観光レビューを利用した観光スポットの観点の自動抽出

小林 らんう † 上野 史 †† 太田 学 ††

† 岡山大学 〒 700–8530 岡山県岡山市北区津島中 3–1–1
†† 岡山大学学術研究院自然科学学域 〒 700–8530 岡山県岡山市北区津島中 3–1–1
E-mail: †pwoy2qot@s.okayama-u.ac.jp, ††{uwano, ohta}@okayama-u.ac.jp

あらまし 観光レビューには観光スポットの特徴を確認できる様々な情報が記載されている．本稿では観光レビューを用いて観光スポットの観点を自動で抽出する手法を提案する．観光スポットの観点は「歴史」や「自然」のように，スポットの特徴を分析する基準になる概念で，多様な観点を用いてスポットを分析することでスポットの特徴を理解しやすくなる．提案手法では，最初に旅行サイトであるじゃらん net の観光レビューから Natural Language API を用いてエンティティを抽出する．それを分散表現に変換した後クラスタリングし，スポットの特徴を反映したエンティティのクラスタを生成する．さらにそれぞれのクラスタの類似語を求めその中から観点を抽出する．実験では岡山県の 30ヵ所の観光スポットからスポット毎に観点を自動抽出し，抽出された観点について定量的に評価した．その結果，30ヵ所の観光スポットから合計 446 の観点が自動で抽出され，その内 78.9%が観点としてふさわしいものだった．

キーワード 観光，情報抽出，レビュー解析

1 はじめに

観光支援サイトや観光支援の研究では観光スポットを観点毎に分析することがある．例えば野本ら [1] は，穴場スポットを発見する研究で，観光スポットに食事，景観，購買，体験，設備，混雑，交通の 7 つの観点を設定し，観光スポットを分析した．また杉浦ら [2] は京都の観光スポットを推薦するシステム「京のおすすめ」を構築するために，アンケートや面接から収集した 7000 件以上の評価表現を上位概念である 137 の評価要因に要約し，さらにそれを「気分」,「体験」,「雰囲気」,「スポットの特徴」という 4 つのカテゴリに分類した．

観点は人が設定することがあるが，設定によっては望んだ分析結果を得られない可能性がある．たとえば,「食事」「自然」「歴史」「土産」の 4 つの観点を設定し観光スポットを分析した時,「金閣寺」などは「歴史」の観点で分析が可能だが,「乗馬体験」や「スキューバダイビング」などの観光客が直接体験する観光スポットは分析することが難しくなる．この場合は，例えば観点「体験」を追加することで，分析を容易にすることができる．

また上記の観点の他に映画やドラマなどの撮影地における「聖地巡礼」や水族館における「イルカ」など，限られたスポットのみに登場する独特な観点も存在する．このような観点はスポットのセールスポイントとなる場合があり，見落とすとスポットの魅力を下げる．

本稿ではこのような問題の解決策として観光レビューを用いて様々な観点を自動で抽出する手法を提案する．観光レビューには実際に観光スポットを訪れた観光客の感想や意見が書かれており，そこには観光スポットの特徴を含む記述が多く含まれる．さらに観点を観光スポット毎に抽出することでそのスポット特有の観点が見つけられ，より多様な観点が集められる．

本稿の構成は以下の通りである．2 節では関連研究を述べる．3 節では本稿で提案する観光スポットの観点の自動抽出手法を述べる．4 節では観光レビューから観点を自動抽出した結果と考察を述べる．5 節ではまとめと今後の課題について述べる．

2 関連研究

観光レビューには様々な情報が含まれており，これを用いた観光スポットの分析や推薦システムの開発はこれまでに多く行われた．

市村 [3] らは観光レビューを利用して観光スポット推薦システムの「旅ゲーター」を開発した．彼らは観光レビューを Doc2Vec[1]や k-means 法を用いてクラスタリングし，各クラスタから観光スポットの特徴を示す重要語を TF-IDF 法を用いて抽出した．その重要語を「旅ゲーター」のインタフェースに表示しユーザに選ばせることで，ユーザの好みを反映した観光スポット推薦システムを構築した．杉本 [4] らは観光レビューから感情を表す語を抽出し，それを用いて観光レビューを喜び，悲しみ，受容，嫌悪，恐れ，怒り，驚き，期待の 8 つの感情に分類し，観光レビューに出現する感情の傾向を確認した．実験の結果，高い精度で分類されることを確認した．

レビューの情報を用いた研究は観光分野のみならず他の分野でも多く行われている．市村 [5] は，ユーザが入力した文と関連のある料理店の情報を表示する Web システム「食探」を開発した．市村は Doc2Vec を用いてユーザが入力した文と類似したレビュー文を探し，日本語係り受け解析器の CaboCha[2]を用いてレビュー文を短い係り受け文に要約した．その要約文の中から料理に関係した文を選別しインタフェースに表示することで，ユーザが手早く料理店の情報を把握できるようにした．

1：https://radimrehurek.com/gensim/models/doc2vec.html
2：https://taku910.github.io/cabocha/

金子 [6] らは美容や健康，医薬品などに関係する商品のレビューから，CaboCha を用いて係受け表現を抽出し，商品の購買意図が含まれるものについて分析した．また抽出した購買意図を構造化し，得られた階層的知識についてその実用性を調査した．小川 [7] は映画「劇場版「鬼滅の刃」無限列車編」と「君の名は。」の 2 つの作品のレビューを用いて，作品の評価に影響を及ぼす要因を分析した．

