc/idr/cookpad/>

³<https://www.nii.ac.jp/dsc/idr/rakuten/>

⁴<https://github.com/neologd/mecab-ipadic-neologd>

⁵<https://api.gnavi.co.jp/PhotoSearchAPI/v3/>

表 1 C データ (クックパッド) に対する検証結果

ID	レビュー文	「料理」	「雰囲気」
1	ヨーグルト入りでしっとりしてて、美味しいカステラでした~v	0.73	0.27
2	ザクザクで美味しかったです！もう常備菓子です~☆	0.73	0.27
3	厚めになっちゃいましたが、カリっとしてて美味しかったです v	0.73	0.27
4	こんな和風のケーキが作りたかった！あっさりして美味しかったです。	0.73	0.27
5	バナナとココアの味がすごく美味しかったですふわふわです♪	0.73	0.27
6	ブラウニー風にちょっとアレンジさせていただきました。	0.73	0.27
7	生地がとろとろなので、マーブル模様にするのが難しい。	0.73	0.27
8	質問ですが生地はこねてしまうといけないのでしょうか？	0.73	0.27
9	つるりと剥けるし美味しいね♪より甘くなる気がするのは私だけ?!	0.29	0.71
10	初めて手作りのですが、そのおいしさに感動してしまいました!!	0.27	0.73

表 2 R データ (楽天トラベル) に対する検証結果

ID	レビュー文	「料理」	「雰囲気」
1	これで仙台で仕事の際によく行く場所からの立地、駅からの立地がともによく程よい価格帯で重宝します	0.27	0.73
2	係り員の対応が最低事務的な割りにはローカル色が出ていためんどくさい態度が垣間見えた	0.27	0.73
3	場所はどこに行くのにも便利で街中へも 15 分くらいでした	0.27	0.73
4	食事も夕食、朝食ともにバイキングで品数も多く楽しみながら食べることができました	0.27	0.73
5	とても快適なお部屋だったのでぜひまた利用したいと思います	0.27	0.73
6	場所もいいし、部屋はきれい、妻も喜んでいました	0.27	0.73
7	シャトルバスがあるから、ロワジュールでも良いかと判断しましたが全然本数の少ないバスなどをシャトルバスなどと表記するな	0.27	0.73
8	些細な問合せを何度かしましたが、あまり感じの良い的確な解答はなかったです	0.27	0.73
9	6月15日に宿泊しましたが、今回は母の闘病の合間の元気なときに行きたいという要望と、父の日をかねて急遽決まったものでした	0.73	0.27
10	ごはんうまいです	0.73	0.27

う料理に対する単語が含まれているため、「料理」に分類されたと考えられる。

3.2 レストランレビューのコンテキスト分類

表 3, 表 4 に, レビューの各文, 各レビューを入力とした分類結果をそれぞれ示す。このとき, それぞれの分類モデルは擬似正解データについても文単位, レビュー単位で学習を行った。以下, 分類結果の C と R をそれぞれ「料理」と「雰囲気」への分類とみなして考察する。

文単位で分類した結果としては, 「料理」に関するレビューでは正しく分類されているものの, 「雰囲気」に対する記述を含むレビューが「料理」として誤って分類されている傾向が見られた。例えば, 表 3 中の文 ID7 では, 「料理」については触れられず, 店内からの景色や利用目的 (友人とのランチ) について記述されているが, 「料理」である確率が高くなっている。文単位で分類モデルを作成した場合, 学習できる単語数が少ないため, 短い文からコンテキストを正しく読み取って分類することが困難である可能性が考えられる。

レビュー単位で分類した結果としては, 正しくコンテキストを分類できている結果も多かったが, 「料理」に対する記述を含むレビューを「雰囲気」として誤って分類する傾向があった。例えば, 表 4 中のレビュー ID10 は, 「食材」や「味覚」に関する単語が含まれていて店内の情報を述べていないが, 「雰囲気」に対しての確率が高くなった。この要因として, R データの中に表 2 中の文 ID4 のようなホテルの朝食ビュッフェや夕食につい

てのレビューが含まれていたことが考えられる。また, 表 4 中のレビュー ID6 では, 文単位での分類結果では「料理」と誤分類された文 (表 3 中の文 ID6) を含んでいるが, レビュー単位の分類では「雰囲気」に対しての確率が高くなった。これらのことから, レビュー単位では文単位よりもコンテキストの分類に参照できる単語数が増加するため, より正しくコンテキスト分類できたと考えられる。

4 おわりに

本稿では, 利用目的に応じたレストラン推薦の実現を目指して, 擬似正解コーパスを用いたレストランレビューのコンテキスト分類を提案した。擬似正解コーパスを用いることでレビューのコンテキストを分類可能である可能性が示された。一方で, 「雰囲気」の擬似正解コーパスに用いた R データ (ホテルのレビュー) 内にも, 「料理」に関する記述が含まれており, 誤判定につながったおそれが確認された。

今後は, あらかじめ擬似正解コーパス内でのクローズドな判定を行い, 誤ったコンテキストへ判定される文を除去したコーパスによって学習を行う。これにより, それぞれのコンテキストとしてより妥当なデータのみを擬似正解として扱った学習が実現されると考える。

参考文献

- [1] 小暮枝里子, 齊藤史哲, 石津昌平. ランダムフォレストの半教師あり学習による「顧客の声」の分類. 日本感性

表 3 文単位でのぐるなびレビューのコンテキスト分類結果.

