学術研究論文

観光の形態の特徴を考慮した将来の訪問国の予測

柴田有基[†1, †2]，石野亜耶[†3]，難波英嗣[†4]，竹澤寿幸[†1]

[†1]広島市立大学大学院情報科学研究科，[†2]現在，株式会社インテージテクノスフィア，

[†3]広島経済大学メディアビジネス学部，[†4]中央大学理工学部

【あらまし】

本研究では，深層学習手法のLSTMを用い，旅行者の過去の訪問国から将来の訪問国を予測する手法を提案する．将来の訪問国を正確に予測するためには，過去の訪問国の位置情報や文化などの基本的な情報に加え，それぞれの国で体験可能な観光の特徴を考慮する必要がある．そこで，本研究では， Wikipediaから獲得した国の情報に加え，各国のソーシャルメディアの解析を通して得られたそれぞれの国の観光の特徴を考慮した将来の訪問国の予測手法を提案する．ソーシャルメディアの一つである旅行ブログエントリのデータセットを用いた評価実験の結果，提案手法のLSTMでベースライン手法より高い，Accuracy@1で26.06%，Accuracy@3で46.31%，Accuracy@5で56.61%，Accuracy@10で70.30%，MRRで0.403が得られた．また，有意水準5%において，観光の形態を考慮することの有効性が確認された．

Predicting the Next Country Considering Features of Tourism Types

Naoki Shibata [1), 2)]，Aya Ishino [3)]，Hidetsugu Nanba [4)]，Toshiyuki Takezawa [1)]

1) Graduate School of Information Sciences, Hiroshima City University

2) Presently with INTAGE TECHNOSPHERE Inc.

3) Faculty of Media Business, Hiroshima University of Economics

4) Faculty of Science and engineering, Chuo University

【Abstract】

We propose a method for predicting traveler's next countries using the LSTM deep learning method. To predict the next country accurately, it needs to consider features of tourism that involve travel experiences, in addition to country information such as location information and culture of each country. Therefore, we propose a method considering country information using Wikipedia and features of tourism types, obtained through analyzing social media with Random Forest. In this paper, we conducted an experiment using LSTM with Wikipedia and tourism types of input data, and obtained the results, which were better than baseline methods, of Accuracy@1, Accuracy@3, Accuracy@5, Accuracy@10, and MRR scores of 26.06%, 46.31%, 56.61%, 70.30%, and 0.403, respectively. Furthermore, the effectiveness of considering tourism types was confirmed at the significance level of 5%.

1. はじめに

現在，COVID-19の影響により，海外旅行を中心とした観光行動が縮小している．しかし，日本交通公社の海外旅行意向調査では，10〜20代の約半数の女性が，COVID-19の収束後，海外旅行に行きたいと回答していることが報告されている[1]．そのため，今後は海外旅行者数が少しずつ回復していくことが期待され，こういった海外旅行を考えている旅行者へのマーケティング活動は非常に重要になると考えられる．

観光データの解析を行うためには，まず，旅行者のデータを収集する必要があるが，その方法として，アンケートや観光ソーシャルメディアを用いる方法が挙げられる．アンケートでは，調べたい内容についての設問を用意することで，欲しいデータを確実に手に入れることができるメリットがある．しかし，大量のデータを集めるためには大きなコストがかかること，また，COVID-19の影響により，対面式のアンケートの収集が難しくなっていることなどのデメリットが存在する．このような現状から，本研究では，観光ソーシャルメディアから収集した，大規模なデータを用いて解析を行うこととする．

観光ソーシャルメディアの解析に関する研究では，観光スポットや観光ルートなどを推薦する研究は既に行われている[2-5]．しかし，多くの研究は，ある都市やある地域などの比較的小さな範囲を対象としているため，世界中の国が対象となる海外旅行者への対応は不十分である．そこで，本研究では，世界中の国について投稿された観光ソーシャ

ルメディアの解析を通し，旅行者の過去の訪問国履歴から将来の訪問国を予測する．将来の訪問国の予測が可能になれば，世界中の国を対象とした海外旅行者に対して，観光地推薦が可能になる．

また，近年は交通網の発達などにより，より観光が生活の一部となっており，従来の娯楽のみを追求する観光だけでなく，例えば，身心の健康の促進を目的とした「ヘルスツーリズム」やスポーツの観戦や体験を目的とした「スポーツツーリズム」など，観光の形態は多様化している．そのため，このような観光の形態に基づいた旅行者の傾向を考慮することができれば，より正確な将来の訪問国の予測が可能になると考えられる．そこで，本研究では，過去の訪問国の位置情報や文化などの基本的な情報に加えて，上記のような観光の形態を考慮し，旅行者の将来の訪問国を予測する．

本論文の構成は以下の通りである．まず，2節では関連研究について述べる．次に，3節では，Wikipediaと観光の形態の特徴を考慮した将来の訪問国の予測について述べる．また，4節では，3節で紹介した手法による評価実験について述べる．そして，最後に5節で本論文のまとめを述べる．

2. 関連研究

2.1 将来の訪問地の予測

将来の訪問地の予測に関する研究では，近年，普及している位置情報を付加した投稿を共有するソーシャルメディア，LBSNs (Location-Based Social Networks) のデータセットを用いた研究を中心に，多く行われている[2-5]．Suら[2]は，ユーザの訪問履歴に加えて，フレンド関係や家が近いユーザの訪問履歴を考慮した予測手法を提案している．また，Baragliaら[3]は，ユーザの特徴とPoI (Point of Interest) の特徴を考慮した機械学習手法のRanking SVM (Support Vector Machine) によって将来訪れるPoIを予測する手法を提案している．これらの研究に対して，本研究では，深層学習手法を用いる点で異なる．

