学術研究論文

観光の形態に基づいた旅行ブログエントリの自動分類と可視化

柴田有基[†1]，篠田広人[†1]，難波英嗣[†2]，石野亜耶[†3]，竹澤寿幸[†1]
[†1] 広島市立大学大学院情報科学研究科
[†2] 中央大学理工学部
[†3] 広島経済大学メディアビジネス学部

あらまし　本研究では，旅行ブログエントリ中のテキストおよび画像情報を用いて 6 種類の観光の形態に分類する手法を提案する．分類の際，さらに，エンティティリンキング技術を用いて，旅行ブログエントリから自動的にリンクされた Wikipedia エントリの情報も併せて用いる．なぜならば，旅行ブログエントリの分類に有益な情報がしばしばリンク先の Wikipedia エントリ内に記述されるからである．本稿では，これらの情報を，深層学習ベースの手法で統合する手法を提案し，実験により，精度 0.768，再現率 0.233，F 値 0.358 を得た．最後に，観光の形態に分類された旅行ブログエントリを地図上にマッピングすることで，観光情報を検索できるシステムを構築した．

キーワード　旅行ブログエントリ，観光の形態，文書分類

Classification and Visualization of Travel Blog Entries Based on Types of Tourism

Naoki Shibata[†1], Hiroto Shinoda[†1], Hidetsugu Nanba[†2], Aya Ishino[†3], Toshiyuki Takezawa[†1]
[†1]Graduate School of Information Sciences, Hiroshima City University
[†2]Faculty of Science and Engineering, Chuo University
[†3]Faculty of Media Business, Hiroshima University of Economics

Abstract　We propose a method for classifying travel blog entries into one or more tourism types among six predetermined types by using textual and image information in each entry. Together with this information, we use Wikipedia entries, which are automatically linked from each travel blog entry by entity-linking technology, because information beneficial for classifying blog entries is often mentioned in Wikipedia entries, and we combine this information by using a deep-learning-based method. We conducted an experiment with a neural network using three types of input data, and obtained precision, recall, and F-measure scores of 0.768, 0.233, and 0.358, respectively. Finally, we constructed a system that enables travelers to look for travel blog entries from a map in terms of tourism type.

Keywords: Travel Blog, Tourism Type, Document Classification

1. はじめに

近年, 観光がより生活に身近な存在となっている．このような流れもあり，従来の娯楽を追求するのみの観光だけではなく，様々なニーズに合わせた観光の形態が多く誕生している．例えば，温泉や登山など，健康の回復や維持，増進につながる観光は「ヘルスツーリズム」, 地域の祭りや着物体験など, その土地の文化を体験することを目的とした観光は「カルチュラルツーリズム」と呼ばれる．多様な観光の形態への取り組みは，これまで観光地として注目されてこなかった地域に対し，新たな雇用の創出や地域の活性化などに貢献するものである．

旅行者が観光の形態に沿った観光体験をしたいと考える場合や，行政が観光の形態を利用して観光政策を考える場合，まず，その地域で体験できる観光の形態の情報を調べる必要がある．しかし，観光の形態に関する情報を収集しようとしても，この観点の検索システムが存在しないため，容易に収集することができない．そこで本研究では，このような情報を自動で収集するため，6 種類の観光の形態を定義し，大量の旅行ブログエントリをこれら 6 種類の観光の形態に自動分類する手法を提案する．旅行ブログエントリを観光の形態に自動分類することが可能になれば，世界各地の観光地でどのような形態の観光が可能か調べることができるほか，特定の形態に基づいた観光地の推薦や旅行計画も可能になる．

ある観光地における旅行者の情報を知るための従来の方法の 1 つに，旅行者に対して直接アンケートを実施する方法がある．アンケートでは，知りたい情報に関する設問を用意することで，欲しい情報が手に入りやすいというメリットがあるが，多くの時間やコストが必要になるというデメリットが存在する. そこで, 近年, ソーシャルメディアである Twitter や Web 上で公開されている旅行記, すなわち旅行ブログエントリを収集・分析するという方法が広まりつつある．特に，旅行ブログエントリには，各観光地における体験談や画像など，詳しくまとまった情報を持つものが多いことから，本研究では旅行ブログエントリを分析の対象とする．

本論文の構成は以下の通りである．2 章では，本研究に関する関連研究について述べる．3 章では，観光の形態の定義, 自動分類の方針について述べる．4 章では，実験内容とその結果，考察について述べる．5 章では，分類した結果の可視化について述べる．6 章で本論文のまとめを述べる．

2. 関連研究

本研究では，テキスト情報と画像情報に加えて，テキストに含まれる単語集合に対し，Wikification[1, 2]を行うことで得られた Wikipedia のエンティティ情報を利用し，旅行ブログエントリを観光の形態に基づいて自動分類する．Wikification とは，エンティティリンキング技術を用いて，テキストに含まれる単語と関連のある Wikipedia のエンティティをリンク付けすることを意味する．本研究では，これらの情報を考慮した深層学習ベースの手法で旅行ブログエントリを分類し，その分類結果を地図上にマッピングすることで観光の形態に関する情報を検索することができるシステムを構築するが，このような研究や取り組みはこれまでにも行われている．

2.1. 観光の形態に関する取り組み

ヘルスツーリズムに関する取り組みとして，河行ら[3]は島根県大田市の自然や特産物を利用し，健康促進を促すヘルスツーリズムを推進している．例えば，島根県大田市の地域資源である国立公園三瓶山を利用したハイキングやワサビなどの地元食材を用いたヘルシー食を提案している．このようなヘルスツーリズムの取り組みは，健康促進による医療費の削減や雇用創出などの観光業による地域の発展につながるとして期待されている．

また，藤井ら[4]は，日本のインフラツーリズムを取り上げた観光ガイドブックを通してインフラツーリズムの魅力を紹介している．このガイドブックでは，群馬県の八ッ場ダムについて取り上げられており，2017 年 8 月までに累計 20 万人が八ッ場ダムを訪れていることが報告されている．また，国土交通省は,「インフラツーリズム PORTAL SITE[1]」を公開しており，ここで日本全国のインフラツーリズムに関する情報を調べることができる．インフラツーリズムという観光の形態に沿って観光地に関する情報を検索することができるシステムを構築するという点では類似しているものの，本研究は世界中の観光地の情報が含まれる旅行ブログエントリを用いているため, 本研究で提案するシステムを用いることで, 日本だけでなく世界各地の観光情報を調べることができる．これらのことからわかるように，多様な目的から構成される観光の形態は，非常に注目されている．本研究では，このような観光の形態に基づいて旅行ブログエントリを分類する．

