観光の形態に基づいた旅行ブログエントリの分類と可視化

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]

概要：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，SCDV を用いた実験により，精度 0.743，再現率 0.217，F 値 0.336 を得た．最後に，観光の形態のから地図上にマッピングされた旅行ブログエントリを検索できるシステムを構築した．

キーワード：観光の形態，旅行ブログ，文書分類，Wikification

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]

Abstract: We propose a method to classify travel blog entries into one of six types of tourism using textual and image information in each travel blog entry. Together with this information, we use Wikipedia entries,*†1 which were automatically linked from each travel blog entry by an entity linking technology, because the useful information for classifying blog entries is often mentioned in Wikipedia entries. We combine this information using a deep-learning-based method, and we conducted an experiment. From the results using SCDV, we obtained precision, recall, and F-measure of 0.743, 0.217, and 0.336, respectively. Finally, we constructed a system that enables travellers to look for travel blog entries in a map in terms of the types of tourism.

Keywords: types of tourism, travel blog, document classification, Wikification

1. はじめに

近年，観光は従来の娯楽を追求するのみだけではなく，様々な形態が誕生し，現在もその多様化は進んでいる．例えば，健康回復や維持，増進につながる観光はヘルスツーリズム，スポーツを体験または観戦することを目的とした観光はスポーツツーリズムと呼ばれる．旅行ブログエントリに対して，このような観光の形態の自動分類が実現すれば，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．また，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と考えられる．そこで本研究では，6 種類の観光の形態を定義し，機械学習を用いて旅行ブログエントリをこれらの観光の形態に自動分類する手法を提案する．

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一つに，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質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かかるというデメリットも存在する．そこで，近年，Twitter などの SNS 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写真など，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類の対象に分析を行う．

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．

†1 広島市立大学
†2 中央大学
†3 広島経済大学

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，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のテキストから自動的にリンク付けされた Wikipedia エントリの情報を利用し，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．さらに，分類結果を地図上にマッピングすることで可視化を行う．なお，テキスト中の単語と Wikipedia の該当ページをリンク付けすることは Wikification[1, 2]と呼ばれている．本研究のように，ブログエントリなどの SNS を対象にした分類はこれまでにも行われている．

テキストを用いた属性推定について，Goot ら[3]は，ツイートのテキストを抽象的な内容に置き換えたもので性別予測を行う手法を提案している．これにより，語彙に依存しなくなり，言語横断な性別予測を可能にしている．また，lyyer ら[4]は，文書に含まれる単語を，Word2Vec を用いてベクトルに変換し，それらを平均したものを分類に用いる手法 DAN(Deep Averaging Networks)を提案している．DAN の概略図を図 2.1 に示す．この図では，“Predator is a masterpiece”のそれぞれの単語を赤い四角で表現されているベクトルに変換し，それらのベクトルを次元ごとに平均することで，緑の四角で表現される新たなベクトルを生成する．この緑のベクトルをいくつかの層で処理し，最後の softmax 層で判定する．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．本研究では，観光形態に基づく旅行ブログエントリの分類実験の提案手法とベースライン手法で，この DAN を用いる．

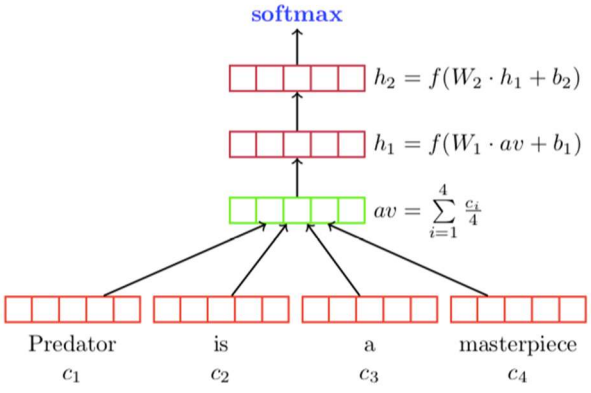

図 2.1: DAN の概略図(lyyer[2]より引用)

DAN の他に文書ベクトルを生成する手法について，Mekala ら[5]は，単語ベクトルを GMM と IDF 値を考慮した新たなベクトルを生成し，生成した単語ベクトルの平均を文書ベクトルとする手法 SCDV(Sparse Composite Document Vectors)を提案している．SCDV の概略図を図 2.2 に示す．この図では，GMM を用いて各クラスタ a-e へ分類を行い，その結果と idf 値を考慮した単語ベクトルを平均することで文書ベクトルを生成している．本研究では，観光形態に基づく旅行ブログエントリの分類実験の提案手法とベースライン手法で，DAN と同様に SCDV を用いる．