また，深層学習を用いた研究では，Liuら[4]は位置情報と時間情報を考慮したRNN (Recurrent Neural Network) による将来の訪問地の予測手法を提案している．また，Kongら[5]は，観光客の訪問地の周期性を考慮したLSTM (Long Short-Term Memory) [6]による予測手法を提案している．本研究では，深層学習手法のLSTMを用いている点で類似しているが，これらの研究では，ある国や地域内のデータを対象としており，本研究は世界中の国を対象としているため異なる．

国の推薦に関する研究では，Majoddiら[7]は，移民のための移住国推薦システムを提案している．また，Johnsonら[8]は，旅行者向けに国ごとの腸チフス感染症の危険性を考慮した予防接種の推薦システムを提案している．これらの研究は世界中の国を対象としている点では類似しているものの，本研究では，観光地の観光情報を研究対象としている点で異なる．

2.2 観光の形態に関する取り組み

温泉やハイキングなど，健康維持を目的とした観光の形態である「ヘルスツーリズム」についての取り組みとして，河行ら[9]は，島根県大田市の自然や特産物を利用した健康促進を促す「ヘルスツーリズム」を推進している．また，ダムや橋など，近代的な建造物を対象とした観光の形態である「インフラツーリズム」についての取り組みとして，藤井ら[10]は，観光ガイドブックを通して日本の「インフラツーリズム」を紹介している．このような取り組みは，観光行動を通した健康促進による医療費の削減やこれまで観光地として注目されていなかった地域での雇用の創出などの利点があり，近年，注目されている．

観光の形態を用いたソーシャルメディアの解析に関する研究として，柴田ら[11]では，世界最大級の旅行ブログサイトであるTravelBlog*1に公開されている大量の旅行ブログエントリを収集し，深層学習ベースの手法を用いてそれぞれの観光の形態に分類する手法を提案している．また，分類の結果から，観光の形態による世界中の観光情報の検索が可能となったほか，分類結果をGoogle Earth上にマッピングすることで，可視化システムを通した地域の魅力発見を可能にしている．この研究で用いられている旅行ブログエントリは，Twitterなどと異なりテキストなどの情報が比較的豊富に掲載されているという特徴がある．また，TravelBlogには，世界中の国について記述された旅行ブログエントリが大量に公開されている．

本研究では，より多くの訪問国履歴データを収集するため，世界中の国について書かれたソーシャルメディアを大量に必要とすることから，柴田らと同じ，TravelBlogの旅行ブログエントリを用いる．また，観光の形態という観点から，それぞれの国について記述された旅行ブログエントリを解析することで，各国の観光の特徴を獲得し，得られた特徴を考慮することでより正確な旅行者の将来の訪問国予測を行う．

3. Wikipedia と観光の形態の特徴を考慮した将来の訪問国の予測

3.1 Wikipedia を用いた各国の基本的な情報の獲得

旅行者が訪問国を決定する際には，各国の位置情報や文化などの基本的な情報を考慮することが考えられる．そこで，本研究では，このような各国の基本的な情報を考慮した将来の訪問国の予測を行う．各国の基本的な情報の獲得

*1 https://www.travelblog.org/

には，Wikipedia を利用する．Wikipedia には，各国のページが作成されており，位置情報や文化，公用語，歴史などの基本的な情報が記述されている．これらの Wikipedia の情報を獲得する方法には，Web 上に公開され，かつ手軽に利用することができる Wikipedia2Vec[12]の事前学習モデルを用いる．Wikipedia2Vec とは，Wikipedia のデータを用いて，単語の分散表現とエンティティの分散表現を同一空間で学習・表現する手法であり，事前学習モデルを用いることで単語やエンティティの分散表現を獲得することができる．本研究では，この Wikipedia2Vec の事前学習モデルから，それぞれの国名に該当するエンティティの分散表現を獲得し，これを予測の際に考慮する．

しかし，アメリカにはハワイ州やアラスカ州，グアム島などが含まれるように，同じ国であっても地理的に離れている場合が存在する．このような場合，各国の基本的な情報に含まれる位置情報や文化などが大きく異なることがある．そこで，本研究ではこの問題に対し，離れていると疑われる国については，著者らで分割すべきかどうかを，位置情報や文化などの側面から議論し，最終的に多数決を行うことで分割の判断を行った．

3.2 観光の形態を用いた観光の特徴の獲得

近年の観光では，旅行者は多様な目的を持っていることが多いため，その訪問先で体験できる観光の特徴は，訪問国の決定の際の重要な判断材料になると考えられる．そこで，本研究では，観光の特徴の一つとして，先行研究の柴田らが参考文献[9-10, 13-15]を参考に定義した，表1中の6つの観光の形態を考慮し，より正確な予測を試みる．ここから，3.2.1節では，Random Forestを用いた手がかり語の獲得について，3.2.2節では，観光の形態の特徴量の獲得について述べる．