観光レビューや旅行ブログの分析には，[3] や [4] のようにレビュー文から特定の単語を抽出する手法がしばしば用いられる．野本 [1] らは，観光レビュー文から Natural Language API[3]のエンティティ感情極性分析を用いて，エンティティとともにそのエンティティの入力テキストにおける感情極性値を抽出した．上原 [9] らは MeCab[4]や TermExtract[5]，独自で作成した辞書を用いて Yahoo!知恵袋 [6]と Yahoo!ブログ [7]から観光地名の共起キーワードを抽出した．本稿では野本らが用いた Natural Language API の中のエンティティ分析を用いて観光レビューからエンティティを抽出する．

レビューなどを特性毎にまとめるために，レビューや抽出した語のクラスタリングを行うことがある．[3] では k-means 法を用いて観光レビューをクラスタリングし，観光レビューを特徴毎にまとめた．田村 [8] らは AppStore の類似したアプリケーションをまとめるために，アプリケーションの機能説明文から MeCab と TF-IDF 法を用いて特徴語を抽出し，x-means 法を用いてクラスタリングした．また淀川 [10] らは生成された文書クラスタにラベルを自動で付与することを試みた．彼らは Support Vector Machine(SVM) を用いてニュースコーパスの文書をクラスタリングし，さらに fastText[8]を用いて各クラスタのラベルの候補となる単語を抽出し分析した．

3　観光スポットの観点の自動抽出手法

3.1　概　　要

本稿で提案する観光スポットの観点の自動抽出手法を図 1 に示す．本手法は旅行サイトから得られる任意の都道府県の観光スポットの観光レビューを用いて観点を抽出する．まず収集した観光レビューからスポット毎にエンティティを抽出する．エンティティのほとんどは人物名，地名などの固有名詞やレストラン，競技場などの普通名詞である．次に，エンティティを分散表現に変換し，それらを Ward 法でクラスタリングする．さらに，生成されたクラスタを代表するクラスタベクトルを算出する．最後にそのベクトルに類似した単語を出力し，その中からスポットの観点となる単語選び，観点として抽出する．エンティティのクラスタは通常複数得られるため，観点もスポット毎に複数得られる．

3：https://cloud.google.com/natural-language?hl=ja
4：http://taku910.github.io/mecab/
5：http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html
6：https://chiebukuro.yahoo.co.jp/
7：Yahoo!ブログは 2019 年にサービスを終了した
8：https://fasttext.cc/

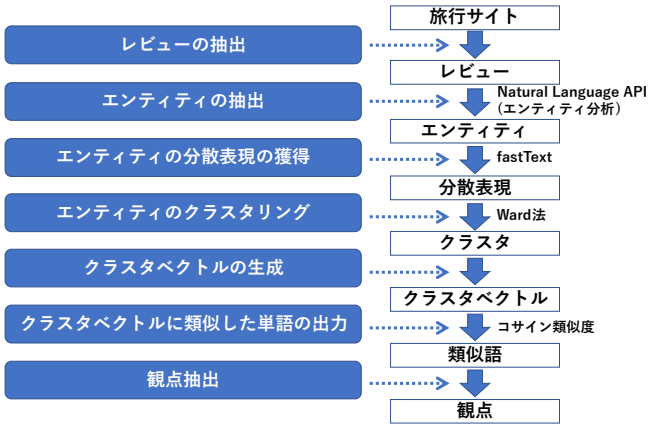

図 1　提案する観点抽出手法の概要

3.2　レビューの取得およびエンティティの抽出

提案手法では，じゃらん net[9]などの旅行サイトから Web スクレイピングを用いて観光スポット毎に観光レビューを収集する．つづいて収集したレビューから，Natural Language API のエンティティ分析 [10]を用いてエンティティを抽出する．Natural Language API は Google Cloud Platform (GCP) が提供する API である．この API では感情分析や構文解析，エンティティ分析などの自然言語処理のタスクを行うことができる．提案手法では，エンティティ分析により観光レビューからエンティティを抽出し，そのエンティティの salience(重要度) の情報を抽出する．salience は [0.0,1.0] の値をとり，1.0 に近いほどエンティティがレビュー中で重要であり，主体となることを示す．さらに，API に渡したレビューから抽出されるエンティティの salience は，すべて足し合わせると 1.0 になる．抽出されるエンティティの数はレビューの長さによって異なるため，平均的な salience の値はレビュー毎に異なる．この問題を回避するため本稿では，レビュー毎にエンティティを抽出するのではなく，スポット毎にレビューをまとめ，スポット毎のレビュー集合からエンティティを抽出する (図 2)．エンティティは後でクラスタリングするため，エンティティ中のノイズを除去する．スポット毎に抽出されたエンティティの内 salience が上位 30%のエンティティを抽出する．また，同じテキスト内で同じエンティティが複数回が抽出されることがある．本稿では複数回登場するエンティティであっても，そのいずれかの salience が上位 30%に入ってなければそのエンティティは抽出しない．一方，取り出した上位 30%のエンティティに重複があれば 1 つを残して削除する．

3.3　エンティティの分散表現の獲得

エンティティの分散表現を獲得する概要を図 3 に示す．3.2 節の方法で抽出したエンティティは fastText を用いて分散表現に変換する．fastText は自然言語処理ライブラリで，単語を分散表現に変換することができる．本稿では，fastText を利用し

9：https://www.jalan.net/
10：https://cloud.google.com/natural-language/docs/analyzing-entities?hl=ja

