地球の歩き方旅行記データセット

大内啓樹 [1,4,*]　進藤裕之 [1]　若宮翔子 [1]　松田裕貴 [1]
井之上直也 [2]　東山翔平 [3,1]　中村哲 [1]　渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学　[2] 北陸先端科学技術大学院大学
[3] 情報通信研究機構　[4] 理化学研究所
* 責任著者：hiroki.ouchi@is.naist.jp
{shindo,wakamiya,yukimat,s-nakamura,taro}@is.naist.jp
naoya-i@jaist.ac.jp　shohei.higashiyama@nict.go.jp

概要

「**地球の歩き方旅行記データセット**」を構築し，学術研究用に無償提供を開始した．本データセットは日本語テキストデータであり，4,500 の国内旅行記と 9,500 の海外旅行記から構成され，全体で 3,100 万単語を超える規模である．これまでは，研究目的で共同利用可能な旅行記データがほとんどなく，各研究者が自前でデータを用意するしかなかった．本データセットの提供によって，公平な手法の比較をはじめとするオープンサイエンスの進展および一層の学術研究の発展が期待できる．本稿では，本データセットの学術的意義，特徴，今後の展望を述べる．

1　はじめに

人間と「場所」の関係性に注目が集まっている．コロナ禍において，ある場所の混雑度（人間の集中）や場所間での往来（人間の移動）に関する情報は，政府・地域自治体・個人といった粒度を問わず，行動の促進/抑制の意思決定に極めて重要である．こうした背景から我々は，人間の行動を「場所」という観点から分析するための方法論を探求している．特に「テキスト」を分析対象データとし，計算機によって，登場人物の行動の舞台背景となる「場所」を読み取り，実世界の地図上に接地することを目指している．その第一歩であり，システムの開発および評価の基盤として，「**地球の歩き方旅行記データセット**（Arukikata Travelogue Dataset）」[1] を構築した．本データセットは，学術研究機関が研究目的で使用するために，株式会社地球の歩き方が提供元となり，国立情報学研究所情報学研究データリポジトリ（IDR）より無料配布されている[1)]．

1）https://www.nii.ac.jp/dsc/idr/arukikata/

なぜ「テキスト」を用いるのか？　人間の位置を把握するだけなら，スマートフォンなどのモバイル端末に搭載された GPS 機能を利用することによって可能である．しかし GPS データから，人間と場所の相互関係を把握することは難しい．例えば，その場所における人間の行為，その場所に対する価値付与，その場所から受ける印象や感覚である．こうした情報は，特に地理学や文化人類学において，人間活動のダイナミクスと環境条件を分析するための情報として極めて重要な役割を担う．この種の情報を含む代表的な資源がテキストである．テキストをうまく構造化し，整理することによって，こうした価値ある情報を引き出せる可能性がある．以上の理由から，テキストデータを対象として採用した．

なぜ「旅行記」を用いるのか？　旅行記を解析対象とする従来研究では，「観光客」や「観光地」，およびそれらの関係という視座から研究の意義を捉え，研究目的を設定することが多い [2, 3, 4]．一方で我々は，それより一段抽象度の高い視座から意義を捉えている．つまり，「人間」と「場所」という実世界の基礎的な構成要素がどのように関わり合っているのか，その関わりはどのようにテキストに描かれるのかという視座である．そのような内容を含む典型的なテキスト（ジャンル）が「旅行記」である．同種の内容が，「小説」「新聞記事」「SNS 投稿」に含まれる場合も少なくない．将来的には，より多様なテキストを対象とすることを見据えつつ，その出発点として「旅行記」を題材に設定した．

本データセットの学術的意義　これまでも旅行記は，テキストマイニングの分析対象テキストとして頻繁に用いられてきた．特に観光情報学分野において，旅行記は各場所・施設の評判分析および種々の情報抽出の題材として用いられてきた [5]．しかし

表 1　旅行記の例.

