

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

大友寛之¹ 大内啓樹^{1,2} 星野智紀¹ 井手佑翼¹ 渡辺太郎¹

¹ 奈良先端科学技術大学院大学 ² 理化学研究所

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

{hoshino.tomoki.ho6,ide.yusuke.ja6,taro}@is.naist.jp

概要

新たなタスク「訪問場所表現グラウンディング」を提案する。文章中の人物が訪れた場所の言語表現を抽出し、地図上にグラウンディングすることを目的とする。これが可能になれば、旅行プランニングなど多様なアプリケーションへの展開が期待できる。その第一歩として、本稿ではデータセット構築のためのアノテーション基準を策定する。実際にアノテーション実験を行い、比較的高い作業員間アノテーション一致率が得られることがわかった。

1 はじめに

古来から人間は地球上のさまざまな場所を訪れ、その記憶を自然言語で書き留めてきた。旅行記や紀行文はその代表であり、訪れた場所での個人的な経験や感想が記述される。これらの文書を計算機によって読み解き、訪問場所に関する言語表現と地図上の位置を紐付け、そこで個人が何を体験したのかを抽出できれば、旅行プランニングや観光地ナビゲーションをはじめ多様なアプリケーションへ有効活用できる情報源になると期待できる。特に昨今のコロナ禍により壊滅的な打撃を受けた観光業の再興のため、位置情報と紐づく主観情報（その場所の体験談や感想）は重要な役割を担う。

本研究では、**訪問場所を表す言語表現を地図上へグラウンディング**することを見据え、そのデータセット構築のためのアノテーション基準を策定することを目的とする。具体的には、旅行ブログを対象テキストとし、著者が実際に訪れた場所（地名や施設名）を複数の作業員にアノテーションしてもらおう。ここで重要なのは、「著者が実際にその場所を訪れたかどうか」の判断である。この判断には周辺文脈を読み解く必要があり、時には明示的に書かれていない情報を推測しないと正確に判断できない事例も少なくない。そこで、作業員間のアノテーショ

ン一致率を算出し、どのような事例でアノテーションの一致/不一致が生じるかを分析する。実験結果として、全体では比較的高い一致率が得られることがわかった。事例分析の結果、個人の経験に関する言及と一般的な言及が同一文中に混在する場合に不一致が起きやすくなることがわかった。

2 訪問場所表現グラウンディング

2.1 問題設定と技術的焦点

本稿で提案するタスク「訪問場所表現グラウンディング」を図1に示す。入力文章中で著者が訪れた場所を表す場所参照表現を抽出（ラベリング）し、地図上の適切な位置（緯度・経度）に紐付けること（グラウンディング）を目的とする。例えば図1において、著者は「西本願寺」と「銀閣寺」を訪れており、それらを地図上にグラウンディングしている。

本タスクの技術的なポイントは、著者が**実際にその場所を訪れたか否か**の判定にある。つまり、著者の行動の事実性を文脈から判断する必要がある。例えば図1中の「京都」は、場所を参照する言語表現であるが、著者の訪問経験を意味しているわけではなく、一般的な事実への言及であるため、地図にはグラウンディングしない。

ここで、あるタイプのテキストには位置情報が付与されていることもあり、わざわざ本タスクを行う必要はないのではないかという指摘もありうる。たしかにTwitterなどのマイクロブログテキストには位置情報が付与されている場合もある。しかしながら、位置情報付きツイートの割合は1パーセントに満たないという報告がある[1]。また、著者が発信している場所と文章中で言及している場所は異なる場合も多く、これは本質的な問題である。これらの問題を鑑みると、場所参照表現を正確に抽出し、地図上に紐付けるためのタスク設計およびデータ整備、システム開発は有用であると言える。