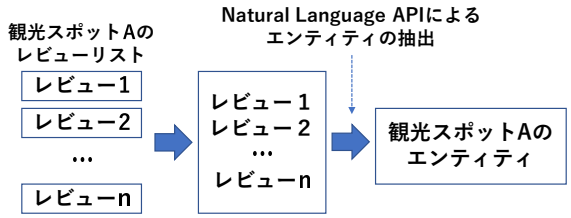

図 2　レビューからのエンティティ抽出の概要

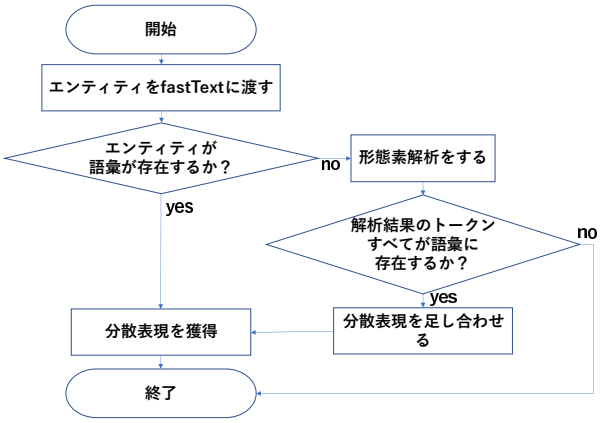

図 3　エンティティの分散表現獲得の概要

て Common Crawl[11] と Wikipedia[12] で事前学習した単語ベクトル [13] を用いる．この事前学習済みの単語ベクトルは 300 次元で，語彙 (単語) 数は 2,000,000 である．エンティティの分散表現は事前学習済み単語ベクトルの語彙にエンティティと同じ単語が存在すれば獲得できる．しかし，エンティティと同じ単語が学習済み単語ベクトルの語彙に存在せず分散表現を獲得できない場合がある．このような場合は図 3 に示すように，形態素解析器の MeCab を用いて形態素解析する．形態素解析を行うと文章や複合語がいくつかの単語 (トークン) に分解される．例えば「観光スポット」は「観光」と「スポット」の 2 つのトークンに分解される．形態素解析したエンティティについては，解析結果のトークンの分散表現を足し合わせて元のエンティティの分散表現とする．例えば「観光スポット」は「観光」と「スポット」の 2 つのトークンに分解されるため，「観光スポット」の分散表現は「観光」の分散表現と「スポット」の分散表現を足し合わせて求める．これでも分散表現を獲得できないエンティティはクラスタリングに利用できないため削除する．

3.4　エンティティのクラスタリング

エンティティをクラスタリングすることで特徴が似ているエンティティを集める．このエンティティは観光スポットのレビューから抽出したものであるため，このエンティティのクラスタは観光スポットの特徴を示している．クラスタリングには，階層的なクラスタリング手法の一つである Ward 法 [11] を用いる．Ward 法は，最初個々のデータをすべて異なるクラスタとし，距離が近いクラスタを順にまとめてクラスタを階層的に統

11：https://commoncrawl.org/
12：https://www.wikipedia.org/
13：https://fasttext.cc/docs/en/crawl-vectors.html

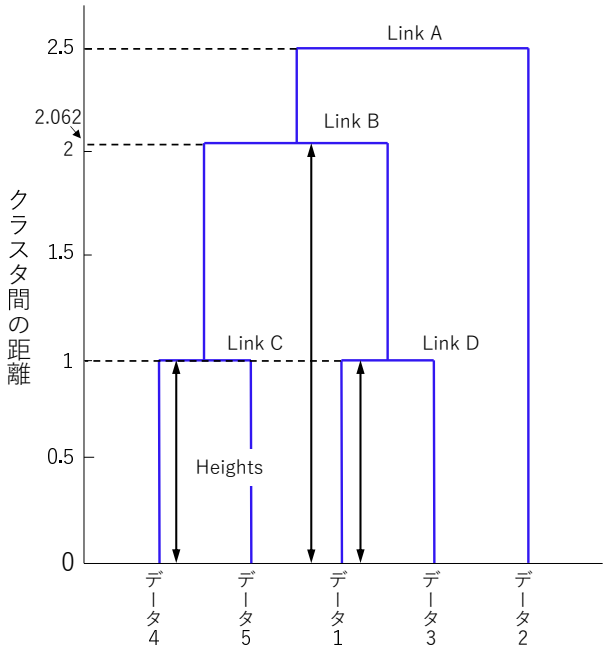

図 4　デンドログラムの例 (MathWorks [12]，一部改変)

合させるクラスタリング手法である．そのため，Ward 法では最終的にデータは一つのクラスタにまとまる．適当なクラスタ数にするには，いずれかの階層のクラスタを抽出する必要がある．クラスタ数の決定には，距離を基準にするものや，不整合係数を基準にするものなど様々な方法があるが，本稿では不整合係数を基準にクラスタ数を決定する．

まず Ward 法でクラスタリングを行った時 k 出力されるデンドログラムの例を図 4 に示す．不整合係数とは，クラスタを統合した時のリンクの不整合度合いを表す値で，この値が高いほど似てないクラスタが統合されていることになる．

不整合係数の計算には参照しているリンクとその 1 つ下の階層のリンクを用いる．例えば図 4 の Link B の不整合係数を計算するには Link B，Link C，Link D の 3 つのリンクを用いる．下の階層にリンクが存在しない場合は不整合係数は 0 となる．図 4 の例では Link C と Link D の不整合係数は 0 となる．式 (1) は k 番目のリンクの不整合係数 $\mathrm{Y}(k)$ を求める式である．

$$\mathrm{Y}(k) = \frac{\mathrm{dist}(k) - \mathrm{mean}(k)}{\mathrm{std}(k)} \qquad (1)$$