2.2. 文書の分散表現

文書の分散表現の作成手法に関する研究として，

[1] https://www.mlit.go.jp/sogoseisaku/region/infratourism/

Iyyer ら[5]は, 文書に含まれるそれぞれの単語の分散表現を平均することで生成される新たな分散表現を文書の分散表現とし，これを入力データとした複数の層から構成されるニューラルネットワークで文書分類を行う手法 DAN (Deep Averaging Networks) を提案している．この手法の特徴は，単純で理解しやすい上，計算時間が短いにも関わらず，語順を考慮する複雑な手法と変わらない精度が出ることである．

また，そのほかの手法として，Shen ら[6]は，文書に含まれる単語の分散表現に簡単な処理を行うことで文書の分散表現を作成する手法 SWEM (Simple Word-Embedding Model) を提案している．具体的には，分散表現の各次元の平均を用いたもの (SWEM-aver) や最大値を用いたもの (SWEM-max)，平均と最大値を結合したもの (SWEM-concat)，n-gram で平均を計算し，それに対し最大値を用いたもの (SWEM-hier) を提案し，Iyyer らの提案する手法 DAN と同様，様々なタスクにおいて，複雑な手法と同等程度の精度を達成している．本研究では，実験の提案手法とベースライン手法において，Shen らが提案するこれらの手法の中から，比較的良い精度が報告されている SWEM-concat を文書の分散表現の作成手法に用いる．

2.3. 観光関連文書の分類

観光関連文書の分類について，Takahashi ら[7]は，Twitter に投稿された旅行ツイートに対し，「観光」，「ビジネス」，「食事」，「購買」に分類する手法を提案している．この手法は，本研究と同様に，観光関連のソーシャルメディアを用いて分類を行っているという点では類似しているものの，分類の観点が異なっているほか，分類の対象が本研究で扱う旅行ブログエントリに対して，ツイートを用いているという点で異なっている．

また, 藤井ら[8]は, 旅行ブログエントリを,「買う」,「食べる」,「体験する」,「泊まる」,「見る」に分類する手法を提案している．この手法については，本研究と同様に旅行ブログエントリを対象としている点で非常に類似している．しかし，分類の観点が旅行者の行動，つまり旅行者が何をやっているかに基づいたものであるため，本研究の「観光の形態」に着目した分類と関連性はあるものの，基本的には別の観点であると考えられる．また，藤井らの分類手法に加えて，本研究で提案する観光の形態に基づく分類が可能になれば，例えば,「広島県」の「カルチュラルツーリズム」で「食べる」に関する情報を調べる，といったように，よりきめ細かい検索が可能になると考えられる．

3. 観光の形態に基づいた旅行ブログエントリの自動分類

3.1. 観光の形態の定義

観光の形態には，厳密な定義がないものや定義が曖昧なため線引きが難しい形態が存在する．そのため，本研究では，参考文献[3, 4, 9, 10, 11]を参考に，6 つの観光の形態を取り上げ，独自に定義付けを行った．観光の形態とその定義，またその具体例を表 1 に示す．本研究では，旅行ブログエントリから旅行者の観光の形態を明らかにするため，英語で書かれた旅行ブログエントリをそれぞれの観光の形態に自動分類する．

ところで，観光の形態には，定義した 6 種類のほかに，ダークツーリズム[12]やコンテンツツーリズム[13]など数多く存在する．これらの多くの観光の形態に対し，本研究では，分類結果から有益な情報が得られること, 旅行ブログエントリの数が 40 件以上あること，この 2 点で議論を行い，どちらも当てはまる観光の形態を本研究で扱う 6 種類とした．

3.2. 観光の形態に基づいた旅行ブログエントリの自動分類

本研究は，3.1 節で示した観光の形態に基づいて，旅行ブログエントリを自動分類する. 本節では, 3.2.1 節で自動分類の方針，3.2.2 節で機械学習を用いた旅行ブログエントリの自動分類について説明する．

3.2.1. 自動分類の方針

自動分類の基本的な方針として，各旅行ブログエントリ中のテキストと画像，テキストに含まれる単語に関連のある Wikipedia の情報を解析し，それらの結果を用いて観光の形態に自動分類する．まず，テキストのみの場合,「世界遺産の原爆ドームを見た」という文章であれば,「世界遺産」という単語が含まれることから，この旅行ブログエントリは，ヘリテージツーリズムであると考えられる．このように，旅行ブログエントリを観光の形態に分類する場合，テキスト情報は重要な判断基準になることが考えられる．

人手で旅行ブログエントリを観光の形態に分類する場合，判断の根拠として，テキスト情報のほかに画像情報が重要な判断材料となることが多い．そのため，自動分類の際に画像を考慮することができれば，テキスト情報のみを利用した分類よりも精度の高い分類の実現が期待される．例えば，旅行者がスキーを体験した場合，これはスポーツツーリズムとなるが，旅行ブログエントリのテキスト中に「スキーをしたかったけど天候が悪くできなかった」という記述がある場合,「スキー」という単語が含まれて

表 1: 観光の形態の定義と具体例

アイコン	観光の形態	定義	例
	インフラ， ハードツーリズム[4]	近代的な建造物や娯楽施設を対象にした観光.	橋，ダム，テーマパーク，ショッピングモール，水族館，博物館，動物園
	ヘルスツーリズム[3]	心身を癒すことや散歩などの軽い運動を通して健康維持を目的とした観光.	宗教的巡礼，温泉，ハイキング，トレッキング
	スポーツツーリズム[9]	スポーツを体験または観戦することを目的とした観光.	MLB，プロ野球，サッカー
	グリーンツーリズム[10]	自然と触れ合うことを目的とした観光.	農業（漁業）体験，フルーツ狩り，ピクニック
	ヘリテージツーリズム[10]	世界遺産や歴史的な建築物を対象にした観光.	世界遺産，国宝，寺，神社，城
	カルチュラルツーリズム[11]	それぞれの地域の生活や文化，民族，伝統などを対象にした観光.	着物体験，神楽，祭り，初詣