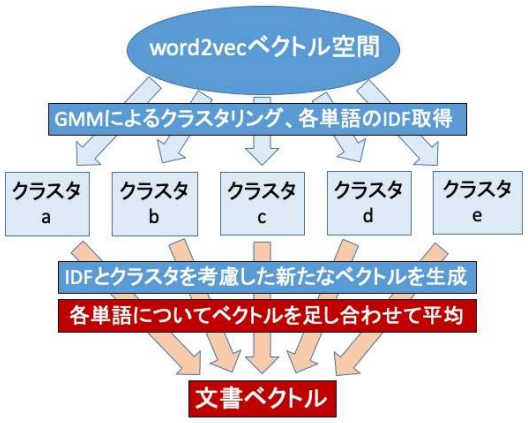

図 2.2: SCDV の概略図

観光に関する文書分類について，Kolomoyets ら[6]や Fujii ら[7]の研究がある．特に，Fujii らは，本研究と同様に旅行ブログエントリを扱っており，「買う」，「食べる」，「体験」，「泊まる」，「見る」に分類する手法を提案している．この分類の観点は，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．そのため，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，「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はないため，使用者によって異なる解釈で用いられる場合がある．本研究では，それぞれの観光の形態に対して，独自に定義した．観光の形態とその定義，またその具体例を表 3.1 に示す．本研究では，表 3.1 の 6 種類の観光の形態に基づいて旅行者の観光の形態を明らかにするため，個々の英語で書かれた旅行ブログエントリを自動分類する．

現在，観光の形態は独自に定義した 6 種類以外にも多く存在する．しかし，今回は分類をする上で有用であると思われ，かつある程度自動分類が可能であろうと思われる 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．

表 3.1: 観光の形態の定義と具体例

観光の形態	定義	例
インフラ，ハードツーリズム	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	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	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	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	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	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

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する Wikipedia の情報を解析し，それらの結果を用いて「観光の形態」を自動分類する．まず，テキスト解析による分類の例を図 3.1 に示す．これはフランスの Mont Saint-Michel を訪れた方が書いたブログの一部である．ブログのテキストから“UNESCO World Heritage Site”という単語が含まれるため，これは「ヘリテージツーリズム」であると考えられる．

> Just last week I spent nine days touring my aunt and cousin around Paris and then we took a side trip to Normandy to visit a very impressive **UNESCO World Heritage Site** site, Mont Saint Michel...

https://www.travelblog.org/Europe/France/Lower-Normandy/Mont-Saint-Michel/blog-490707.html より引用

図 3.1: テキストによる分類の例

ここで，テキストのみを対象とした場合，もしブログエントリ中に「スキー」という表現が存在しても「スキーをしたかったけどできなかった」という記述であれば，ブログ著者は実際にスキーをしていないことになる．これに対し，スキーの画像がブログエントリ中にあれば，ブログ著者は確実にスキーをしていると判断できる．図 3.2 はスキーをしたと思われる観光客が書いたブログエントリ中の画像の例である．この画像からはスキーをしている様子が読み取れるため，画像のみでスポーツツーリズムであると判断できる．このように分類の際に画像に写っているものが重要な判断材料になることが考えられる．そこで，本研究では，画像からの物体検出に Google Cloud Vision API*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．実際に，図 3.2 の画像に対する Google Cloud Vision API の解析結果の一部を表 3.2 に示す．表 3.2 より，スキー関連の単語が多く並んでいることから，これは「スポーツツーリズム」であると判断できる．このように，ブログエントリを分類する際には，テキスト情報に加えて画像情報も分類の際に重要な情報になると考えられる．

https://www.travelblog.org/Europe/Switzerland/South-West/Geneva/blog-1033087.html より引用

図 3.2: 旅行ブログエントリの画像の例

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，テキストや画像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．具体的な例として，図 3.3 に外部知識が必要な例を示す．

これは，イギリス・スコットランドの Forth Bridge を訪れた観光客が書いたブログの一部である．このブログエントリを観光の形態に分類する場合，Forth Bridge が世界遺産のため，これは「ヘリテージツーリズム」となる．しか

* https://cloud.google.com/vision/?hl=ja

表 3.2: 図 3.2 の画像に対する
Google Cloud Vision API の解析結果の一部

ラベル	スコア
Skiing	0.99
Sports	0.99
Winter Sport	0.99
Snow	0.98
Ski	0.98
Cross-country Skiing	0.97
Norbic Skiing	0.97
Ski Equipment	0.96