ここで $\mathrm{dist}(k)$ は参照しているリンクの高さで，$\mathrm{mean}(k)$ は計算に用いた全リンクの高さの平均値，$\mathrm{std}(k)$ は計算に用いた全リンクの高さの標準偏差である．例えば図 4 の Link B から Link D までの高さの平均値は 1.353，標準偏差は 0.613，また Link B の高さは 2.062 である．これらの値と式 (1) から $\mathrm{Y(B)} = (2.062 - 1.354)/0.613 = 1.155$ となる．

不整合係数を基準にクラスタ数を決定するためには不整合係数の閾値を設定する．あるリンクとその下の全階層のリンクの不整合係数が閾値未満であればそのリンクと下の全階層を同一クラスタとする．本稿では不整合係数の閾値を，全リンクの不整合係数の 最大値× 0.99 に設定する．例えば図 4 では全てのリンクの内，Link B の不整合係数が最も高いため，不整合係数

こと，事，もの，物，観光，市，県，市内，県内，すべて，全て，ところ，所，場所，旅行，何，ある，いる，東，西，南，北，右，左，左右，上，下，東西南北，東側，西側，南側，北側，側，前，右側，左側，前側，後ろ側，後ろ，後，途中，近く，遠く，そこ，それ，ここ，どこ，あれ，これ，名前，さん，とき，地域，方面，方，一緒，客，思う，他，スポット，感じ，考え，中，うち，最初，最後，明日，昨日，私，君，僕，あなた，次回，方々，今回，今度，大体，ほとんど，多く，たち，達 (計 81 語)

図 5　類似語とはしないストップワードド

の閾値は Y(B) × 0.99 となる．なお Y(B) は閾値より大きく，Y(A) と Y(C)，Y(D) は閾値より小さい．この時，データ 4 とデータ 5，データ 1 とデータ 3，データ 2 が各クラスタとなる．

3.5　エンティティクラスタからの観点抽出

3.5.1　類似語リストの生成

3.4 節の方法で生成したクラスタのクラスタベクトルを求め，そのクラスタベクトルから類似語リストを生成する．本稿ではクラスタに属するエンティティの分散表現の平均値をクラスタベクトルとする．そのクラスタベクトルに類似した単語を 3.3 節で説明した事前学習済み単語ベクトルの語彙から出力する．類似語リストにはクラスタベクトルとのコサイン類似度が上位 3 件の類似語が入る．また，エンティティの数が少ないクラスタ (小さいクラスタ) から抽出される観点は，レビューでの言及が少ない．そのため本稿ではクラスタに含まれるエンティティ数が 5 未満のクラスタについて類似語リストを生成しない．

類似語は観点の候補となるため,「の」や「も」などの助詞や,「岡山」や「倉敷」などの地名はふさわしくない．そのため本稿では形態素解析器の MeCab と Juman++[14]を用いて品詞が助詞や地名などの類似語を削除する．出力する類似語を選別する際，MeCab と Juman++の 2 つの形態素解析器を用いる理由は，類似語選別の精度を向上させるためである．

以下は本稿で用いた類似語出力の条件であり，出力するすべての類似語はストップワード (図 5) に含まれないことが前提である．ただし MeCab は M，Juman++は J で表記している．

1　**「M の解析結果のトークン数が 2 以上である」**
2　**「M の解析結果のトークン数が 1 である」** かつ
　　「M と J の品詞判定のいずれかが『名詞』である」 かつ
　　「M の詳細品詞判定が『地域』ではない」かつ
　　「J の詳細品詞判定が『地名』ではない」かつ
　　「J の品詞判定が『未定義語』ではない」
3　**「M と J の品詞判定がいずれも『動詞』である」**
4　**「M と J の品詞判定がいずれも『形容詞』である」**
5　**「M と J の品詞判定がいずれも『形容動詞』である」**

3，4，5 の条件に当てはまる場合，原形を類似語として抽出する．また図 5 のストップワードは本稿の第一著者が観点としてふさわしくないと判断した単語のリストである．

類似語リストにはクラスタベクトルとの類似度が上位 3 件の類似語を出力するが，上記の類似語出力条件により削除された類似語が存在する場合，次に類似度が高い類似語を出力する．なお，削除した類似語が 3 つ以上であればそのクラスタは観点候補としてふさわしい類似語を出力しないと判断し，類似語を出力しない．

3.5.2　接尾語からの複合語生成

3.5.1 節で生成した類似語リストに接尾語が含まれていれば他の類似語と複合語を生成し，その複合語を類似語リストに追加する．また接尾語は類似語リストから削除する．接尾語が含まれているかどうかは MeCab の N-Best 解出力機能を用いる．ここで N-Best 解とは，形態素解析結果の候補を N 番まで示したものである．本稿では 5 番目までの候補の中に接尾語と判定したものが一つでも存在すればその単語は接尾語とする．生成した複合語の分散表現は，結合前の 2 単語の分散表現の和とする．例えば類似語リストに「駐車」と「場」を含む場合は「場」が接尾語となり「駐車場」を新たに類似語リストに追加し，接尾語の「場」は類似語リストから削除する．この時「駐車場」の分散表現は「駐車」の分散表現と「場」の分散表現を足し合わせたものとなる．また，「海風」と接尾語の「風」から「海風風」のような複合語を生成しないように，接尾語が他の類似語に含まれていれば複合語を生成しない．そのため，類似語リストに残る類似語は必ずしも 3 つとは限らない．