3.2.1 Random Forest を用いた手がかり語の獲得

柴田らの研究では，収集した旅行ブログエントリを，機械学習手法を用いてそれぞれの観光の形態に分類することで，解析を行っている．分類実験に用いられたデータセットの内訳を表2に示す．このデータセットは，英語で書かれた旅行ブログエントリを対象に，英語の読み書きに不自由がない大学生と著者らで1,909件の旅行ブログエントリを分類したものである．各観光の形態の旅行ブログエントリの件数の合計が，分類した旅行ブログエントリの総数1,909件より小さくなっているのは，一つの旅行ブログエントリに対して複数の観光の形態が付与されることや一つも付与されない旅行ブログエントリが含まれるためである．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，一つも付与されない旅行ブログエントリは1,182件であった．本研究では，このデータセットを用い，旅行ブログエントリの解析を行う．

表 2: 人手で分類した分類結果の内訳

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

旅行ブログエントリを用いた各国の観光の形態の特徴を獲得する方法の一つとして，国ごとに集めた大量の旅行ブログエントリを分類し，その結果からそれぞれの国の観光の形態の割合を用いることで，例えば，「世界遺産が有名なエジプトでは，ヘリテージツーリズムの割合が大きい」などの特徴量を獲得する，といったことが考えられる．しかし，TravelBlogでは，小さな国や有名ではない国は投稿されている旅行ブログエントリの数が少ない場合があり，こういった場合，十分な特徴量を獲得することが難しくなる

表 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[10]	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[9]	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[13]	スポーツを体験または観戦することを目的とした観光.	野球，サッカー，オリンピック
グリーンツーリズム[14]	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[14]	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[15]	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

ことが予想される．そこで，本研究では，旅行ブログエントリから観光の形態の特徴を獲得する方法の一つとして，表2のデータセットに含まれる旅行ブログエントリのテキストデータから，各観光の形態に関連する手がかり語を獲得し，この手がかり語を用いた国ごとの特徴量の獲得を行う．

手がかり語の獲得には，機械学習手法のRandom Forest[16]を用いる．Random Forestとは，複数の決定木を用いてアンサンブル学習を行う機械学習アルゴリズムであり，分類や回帰を行うことが可能であるが，加えて学習を通して得られた入力データの各要素の重要度を獲得することができる．そこで，本研究では，表2のデータセットの1,909件すべての旅行ブログエントリのテキストデータをRandom Forestの学習に用い，旅行ブログエントリをそれぞれの観光の形態へ分類する2値分類器の学習結果から，重要度が高い単語を手がかり語として獲得する．

学習を行う際のテキストデータは，前処理としてRNNTagger[17]による品詞タグ付けおよび名詞以外の除去を行い，Bag-of-wordsモデルで分散表現に変換する処理を行った．また，Random Forestの実装には，プログラミング言語のPython3および機械学習ライブラリであるscikit-learn*2のRandom Forest Classifier*3を用いた．Random Forest Classifierのパラメータについては，実行毎に結果が変わることを防ぐため，random_stateを0とし，そのほかのパラメータについては，scikit-learnであらかじめ設定されている値を採用した．重要度については，scikit-learnでは，以下の式(1)で表されるジニ不純度が用いられている．

$$G(k) = 1 - \sum_{i=1}^{J} {p_i}^2 \qquad (1)$$

$G(k)$はノードkにおけるジニ不純度，Jはラベルの種類の数，p_iはラベルの割合を表す．本研究では，Random Forestを用いて，各観光の形態で2値分類の学習を行うため$J = 2$，$p_1 + p_2 = 1$となる．重要度は，特徴量ごとに分割することで，どの程度ジニ不純度を下げることができたかを求め，最後に各特徴量の重要度の合計が1になるように正規化される．

Random Forestで獲得した重要度の分布と上位10単語をそれぞれ表3，4に示す．表3では「グリーンツーリズム」における重要度の分布を表しているが，総単語数17,394に対して，半分以上の10,593単語は重要度が0.000となっており，そのほかの単語もほとんどが重要度0.002未満となっている．このことから，重要度を用いることで単語数が削減可能であることがわかる．また，表4中の獲得した手がかり語を見てみると，正例が少ない「スポーツツーリズム」と「カルチュラルツーリズム」では，いくつか観光の形態を表しているとは言えない単語が含まれているが，「インフラ，ハードツーリズム」では，博物館を表す“museum”やショッピングモールを表す“mall”が含まれていることがわかり，そのほかの観光の形態でも，ある程度観光の形態を表す単語を獲得できている．本研究では，ここで得られた重要度が高い単語を手がかり語として利用し，観光の形態の特徴量を獲得する．

表3: 「グリーンツーリズム」における重要度の分布

重要度	単語数
0.000	10,593
0.000 ～ 0.002	6,749
0.002 ～ 0.004	33
0.004 ～ 0.006	14
0.006 ～ 0.008	4
0.008 ～ 0.010	1
0.010 ～ 1.000	0
総単語数	17,394

3.2.2 観光の形態の特徴量の獲得

手がかり語を用いた特徴量の作成を行うため，本研究では，TravelBlogから国ごとに5,500件以上の旅行ブログエントリを収集した．なお，比較的小さい国や発展途上国などでは，投稿されている旅行ブログエントリの数が5,500件に満たない国もある．こういった国については，公開されている旅行ブログエントリ分のみの収集を行った．収集する旅行ブログエントリを5,500件とした理由については，十分に特徴量を獲得するため，表2のデータセットで分類対象と

