

旅行記中の場所に対する訪問状態の予測

大友寛之¹ 東山翔平^{2,1} 大内啓樹^{1,3}

山本和太郎¹ 井手佑翼¹ 進藤裕之¹ 渡辺太郎¹

¹ 奈良先端科学技術大学院大学 ² 情報通信研究機構 ³ 理化学研究所

{otomo.hiroyuki.ob7,hiroki.ouchi,yamamoto.aitaro.xv6}@is.naist.jp

{ide.yusuke.ja6,shindo,taro}@is.naist.jp, shohei.higashiyama@nict.go.jp

概要

本研究では、文章中の「場所」に書き手が訪れたかどうかを推定する問題に取り組んだ。まず、旅行記に訪問状態のアノテーションを行い、小規模なデータセットを構築した。複数人でアノテーションを行った結果、作業員間で高い一致率が確認された。さらに、構築したデータセットを用いて訪問状態を予測するモデルを構築し、その性能を評価した。分析の結果、モデルは訪問状態の手がかりとなるトークンに注意を向けていることがわかった。

1 はじめに

人がある場所を訪れ、その場所での経験を綴った文章がウェブ上に蓄積されている。こうした情報を整理することによって多様な応用への道が拓ける。例えば、観光地を実際に訪れた観光客の感想を自動で整理できれば、観光地推薦や旅行計画支援への応用が期待できる。

そこで本研究では、人が訪れた場所の情報を文章から抽出する問題に取り組む。特に、書き手の行動を正確に読み解き、場所に対する訪問状態（訪れた、訪れていない、訪れる予定、など）を判定することに焦点を当てる。例えば「近鉄奈良駅に到着！」という実体験に関する記述があった場合、書き手は「近鉄奈良駅」を実際に訪れたと判断できる。一方、「JR 奈良駅は近鉄奈良駅から少し離れたところにあります」のような記述の場合、書き手がこれらの場所を訪れたことを必ずしも意味しない。

人が訪れた場所の情報は、GPS データを活用することでも取得できる。しかし文章には、場所を訪れた状況や目的、感想など、GPS では取得困難な情報も含まれる。将来的には、こうした情報も同時に取得することを見据え、文章を解析対象とする。

まず、訪問状態予測モデルの学習・評価のため、

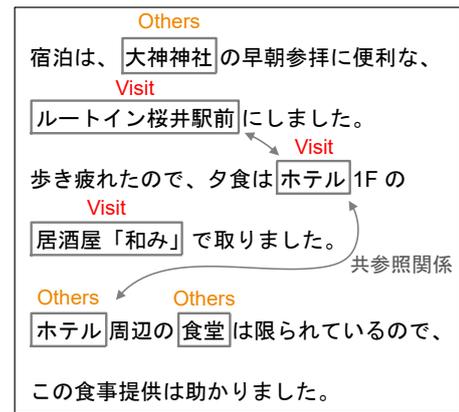


図1 アノテーションの概略図

場所を参照する表現（場所参照表現）に対する訪問状態のアノテーション方針を策定し、複数の作業員でアノテーションを行った。作業員間一致率は F1 値 0.8 前後と比較的高い値となった。

次に、そのアノテーションを用いて、訪問状態予測モデルの構築・評価を行った。モデルの予測精度は F1 値 0.6 程度であり、作業員の一致率を考えると改善の余地があることが示唆された。より詳しい分析として、モデルが訪問状態を予測する際に、入力文中のどのようなトークンに注意を向けているか調査した。その結果、例えば「到着」のように、訪問に関連するトークンに注意を向けており、人間の直感的な判断と類似する可能性が示唆された。

2 旅行記データセットの構築

2.1 アノテーションの概要

本研究では、国内外の実際の旅行についての旅行記である「地球の歩き方旅行記データセット」[1, 2]の一部の国内旅行記事に対して、図1に示すように、場所参照表現の同定（図中の枠囲み文字）、場所参照表現の言及間の共参照関係の付与（矢印）、訪問状態のラベル付与（赤・橙色文字）を行った。

表 1 場所参照表現の言及に対する訪問状態の種類

ラベル	定義
1. Visit	参照先の場所を訪問した
2. VisitPossibly	参照先の場所を訪問した可能性が示唆されている
3. PlanToVisit	当該旅行中に、参照先の場所を訪問する予定であると言及している
4. See	参照先の場所を視認した
5. Visit-Past	1 から 4 のいずれにも該当せず、当該旅行前に、参照先の場所を訪問したことに言及している
6. Visit-Future	1 から 4 のいずれにも該当せず、当該旅行後に、参照先の場所を訪問する意思があると言及している
7. Others	1 から 6 のいずれにも該当しない