し，図 3.1 のようにテキスト中に「Forth Bridge」が世界遺産であるという情報がない．また，画像解析結果からも情報は得られない．これを正しく「ヘリテージツーリズム」と判断するには，外部知識が必要になると考えられる．これについては，Google Cloud Natural Language API*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 Forth Bridge 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．そこで，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中の物体に加えて，外部知識として Wikification の結果から Wikipedia の情報を与えることで，より精度の高い分類の実現を目指す．

The crossing between North and South Queensferry in central Scotland is a unique tourist attraction and hosts some of Britain’s busiest transport structures — The Forth Bridge, Forth Road Bridge and Queensferry Crossing.

https://www.travelblog.org/Europe/United-Kingdom/blog-1025495.html より引用

図 3.3: 外部知識が必要ブログな例

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する．次に，Wikification の結果から，リンク付けされた Wikipedia の abstract(最初の段落)に含まれる単語を抽出する．そし

表 3.3: 図 3.3 の画像に対する
Google Cloud Vision API の解析結果の一部

ラベル	スコア
Bridge	0.98
Cantilever Bridge	0.93
Landmark	0.91
Fixed Link	0.88
Girder Bridge	0.83
Architecture	0.82

て，物体検出の結果として得られた単語集合と Wikipedia の abstract から得られた単語集合，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の構築をする．分類器の概略図を図 3.4 に示す．本研究では，図 3.4 のように旅行ブログエントリ中のテキストと画像，リンク付けされた Wikipedia の該当ページの abstract に含まれる単語集合のそれぞれの入力ごとに処理をし，最後に統合することで分類を行う形式とする．

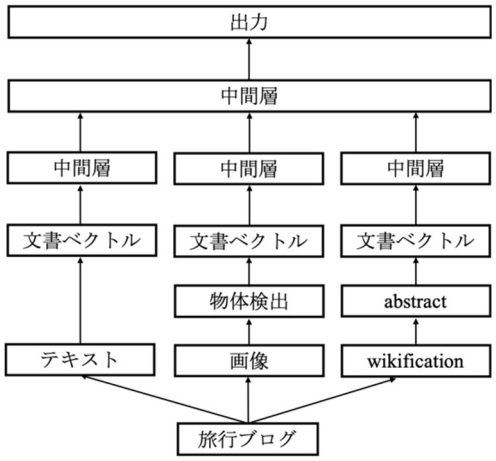

図 3.4: 分類機の概略図

観光の形態には様々な種類があるが，地域によって過去の歴史や土地柄などの理由から，偏りがあると推測される．例えば，広島県廿日市市では世界遺産の「厳島神社」が有名であることから，観光の形態の一つである「ヘリテージツーリズム」が多くなると考えられる．こうした情報を明らかにすることにより，各訪問地の特徴を活かした観光施策を行うことができると思われる．また，世界遺産の「日光の社寺」がある栃木県日光市も同様の傾向になるようであれば，似た観光地同士で観光施策を参考にできることも期待される．

4. 実験

本実験で使用する旅行ブログエントリでは，3 章で述べ

* https://cloud.google.com/natural-language/?hl=ja

た TravelBlog のテキストデータと画像データを用いる．

4.1 実験条件

【実験に用いるデータ】

2,017 件の旅行ブログエントリを実験に用いた. 各ブログエントリに対し，観光の形態 6 種類のカテゴリに人手で分類し，これらを機械学習の際の訓練用および評価用データとして用いた. 人手で分類した結果の内訳を表 4.1 に示す. なお，一つの旅行ブログエントリには複数の観光の形態が付与されることや逆に一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，9 つの件数の総和が 2,017 より小さくなっている．

【実験条件】

単語の分散表現には，Google が提供している 2 種類の事前学習モデル，Google News を対象に作られた 300 次元の Word2vec のモデル*と Wikipedia と BookCorpus を対象に作られた 1024 次元の BERT[8]のモデル†を用いる．画像認識には, 3.2.1 節で説明した Google Cloud Vision API を用いる．DAN では, 入力データの前処理としてあらかじめ stopword‡を削除したものを用いる．Wikification の結果を考慮した分類については，Wikipedia の abstract から分散表現を得る手法とリンクから抽出したエンティティ名から分散表現を得る手法の 2 種類を行う．なお，エンティティ名の分散表現には，Wikipedia2Vec[9]の事前学習モデル§を用いる．