いるものの，ブログ著者は実際にはスキーをしていないため，スポーツツーリズムとは言えない．これに対し，もしテキストからブログ著者がスキーをしたかどうか判定できない場合であっても，スキーをしている様子や雪などのスキー関連の物体が画像中に含まれていることがわかれば，ブログ著者はスキーを体験した可能性が高いと推測できるため，画像中に写っているものは，旅行ブログエントリを観光の形態に分類する際に重要な判断材料になると考えられる．そこで，本研究では画像に対して物体検出を行い，得られた物体の単語集合を分類に用いる．画像からの物体検出には，Google Cloud Vision API[2]を用いる．Google Cloud Vision API では，画像を数千のカテゴリに分類することや物体・顔検出などを行うことができる．本研究では，分類器の入力データに，この Google Cloud Vision API の物体検出結果から得られた単語集合を用いる．

人手で分類を行う際に根拠となる情報源は，上記で述べたテキストや画像のどちらかとなるが，旅行ブログエントリの内容によっては，これらの情報から読み取ることができたものに関する外部知識が必要になることがある．例えば，「原爆ドームを見た」という文章では，「世界遺産」という単語が含まれていないことから，ヘリテージツーリズムに正しく分類することが難しいと考えられる．この場合，正しく分類するためには，原爆ドームが世界遺産であるという情報が必要になる．そこで本研究では，Wikification の技術を用いて Wikipedia の情報を利用することを考える．Wikification には，Google Cloud Natural Language API[3]を利用する．Google Cloud Natural Language API では，テキストを API 経由でクラウドに送ると，形態素解析，構文解析，固有表現抽出に加え，Wikification も行われることから，リンク先の Wikipedia の該当ページから原爆ドームが世界遺産であるという情報が得られることが期待される．このことから，本研究では，旅行ブログエントリに含まれるテキストと画像に対する物体検出の結果に加えて，外部知識として Wikification の結果から得られた Wikipedia の情報を分類の際に考慮することで，より精度の高い分類の実現を目指す．

3.2.2. 機械学習を用いた旅行ブログエントリの自動分類

本研究では，まず，分類対象の旅行ブログエントリに含まれる画像に対し，画像認識技術を用いて物体検出を行う．次に，Wikification を行い，リンク付けされた Wikipedia のエンティティ名を抽出する．そして，物体検出の結果として得られた単語集合と Wikification から得られた Wikipedia のエンティティ名の集合，旅行ブログエントリのテキストを入力とし，それぞれの入力データを考慮した分類器を構築する．分類器の概略図を図 1 に示す．本研究では，旅行ブログエントリ中のテキストから得られた単語集合，画像に対して物体検出をすることで得られた単語集合およびテキストからリンク付けされた Wikipedia のエンティティ名の集合の 3 つのそれぞれの入力データに対し，まず文書の分散表現を作成

[2] https://cloud.google.com/vision/?hl=ja
[3] https://cloud.google.com/natural-language/?hl=ja

する. 次に, それぞれの分散表現をニューロン数 300, 100 から構成される 2 層の中間層に通し，これらの出力から得られた各 100 次元の分散表現を結合することで 300 次元の新たな分散表現を獲得する．そして最後に，この分散表現をニューロン数 300 から構成される 1 層の中間層に通し，その結果から分類結果を出力する．なお，本研究では，旅行ブログエントリを複数の観光の形態に分類する手法として，2 値分類器を観光の形態と同じ数用意することで実現する．そのため，例えばヘリテージツーリズムの分類器の場合, 出力は「ヘリテージツーリズムである」と「ヘリテージツーリズムではない」の 2 つになる．

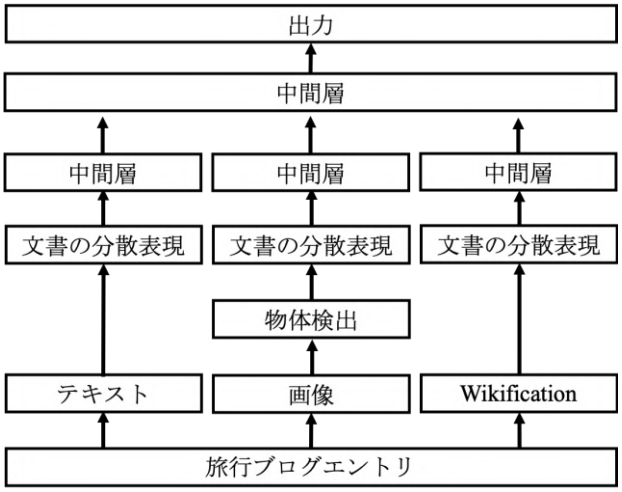

図 1: 分類器の概略図

旅行ブログエントリの内容によっては，テキストや画像に対しての物体検出の結果などのそれぞれの入力データに含まれる単語数に偏りがある．そのため，あらかじめ結合したものを入力データとする場合，考慮する情報の重みに偏りが生じると考えられる．例えば，テキスト中の単語数が非常に多く，そのほかの画像に対する物体検出や Wikification から得られた単語数が非常に少ない場合，これらの入力データを結合するとほとんどテキスト情報のみになることが推測される．このことから，本研究では，図 1 のようにそれぞれの入力データごとに中間層で処理を行う構成の分類器を採用した．

4. 実験

4.1. 実験条件

【実験に用いるデータ】

旅行ブログエントリの収集には，TravelBlog[4]を用いる．TravelBlog とは，世界最大規模の旅行ブログサイトであり，様々な言語で書かれた旅行ブログエントリが掲載されている．本研究では，この TravelBlog から，2006 年 4 月 29 日～2014 年 9 月 4 日に投稿された約 240,000 件の英語で書かれた旅行ブログエントリの収集を行い，ここからランダムに抽出した 1,909 件を実験に用いた．なお，TravelBlog では，旅行ブログエントリを投稿する際に，訪問地に関する情報をあらかじめ決めて投稿する仕様となっている．その情報から抽出した 1,909 件の地域別の割合は，North-America: 51.9%，Asia: 39.1%，Africa: 8.7%，その他: 0.2%であった．地域によって偏りがあるが，これは TravelBlog のユーザの偏りが原因の 1 つであると考えられる．また，機械学習に用いる教師データの作成については，英語の読み書きが不自由なくできる大学生 1 人に依頼し，6 種類の観光の形態に人手で分類を行った．判定者が判定に迷った場合は，著者らと議論し，多数決を行うことで判定を行った．