場所参照表現・共参照関係 訪問状態のラベルを付与する前に、文章に出現する場所参照表現のスパンを同定した。場所参照表現として認定する対象は、「近鉄奈良駅」のような固有名詞に加え、「駅」のような一般名詞も含む。また、場所参照表現の言及間に対して共参照関係の情報の付与を行った。これにより、訪問状態予測の評価において、言及単位だけではなく、共参照関係にある言及のクラスタ（以下、共参照クラスタ）単位での評価が可能となる。

訪問状態 訪問状態のアノテーションは、場所参照表現の言及単位と共参照クラスタ単位の二通りのパターンで行った。共参照クラスタ単位の訪問状態は、文章の書き手が最終的にその場所を訪問したかどうかに対応する。言及単位の訪問状態は、該当する言及の時点で書き手がその場所を訪問していたかどうかに対応し、共参照クラスタの訪問状態の根拠にもなり得る。例えば、「久々の近鉄奈良駅です。」のような言及を根拠として、書き手が「近鉄奈良駅」に訪れたという判断が成立し得る。

表 1 に、言及に対する訪問状態の種類を示す。各ラベルの事例は付録 A に記載する。各言及に対してアノテーションする際は、言及が出現する文内情報を基にラベルを付与する方針を設けた。なお、「訪れた」のような述語に対してではなく、「近鉄奈良駅」のような名詞句に対して訪問状態のラベルを付与した理由は、「近鉄奈良駅へ。」のように、述語の省略がしばしば起こるためである。

表 2 に、共参照クラスタに対する訪問状態の種類を示す。各共参照クラスタに対してアノテーションする際は、記事全体を考慮してアノテーションを行う方針とした。そのため、共参照クラスタに含まれ

表 2 共参照クラスタに対する訪問状態の種類

ラベル	定義
Visit	参照先の場所を訪問した
Unknown	参照先の場所を訪問した可能性があるが、確定しがたい
NotVisit	参照先の場所を訪問していない

表 3 訪問状態のラベルの作業員間一致率

	micro F1	macro F1	kappa κ
言及単位	0.817	0.819	0.743
共参照クラスタ単位	0.782	0.696	0.616

る言及に Visit が付与されていなくても、共参照クラスタに対して Visit が付与される場合がある。

2.2 構築したデータセットの統計

2.1 節のアノテーション方針を基に、旅行記 5 記事に対して著者らでアノテーションを行い、486 文からなる旅行記データセットを構築した。場所参照表現の言及数は 684 件含まれ、各共参照クラスタ内の言及数は平均 1.52、標準偏差 1.06 であった。

また、5 記事に含まれる一部の言及および共参照クラスタに対して、各文につき 2 名の作業員でアノテーションを行い、訪問状態のラベルの作業員間一致率を測定した。一致率測定の対象となった言及の数は 180、共参照クラスタの数は 124 である。一致率評価には、F1 値のマイクロ平均値 (micro F1)、F1 値のマクロ平均値 (macro F1)、Cohen's Kappa (κ) [3] を用いた。表 3 に示すように、言及単位では micro F1, macro F1 とともに 0.82 程度という比較的高い一致率であり、共参照クラスタ単位ではそれらを下回る一致率となった。一致率の詳細は付録 B に示す。

3 実験設定

3.1 正解スパン設定

一つ目の正解スパン設定では、入力文と、その文中に出現する場所参照表現のスパンが所与のもと、各スパンの訪問状態を予測する。本設定における訪問状態の予測方法は以下の通りである。まず、文に含まれるトークン列を $\{X_1, X_2, \dots, X_n\}$ として、事前学習済みエンコーダ BERT [4]¹⁾ を用いて各トークンに対応するベクトル表現 $\{H_1, H_2, \dots, H_n\}$ を得る。文中に出現する場所参照表現 m のトークン列

1) <https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking>

を $\{H_i, H_{i+1}, \dots, H_{i+k}\}$ としたとき、各トークンに対応するベクトル表現の平均を取ることで、 m のベクトル表現 H' を得る。そして、 H' をフィードフォワードネットワークに入力し、訪問状態を予測する。

言及単位での評価に加えて、言及間の共参照関係を所与として共参照クラスタ単位での評価も行った。言及単位でのモデルの予測を基に、次のルールを用いて共参照クラスタ単位での予測を行った。

1. 共参照クラスタ内の少なくとも一つの言及に対して Visit と予測していれば、Visit とする。
2. Visit と予測した言及が一つもなく、VisitPossibly もしくは PlanToVisit と予測した言及が一つでもあれば、Unknown とする。
3. Visit, VisitPossibly, PlanToVisit と予測した言及が一つもない場合、NotVisit とする。