表 4.1: 人手で分類した結果の内訳

観光の形態	件数
1. インフラ，ハードツーリズム	168
2. ヘルスツーリズム	125
3. スポーツツーリズム	57
4. グリーンツーリズム	453
5. ヘリテージツーリズム	196
6. カルチュラルツーリズム	49
判定したブログエントリの総数	2,017

【評価尺度】

実験では，5 分割交差検定を行い，評価尺度には，精度・再現率・F 値を用いる．また，これらを算出するにあたって，観光の形態ごとのデータ件数の偏りを考慮するため，micro 平均を採用する．また，本研究では，多クラス分類を行う方法として，それぞれの観光形態での 2 値分類を拡張することで実現する．

【比較手法】

本実験では，以下に示す 2 種類の提案手法と 6 種類のベースライン手法で実験を行った．なお，入力データについて，“txt”はテキストに含まれる単語集合を，“img”は画像から物体検出の結果として得られた単語集合を，“wiki(abst)”は Wikification の結果から得られた Wikipedia エンティティの abstract に含まれる単語集合を，“wiki(entity)”は Wikification の結果から抽出した Wikipedia エンティティ名の集合を表す.

提案手法

- SCDV(txt+img+wiki(abst), word2vec model): 3 つの入力データに対し, SCDV を用いて文書ベクトルを作成し，そのベクトルを入力とするニューラルネットワークで分類を行う．epoch 数は 30 とする．
- DAN(txt+img+wiki(abst), word2vec model): 3 つの入力データに対し，Word2Vec でベクトルに変換し，DAN を用いて分類を行う．epoch 数は 20 とする．
- DAN(txt+img+wiki(entity), word2vec model): テキストの単語集合と画像解析結果の単語集合を Word2Vec のモデルでベクトル化する．また，Wikipedia のリンクからエンティティ名を抽出し，Wikipedia2Vec のモデルでエンティティベクトルを獲得する．これら 3 つのベクトル集合を入力データとし，DAN を用いる．epoch 数は 20 とする．

ベースライン手法

- SCDV(txt+img, word2vec model): 2 つの入力データに対して，SCDV を用いて文書ベクトルを作成し，そのベクトルを入力とするニューラルネットワークで分類を行う．epoch 数は 30 とする．
- DAN(txt+img, BERT model): 2 つの入力データに対し，BERT のモデルから得られたベクトルに変換し，DAN を用いて分類を行う．epoch 数は 10 とする．
- DAN(txt+img, word2vec model): 2 つの入力データに対し，Word2Vec のモデルから得られたベクトルに変換し, DAN を用いて分類を行う. epoch 数は 20 とする．
- SCDV(txt, word2vec model): テキストに含まれる単語集合に対し，SCDV を用いて文書分類を作成し，それを入力とするニューラルネットワークで分類を行う. epoch 数は 20 とする．
- DAN(txt, BERT model): テキストに含まれる単語集合に対し，BERT のモデルから得られたベクトルに変換し, DAN を用いて分類を行う. epoch 数は 10 とする．
- DAN(txt, Word2Vec model): テキストに含まれる単語集合に対し，Word2Vec のモデルから得られたベクトルに変換し，DAN を用いて分類を行う．epoch 数は 10 とする．
- SVM(txt): テキストに含まれる単語集合に対し，Bag-of-Words で生成したベクトルを入力とする．カーネル

* https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing
† https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-24_H-1024_A-16.zip
‡ http://xpo6.com/list-of-english-stop-words/
§ http://wikipedia2vec.s3.amazonaws.com/models/en/2018-04-20/enwiki_20180420_300d.pkl.bz2