さらに本稿では「動物犬」や「愛称声」のように日本語の語彙に存在しない複合語の生成を防ぐために，Wikipedia の検索結果に現れない語を削除する．検索した結果その単語の記事が存在すれば，日本語の語彙にあると判断する．また Wikipedia での検索結果にその記事が存在しない場合は，検索結果の 1 ページ (最大 20 記事) のスニペットにその検索単語が含まれるかどうかを調べる．検索結果に含まれる件数が 10 件以上であればその単語は日本語の語彙にあると判断する．

3.5.3　類似語リストからの観点抽出

類似語リストが存在するクラスタについては，その類似語リストから類似語を 1 つ選択し観点として抽出する．類似語リストから観点を選ぶ方法は，まず，各類似語の分散表現と，該当クラスタに含まれるエンティティの分散表現とのコサイン類似度を算出し，平均値を求める．この類似度の平均値が最も高い類似語を観点として抽出する．

4　観点の自動抽出の実験

4.1　データセット

観光スポットの観点の自動抽出の実験に用いるデータは，じゃらん net から 2022 年 7 月 5 日時点の岡山県の「7 月にオススメランキング」の内上位 30ヵ所のレビューを取得したものである．この 30 スポットの内，最もレビュー数が多かったスポットは 3,143 レビューの「倉敷美観地区」であり，レビュー数が最も少なかったスポットは 116 レビューの「塩釜冷泉」である．全スポットのレビューの合計は 19,307 である．また，最も多くのエンティティが抽出されたスポットは「岡山後楽園」で，その数は重複を除いて 494 ある．最もエンティティが少なかったスポットは「石山寺（岡山県津山市）」で，その数は重複を除

14：https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN%2B%2B

表 1　観点の 5 段階評価の基準

評価段階	詳細	観点の例
A	具体的な観点になりうる	「ビール」,「ウサギ」など
B	観点としてふさわしい	「飲み物」,「動物」など
C	どちらとも言えない	「おすすめ」,「企業」など
D	観点としてふさわしくない	「手前」,「罪悪感」など
E	日本語の語彙にない語または非自立語	「宿店」,「さ」など

いて 85 ある．スポット毎に重複を除いて抽出されたエンティティは全スポットの合計で 8,684 あった．

4.2　実験内容および評価内容

実験では 4.1 節で説明した岡山県の 30 スポットのレビューから提案手法によりスポット毎に観点を抽出する．そして抽出した観点について定量的に評価する．抽出した観点が観点としてふさわしいか，またふさわしくない観点はどのようなものかを分析する．定量的評価として，抽出した観点を A(具体的な観点になりうる)，B(観点としてふさわしい)，C(どちらとも言えない)，D(観点としてふさわしくない)，E(日本語の語彙にない語または非自立語) の 5 段階で評価する．5 段階の評価の一覧を表 1 にまとめる．また，この評価は本稿の第一著者が行う．

表 1 に示すように，評価基準 A は「ビール」や「ウサギ」のような具体的な単語でこれらのほとんどは上位概念が存在する．例えば「ウサギ」は「動物」，「ビール」は「飲み物」などが存在する．評価基準 B は観点としてふさわしいものでり，その観点で観光スポットを分析できると判断したものである．また評価基準 B の観点が評価基準 A の観点と異なる点は，B の観点は A の観点より具体性に欠ける点である．例えば先ほど説明した「飲み物」や「動物」も評価 B になる．その他には「時間」や「雰囲気」，「敷地」などがある．「時間」は「待ち時間が長い」や「体感時間が短い」のように派生し，「雰囲気」は「穏やか」や「良い雰囲気」，「敷地」は「敷地が広い」や「敷地が狭い」などの派生があり得る．評価段階 C は「おすすめ」，「企業」などのように考え方によっては観点としてみなせる単語である．評価段階 D は「手前」や「罪悪感」などのように観光スポットの観点としてはふさわしくない単語である．評価段階 E の単語は「宿店」のように日本語の語彙にない単語や，「さ」や「の」のようにその単語単体では意味が分からない非自立語である．

また本稿では抽出した観点に含まれる固有表現にも注目する．固有表現は具体的なものが多く，観光スポットの特有の特徴をとらえていることが多いためである．

4.3　観点の自動抽出の比較手法

3.5 節で説明した通り，提案手法ではレビュー文から生成したエンティティのクラスタから，クラスタベクトルを求め，そのクラスタベクトルから類似語リストを生成する．またその類似語リストから観点を抽出する．比較実験では類似語リストから観点を抽出する代わりに，エンティティのクラスタからクラスタベクトルに最も類似したエンティティを一つ観点として抽出し，その結果を示す．ここで，抽出する観点が 3.5.1 節で説明した類似語出力条件に該当しなければ，クラスタベクトルと 2 番目に類似したエンティティを観点として抽出する．この処理を繰り返し，クラスタベクトルとの類似度が上位 3 のエンティティまでに観点を抽出できなかった場合，そのクラスタの観点は抽出しない．比較手法によって抽出した観点についても提案手法と同様に表 1 の評価を行い，さらに固有表現も数える．

4.4　実験結果と考察

岡山県の 30 スポットについて提案手法を用いて観点を抽出した結果の一部を表 2 に，比較手法を用いて観点を抽出した結果の一部を表 3 に示す．また提案手法で抽出した観点に対する評価の結果を表 4 に，比較手法で抽出した観点に対する評価の結果を表 5 に示す．表 2 と表 3 では各スポットの名前とその横に抽出した観点の一覧をまとめており，表 4 と表 5 では表 1 に示した評価 A から E の観点の数とその割合に加え，それに含まれる固有表現の数とその割合をまとめている．