会津若松へ向かう磐越西線の接続を考慮して選んだやまびこ 203 号は E5 系での運転でした。何度も乗っている E2 系よりも座席が広く感じ、快適な移動でした。
・・・
会津若松駅から快速あいづ 4 号に乗車し、郡山へ向かいました。
会津若松は晴れ間がありましたが、山を上るにつれて雲が増えて行き、途中から雨が降り出しました。
天気予報通りでしたが、今回の旅行は暖かい 2 日間で移動時間を除いて雨に降られることがなかったのはラッキーでした。

表 2　旅スケジュールの例.

```
1 日目　2021 年 10 月 15 日 (金)
  05:40 - 05:50　自宅
  05:50 - 05:53　最寄駅
   ・・・
  15:33 - 23:59　庄助の宿　瀧の湯

2 日目　2021 年 10 月 16 日 (土)
  00:00 - 09:25　庄助の宿　瀧の湯
  09:25 - 09:32　東山温泉入口
                 (瀧の湯前) バス停
   ・・・
  17:33 - 17:38　最寄駅
  17:38 - 17:45　自宅
```

ながら，それぞれの研究者がウェブ上の旅行記投稿サイトなどから独自に取得した旅行記データを用いることが多く[2)]，研究の再現や実験結果の公平な比較分析が困難であった．本旅行記データセットを一定条件下での利用機会についてオープン化することによって，大須賀ら [6] が指摘するように，大学等の研究者が実社会のデータや実用性の高いデータを使用できるだけでなく，使用したデータセットが特定可能となることにより，**研究の透明性・再現性が担保**され，他の研究との比較も格段に容易となる．ひいてはオープンサイエンスを進展させ，研究の知見の蓄積が加速することで，一層の学術研究の発展が期待されるため，将来にわたって学術的意義は大きいと考えられる．

2　本データセットの構成と特徴

本データセットは，2007 年 11 月から 2022 年 2 月までの期間に，株式会社地球の歩き方が運営するサイト[3)]上の旅行記投稿サービス[4)]に投稿された情報に基づいている．具体的には，ユーザが書き記した「**旅行記**」とその旅程を記した「**旅スケジュール（旅スケ）**」から構成される．旅行記として，国内旅行記に加え，海外旅行記も含まれる．

2)　一般ユーザの書いたコンテンツである旅行記は，著作権に関わる諸事項をクリアしなければ再配布できないことが主な原因として挙げられる．
3)　https://www.arukikata.co.jp/
4)　https://tabisuke.arukikata.co.jp/（2022 年 3 月で本サービスは終了している）

2.1　国内・海外旅行記

表 1 は旅行記の実例を示している．一般的に旅行記は，著者（ユーザ）の視点から書かれた一人称視点の文章である．読者は著者の視点を借りて，著者の辿った旅路やその過程で見た景色などを疑似体験することができる．旅行記の各記事はある程度まとまった分量であるため，著者の行動の系列，場面展開，共参照関係をはじめとする文間・段落間にまたがる談話的要素も多い．その点は，ツイッターの投稿を代表とする SNS の短い文章とは異なる特徴と言える．典型的な記述内容として，人間（著者）の 1 日の（旅行）行動が描かれており，行動の時系列的な分析にも適したテキストであると言える．その他にも，場所や風景の描写，各場所や全旅程についての感想をはじめ多彩な内容が含まれるため，種々の応用へつながる可能性がある．

2.2　旅スケジュール

表 2 は旅スケの実例を示している．主な構成要素のひとつは，著者が滞在した「場所」である．例えば，表 2 中の「**自宅**」「**最寄駅**」「**庄助の宿 瀧の湯**」などが場所に該当する．旅スケの入力は任意であり，本文中に記載のある「場所」が旅スケにも記載されているとは限らない．逆に，旅スケに記載されている「場所」でも本文中には記載がない場合もある．もうひとつの主な構成要素は，著者が各場所に滞在した「時間帯」である．例えば，表 2 中の「05:40 - 05:50」「05:50 - 05:53」などが時間帯に該当する．時間帯に関する情報は，著者が自由に記述できるため，「朝」や「夕刻」といった粗い粒度のものもある．以上の情報を踏まえると，人間の 1 日の行動を「場所」に加え「時間」の観点から分析するためにも本データセットは利用可能である．

2.3　データの記述統計

前述したように，本データセットの旅行記は「国内旅行記」と「海外旅行記」に分けられ，さらに「旅

表 3　本データセットの記述統計.