表 2: 人手で分類した結果の内訳

観光の形態	件数
インフラ，ハードツーリズム	156
ヘルスツーリズム	116
スポーツツーリズム	54
グリーンツーリズム	421
ヘリテージツーリズム	177
カルチュラルツーリズム	40
分類した旅行ブログエントリの総数	1,909

人手で分類した結果の内訳を表 2 に示す．6 種類の観光の形態における旅行ブログエントリの総和が 1,909 ではないが，これは 1 つの旅行ブログエントリに，複数の観光の形態が付与されることや逆に 1 つも付与されないこともありうるという設定にしているためである．なお，複数の観光の形態が付与された旅行ブログエントリは 209 件，1 つも付与されなかった旅行ブログエントリは 1,182 件であった．複数の観光の形態が付与された旅行ブログエントリの内容を見てみると，1 つの投稿に「着物体験」,「花見」および「銀閣寺の見学」など複数の観光体験や複数日程の出来事を 1 つの投稿にまとめて記述しているものが多く見受けられた．また，1 つも付与されなかった旅行ブログエントリでは，旅行とは関連のない投稿や旅行に行く直前の投稿が多く確認された．

【実験条件】

プログラミング言語には Python3 を用い，分類器の実装を行う．提案手法とベースライン手法に用いる機械学習のニューラルネットワーク (以下 MLP) とサポートベクターマシーン (以下 SVM) には，そ

[4] https://www.travelblog.org/

れぞれ Chainer と scikit-learn を用いて実装を行う．入力データで用いる単語の分散表現については，テキストと Google Cloud Vision API で得られた画像解析結果で得られる単語集合に対しては，約 240,000 件の TravelBlog に含まれる約 3 億単語で学習した 300 次元の Word2Vec モデルを用いる．また，Google Cloud Natural Language API による Wikification で得られた Wikipedia のエンティティ名の集合に対しては，Wikipedia2Vec[14]の事前学習モデル[5]を用いることで，分散表現を獲得する．なお，テキスト中の単語集合に対しては，分類に有効であると考えられる名詞のみを入力データとして採用する．また，1,909 件の旅行ブログエントリに含まれる画像の数の平均は，1.914 枚，1 つ以上の観光の形態が付与された旅行ブログエントリに含まれる画像の数の平均は，3.558 枚であった．

【評価尺度】

観光の形態ごとに構築した 2 値分類器で分類を行い，その結果に対して精度，再現率および F 値を算出することで評価を行う．評価値の算出には，まず，各観光の形態で 5 分割交差検定を行い，その結果から，各観光の形態の正例の数のばらつきを考慮するため，micro 平均を用いて全ての観光の形態に対する評価値を算出する．それぞれの観光の形態の 5 分割後の数を表 3 に示す．本研究では，これらのデータを用いて各観光の形態ごとに 5 分割交差検定を行う．なお，本研究で扱う TravelBlog には，700,000 件以上のブログエントリが 2013 年 2 月時点で投稿されている．そのため，旅行ブログエントリの数は十分に存在すると考え，本研究では精度を重視し，分類器のパラメータの調整および実験結果の評価を行う．

【比較手法】

本実験では，下記の 2 種類の提案手法と 5 種類のベースライン手法で実験を行った．なお，noun，img および wiki はそれぞれ，テキスト中の名詞，物体検出の結果から得られた単語集合および Wikification で得られたエンティティ名の集合を表す．

提案手法

- MLP(noun+img+wiki): テキスト中の名詞，物体検出の結果から得られた単語集合および Wikification で得られたエンティティ名の集合の 3 つの入力データに対し，SWEM-concat を用いて文書ベクトルをそれぞれ作成し，これらのベクトルを入力データとするニューラルネットワークで分類を行う．
- MLP(noun+img): テキスト中の名詞および物体検出の結果から得られた単語集合の 2 つの入力データに対し，SWEM-concat を用いて文書ベクトルをそれぞれ作成し，これらのベクトルを入力データとするニューラルネットワークで分類を行う．

ベースライン手法

- MLP(noun), MLP(img), MLP(wiki): テキスト中の名詞，物体検出の結果から得られた単語集合および Wikification で得られたエンティティ名の集合から，それぞれ 1 つの入力データに対して，SWEM-concat を用いて文書ベクトルを作成し，そのベクトルを入力データとするニューラルネットワークで分類を行う．
- SVM(noun): テキスト中の名詞に対し，Bag-of-Words で生成したベクトルを入力データとする SVM で分類を行う．カーネル関数などのパラメータについては，各分類器に対し，精度優先の grid search を行い，その値を採用する．
- SVM(img): 物体検出の結果から得られた単語集合に対し，Bag-of-Words で生成したベクトルを入力データとする SVM で分類を行う．カーネル関数などのパラメータについては，各分類器に対し，精度優先の grid search を行い，その値を採用する．

表 3: 各観光の形態の 5 分割後の内訳

観光の形態	①	②	③	④	⑤	合計
インフラ，ハードツーリズム	14	28	33	39	42	156
ヘルスツーリズム	31	24	18	27	16	116
スポーツツーリズム	10	13	7	11	13	54
グリーンツーリズム	85	89	67	92	88	421
ヘリテージツーリズム	9	20	45	51	52	177
カルチュラルツーリズム	0	3	11	12	14	40
合計	149	177	181	232	225	964

[5] http://wikipedia2vec.s3.amazonaws.com/models/en/2018-04-20/enwiki_20180420_300d.pkl.bz2