表 4: Random Forest を用いて獲得した重要度の上位 10 単語

観光の形態	重要度の上位 10 単語
インフラ，ハードツーリズム	museum, mall, show, city, exhibit, picture, cuisine, building, tower, shop
ヘルスツーリズム	hike, mountain, trail, hiker, glacier, hiking, river, way, rock, cloud
スポーツツーリズム	game, stadium, onsen, ski, baseball, rafting, match, kayak, timewwwooohhhooo, mountain
グリーンツーリズム	river, mountain, rock, park, boat, view, day, valley, road, town
ヘリテージツーリズム	temple, city, site, century, street, emperor, statue, wall, complex, palace
カルチュラルツーリズム	Tribe, absurdity, pilgrim, prat, koreum, hellooooo, texaswillie, festival, transfer, dukk

*2 https://scikit-learn.org/stable/
*3 https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

なった1,909件の2倍より多い5,000件を目安にし，英語で書かれた旅行ブログエントリのみを対象とするため，5,000件より少し多い5,500件とした．

手がかり語による特徴量の作成方法については，ある国について投稿された旅行ブログエントリを収集し，それぞれの手がかり語が出現する割合を算出する．例えば，日本について投稿された旅行ブログエントリを100件収集し，100件中30件に “temple” が含まれる場合，日本の “temple” についての特徴量は0.3となる．このように，それぞれの手がかり語で割合を求め，これをその国の特徴量とし，Wikipedia2Vecから獲得した分散表現に加えて予測の際に考慮する．

4. 評価実験

4.1 実験設定

【実験に用いるデータ】

訪問履歴データの収集には，TravelBlogの37,371個のアカウントを対象とし，それぞれの訪問国履歴データを収集することでデータセットを構築した．時系列データを作成するための旅行日の判定には，旅行ブログエントリ内に記述されているブログ著者が設定した日付を用いた．また，TravelBlogでは，ブログエントリを投稿する際にあらかじめ訪問国を設定する場合があり，設定した場合，その内容はURLに反映される．本実験では，このURLから国名を抽出し，これを訪問国とした．

表 5: 実験に用いるデータセットの内訳

訓練データのアカウント数	7,684	9,604
検証データのアカウント数	960	
評価データのアカウント数	960	
アカウントごとの訪問国数の平均		14.1

TravelBlogでは，「旅行1日目」，「旅行2日目」のように，短期間で同一国についての投稿が複数存在するブログ著者が多く見られた．そのため，60日以内に連続して同じ訪問国の履歴データが存在する場合，これらの投稿をまとめる処理を行った．また，過去の訪問国履歴から将来の訪問国を予測するためには，各アカウントの訪問国の数がある程度必要となる．そのため，訪問国が6か国以上の訪問国履歴が存在するアカウントのみに絞った．本研究では，これらの前処理を行い，ブログデータの日付が2002年4月2日から2020年10月31日，訪問国が198か国とそのほかの38の地域から構成されるデータセットを構築した．データセットの内訳を表5に示す．アカウント数は9,604個となり，これを8:1:1に分割することで，訓練・検証・評価データとした．また，アカウントごとの訪問国数の平均は14.1となった．

データセットには，2002年4月2日から2020年10月31日の旅行ブログエントリデータが含まれているが，オリンピックなどの世界規模のイベントの有無や国際情勢によっては，局所的な観光客の増減があると考えられる．そこで，夏季オリンピックが開催された2008年の中国，2012年のイギリス，2016年のブラジルの三つの国を例に，世界規模のイベントによってデータセット中の旅行ブログエントリの件数に偏りが生じるかを検証するため，それぞれの年および国の割合を求めた．割合の算出方法については，2012年のイギリスを例にすると，2012年ではデータセット内に8,112件の旅行ブログエントリデータが存在し，イギリスに関する旅行ブログエントリは456件であったため, 2012年のイギリスの割合は$456/8{,}112 \fallingdotseq 0.056$となる．データセット中の各年に対する国の割合を求めた結果を図1に示す. 横軸は年を，縦軸は割合を表している．図1より，それぞれの国のオリンピックが開催された年では，顕著な割合の上昇は確認できなかった．ここから，国単位のデータでは，年ごとに大きな割合の変化が現れにくいことが考えられる．また，2020年では，中国とイギリスで割合が下がっていることが確認できる．これは，COVID-19が旅行者の訪問国の決定に影響したことが考えられる．

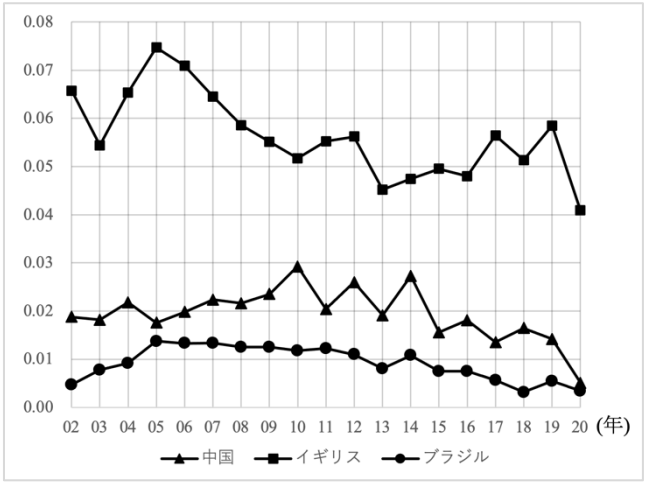


図1: 各年に対するデータセット中の国の割合

【実験条件】

プログラミング言語にはPython3を用い，各実験手法の実装を行う．また，提案手法とベースライン手法のLSTMには，オープンソースの深層学習ライブラリであるKeras*4を用いる．LSTMの主なパラメータについては，最適化アルゴリズムをAdam，バッチサイズを2,048，学習率を0.001，LSTM層を1層，出力ユニット数を入力ユニット数と同じ数とし，検証データによる実験からepoch数を決定する．