	国内旅行記		海外旅行記		全旅行記
	旅スケ付き	旅スケなし	旅スケ付き	旅スケなし	
記事数	3,153	1,519	6,419	3,188	14,279
段落数	76,307	16,412	188,908	58,700	340,327
文字数	5,878,704	1,541,124	19,273,201	4,870,061	31,563,090
単語数	3,568,354	928,936	10,950,950	2,785,494	18,233,734
固有表現数	304,606	71,598	928,487	227,191	1,531,882
地名・施設名数	95,282	25,455	268,417	71,830	460,984
段落数/記事	24.2	10.8	29.4	18.4	23.8
文字数/記事	1864.4	1014.5	3002.5	1527.6	2210.4
単語数/記事	1131.7	611.5	1706.0	873.7	1276.9
固有表現数/記事	96.6	47.1	144.6	71.2	107.2
地名・施設名数/記事	30.2	16.7	41.8	22.5	32.2

スケ付き」のものと「旅スケなし」のものに分けられる．日本語自然言語処理オープンソースライブラリ GiNZA[5]で各記事を解析し，単語分割および固有表現抽出を行なった．解析した全ファイルについての記事数，段落数，文字数，単語数[6]，固有表現数，地名・施設名数[7]を表 3 に記載する．「*/記事」は 1 記事あたりの平均値を表す．例えば，「段落数/記事 = 24.2」は 1 記事あたりの段落数が平均で 24.2 となることを表す．本データセットの特徴として，各記事の平均文字数が 2,000 文字を超え，まとまった分量であることがわかる．また，概算で 30 個以上の地名・施設名が各記事に出現していることもわかる．この点から，本データセットには多くの「場所」が含まれ，「場所」に関する分析をする上で望ましい性質を持つと言える．

2.4　国内旅行記のカバーする都道府県

図 1 に，各都道府県へ言及している国内旅行記記事数の分布を示す[8]．ひとつの特徴は，すべての都道府県がカバーされている点である．そのため，各都道府県の旅行行動の傾向分析も可能である．これ

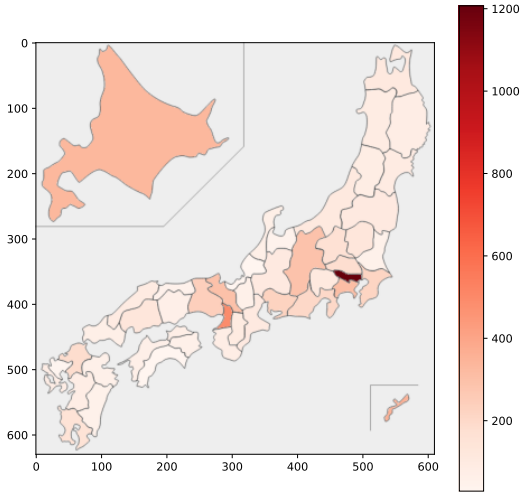

図 1　各都道府県へ言及している国内旅行記記事数の分布.