4.2.　実験結果と考察

4.1 節で説明したそれぞれの提案手法とベースライン手法による実験結果を表 4 に示す．表 4 より，精度ではベースライン手法の SVM(noun)と提案手法の MLP(noun+img+wiki)でそれぞれ 0.761, 0.768 が得られた．この実験結果から，精度では明らかな有意差が確認できなかったものの，再現率ではベースライン手法の SVM(noun)で得られた 0.156 に対して，提案手法の MLP(noun+img+wiki)では 0.233 が得られた．このことから，再現率が向上したという点で提案手法の有効性が示されたと言える．また，提案手法の MLP(noun+img)では，再現率と F 値において，全ての手法の中で最も高い値 0.274，0.396 が得られた．これらのことから，提案手法の複数の入力データを考慮することの有効性が示されたと考えられる．

表 4: 旅行ブログエントリの分類結果

手法	精度	再現率	F 値
MLP(noun+img+wiki) (proposed)	**0.768**	0.233	0.358
MLP(noun+img) (proposed)	0.717	**0.274**	**0.396**
MLP(noun)	0.669	0.170	0.271
MLP(img)	0.739	0.232	0.354
MLP(entity)	0.676	0.104	0.180
SVM(noun)	0.761	0.156	0.258
SVM(img)	0.591	0.198	0.297

ところで，MLP(img)と MLP(noun+img)を比較すると，精度が下がったものの，トレードオフの関係である再現率と F 値が上がっていることがわかる．これは分類の際に考慮する情報の種類がテキスト中の名詞の分多くなったため，拾うことができる旅行ブログエントリの件数が増えたと考えられる．また，MLP(noun+img+wiki)に関しては，ベースライン手法の SVM(noun)とは同等程度の精度となったものの，そのほかのベースライン手法や提案手法より高い精度が得られた．これは，テキスト情報と画像情報に加え，Wikipedia の情報が加わったことにより，分類を行う際に判断根拠となりうる補助的な情報が加わったことで，精度が向上したと考えられる．

ここで，考慮する入力データの種類を増やすことで，結果に変化があった旅行ブログエントリについて実際に見てみる．まず，ベースライン手法の MLP(noun)と提案手法の MLP(noun+img)をスポーツツーリズムで調査を行った．実際に，正しく分類できるようになった旅行ブログエントリを見てみると，野球観戦をしたという内容の旅行ブログエントリが複数存在しており，これらの旅行ブログエントリに含まれる画像の物体検出の結果を確認してみると，野球や野球場という単語が多く検出できていることがわかった．また，その中から，ある旅行ブログエントリの 1 つを詳しく見てみると，テキスト中の名詞の数 321 に対し，“baseball”という単語が 1 つしか含まれていなかったが，画像解析結果だけに注目すると，多くの野球関連の単語が抽出できていたことが確認できた．このようにテキスト情報だけでなく画像情報を用いることで，考慮する情報の種類が増え，より正確な分類が可能になっていると考えられる．

次に，提案する 2 つの手法，MLP(noun+img)と MLP(noun+img+wiki)の結果から，Wikification を考慮することで正しく分類できるようになった旅行ブログエントリの内容を見てみる．インフラ，ハードツーリズムの結果の変化に対して調査をしたところ，博物館を訪れた旅行者が書いた旅行ブログエントリにおいて Wikification の結果を考慮することで分類できるようになったものがあった．実際に，この旅行ブログエントリの中身を見てみると，テキスト中に博物館を表す単語が含まれているほか，博物館の展示物と思われる画像も多く含まれていた．しかし，これらの画像に対しての物体検出の結果を確認したところ，物体検出の結果から，これが博物館の展示物であると推定することが難しいことがわかった．また，実際に人が画像を見た場合であっても，これをもとに博物館と判断するのは難しいものであった．よって，これらのことから，画像の情報を追加したことで，「インフラ，ハードツーリズムではない」と分類してしまったと考えられる．これに対し，Wikification の結果を見てみると，博物館の固有名詞をしっかりと抽出できていることがわかった．このことから，入力データのテキストと画像解析結果に Wikification の結果を追加することで，分類の際に考慮する情報を補完することができたと考えられる．

ここで，1 つの入力データのみを利用した分類に注目すると，画像に対しての物体検出で得られた単語集合を用いた MLP(img)で比較的高い精度 0.739 が得られていることがわかる．しかし，物体検出の結果は，テキストから得たテキスト情報と Wikipedia の情報とは異なり，全て正しく検出できているとは限らない．そこで，実際にどのような単語が分類に有効であったかどうかを調査するため，ヘリテージツーリズムのデータを用いて，正例との単語の共起頻度と MLP(img)による分類精度を用いて，どのような単語が分類に有効であったかどうかを検証する．MLP(img)による分類精度での検証については，1,909

件の旅行ブログエントリに含まれる画像から物体検出で得られた単語集合に対して，1 つずつ単語を抜粋し，分類精度を確認していく．これにより，ある単語を抜粋することで分類精度が下がった場合，その単語は分類の際に重要な単語であると言える．

表 5: ヘリテージツーリズムに対する単語の共起頻度

単語	検出された件数	
	正例	負例
of	70	222
building	55	64
site	55	25
tourism	53	61
temple	51	13

まず，物体検出で得られた単語集合の中から，ヘリテージツーリズムの旅行ブログエントリに共起した単語，上位 5 つとそれに対応する旅行ブログエントリの正例･負例の数を表 5 に示す. 表 5 中の単語，“of”については，物体検出の結果を確認すると，船やバスなどの交通機関表す“mode of transport”で多く検出されていた. また，正例と負例の数に注目すると，“of”, “building”および“tourism”については，正例・負例問わず多くの旅行ブログエントリから検出されていたが，“site”と“temple”については，負例に比べて正例に出現する数が多いことがわかる．この 2 つの単語から特に正例で多く検出された“temple”について調べてみると，京都の金閣寺や銀閣寺，カンボジアのアンコールワットなどのアジア地域の世界遺産を写した画像から多く検出されていることがわかった．このように，物体検出から得られる単語を考慮することは，アジア地域におけるヘリテージツーリズムを分類する際に有効であることが考えられる．