3.2 End-to-End 設定

二つ目の End-to-End 設定では、場所参照表現の同定および訪問状態の予測を同時に行う。系列ラベリングに基づく一般的な固有表現抽出と同様に、トークンごとに B-Visit や I-Visit のような BIO 形式のタグを予測する。

旅行記データセットは、場所参照表現の同定を学習するにはデータ量が不十分である可能性がある。そこで、場所参照表現の同定精度向上を目的として、拡張固有表現タグ付きコーパス [5] を用いて事前に BERT をファインチューニングした。本実験では、同コーパスに付与されている「関根の拡張固有表現階層」[6] (Version 7.1.0) の固有表現ラベルのうち、地名・施設名のラベルが付与された固有表現（つまり、場所参照表現）のみを対象とし、同コーパスを学習用と評価用に分割した²⁾。ファインチューニングの結果、評価データでの場所参照表現の同定精度 (F1 値) は 0.829 であった。

3.3 データセット

訪問状態予測の実験には、本研究で構築した旅行記データセットを用いた。正解スパン設定では、前述の通り、言及間の共参照関係を所与とし、共参照単位での評価も行う。そのため、共参照クラスタ内の言及を含む文の集合単位で旅行記データセットをおよそ 8 対 1 対 1 の比率で訓練、開発、評価データに分割した。End-to-End 設定では共参照単位での評

2) 学習 (ファインチューニング) データに含まれる文の数は 49,440、場所参照表現の言及数は 13,242 個であった。

表 4 正解スパン設定の結果

	正解率	適合率	再現率	F1 値
言及単位：7 ラベル	0.704	0.645	0.635	0.601
言及単位：2 ラベル	0.761	0.771	0.756	0.756
共参照クラスタ単位	0.511	0.429	0.390	0.375

表 5 End-to-End 設定の結果

	適合率	再現率	F1 値
言及単位：7 ラベル	0.492	0.451	0.471
言及単位：2 ラベル	0.657	0.648	0.652

価を行わない³⁾ため、文単位で分割した。

また、言及単位での評価において、アノテーションされたラベルをそのまま使用する 7 ラベル設定に加えて、訪問の有無に注目してラベルをグルーピングした 2 ラベル設定での評価も行った。具体的には、Visit, VisitPossibly, PlanToVisit を Visit に、他のラベルを NotVisit にグルーピングした。

4 実験結果

表 4 に、正解スパン設定の結果 (各指標はマクロ平均) を示す。表 3 に示した作業員間一致率 (macro F1) と比べ、7 ラベル設定の F1 値は 0.2 程度低い。原因として、旅行記データセットの学習データが少量かつラベルが不均衡なデータであったため、モデルが十分に学習できなかったことが考えられる。また、言及単位よりも共参照クラスタ単位での予測精度が低い理由として、記事内の大域的な内容を基に共参照クラスタの訪問状態が決まるケースに対し、3.1 節のルールでは対応できなかったことが挙げられる。例えば、「奈良県」を訪問したと明示的に言及せず、「近鉄奈良駅」を訪問したと言及している場合、「近鉄奈良駅」が位置する「奈良県」にも訪問したと推論できるが、使用したルールでは正しく予測することができない。

また、End-to-End 設定の結果 (各指標はマイクロ平均) は表 5 の通りとなった。7 ラベル設定よりも 2 ラベル設定のほうが F1 値が 0.18 程度高い理由は、ラベルをグルーピングすることで、事例数が少ないラベルが減り、データの不均衡性がある程度解消されたためと考えられる。

5 分析

本節では、正解スパン設定の訪問状態予測モデルに対する分析結果を報告する。

3) 場所参照表現の共参照解析は本研究のスコープ外のため。

表6 Integrated Gradientsに基づく寄与度. 下線部は訪問状態の予測対象となるスパンを表す. (例文は [1] より引用)

	正解	予測 (確率)	入力文
(a)	Visit	Visit (0.92)	[CLS] 別府 駅 に 到着 しまし た 。 [SEP]
(b)	See	See (0.67)	[CLS] 展望 塔 から は 遠く 日本 海 に 浮かぶ 利 ##尻 ##富 ##土 (利 ##尻 島) を 眺望 し ます 。 [SEP]
(c)	Visit	Other (0.95)	[CLS] 宿泊 は 、 大神 神社 の 早朝 参拝 に 便利 な 、 ルート ##イン 桜井 駅前 に しました 。 [SEP]

表7 正解スパン設定の結果 (場所参照表現マスク時)