Wikipedia2Vecの事前学習モデルには，Web上に公開されている英語のモデルで最も軽量な100次元のモデル*5を用

*4 https://keras.io/
*5 http://wikipedia2vec.s3.amazonaws.com/models/en/2018-04-20/enwiki_20180420_nolg_100d.pkl.bz2

いる．これは，次元数の異なるいくつかのモデルで予備実験を行ったところ，実験結果への影響がほとんど確認できなかったためである．また，各国の観光の形態の特徴量を獲得するため，表2の観光の形態のデータセットに含まれる旅行ブログエントリのテキストデータから，Random Forestを用いて収集した，各観光の形態の重要度上位10単語を選択し，重複している単語を除いた合計55単語で国ごとの旅行ブログエントリに含まれる割合を求めた．なお，使用する手がかり語を重要度の上位10単語とした理由は，手がかり語上位10, 20単語のそれぞれで予備実験を行ったところ，上位10単語の方が良い結果が得られたためである．また，テキストの前処理には，まず，fastText[18]の言語判定モデルによって，英語と判定された旅行ブログエントリのみに絞り，その後， RNNTaggerを用いた品詞タグ付けを行うことで，名詞のみとした．

【評価尺度】

本実験では，表5のデータセットを用い，直前の5か国を入力データとし，次の訪問国を予測することを繰り返すことで実験の訓練・検証・評価を行う．評価尺度には，Accuracy@k (k = 1, 3, 5, 10) およびMRR (Mean Reciprocal Rank) を用いる．Accuracy@kとは，Top-k accuracyと呼ばれる評価指標を指し，予測の確率上位k件に正解が含まれる場合は1, 正解が含まれない場合は0として評価を行う．また，MRRについては，予測結果に正解が n 番目にある場合，その逆数である $1/n$ を評価値とする．なお，本研究では，データセットに含まれる198か国とそのほかの38の地域の合計236個の要素から成る予測結果を用いて，これらの評価値を算出する．

【実験手法】

本実験では，下記の1種類の提案手法と8種類のベースライン手法で実験を行った．なお，LSTMなどはパラメータの初期値がランダムに設定されるため，実行のたびに結果が変わる．そのため，本実験では，10回の実行および評価値の算出を行い，各評価値の平均をその手法の実験結果として採用する．

ベースライン手法

- Random: 無作為に選んだ国を予測結果とする．
- n-gram (n = 4): 訓練データを用いてn-gramモデルを構築する．本実験では，予備実験で最も良い結果が得られたn = 4の結果を報告する．
- Distance: 最後に訪れた国から距離が近い国を予測結果とする．位置情報データの収集については，アマノ技研が公開する世界の首都の位置データ*6を用いて198か国の首都の位置情報を収集し，そのほかの38の地域はPythonライブラリのGeocoder*7を用いて収集した．また，二つの位置情報間の距離の算出には，Pythonライブラリの GeoPy*8を用いた．
- Most Popular: 訓練データで頻出する国を予測結果とする．上位10か国は，アメリカ，イギリス，フランス，タイ，ドイツ，オーストラリア，イタリア，スペイン，アルゼンチン，カナダの順番であった．
- StarSpace: Facebookが公開している汎用ニューラルモデルのStarSpace[19]を用い，協調フィルタリングの問題として，学習・予測を行う．主なパラメータについては，学習率を0.001，次元数を100次元とする．
- Wikipedia2Vec: 五つのそれぞれの訪問国をWikipedia2Vecの事前学習モデルを用いてエンティティの分散表現に変換し，これらの分散表現の平均とコサイン類似度が高い国を予測結果とする．
- LSTM(Wiki): Wikipedia2Vecの事前学習モデルからそれぞれの国や地域のエンティティの分散表現を獲得し，これを入力データとしたLSTMにより予測を行う．LSTM層は1層とし，出力ユニット数を入力データの分散表現と同じ100とする．

提案手法

- LSTM(Wiki+Type): LSTM(Wiki) に対して，観光の形態の手がかり語で得られた割合の55次元を加えることでLSTM(Wiki) を拡張し，予測を行う．LSTM層は1層とし，出力ユニット数を入力データの分散表現と同じ155とする．

【LSTM(Wiki+Type) について】

Wikipedia2Vecの事前学習から獲得した分散表現および各国の旅行ブログエントリから観光の形態のそれぞれの手がかり語の割合を算出することで獲得した分散表現を入力データとし，これら二つの情報を考慮した予測を行う．LSTM(Wiki+Type) の詳細として，予測の例を図2に示す．この例では，10か国目を予測する様子を表しており，直前の5か国目から9か国目を入力データとしている．入力データの処理の手順については，まず，入力データとなる国をWikipedia2Vecの事前学習モデルおよび観光の形態の手がかり語の割合を用いて，それぞれ100次元と55次元の分散表現に変換し，これら二つの分散表現を結合することで得られる新たな155次元の分散表現を国の分散表現とする．そして，得られた国の分散表現をLSTMの入力とし，学習・予測を行う．予測結果の求め方については，$f(x) = 1/(1 + e^{-x})$ で表されるシグモイド関数を採用し，国の数だけ用意した236個の出力値が大きい国から順位づけを行うことで予測結果とする．本実験では，この図2のように，過去の五つの訪問国を入力データとし，次の訪問国を予測することを繰り返す．なお，括弧内の数字は分散表現の次元数を表して