は，旅行先として「東京」が最も多く選ばれることを必ずしも意味しない．旅行記には，目当ての旅行先に行くまでの過程を含むものも多い．そのため，旅行の出発地点としての「東京」，および，旅行先への中継地点としての「東京」が記述されることも多く，それらを総合した結果として「東京」が最も多い結果となっている．表 4 は，各都道府県に言及している記事数のランキングである．「東京」に関する記事は 1,206 記事で最も多い．一方で，最も少ないのが「福井」の 29 記事である．

5)　ja_ginza_electra バージョン 5.1.2 を使用．詳しくは次のページを参照：https://github.com/megagonlabs/ginza.
6)　単語数は「単語の定義（単語分割基準）」に依存して変わる．本データセットでは GiNZA の単語分割モード C（長単位）を使用して単語分割し，その結果をカウントした．
7)　GiNZA で抽出した固有表現の中でも，地名・施設名に該当する LOC，GPE，FAC のいずれかと認識されたものをカウントした．
8)　著者（ユーザ）の自己申告のため，概算として考えた方がいい点に注意．また，記事によっては複数の都道府県へ言及しているため，国内旅行記全記事数 4,672 よりも都道府県への言及記事数の方が多くなる点にも注意.

表 4　都道府県のランキング．各都道府県に言及している記事数と国内旅行記全記事（4,762）における割合を表示．

都道府県名	記事数	割合
東京	1,206	25.81%
大阪	492	10.53%
沖縄	353	7.56%
北海道	336	7.19%
神奈川	313	6.70%
京都	299	6.40%
長野	295	6.31%
兵庫	242	5.18%
愛知	231	4.94%
千葉	219	4.69%

表 5　国・地域のランキング．各国・地域に言及している記事数と海外旅行記全記事（9,607）における割合を表示．

国・地域名	記事数	割合
アメリカ	732	7.62%
韓国	708	7.37%
フランス	697	7.26%
中国	692	7.20%
台湾	611	6.36%
ドイツ	577	6.01%
タイ	554	5.77%
イタリア	548	5.70%
スペイン	413	4.30%
スイス	411	4.28%

2.5　海外旅行記のカバーする国・地域

海外旅行記は，世界 150 以上の国と地域をカバーしている．そのため，各国・地域への旅行行動の傾向分析なども可能である．表 5 は，各国・地域に言及している記事数のランキングである．国内旅行記とは異なり，突出して記事数の多い国・地域はない点が特徴である．また，上位の国・地域は，日本政府観光局による「日本人旅行者の国別訪問者数[9)]」と概ね一致する傾向が見られる．

3　関連研究

学術目的で利用可能な現代語旅行記データセットは非常に少ない．数少ない例外である現代語旅行記データセットを表 6 に示す．Diachronic News and Travel コーパス[10)] [7] は，2 つの時代区分（1862-1939 と 1998-2017）の 3 つの分野（ニュース，旅行報告，旅行ガイド）の英語テキストを収録している．そのうち，表 6 には「現代（1998-2017）」の「旅行報告」のテキストの情報を記載している．The SpaceBank Corpus[11)] [8] は空間情報をタグ付けしたコーパスであり，SemEval-2015 Task 8 [9] でも利用された．表 6 には，旅行ブログ「Ride for Climate」のエントリから構築したサブセットの情報を記載している．KNB コーパス[12)] [10] は，4 つの分野（「京都観光」「携帯電話」「スポーツ」「グルメ」）の日本語テキストから構成される．形態素，係り受け，格・省略・照応，固有表現などの言語情報がアノテーションされている．表 6 では，「京都観光」分野のテキストの情報を記載している．これらのデータセットは人手アノテーションを含むため，単純に我々のデータセットと「量」の観点で比較できない．将来的には我々のデータセットにも多彩なアノテーションを施し，より広範な応用へつなげていく予定である．

表 6　既存の旅行記データセット．

	言語	記事数	単語数
Diachronic News and Travel	英語	23	30,747
The SpaceBank Corpus	英語	44	21,048
KNB コーパス	日本語	91	24,900