	正解率	適合率	再現率	F1 値
7ラベル設定	0.704	0.645	0.635	0.601
+マスク	0.676	0.600	0.650	0.564
2ラベル設定	0.761	0.771	0.756	0.756
+マスク	0.761	0.789	0.754	0.751

Integrated Gradients モデルが訪問状態を予測する際、入力文中のどのトークンに対して注意を向けているかを調べた。具体的には、Integrated Gradients [7] (captum [8] で実装されている Layer Integrated Gradients⁴⁾) を基に、BERT の埋め込み層における予測への寄与度をトークン位置ごとに計算した。表6は、分析結果の可視化であり、寄与度が高いトークンほど濃くハイライトされている。訪問状態を正しく予測できた(a)と(b)を見ると、訪問や視認に関連する動作性名詞「到着」、「眺望」やそれらに付随する助動詞「ます」、副詞「遠く」などの寄与度が高いことがわかる。一方、訪問状態を正しく予測できていない(c)では、文中のトークン全体への寄与度が高くなっていることがわかる。訪問状態の判定のために文全体を読み解く必要があるような例で、適切なトークンに注目して正しく予測することが難しかった可能性がある。

場所参照表現のマスク 正解スパン設定において所与とした場所参照表現の表層情報が、モデルの予測にどのように影響を与えているかを調べた。具体的には、訪問状態の予測対象となる場所参照表現の各トークンを [MASK] に置き換え、学習および評価を行った。言及単位での評価結果を表7に示す。7ラベル設定では場所参照表現をマスクすることで正解率が0.03程度低くなっているが、2ラベル設定ではマスクしても同等の結果となった。以上より、詳細なラベルを予測する設定では表層情報も手がかりとして有用であることが示唆された。

4) <https://captum.ai/api/layer.html#layer-integrated-gradients>

6 関連研究

場所参照表現抽出 [9, 10, 11] は、固有表現抽出の特殊なケースであり、実世界上の特定の地点を指す表現を文章中から抽出することを目的としている。これら従来研究では、文章の書き手が訪問したかという観点での解析は行っていない。

Liら [12] は、ツイートを対象に、言及単位での場所参照表現の抽出とその訪問状態(訪れた、現在いる、訪れる予定である)の予測を行った。同一の場所を指す場所参照表現が、1ツイート内に異なる訪問状態として複数回出現することは稀と考えられるため、本研究のように共参照クラスタ単位での解析を行うには、長さのある文章が適している。

大友ら [13] は、KNB コーパス [14] に対して、場所参照表現とその訪問状態のアノテーションを行った。アノテーション対象となった記事は、京都観光をテーマに大学生により執筆されたブログ記事である。そのため、書き手の属性や記述内容に偏りがあり、観光地の一般的な説明や日常的な出来事の描写など、旅行以外の記述も多く含まれる。

7 おわりに

本研究では、人が訪れた場所の情報を文章から抽出する問題に取り組んだ。場所参照表現に訪問状態を付与したデータセットを構築し、訪問状態予測の実験を行った。モデルの予測精度は作業員間一致率を下回り、学習データが少量かつ不均衡であることが難しさの原因であったと考えられる。また、分析の結果、文全体を読み解く必要がある事例において、人間と異なり、単純なBERTベースのモデルでは適切な判定が難しいという示唆が得られた。今後、アノテーションデータの増量や、場所に紐づいた意味的な情報の抽出に取り組む予定である。

謝辞

本研究は JSPS 科研費 JP22H03648 の助成を受けたものである。

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] 大内啓樹, 進藤裕之, 若宮翔子, 松田裕貴, 井之上直也, 東山翔平, 中村哲, 渡辺太郎. 地球の歩き方旅行記データセット. 言語処理学会第 29 回年次大会, 2023.
- [3] Jacob Cohen. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, Vol. 20, pp. 37–46, 1960.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [5] 橋本泰一, 乾孝司, 村上浩司. 拡張固有表現タグ付きコーパスの構築. 情報処理学会研究報告自然言語処理, Vol. 2008, No. 113 (2008-NL-188), pp. 113–120, 2008.
- [6] Satoshi Sekine. Extended named entity ontology with attribute information. In **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)**, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).
- [7] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Ax-omatic attribution for deep networks. In **Proceedings of the 34th International Conference on Machine Learning - Volume 70**, ICML'17, pp. 3319–3328. JMLR.org, 2017.
- [8] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch, 2020. arXiv: 2009.07896 [cs.LG].
- [9] Koji Matsuda, Akira Sasaki, Naoaki Okazaki, and Kentaro Inui. Annotating geographical entities on microblog text. In **Proceedings of The 9th Linguistic Annotation Workshop**, pp. 85–94, Denver, Colorado, USA, June 2015. Association for Computational Linguistics.
- [10] Zongcheng Ji, Aixin Sun, Gao Cong, and Jialong Han. Joint recognition and linking of fine-grained locations from tweets. In **Proceedings of the 25th International Conference on World Wide Web**, WWW '16, p. 1271–1281, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.
- [11] Pei Chen, Haotian Xu, Cheng Zhang, and Ruihong Huang. Crossroads, buildings and neighborhoods: A dataset for fine-grained location recognition. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3329–3339, Seattle, United States, July 2022. Association for Computational Linguistics.
- [12] Chenliang Li and Aixin Sun. Fine-grained location extraction from tweets with temporal awareness. In **Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval**, SIGIR'14, p. 43–52, New York, NY, USA, 2014. Association for Computing Machinery.
- [13] 大友寛之, 大内啓樹, 星野智紀, 井手佑翼, 渡辺太郎. 訪問場所表現グラウンディングのためのアノテーション. 言語処理学会第 28 回年次大会, 2022.
- [14] 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. 2, pp. 175–201, 2011.