*6 https://amano-tec.com/data/world.html
*7 https://geocoder.readthedocs.io/
*8 https://geopy.readthedocs.io/en/stable/

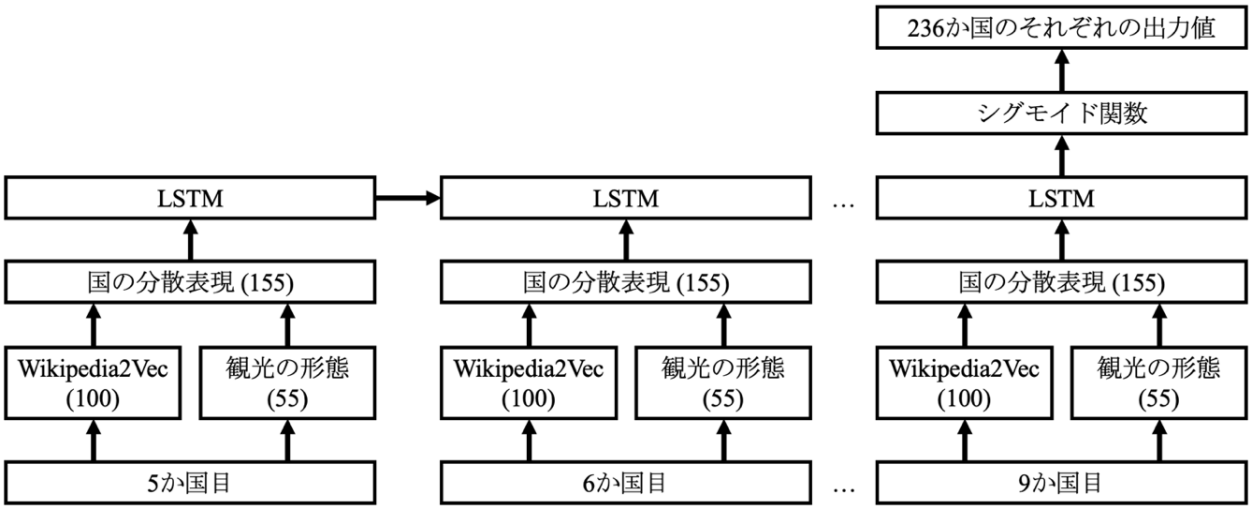


図2: LSTM(Wiki+Type) による10か国目の予測の例

いる．

4.2 実験結果と考察

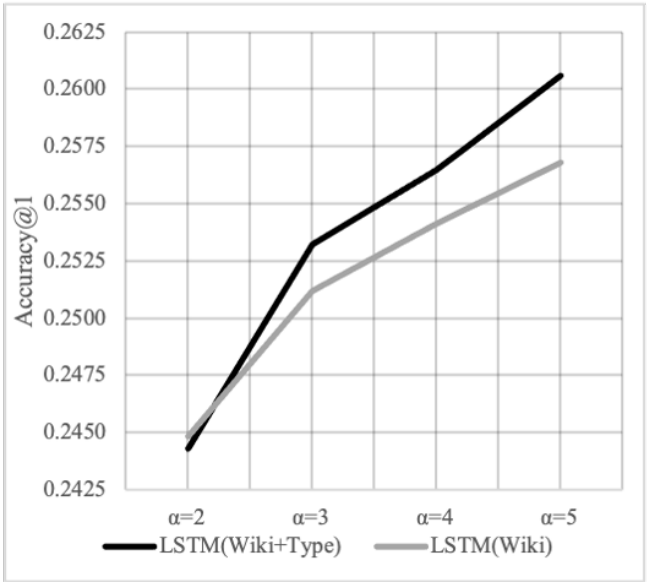
将来の訪問国予測の評価実験の結果を表6に示す．実験結果から，ベースライン手法のLSTM(Wiki) との差が小さいものの，提案手法のLSTM(Wiki+Type) で，すべての評価指標において最も高い値が得られた．また，LSTMを用いた手法とそのほかの手法を比較すると，LSTMの方が高い値が得られた．このことから，将来の訪問国の予測にLSTMが有効であると言える．そのほか，LSTM以外のベースライン手法では，StarSpaceとWikipedia2Vecで比較的良い値が得られた．

提案手法のLSTM(Wiki+Type) とベースライン手法のLSTM(Wiki) の差が小さいため，提案手法の有効性を統計的な側面とシステムとしての実用性の二つで検証する．まず，統計的な有意性について，有意水準5%における検定を行った．検定に利用する評価データを増やすため，それぞれの手法の評価値の算出を100回ずつ行った．検定の結果，すべての評価指標でLSTM(Wiki) とLSTM(Wiki+Type) の違いが確認された．また，実用性を検証するため，実験では旅行者が次に訪れる1か国を正解として扱っているが，その後に訪れている2か国を加えた計3か国の中に，予測した国が含まれている確率を調べた．調査の結果，ベースライン手法のLSTM(Wiki) では27.17%であったのに対し，提案手法のLSTM(Wiki+Type) ではより高い27.54%であった．これらのことから，旅行者の将来の訪問国を予測するにあたって，Wikipediaに加えて，観光の形態の特徴を考慮することが有効である可能性を示唆している．