4.4.1　提案手法を用いた観点抽出結果の考察

表 2 のスポット「倉敷アイビースクエア」の「レンガ」やスポット「倉敷美観地区」の「デニム」や「町並み」，スポット「ブラジリアンパーク鷲羽山ハイランド」の「サンバ」，スポット「大原美術館」の「ピカソ」など実際調べてみないと分からない具体的な情報が観点として抽出された．実際の真偽が不明の観点もあるが，これらは有用な観点の可能性が高い．また，「おかやまフォレストパーク　ドイツ森」の「屋台店」やスポット「ブラジリアンパーク鷲羽山ハイランド」の「絶叫系」と「異国風」，スポット「吉備津彦神社」の「山道」のようにスポットの特徴を捉えられている複合語の観点が抽出されている．

3.5 節で説明したように Wikipedia を用いて不適切な複合語を除去したにもかかわらずスポット「神庭の滝」の「宿店」やスポット「池田動物園」の「バス車」や「子供手」のように日本語の語彙にない複合語も抽出された．これは，例えば「宿店」は「新宿店」,「子供手」は「子供手当」,「バス車」は「バス車庫」に含まれるためである．この問題を解決するためにはオンラインの国語辞典を用いたり，または Wikipedia の記事を形態素解析し，上記のような包含関係によって登場した複合語は除去したりするなどの処理が必要である．また，スポット「岡山後楽園」の「さ」やスポット「倉敷アイビースクエア」の「の人」，スポット「吉備津神社」の「て」なども抽出されている．これらの観点が抽出された理由は Juman++ または MeCab のいずれかで「名詞」と判定され，かつ「地名」でないためである．この問題を解決するにはひらがな 1 文字の単語を類似語として出力しないなどの方法が挙げられる．

4.4.2　提案手法と比較手法の比較

比較手法を用いた実験では，エンティティがそのまま観点となるため提案手法より具体的な観点が抽出された．例えば，スポット「おかやまフォレストパーク ドイツの森」の提案手法の観点は表 2 から「チーズ」であるのに対し，比較手法では表 3 から観点「チーズたっぷり」となった．またスポット「おかやまフォレストパーク ドイツの森」の提案手法の観点は表 2 から「体験」であるのに対し，比較手法では表 3 から観点「ソーセージ手作り体験」となった.「チーズ」のように既に具体的な

表 2　提案手法を用いた観点の自動抽出結果の例

スポット	観点
おかやまフォレストパーク　ドイツの森	チーズ, 飲み物, ビール, 雰囲気, アトラクション, 雨上がり, バーベキュー, ウサギ, 花, 態度, 席, 屋台店, 大人, 敷地, パン, 体験, 動物, 子供, 同伴
ブラジリアンパーク鷲羽山ハイランド	パスポート, テーマパーク, スタンディングコースター, イベント, ストレス, ダンス, 子ども, サンバ, スカイ, フード, 大会, 魅力, 異国風, 子供, 地元, 施設, 乗り物, 事故車, 景色, 斜面, 動物, 罪悪感, 絶叫系, 乗る, 入園料, 点検
倉敷アイビースクエア	レストラン, テーブル, 雰囲気, カフェ, の人, レンガ, 夕食, 写真, 景色, 受付, ホテル, 建物, 売店, 美観, 工場, 宿泊, 館, 施設
倉敷美観地区	カンジ, ランチ, バウムクーヘン, お菓子, カメラ, カード, デニム, 町並み, 景色, 子供, 小物, 部屋, 映画, 駅, 夕方, 周り, 有数, 前回, 川船, ネコ, 駐車場, 土産屋, 時間, 町並み, 美観
吉備津彦神社	感覚, 笑顔, お参り, レンタサイクル, 伝統, 愛称, 友人, 山道, 伝説, 神主, 鬼, 駅, 神社
吉備津神社	て, 喜び, 時期, ご利益, 横, 地方, 娘, 関係, 写真, 回廊, 桜, 神, 山道, 境内, 敷地, 建物, 神社
大原美術館	美術, 西洋, 西洋, ピカソ, 雰囲気, レストラン, 大地, 地元, 夕方, 大勢, 館, 画家, 展示, 作品, 土産, 絵画, 中庭, 入場
岡山後楽園	庭園, 庭園, さ, 整備, 訪問, 時間, バス, 抹茶, 春, 桜, 紅葉, 鶴, 鯉池, つつじ, メダカ, 見どころ, 夕暮れ, バス停, ガイド, 時期, 価値, 敷地, 川沿い, 茶店, 風情
池田動物園	動物園, 迫力, 距離, 気持ち, イベント, バス停, ウサギ, 鹿, バス車, 子供手, 子供園
備中松山城	猫. 城郭. 岩. 景色. 駅道. タクシー. 天守閣. 遊歩道. トイレ. タクシー. 若い人. 山奥. 風情. 滞在. 駐車場. 名所. 登城. 運動. 天守. 城. 戦国
神庭の滝	湯原, 岩石, 川, 子供, 山道, 宿店, 風景, 夏, 見どころ, 遊歩道, 想像, 迫力, 間数, 禁止, 猿, さ, 駐車, 子供