関数には，RBF カーネル，ソフトマージンのパラメータ C は，C=10 とする．

- fastText(txt): fastText[10]とは，Facebook が提供する単語のベクトル化とテキスト分類をサポートした機械学習のライブラリである．計算時間が短く，精度が高いう特徴がある．パラメータについては，次元数は 300，単語は bi-gram，epoch 数は 60 とする．
- SCDV(img, word2vec model):画像解析結果に含まれる単語集合に対し，SCDV を用いて文書分類を作成し，それを入力とするニューラルネットワークで分類を行う．epoch 数は 10 とする．
- DAN(img BERT model): 画像解析結果に含まれる単語集合に対し，BERT のモデルから得られたベクトルに変換し，DAN を用いて分類を行う．epoch 数は 30 とする．
- DAN(img, word2vec model): 画像解析結果に含まれる単語集合に対し，Word2Vec のモデルから得られたベクトルに変換し，DAN を用いて分類を行う．epoch 数は 10 とする．
- SVM(img): 画像解析結果に含まれる単語集合に対し，Bag-of-Words で生成したベクトルを入力とする．カーネル関数には，RBF カーネル，ソフトマージンのパラメータ C は，C=100 とする．
- DAN(wiki(abst), word2vec model): Wikipedia の abstract に含まれる単語集合に対し，Word2Vec のモデルから得られたベクトルに変換し，DAN を用いて分類を行う．epoch 数は 30 とする．
- DAN(wiki(entity), wikipedia2vec): Wikipedia のリンクからエンティティ名を抽出し，Wikipedia2Vec のモデルでエンティティベクトルを獲得する．このベクトルを入力とした DAN で分類を行う．epoch 数は 10 とする．

4.2 実験結果と考察

4.1 節で説明したそれぞれのベースライン手法と提案手法による実験結果を表 4.2 に示す．これより，精度では画像を考慮した SVM で最も高い値 0.783 を得ている．しかし，再現率が 0.165 と低いため，分類漏れが多いことがわかる．再現率と F 値については，テキストを入力とした SVM で最も高い値 0.273，0.385 を得た．また，テキスト，画像解析結果，Wikipedia の abstract のそれぞれの入力の組み合わせごとに確認すると，入力データを増やしていくことで再現率が下がる場合があるものの，入力データを増やしていくことで精度が上がっていることが確認できた．このことから，ブログエントリを分類する際に，複数の入力データを考慮することの有効性が示された．

ところで，提案手法である 3 つの入力データを考慮した SCDV(txt+img+wiki(abst), word2vec model)について考察する．SCDV が誤ってグリーンツーリズムに分類したブログエントリの例を図 4.1 に示す．このブログエントリはテキストや画像から自然に触れ合うという内容が確認できないため，グリーンツーリズムではないと判断できる．SCDV で誤って分類した原因について，テキストや画像解析結果の中に，‘River’など自然を連想させる単語がいくつか存在していることがわかる．しかし，テキストでは，‘will’や‘hope’といった単語が存在することから，まだ体験していないことが読み取れる．そのため，SCDV ではできていない，単語の並びを考慮することや観光の形態と関連があると思われる単語の周辺語に注意を向けさせることで，更なる精度向上が期待できる．

We will have travelled 2000 miles and transited numerous locks, passed many tows pushing barges and dodged River debris.
We hope to spot a myriad species of birds and wildlife, …

https://www.travelblog.org/North-America/blog-895066.html より引用

図 4.1: 誤ってグリーンツーリズムに分類してしまったブログエントリの例

表 4.3: 図 4.1 の画像に対する Google Cloud Vision API の解析結果の一部

ラベル	スコア
Mode of transport	0.88
Tree	0.86
Recreation	0.81
Plant	0.79

5. 可視化

4 章で得た分類結果をもとに，本研究では，地図上に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形態がわかるようになると考えられる．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) 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．

表 4.2: 観光の形態に基づくブログエントリの分類結果(micro 平均)

	手法	precision	recall	F-measure
提案手法	SCDV(txt+img+wiki(abst), word2vec model)	0.743	0.217	0.336
	DAN(txt+img+wiki(abst), word2vec model)	0.729	0.191	0.302
	DAN(txt+img+wiki(entity), word2vec model)	0.718	0.202	0.315
ベースライン手法	SCDV(txt+img, word2vec model)	0.729	0.227	0.347
	DAN(txt+img, BERT model)	0.649	0.272	0.383
	DAN(txt+img, word2vec model)	0.701	0.202	0.314
	SCDV(txt, word2vec model)	0.639	0.170	0.268
	DAN(txt, BERT model)	0.573	0.231	0.330
	DAN(txt, word2vec model)	0.596	0.238	0.340
	SVM(txt)	0.656	**0.273**	**0.385**
	fastText(txt)	0.461	0.129	0.202
	SCDV(img, word2vec model)	0.725	0.140	0.235
	DAN(img, BERT model)	0.661	0.228	0.339
	DAN(img, word2vec model)	0.701	0.159	0.259
	SVM(img)	**0.783**	0.165	0.273
	DAN(wiki(abst), word2vec model)	0.539	0.020	0.038
	DAN(wiki(entity), wikipedia2vec)	0.551	0.189	0.282