本実験では，過去の五つの訪問国履歴を用いて次の訪問国を予測したが，LSTMは入力データの時系列の長さが結果に影響することが報告されている[20]．そのため，考慮する訪問国は結果を左右する重要な要素であると考えられる．そこで，考慮する訪問国の数を2〜4に変化させた場合のLSTM(Wiki) とLSTM(Wiki+Type) のそれぞれの評価値の変化を追加で調査した．考慮する訪問国の数の変化に伴うAccuracy@1およびMRRの変化を図3に示す．図中の直線は，黒色が提案手法のLSTM(Wiki+Type) を，灰色がベースライン手法のLSTM(Wiki) を表しており，横軸のαは考慮する訪問国の数，縦軸はその評価値の値を表している．それぞれの評価値の算出には，表5の8:1:1に分割したデータセットで訓練・評価・検証を用い，LSTM(Wiki+Type) およびLSTM(Wiki) で同様の条件下で実験を行っている．図3を確認すると，考慮する訪問国の数を2から5へと増やしていくことで，それぞれの評価値が良くなっていることがわかる．これは，考慮する情報量が増えたことで，より正確な予測を可能にしていることが考えられる．また，2か国を考慮し

表 6: 将来の訪問国予測の評価実験の結果

	手法	Accuracy@1	Accuracy@3	Accuracy@5	Accuracy@10	MRR
ベースライン手法	Random	0.44%	1.31%	2.15%	4.20%	0.026
	n-gram (n = 4)	3.75%	8.75%	12.71%	19.43%	-
	Distance	6.60%	25.78%	37.03%	50.05%	0.209
	Most Popular	7.85%	16.01%	22.24%	37.29%	0.170
	Wikipedia2Vec	12.36%	25.12%	33.83%	49.33%	0.236
	StarSpace	12.76%	25.90%	34.90%	49.72%	0.238
	LSTM(Wiki)	25.68%	46.00%	56.22%	70.24%	0.401
提案手法	LSTM(Wiki+Type)	26.06%	46.31%	56.61%	70.30%	0.403



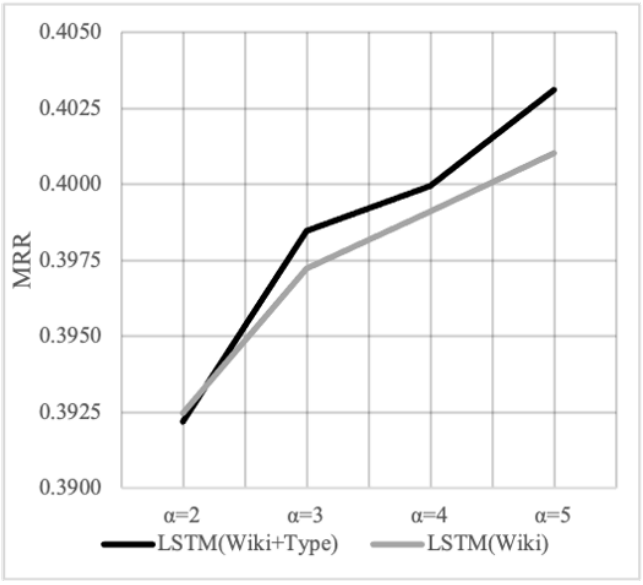


図3: 考慮する訪問国の数の変化に伴うAccuracy@1およびMRRの変化

た場合では，評価値のAccuracy@1およびMRRでほとんど差は確認できなかったものの，考慮する訪問国の数が3以上になると，提案手法のLSTM(Wiki+Type) の方が良い値となっている．このことから，手がかり語を用いて観光の形態を考慮することが将来の訪問国の予測に良い影響を与えていることが考えられる．

表 7: Wikipedia2Vec と観光の形態の手がかり語のエジプトに対するコサイン類似度の上位 5 か国

Wikipedia2Vec		観光の形態	
国名，地域名	類似度	国名，地域名	類似度
リビア	0.874	カンボジア	0.968
チュニジア	0.870	インド	0.967
ヨルダン	0.863	インドネシア	0.963
シリア	0.860	ミャンマー	0.963
スーダン	0.837	ベトナム	0.962

ここで，実際に，観光の形態の手がかり語から獲得した分散表現がどのような特徴を獲得しているのかについて，コサイン類似度を用いてWikipedia2Vecとの比較を行う．まず，ピラミッドなどの世界遺産が有名なエジプトを例に，Wikipedia2Vecおよび観光の形態の手がかり語のそれぞれの分散表現におけるコサイン類似度の上位5か国とその類似度を表7に示す．Wikipedia2Vecの上位5か国に注目すると，中東や北アフリカなどのエジプトと比較的距離が近い国が並んでいる．この結果に対して観光の形態では，TravelBlogに世界遺産の仏教遺跡に関する投稿が多く見られるインドネシアやアンコール・ワットに関する投稿が多く見られるカンボジアなど，「ヘリテージツーリズム」が有名であると思われる国が並んでいる．特に，インドネシアでは，表2の人手で観光の形態へ分類を行ったデータセットに含まれていなかったため，獲得した手がかり語によって，「ヘリテージツーリズム」の特徴を捉えることができていると考えられる．

表 8: Wikipedia2Vec と観光の形態の手がかり語のシンガポールに対するコサイン類似度の上位 5 か国

Wikipedia2Vec		観光の形態	
国名，地域名	類似度	国名，地域名	類似度
マレーシア	0.883	香港	0.970
香港	0.864	クウェート	0.964
ブルネイ	0.822	アラブ首長国連邦	0.962
インドネシア	0.788	マレーシア	0.959
台湾	0.786	マカオ	0.959