A 訪問状態の各ラベルの事例

表 8 言及に対する訪問状態のラベルと事例 ([1] より引用). 下線部はラベル付与の対象となるスパンを表す.

ラベル	事例
Visit	昼食は定番の回轉寿司和樂小樽店で握り寿司を頂きました。
VisitPossibly	昼食は西ノ京駅から徒歩の「蕎麦切りよしむら」で取りました。
PlanToVisit	拝観券を購入して国宝の東金堂へ。
See	朝、京都へ向かう東海道新幹線の車窓から、頂上付近に雲をまとった富士が良く見えました。
Visit-Past	数年前に訪れたフランス・ルルドの泉を思い出しました。
Visit-Future	-
Others	三輪山登拝口はコロナ禍のため、立ち入り禁止になっていました。

表 8 は、言及に対する訪問状態の各ラベルの事例である。Visit-Future については、本研究でアノテーション対象となった旅行記 5 記事において、付与された事例が確認されなかった。なお、各ラベルの定義は表 1 の通りである。

B 訪問状態のアノテーション一致率の詳細

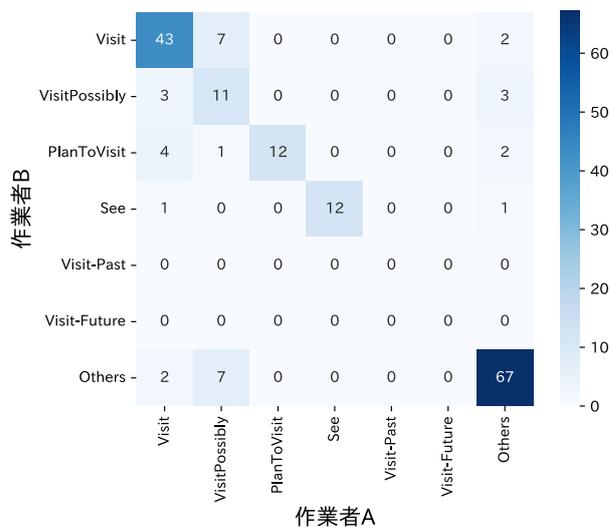


図 2 各言及へ訪問状態を付与した際の混同行列

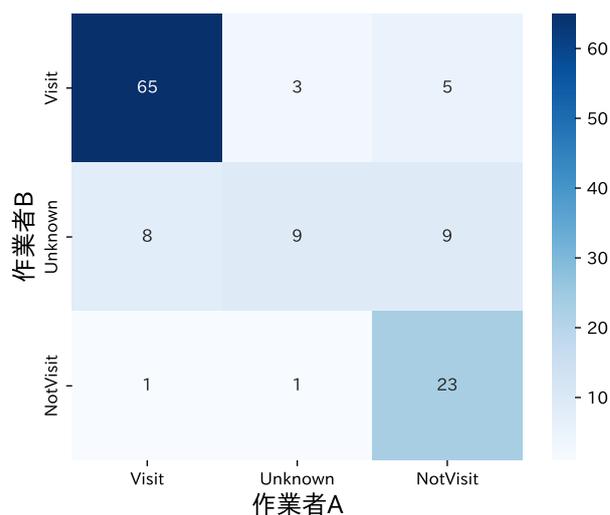


図 3 各共参照クラスタへ訪問状態を付与した際の混同行列

図 2 は、複数の作業者によって、言及へ訪問状態のラベルを付与した際の混同行列を表している。また、図 3 は共参照クラスタに関する混同行列を表している。