表 3　比較手法を用いた観点の自動抽出結果の例

スポット	観点
おかやまフォレストパーク　ドイツの森	チーズたっぷり, 飲み物, ビール, 雰囲気, バイキングレストラン, 雨上がり, バーベキュー, おもしろ自転車, 小さなお子, できたて, 接客態度, テラス席, 店, 大人分, 食事場所, パン, ソーセージ手作り体験, 動物, 子供たち, 子ども連れ, 岡山市
ブラジリアンパーク鷲羽山ハイランド	チケット, テーマパーク, メリーゴーランド, イベント, 待ち時間, ダンス, ジャングルカフェ, サンバカーニバル, スカイレール, ブラジルフード, ビンゴ大会, 魅力, 留学生風, 子人, 外側, 地元, 施設, 中乗り物, 車, 景色, 斜面, 岡山県民, ふれあい動物園, おしゃれ感, 絶叫系, 乗り物乗り放題, 入園料, 施設点検
倉敷アイビースクエア	レストラン. ロビー. 町並み. キャンドルショップ. 女性スタッフ. レンガ壁. 朝食バイキング. 写真スポット. 景色. 閉館. クラッシックホテル. 店. 売店. 美観地区. 倉敷駅. 紡績工場跡地. 宿泊施設メイン. 記念館. 見学施設
倉敷美観地区	カンジ, ランチ, トートバッグ, アイス, バッグ, ポストカード, デニムストリート, 町並み, 景色, 子供, 小物, 部屋, 映画, 駅, 夕方, 周り, 老舗, アーケド通り, 楽しみ方, 遊覧川舟, ネコカフェ, 駐車場, お土産物屋, 時間, 途中商店街, にぎわう美観地区, 倉敷ジーンズストリート
吉備津彦神社	感覚. 人柄. 吉備津. お参り. レンタサイクル. 伝統. 桃太郎さん. 夫婦. 陶器市. 桃太郎伝説. 神主. 鬼. 駅. 備前の国の神社
吉備津神社	した神社. 方たち. 喜び. 時期. パワースポット. 左手. 地方討伐. 娘. 関係. 写真. 回廊沿い. 桜. 雷. 道. 境内. 敷地. 建築物. 吉備津神社
大原美術館	現代美術, 西洋絵画, 私立西洋近代美術館, 意味, ピカソ, 雰囲気, グッズ, コロナ禍, 地元人, 夕方, 大勢, 倉敷美観地区, オリエント館, 芸術家, 展示, 作品, 鑑賞者, 絵画, 中庭, 入場チケット
岡山後楽園	岡山城と後楽園, 日本三庭園, 庭園, 素晴らしさ, 整備, 訪問, 家族旅行, 開園時間, 観光ボランティア, 抹茶カフェ, 春夏, 桜, 紅葉時期, タンチョウ鶴, 小川, 池, ツツジ, メダカ, おすすめスポット, 夕暮れ, バス停, ボランティアガイド, 時期, 価値, 気, 敷地, 川, 茶店屋, 夜風
池田動物園	動物園, 迫力タップリ, 移動距離, 気持ち, イベント, バス停, ホワイトライオン, 鹿コーナー, ほか, 車, 手ぶり, 動物達
備中松山城	猫城主, 城郭, 岩, 景色, 幅広道, シャトルタクシー, 天守閣, 遊歩道, トイレ, マイカー, 手前, 高校生たち, 山奥, 風情, 帰り, さんじゅーろー, 城見橋公園駐車場, 観光案内, 登城, 運動靴, 現存天守, 日本三大山城, 城自体, 武家屋敷どおり, 備中松山城
神庭の滝	滝そのもの, 韮山高原探訪のあと湯原インターチェンジ, 岩, 川魚, 子供, 山道, 店, 景色, 夏, 雰囲気, 見どころ, 遊歩道, 想像, 迫力, 数, 禁止, 楽しみのお猿, 美しさ, 見学料金, 観光客, 私達家族

観点となっているものも，比較手法では「チーズたっぷり」のように情報が追加されて抽出される場合が多かった．

本稿では抽出した観点の中に含まれる固有表現の数も数えた．表 4 から提案手法では抽出した全観点 446 の内 7 の観点が固有表現であり，さらにその全てが評価 A に相当する観点であった．この固有表現には例えばスポット「大原美術館」の「ピカソ」や神庭の滝の「湯原」などがある．一方表 5 の比較手法では，抽出した全観点 516 の内 53 の観点が固有表現であり，その内 42 が評価 A，11 が評価 D となった．評価 A となった観点には例えばスポット「備中松山城」の「さんじゅーろー」や

表 4　提案手法で抽出した観点の評価結果

	全評価の数	A　(割合)	B　(割合)	C　(割合)	D　(割合)	E　(割合)
観点	446	238　(0.534)	114　(0.256)	34　(0.076)	38　(0.085)	22　(0.049)
		352(0.789)		94(0.211)		
固有表現	7	7　(1.000)	0　(0.000)	0　(0.000)	0　(0.000)	0　(0.000)
		7(1.000)		0(0.000)		
固有表現の割合	0.016	0.030	0.000	0.000	0.000	0.000

表 5　比較手法で抽出した観点の評価結果

	全評価の数	A　(割合)	B　(割合)	C　(割合)	D　(割合)	E　(割合)
観点	516	289　(0.560)	103　(0.200)	31　(0.060)	63　(0.122)	30　(0.058)
		392(0.760)		124(0.240)		
固有表現	53	42　(0.792)	0　(0.000)	0　(0.000)	11　(0.208)	0　(0.000)
		42(0.792)		11(0.208)		
固有表現の割合	0.103	0.145	0.000	0.000	0.175	0.000