次に，マリーナベイ・サンズなどの近代的な建造物が有名なシンガポールを例に，Wikipedia2Vec および観光の形態の手がかり語のそれぞれの分散表現におけるコサイン類似度の上位 5 か国とその類似度を表 8 に示す．Wikipedia2Vec では表 7 と同様に，距離が近い国を中心に類似度が高くなっているが，観光の形態では，TravelBlog に大きなビルや夜景についての画像が多く投稿されているアラブ首長国連邦やクウェート・タワーが観光地として知られているクウェートなど，「インフラ，ハードツーリズム」が有名であると思われる国が並んでいる．また，これらの国も表 2 のデータセットには含まれていない．これらの例のように，観光の形態を考慮することで Wikipedia から獲得できなかった観光の特徴を捉えることが可能になったことで，より正確な予測が可能になったと考えられる．

5. おわりに

本研究では，TravelBlogに公開されている大量の旅行ブログエントリを収集し，訪問国履歴データセットを構築した．また，構築したデータセットを用い，過去の訪問国の基本

的な情報および観光の特徴を考慮した，LSTMによる将来の訪問国の予測手法を提案した．各国の基本的な情報には，Wikipediaを採用し，Web上に公開されているWikipedia2Vecの事前学習モデルから分散表現を獲得した．各国の観光の特徴については，まず，観光の形態のデータセットを用いてRandom Forestによる分類学習を行い，学習の結果から得られた重要度の高い単語を手がかり語として収集した．そして，国ごとに集めた旅行ブログエントリから，それぞれの手がかり語の出現する割合を求めることで，観光の形態の特徴を考慮した分散表現を作成した．

上記の二つの分散表現を結合することで得られる新たな分散表現を入力データとしたLSTMの有効性を検証するため，いくつかのベースライン手法を加えた評価実験では，評価指標のAccuracy@k (k = 1, 3, 5, 10) およびMRRにおいて，すべての評価指標でWikipediaおよび観光の形態の特徴を考慮したLSTMが最も高い値を得た．また，有意水準5%による検定の結果，観光の形態を考慮することの有効性が確認された．

本研究では，各国の観光の特徴を獲得するため，英語で書かれた旅行ブログエントリを観光の形態に基づいて解析した．今後の発展では，英語だけでなく，中国語や日本語など，そのほかの言語で書かれた旅行ブログエントリも解析に用いることで，より多角的な観光の特徴の獲得を試みることを検討している．

参考文献

[1] 公益財団法人日本交通公社 : 新型コロナウイルス感染症流行下の日本人旅行者の動向（その 4），https://www.jtb.or.jp/wp-content/uploads/2020/08/covid-19-japanese-tourists-4_JTBF20200730.pdf (2020)

[2] Y. Su, X. Li, W. Tang, J. Xiang and Y. He, “Next Check-in Location Prediction via Footprints and Friendship on Location-Based Social Networks”, Proceedings of the 19th IEEE International Conference on Mobile Data Management (MDM’18), pp. 251-256, (2018)

[3] R. Baraglia, C. I. Muntean, F. M. Nardini and F. Silvestri, “LearNext: Learning to Predict Tourists Movements”, Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM’13), pp. 751-756 (2013)

[4] Q. Liu, S. Wu, L. Wang and T. Tan, “Predicting the Next Location: A Recurrent Model with Spatial and Temporal Context”, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), pp. 194-200 (2016)

[5] D. Kong and F. Wu, “HST-LSTM: A Hierarchical Spatial-Temporal Long-Short Term Memory Network for Location Prediction”, Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18), pp. 2341-2347 (2018)

[6] S. Hochreiter and J. Schmidhuber, “LONG SHORT-TERM MEMORY”, Neural Computation, Vol. 9, No. 8, pp. 1735-1780 (1997)

[7] A. E. Majjodi, M. Elahi, N. E. Ioini and C. Trattner, “Towards Generating Personalized Country Recommendation”, Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP’20), pp. 71-76 (2020)

[8] K. J. Johnson, N. M. Gallagher, E. D. Mintz, A. E. Newton, G. W. Brunette and P. E. Kozarsky, “From the CDC: New Country-Specific Recommendations for Pre-Travel Typhoid Vaccination”, Travel Medicine, Vol. 18, No. 8, pp. 430-433 (2011)

[9] 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)

[10] 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)

[11] 柴田有基，篠田広人，難波英嗣，石野亜耶，竹澤寿幸：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，観光と情報，Vol. 16, No. 1, pp. 49-61 (2020)

[12] I. Yamada, A. Asai, J. Sakuma, H. Shindo, H. Takeda, Y. Takefuji, and Y. Matsumoto, “Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia”, arXiv:1812.06280 [cs.CL] (2018)

[13] 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)

[14] 山下晋司：観光学キーワード，有斐閣 (2011)

[15] 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)

[16] L. Breiman, “Random Forests”, Machine Learning, Vol. 45, No. 1, pp. 5-32 (2001)

[17] H. Schmid, “Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts”, Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019), pp. 133-137 (2019)

[18] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, “Bag of Tricks for Efficient Text Classification”, arXiv:1607.01759 [cs.CL] (2016)

[19] L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, J. Weston, “StarSpace: Embed All The Things!”, arXiv:1709.03856 [cs.CL] (2017)

[20] S. J. Kim, S. H. Kim, H. M. Lee, S. H. Lim, G. -Y. Kwon and Y. -J. Shin, “State of Health Estimation of Li-Ion Batteries Using Multi-Input LSTM with Optimal Sequence Length”, Proceedings of the 29th International Symposium on Industrial Electronics (ISIE), pp. 1336-1341 (2020)