ID	レビュー文	C である確率	R である確率
1	とてもヘルシーな和食です。上品な味付けで体にも優しいそう。	0.73	0.27
2	見た目も綺麗に盛り付けされていました。	0.73	0.27
3	色々な味が楽しめたのも良かったです。	0.73	0.27
4	ご飯はおかわり自由だそうです、お替り無しで女性の私には丁度良かったです。	0.73	0.27
5	是非また伺いたいです。	0.73	0.27
6	この時期にあわせた「水無月」と命じた「ゴマ豆腐」をアレンジした一品は味・見た目よし！	0.73	0.27
7	3日前の予約でしたが、京都タワーを真正面に見えるテーブル席を取って戴き、名古屋や亀岡から来た、高校時代の同級生との1年ぶりのランチで、皆満足して暮れたので、とても嬉しかったです。	0.73	0.27
8	京都駅・伊勢丹8階にある「加賀屋」	0.41	0.59
9	石川県の高級旅館「加賀屋」の系列店。味よし・眺めよし（京都駅北側を見下ろす）従業員の気くばりよし！	0.28	0.72
10	窓からの眺望も良く落ち着いた店内でゆっくりお食事が楽しめました。	0.27	0.73
11	さすがに、和倉温泉の老舗旅館『加賀屋』さんの、レストラングループの1店舗だけあり、京都駅ビルの中でも、眺めの良い一角にありました。	0.47	0.53

表 4 レビュー単位でのぐるなびレビューのコンテキスト分類結果.

ID	レビュー	C である確率	R である確率
1	濃厚で食後のデザートにピッタリでした。	0.73	0.27
2	グリルした白身魚が、たくさんの焼き野菜に囲まれて出てきました。優しい味でほっこりしました。	0.73	0.27
3	本日のデザートは、こんな感じでお知らせ。	0.73	0.27
4	オリジナルのノンアルコール。酸味の調整に、とガムシロップを置いていってくれましたが、無しでも甘みあり、美味しいです。	0.73	0.27
5	とてもヘルシーな和食です。上品な味付けで体にも優しいそう。見た目も綺麗に盛り付けされていました。色々な味が楽しめたのも良かったです。ご飯はおかわり自由だそうです、お替り無しで女性の私には丁度良かったです。窓からの眺望も良く落ち着いた店内でゆっくりお食事が楽しめました。是非また伺いたいです。	0.27	0.73
6	京都駅・伊勢丹8階にある「加賀屋」。石川県の高級旅館「加賀屋」の系列店。味よし・眺めよし（京都駅北側を見下ろす）従業員の気くばりよし！この時期にあわせた「水無月」と命じた「ゴマ豆腐」をアレンジした一品は味・見た目よし！さすがに、和倉温泉の老舗旅館『加賀屋』さんの、レストラングループの1店舗だけあり、京都駅ビルの中でも、眺めの良い一角にありました。3日前の予約でしたが、京都タワーを真正面に見えるテーブル席を取って戴き、名古屋や亀岡から来た、高校時代の同級生との1年ぶりのランチで、皆満足して暮れたので、とても嬉しかったです。また、友人や家族と訪れたいです。	0.27	0.73
7	最高ーっ！4月17日また行きます w 料理も…雰囲気も…店長のノリも…店員の対応も…価格も…全て☆3つ (o/) 大阪にもぜひ出店してほしい w	0.27	0.73
8	修学旅行の自由行動、「京料理が食べたい！」ということで女子5人の女子会 (i 修学旅行生ということで予算は少なめでしたが値段以上のサービスをいただきました！大将も店員さんもすごくよくしてくれて、楽しい思い出作らせてもらいました、最高です!!	0.27	0.73
9	JR 南草津駅から歩いてすぐ。近くには Times などのコインパーキングもあります。	0.27	0.73
10	色々な食材が少しずつ盛りつけられていて、目にも楽しく、味覚も様々で楽しいです。	0.27	0.73

工学会論文誌, Vol. 17, No. 5, pp. 537–545, 2018.

- [2] 岡田真, 竹内和広, 橋本喜代太. 利用状況を考慮したカスタマーレビューの自動分類における評価文型の有効性に関する調査. 人工知能学会全国大会論文集, Vol. 28, pp. 1–3, 2014.
- [3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017.
- [4] 平川遼汰, 宮崎佳典, 谷誠司. 日本語例文自動分類による cefr 読解指標推定支援 web アプリケーションの開発. 第 80 回全国大会講演論文集, No. 1, pp. 635–636, 2018.
- [5] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431. Association for Computational Linguistics, 2017.
- [6] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto.

Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, 2004.