次に，上記で述べたような物体検出の結果から得られた単語を分類器が上手く学習しているか調べてみる．実験で用いた 1,909 件の旅行ブログエントリに含まれる画像に対して物体検出を行った結果から，1,353 種類の単語が確認できた．そこで，これらの単語から 1 単語を抜粋した実験データを用い，MLP(img)の分類精度を確認するという作業を 1,353 単語分繰り返し行うことで検証を行った．分類精度がより下がった単語 5 つ，その条件下での分類精度およびその単語が含まれる旅行ブログエントリの正例・負例の数を表 6 に示す．この表から，“forest”を抜粋した場合が最も精度が下がっていることがわかる．そのため，“forest”は分類の際に重要な単語であることが考えられるが，正例に含まれている件数が少なく，逆に負例に含まれる件数の方が多くなっている．これにより，我々の期待した学習ではなかったものの，“forest”が出現した場合，負例になる可能性が高いという潜在的な傾向を分類器が学習したことが推測される．また，“forest”は観光の形態の中でもグリーンツーリズムに関連があると考えられる．そこで，実際に物体検出の結果に“forest”が含まれる旅行ブログエントリを見てみると，山などの自然を体験しているグリーンツーリズムの旅行ブログエントリが多く確認された．これらのことから，山を対象としたグリーンツーリズムを体験する旅行者はヘリテージツーリズムを体験しない傾向にあると考えられる．

表 6: 抜粋した単語と MLP(img)による分類精度

抜粋した単語	精度	検出された件数	
		正例	負例
forest	0.641	6	44
commercial	0.655	3	11
terrestrial	0.658	1	0
fireplace	0.665	2	0
glass	0.667	1	2
(抜粋しなかった場合)	0.778	-	-

5. 分類結果の可視化

4 章で得られた分類結果をもとに，本研究では地図上に旅行ブログエントリをマッピングすることで可視化を行う．これによって，直感的に観光地ごとの観光の形態の特徴がわかるようになると期待される．旅行ブログエントリの位置情報の取得には，旅行ブログエントリに含まれる画像からほとんど取得できなかったため，Google Cloud Vision API の解析結果を用いた．Google Cloud Vision API では，物体検出のほかに，ランドマーク検出もでき，ランドマークの名称やランドマークの位置情報を取得することができる．本研究ではここから得られた位置情報を旅行ブログエントリの位置情報とし，分類結果の可視化を行った．可視化の手順は以下の通りである．

(1) 旅行ブログエントリを収集する．
(2) 収集した旅行ブログエントリからテキストと画像を抽出する．
(3) Google Cloud Vision API を用いて，画像の解析を行い，物体検出・位置情報の推定を行う．
(4) 旅行ブログエントリを観光の形態に自動分類する．
(5) 画像解析結果で得られた，緯度･経度をもとに，Google Earth を用いて地図上にマッピングを行う．複数の画像から位置情報が抽出できた場合，最初に抽出できた位置情報を採用する．

今回は，可視化を行うにあたって，TravelBlog から無作為に約 240,000 件の旅行ブログエントリを収集した．また，この旅行ブログエントリの中から Google Cloud Vision API により，画像から位置情報を推定することができた 24,023 件の旅行ブログエントリに対し，分類を行い，その結果をもとに Google Earth にマッピングをすることで可視化を行った．分類には，人手で分類した表 2 のデータを用いて学習を行い，作成したモデルで分類を行った．

可視化の結果を図 2 に示す．これはエジプト上空から見たシステムの画面である．このシステムでは，それぞれの観光の形態に分類された旅行ブログエントリを表 1 で示すアイコン[6, 7, 8, 9, 10, 11]を用いて表している．また，マッピングされているこれらのアイコンをクリックすると，旅行ブログエントリの URL が現れ，これをさらにクリックすることで，対応する旅行ブログエントリの内容を見ることができる．実際に，図 2 のシステム画面を見ると，ナイル川周辺でヘリテージツーリズムを表す緑色の旗付きの建物のアイコンが多く存在することがわかる．そこで，これらの旅行ブログエントリの内容を調べてみると，ピラミッドなどの有名な世界遺産について記述されていることが確認できた．また，内陸側にグリーンツーリズムのアイコンがあったため，調べてみると，“white desert”という観光地について書かれた旅行ブログエントリであることがわかった．この観光地は，旅行ブログエントリによると，「white desert へは 4 時間の車での移動が必要だったが，そこは非常に魅力的な場所であった」という内容が書かれていた．このように，あまり知られていない観光地であっても，このシステムを利用することで，観光の形態に基づいた情報を調べることができる．

本システムでは，観光地の情報を調べることができるほかに，Google Earth の機能を用い，現在地やアイコン同士を繋ぐことで，移動経路の検索も行うことも可能となる．例えば，広島でヘリテージツーリズムを体験したいというユーザがこのシステムを利用した場合，原爆ドームと厳島神社にアイコンが存