(4) Google Cloud Natural Language API を用いて，テキストから Wikification を行う．
(5) Wikification で得られたリンクから，Wikipedia エンティティの情報を獲得する．
(6) 得られたテキストと画像解析結果，Wikipedia エンティティの情報から，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．
(7) 画像解析結果で得られた，緯度・経度をもとに，Folium*を用いて地図上にマッピングを行う．なお，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，人手で分類した 2,017 件の旅行ブログエントリの中から，位置情報が抽出できたものを対象に Folium を用いて可視化を行った．人手で分類した旅行ブログエントリの可視化の結果を図 5.1 に示す．それぞれのピンの色については，黒色が「インフラ，ハードツーリズム」，緑色が「グリーンツーリズム」，紫色が「スポーツツーリズム」，青色が「グリーンツーリズム」，橙色が「ヘリテージツーリズム」，赤色が「カルチュラルツーリズム」を表している．図 5.1 の可視化の結果を見てみると，エジプト周辺で橙色のピンであるヘリテージツーリズムのみであることがわかる．そこで，旅行ブログエントリの内容を見てみると，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．今回，扱った旅行ブログエントリの数は少ないものの，地域の新たな魅力の発見や観光形態の偏りが見えることで，観光政策などに役に立つと考えられる．

図 5.1:人手で分類した旅行ブログエントリの可視化の結果

6. おわりに

本研究では，ブログエントリ中のテキストと画像，Wikification による Wikipedia の情報を考慮し，旅行ブログエントリに対して，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．画像に対しては，Google Cloud Vision API を用いて，画像中に含まれる物体を検出し，その

* https://python-visualization.github.io/folium/

結果を分類の素性に用いた．また，Wikipedia の情報に対しては，Google Cloud Natural Language API でリンク付けを行い，リンク先の Wikipedia の abstract に含まれる単語集合を分類の素性に用いた．実験の結果，提案手法の 3 つの入力データを考慮した SCDV では，精度ではベースライン手法の画像解析結果を入力とした SVM に 0.040 ポイント劣るものの，再現率と F 値では SVM より良い値が得られた．ところで，テキスト，画像解析結果，Wikification による Wikipedia の abstract のそれぞれの入力データの組み合わせに注目すると，SCDV と DAN のそれぞれで，入力データの組み合わせの数を増やすと結果が改善されていることが確認されたことから，文書分類における複数の入力データを考慮することの有効性が示された．今後は，アンサンブル学習を用いて，複数の分類結果を統合する手法について検討する．

参考文献

[1] Rada Mihalcea and Andras Csomai.. Wikify! Linking Documents to Encyclopedic Knowledge. The ACM Conference on Information and Knowledge Management, 2007, 233-242.
[2] Yugo Murawaki and Sinsuke Mori.. Wikification for Scriptio Continua. the 10th Edition of the Language Resources and Evaluation Conference, 2016, 1346-1351.
[3] Rob van der Goot, Nikola Ljubesic, Ian Matroos, Malvina Nissim, and Barbara Plank.. Bleaching Text: Abstract Features for Cross-lingual Gender Prediction. the Association for Computational Linguistics, 2018.
[4] Mohit lyyer, Varun Manjunatha, Jordan Boyd-Grader, and Hal Daume.. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. the Association for Computational Linguistics, 2015.
[5] Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape and Harish Karnick.. SCDV: Sparse Composite Document Vectors using soft clustering over distributional representations. EMNLP, 2017, 659-669.
[6] Yuliya Kolomoyets and Astrid Dickinger.. A Text Mining Approach to Measuring and Predicting Perceived Service Quality from Online Chatter. The 26th Annual eTourism Conference, ENTER, 2019.
[7] Kazuki Fujii, Hidetsugu Nanba, Toshiyuki Takezawa, Aya Ishino, Manabu Okumura, and Yohei Kurata.. Travellers' behaviour analysis based on automatically identified attributes from travel blog entries. Workshop of Artificial Intelligence for Tourism, PRICAI, 2016.
[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova.. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805v2 [cs.CL], 2018.
[9] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda and Yoshiyasu Takefuji.. Wikipedia2Vec: An Optimized Tool for Learning Embeddings of Words and Entities from Wikipedia. arXiv: 1812.06280v2 [cs.CL], 2018.
[10] Armand Joulin, Eduard Grave, Piotr Bojanowski, and Tomas Mikolov.. Bag of Tricks for Efficient Text Classification. arXiv:1607.01759v3 [cs.CL], 2016.