スポット「倉敷美観地区」の「倉敷ジーンズストリート」などがある．評価 D となった観点には例えばスポット「吉備津神社」の観点「吉備津神社」などがある．

表 4 から提案手法では 446 の 78.9%である 352 がふさわしい観点であり，表 5 から比較手法では 516 の 76.0%である 392 の観点がふさわしい観点であった．この結果から提案手法は比較手法に比べて少し高い精度でふさわしい観点が抽出できたが，比較手法の方が多くの観点を抽出した．また前述から比較手法で抽出した観点の方がより具体的である．具体的な観点を用いてスポットを分析する場合は，スポットを限定的に分析することは可能であるが幅広いスポットを分析することはできない．例えば,「ソーセージ手作り体験」の観点から観光スポットを分析しても，ほとんど情報を得られない．一方抽象度の高い「体験」の観点を用いることで多くのスポットをこの観点で分析でき，その中には「ソーセージ手作り体験」の特徴を持つスポットも含まれる．このような抽象度が高い観点を抽出できるのが提案手法の利点である．

評価段階 E は辞書などを用いることで削減できる見込みがあるが，評価段階 C や評価段階 D を削減し観点自動抽出の精度を上げるには，エンティティの抽出方法や単語ベクトルの学習データなどをさらに検討する必要がある．

5 ま と め

本稿では観光レビューから観光スポットの観点を自動抽出する手法を提案した．提案手法を用いた実験では，旅行サイトのじゃらん net の観光レビューからエンティティを抽出し，Natural Langual API を利用してその分散表現を求めた．次に分散表現を用いてエンティティをクラスタリングし，生成されたクラスタのクラスタベクトルに類似した単語を出力した．観光スポット毎にこのような類似語のリストを作成し，その中から観点を抽出した．

岡山県の 30 スポットのレビューから観点を自動抽出した結果，スポット「大原美術館」の「ピカソ」やスポット「ブラジリアンパーク鷲羽山ハイランド」の「異国風」や「絶叫系」，スポット「岡山後楽園」の「桜」や「紅葉」などのように適切な観点が抽出される一方，スポット「神庭の滝」の「宿店」，スポット「倉敷美観地区」の「前回」などのように不適切な観点も少ないながらも抽出された．

比較実験では類似語ではなくエンティティから観点を抽出し，提案手法を用いた実験との比較をした．提案手法で抽出した 446 の観点の内観点としてふさわしいものは 352 で割合としては 78.9%となった．一方比較手法で抽出した観点は 516 でその内観点としてふさわしいものは 392 で割合としては 76.0%となり，提案手法と比較手法にあまり大きな差は見られなかった．しかし，提案手法は比較手法に比べてより抽象的な観点が多く抽出された．このことから，提案手法は比較手法に比べて観光スポットを幅広く分析する観点を抽出できる．

今後は複数回抽出されたエンティティには重みを付与してクラスタリングする方法を用いて，提案手法との比較をしたい．その他にエンティティのクラスタリングに利用する Ward 法の不整合係数の閾値や，類似語リストの作成の際設定する閾値 (クラスタ内の最小エンティティ数や出力類似語の数など) についてさらに検討する必要がある．

文　献

[1] 野本輝，上野史，太田学，“観光レビュー文を用いた穴場スポットの発見，”DEIM 2022 B43–3，2022.

[2] 杉浦孔明，岩橋直人，芳賀麻誉美，堀智織，“観光スポット推薦アプリ「京のおすすめ」を用いた長期実証実験，”観光情報学会誌「観光と情報」，vol. 10，pp. 15–24，2014.

[3] 市村哲，上石萌恵，堀口莉里花，“口コミ解析と好み診断に基づいた旅行先推薦，”情報処理学会 グループウェアとネットワークサービス研究会報告，vol. 2020-GN-110，no. 11，pp. 1–8，2020.

[4] 杉本祐介，水野忠則，菱田隆彰，“口コミに含まれる感情語を利用した観光地分類の検討，”マルチメディア，分散協調とモバイルシンポジウム 2014 論文集 (DICOMO2014)，pp. 1345–1350，2014.

[5] 市村哲，“口コミから美味しい料理店を手早く探すシステム，”情報処理学会論文誌，vol. 61，no. 11，pp. 1748–1756，2020.

[6] 金子貴美，村上浩司，石野亜耶，“商品レビューからの購買意図の抽出と自動階層化，”人工知能学会全国大会論文集，vol. 30，no. 2P1-13in1，pp. 1–4，一般社団法人 人工知能学会，2016.

[7] 小川哲司，“テキストマイニングとネットワーク分析を用いた映画評価の要因分析，”経済経営論集，vol. 29，no. 2，pp. 26–35，2022.

[8] 田村亮介，白石絵里奈，浅沼爽汰，藤田和成，町田翔，白井聡一，延澤志保，“類似アプリ比較のためのレビュー自動分析，”情報処理学会 第 81 回全国大会講演論文集，pp. 125–126，2019.
[9] 上原尚，嶋田和孝，遠藤勉，“Web 上に混在する観光情報を活用した観光地推薦システム，”電子情報通信学会技術研究報告書，vol. 112，no. 367，pp. 13–18，2012.
[10] 淀川翼，加登一成，伊東栄典，“単語の分散表現を用いた文書クラスタのラベル推定，”人工知能学会第 49 回セマンティックウェブとオントロジー研究会 (SIG-SWO)，vol. 49，no. 3，pp. 03-01 – 03-05，2019.
[11] Ward，J. R. Jr.，“Hierarchical Grouping to Optimize an Objective Function，”Journal of the American Statistical Association，vol. 58，no. 301，pp. 236–244，1963.
[12] MathWorks，階層クラスタリング，https://jp.mathworks.com/help/stats/hierarchical-clustering.html