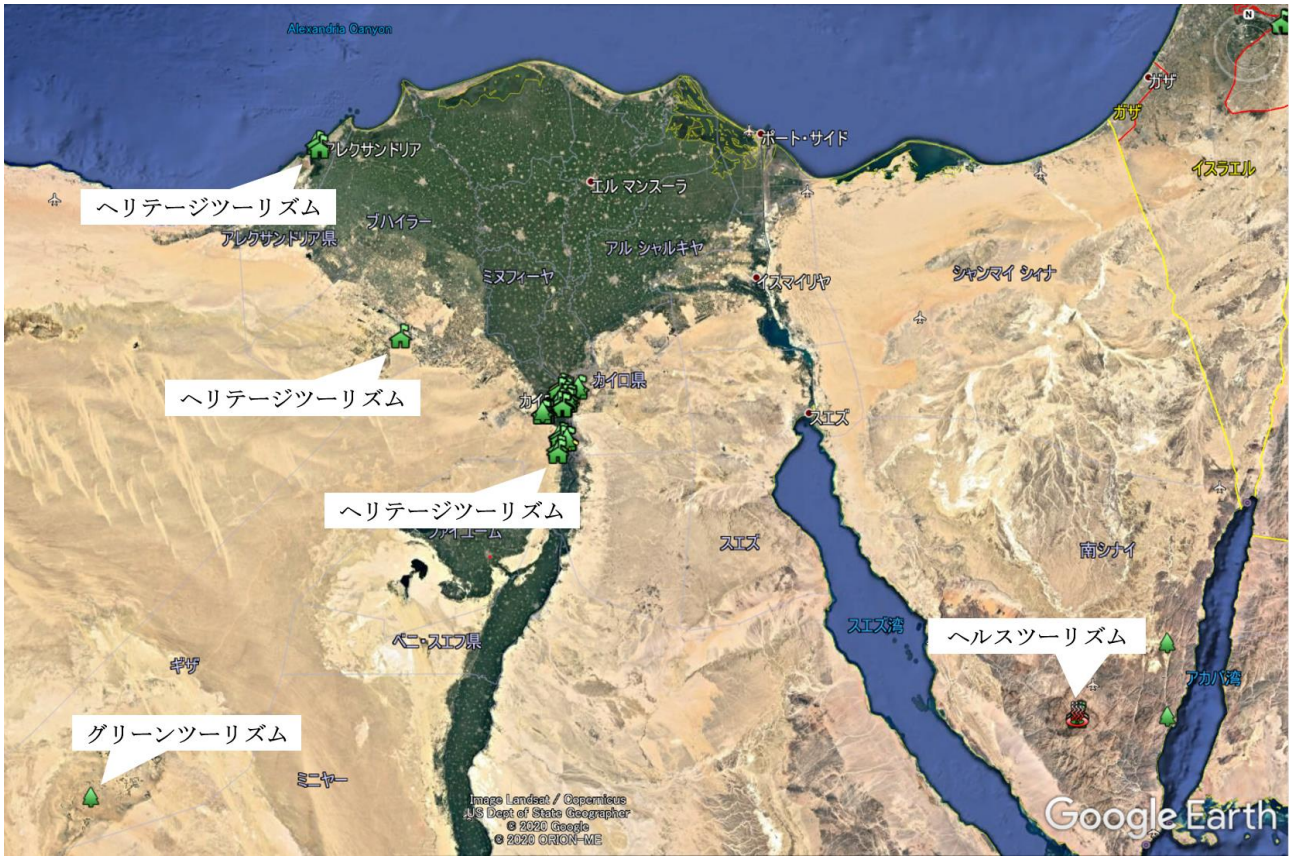

図 2: エジプト上空から見たシステム画面

[6] http://maps.google.co.jp/mapfiles/ms/icons/plane.png
[7] http://maps.google.co.jp/mapfiles/ms/icons/hotsprings.png
[8] http://maps.google.co.jp/mapfiles/ms/icons/cycling.png
[9] http://maps.google.co.jp/mapfiles/ms/icons/tree.png
[10] http://maps.google.com/mapfiles/kml/shapes/ranger_station.png
[11] http://maps.google.com/mapfiles/kml/shapes/campfire.png

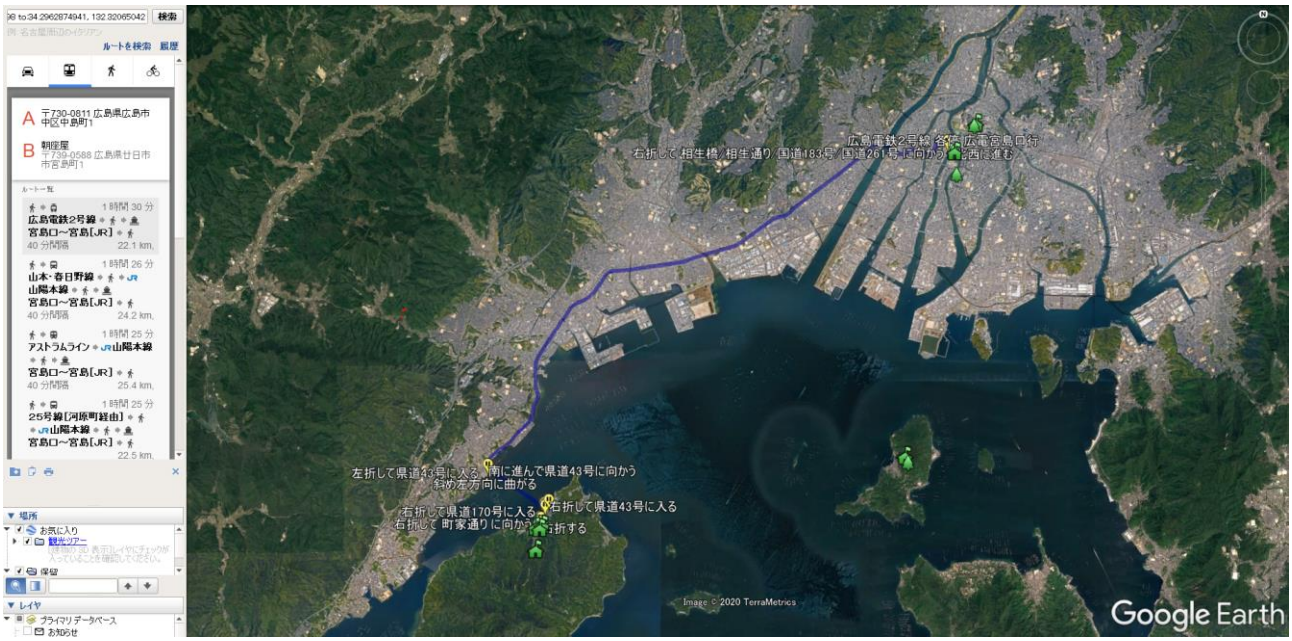

図 3: 原爆ドームから厳島神社までの経路検索の例 (広島上空から見たシステム画面)

在していることが確認できる．そこで，これらのどちらの場所にも訪問する場合，両方のピンをクリックすることで，システム上で使用する乗り物や所要時間などの経路の検索を行うことができる．図 3 は実際に原爆ドームと厳島神社の経路を検索した際に，広島上空から見たシステム画面である．この画面には，原爆ドームから厳島神社までの経路が線で結ばれているほか，画面左に利用する公共交通機関や所要時間などが書かれている．このように，Google Earth を用いた本システムは, 旅行ブログエントリの情報を調べるだけでなく，調べた情報から観光の形態に基づいた旅行計画の作成にも利用することができると考えられる．