4　おわりに

「地球の歩き方旅行記データセット」について，学術的意義と特徴を中心に述べてきた．本節ではこれからの展望を述べる．本データセットの生テキストに有用な情報を付与していく予定である．特に，(1) 場所に関する言語表現の情報と (2) 人間と場所の相互作用に関わる情報が挙げられる．(1) に関しては，地名や施設名などの固有表現だけでなく，「この店」「レストラン」などの一般名詞句も含め，場所に関する言語表現を総合的にカバーする予定である．また実世界との接続を見据え，地図座標や地理データベースとの紐付けも視野に入れている．(2) に関しては，その場所における人間の「行動」「思考」「感情」に関する情報を付与する予定である．それらの情報を場所の情報と併せて抽出できるようなツールを開発して各種応用につなげる．直接的な応用として，旅行者の移動行動分析，観光地のトレンド分析，穴場の観光スポットの発掘，旅行計画・推薦への利活用が考えられる．以上のように，多様な展開が考えられる．我々のみならず，幅広い分野の研究者に本データセットを利用していただき，独創的な研究開発を進めていただけたら幸いである．

9）https://www.jnto.go.jp/jpn/statistics/20220610_4.pdf
10）https://github.com/tommasoc80/DNT.
11）https://alt.qcri.org/semeval2015/task8/index.php?id=data-and-tools.
12）https://nlp.ist.i.kyoto-u.ac.jp/kuntt/.

謝辞

本研究は JSPS 科研費 JP22H03648，JST さきがけ JPMJPR2039 の助成を受けたものです．また，データセットの構築・提供にあたり，株式会社地球の歩き方の上原康仁氏と国立情報学研究所の大須賀智子氏，大山敬三氏から多大なご協力をいただいたことを深謝します．

参考文献

[1] 株式会社地球の歩き方. 地球の歩き方旅行記データセット, 2022. 国立情報学研究所情報学研究データリポジトリ. (データセット). https://doi.org/10.32130/idr.18.1.

[2] Qiang Hao, Rui Cai, Xin-Jing Wang, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang. Generating Location Overviews with Images and Tags by Mining User-Generated Travelogues. In **Proceedings of the 17th ACM International Conference on Multimedia**, pp. 801–804. Association for Computing Machinery, 2009.

[3] Qiang Hao, Rui Cai, Changhu Wang, Rong Xiao, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang. Equip Tourists with Knowledge Mined from Travelogues. In **Proceedings of the 19th International Conference on World Wide Web**, pp. 401–410. Association for Computing Machinery, 2010.

[4] Yanwei Pang, Qiang Hao, Yuan Yuan, Tanji Hu, Rui Cai, and Lei Zhang. Summarizing Tourist Destinations by Mining User-Generated Travelogues and Photos. **Computer Vision and Image Understanding**, Vol. 115, No. 3, pp. 352–363, 2011.

[5] Gary Akehurst. User generated content: the use of blogs for tourism organisations and tourism consumers. **Service Business**, Vol. 3, No. 1, pp. 51–61, 2009.

[6] 大須賀智子, 大山敬三. 情報学研究データリポジトリ IDR における研究用データセット共同利用の取り組み. 情報処理学会論文誌デジタルプラクティス（DP）, Vol. 2, No. 2, pp. 47–56, apr 2021.

[7] Tommaso Caselli and Rachele Sprugnoli. DNT: un Corpus Diacronico e Multigenere di Testi in Lingua Inglese. In **AIUCD2021 - Book of Abstracts, Quaderni di Umanistica Digitale.**, 2021.

[8] James Pustejovsky and Zachary Yocum. Capturing Motion in ISO-SpaceBank. In **Proceedings of the 9th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation**, pp. 25–34, 2013.

[9] James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. SemEval-2015 Task 8: SpaceEval. In **Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)**, pp. 884–894, 2015.

[10] 橋本力, 黒橋禎夫, 河原大輔, 新里圭司, 永田昌明. 構文・照応・評価情報つきブログコーパスの構築. 自然言語処理, Vol. 18, No. 3, pp. 175–201, 2011.