6. おわりに

本研究では，旅行ブログエントリ中のテキスト情報，画像情報および Wikipedia の情報を考慮した機械学習による分類器を構築し，定義した 6 種類の観光の形態に自動分類する手法を提案した．テキスト情報については，分類に有効であると考えられる名詞のみを抽出し，これを入力データに用いた．画像情報については，Google Cloud Vision API を用いて，画像中に含まれる物体を検出し，その結果から得られる単語集合を分類器の入力データに用いた. また，Wikipedia の情報については，Google Cloud Natural Language API で Wikification を行い，リンク先の Wikipedia のエンティティ名の集合を分類器の入力データに用いた．観光の形態に基づいた旅行ブログエントリの分類実験の結果，精度では，ベースライン手法の SVM(noun) で 0.761，提案手法の MLP(noun+img+wiki)で 0.768 と同等程度の値であったが，再現率では，SVM(noun)の 0.156 に対して，MLP(noun+img+wiki)で 0.233 が得られたことから，提案手法の有効性が確認された．また，再現率と F 値では，提案手法の MLP(noun+img)で最も高い値 0.274，0.396 が得られた．これらの結果から，旅行ブログエントリを観光の形態に分類する際に，複数の入力データを考慮することの有効性が示された．

また，本研究では，自動分類で得られた結果を用い，Google Earth 上にマッピングを行うことで分類結果を可視化するシステムを構築した．このシステムを用いることで, 観光の形態に沿った検索ができ，観光の形態という新たな観点から観光地の魅力を発見することが可能となった．さらに，Google Earth では経路の検索も可能であることから，旅行計画の作成にも応用できると考えられる．

参考文献

[1] R. Mihalcea and A. Csomai, “Wikify! Linking Documents to Encyclopedic Knowledge”, Proc. of the ACM Conference on Information and Knowledge Management, pp. 233-242 (2007)

[2] Y. Murawaki and S. Mori, “Wikification for Scriptio Continua”, Proc. of the 10th Edition of the Language Resources and Evaluation Conference, LREC, pp. 1346-1351 (2016)

[3] 河行茜，木下藤寿，“島根おおだ健康ビューローの取り組み”，生涯スポーツ実践研究年報 ：鹿屋体育大学生涯スポーツ実践センター所報，Vol. 17，pp. 28-35 (2019)
[4] 藤井千賀子，茂木直美，林由利子，柳沼しほ，“インフラツーリズムガイド 2018”，芸文社 (2018)
[5] M. Iyyer, V. Manjunatha, J. Boyd-Grader, and H. Daume III, “Deep Unordered Composition Rivals Syntactic Methods for Text Classification”, Proc. of the Association for Computational Linguistics, ACL (2015)
[6] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, and L. Carin, “Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms”, Proc. of the Association for Computational Linguistics, ACL (2018)
[7] K. Takahashi, D. Kato, M. Endo, T. Araki, M. Hirota, and H. Ishikawa, “Analyzing Travel Behavior using Multi-label Classification from Twitter”, Proc. of the 9th International Conference on Management of Digital EcoSystems, MEDES’17 (2017)
[8] 藤井一輝，難波英嗣，竹澤寿幸，石野亜耶，奥村学，倉田洋平，“旅行者の行動分析のための旅行ブログエントリの属性推定”，観光と情報，Vol.13，No.1，pp. 83-96 (2017)
[9] 高橋義雄，原田宗彦，岡星竜美，工藤康宏，二宮浩彰，松岡宏高，山下玲，青木淑浩，“スポーツツーリズム・ハンドブック”，学芸出版社 (2015)
[10] 山下晋司，“観光学キーワード”，有斐閣 (2011)
[11] 後藤和子，“観光と地域経済 −文化観光の経済分析を中心に−”，地域経済学研究，Vol. 34，pp. 41-47 (2018)
[12] M. Foley and J. LeMLPon, “Editorial: Heart of Darkness”, Proc of International Journal of Heritage Studies, Vol. 2, pp. 195-197 (1996)
[13] 岡本亮輔，“聖地巡礼”，中央公論新社 (2015)
[14] I. Yamada, A. Asai, H. Shindo, and H. Takeda, “Wikipedia2Vec: An Optimized Tool for Learning Embeddings of Words and Entities from Wikipedia”, arXiv preprint arXive:1812.06280v2 (2018)

柴田 有基 （非会員）
1997 年生. 2015 年広島市立大学情報科学部知能工学科卒業．現在，広島市立大学大学院情報科学研究科博士前期課程在学中.

篠田 広人 （非会員）
1994 年生. 2018 年広島市立大学情報科学部知能工学科卒業．2020 年広島市立大学大学院情報科学研究科知能工学専攻博士前期課程修了.

難波 英嗣 （正会員）
1972 年生. 1996 年東京理科大学理工学部電気工学科卒業．2001 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了．博士(情報科学)．2001 年日本学術振興会特別研究員．2002 年東京工業大学精密工学研究所助手. 同年広島市立大学情報科学部講師. 2010 年広島市立大学大学院情報科学研究科准教授．2019 年中央大学理工学部教授．現在に至る．テキストマイニング，情報検索，テキスト要約に関する研究に従事．観光情報学会，情報処理学会，人工知能学会，言語処理学会会員.

石野 亜耶 （正会員）
2009 年広島市立大学情報科学部知能情報システム工学科卒業．2011 年広島市立大学大学院情報科学研究科博士前期課程修了．2014 年同大学大学院情報科学研究科博士後期課程満期退学．同年同大学大学院にて博士号（情報科学）取得. 同年広島経済大学経済学部ビジネス情報学科助教．2017

年同大学経済学部ビジネス情報学科准教授．2019 年同大学メディアビジネス学部ビジネス情報学科准教授．現在に至る．テキストマイニング，観光情報処理に関する研究に従事．観光情報学会，情報処理学会，人工知能学会，言語処理学会会員．

竹澤 寿幸 （正会員）
1961 年生. 1984 年早稲田大学理工学部電気工学科卒業．1989 年早稲田大学大学院理工学研究科博士後期課程修了．工学博士．1987 年早稲田大学情報科学研究教育センター助手．1989 年（株）ATR 自動翻訳電話研究所研究員. 音声対話翻訳の研究開発に従事. 2007 年より広島市立大学大学院情報科学研究科教授．現在に至る．音声対話や観光情報学の研究と教育に従事．観光情報学会，電子情報通信学会，情報処理学会，人工知能学会，日本音響学会，言語処理学会会員．