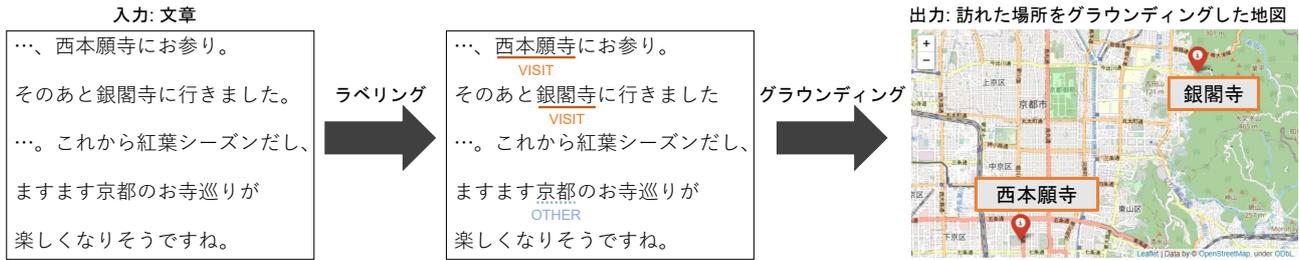


図1 訪問場所表現グラウンディングの概略図

2.2 先行研究との関係

関連する研究分野の一つに「場所参照表現解析 (Geoparsing, Toponym Resolution)」 [2, 3, 4, 5, 6] が挙げられる。このタスクでは、テキスト中に現れる場所参照表現を抽出し、各表現に位置情報（緯度・経度）を付与する。典型的には位置情報付き地名・施設名辞書を用いることを前提とし、該当する項目を場所参照表現で辞書引きして位置情報を取得するため、エンティティリンキング (Entity Linking) [7] の一種として捉えることもできる。これらの基本設定に加え、本研究では著者が実際にその場所を訪れたかどうかを問題とする。この新たな視点の導入によって、場所参照表現の周辺文脈をより正確に読み解かなければタスクを解くことはできない。特に、前述したような図1中の「京都」の例のように、事実性（時制、アスペクト、モダリティ）を正確に解析し、一般的な言及なのか著者の実体験に関する言及なのかを峻別する必要がある。このような点から、従来の場所参照表現解析と比べ、本タスクでは**意味解析の要素が色濃く要求される**点が特色である。

もう一つの関連研究分野として「経験マイニング (Experience Mining)」 [8, 9, 10] が挙げられる。このタスクでは、商品やサービスなどの事物の利用に関する個人の経験を文章から抽出する。経験マイニングでも事実性解析などの意味解析が重要な役割を担う点は我々の提案タスクと共通する。しかしながら、我々の提案タスクは経験マイニングには収まらない問題を含んでいる。それは地図へのグラウンディングである。**テキストの世界と地図の世界を繋ぐマルチモーダル処理を目的とする**点が明確に異なる。将来的には、訪問場所における個人の経験や感想を地図上に重ね合わせて表示することによって、拡張地図 (Augmented Map) の生成を目指す。そのため第一歩として、本稿では訪問場所参照表現のアノテーションを行う。

3 アノテーション実験

本節では、訪問場所参照表現の人手によるアノテーション実験を行い、作業員間アノテーション一致率をはじめとする種々の分析を行う。

3.1 アノテーションの概要

アノテーション作業として、(1) 場所参照表現の抽出と (2) その場所に訪れたと文中で言及しているか否かを表すラベルを付与する。一つ目の作業では、文章中に現れる地名・施設名などの場所参照表現のテキスト範囲を文字単位で同定する。二つ目の作業では、同定済みの場所参照表現の指し示す場所を、文章の著者が実際に訪れ、かつ、文中で訪れていることに言及している場合は VISIT のラベルを付与し、それ以外は OTHER のラベルを付与する。

3.2 アノテーションの手順

2名のアノテーション作業員が実際の作業を行った。データとして、KNB コーパス [11] を利用した。KNB コーパスは 249 記事、4,186 文から構成される解析済みブログコーパスであり、係り受けや固有表現などのアノテーション情報が含まれる。また、KNB コーパスに含まれる記事は、京都観光、携帯電話、スポーツ、グルメのいずれかのテーマで執筆されている。本稿では京都観光に属する 81 件の記事の内、「旅行記」に該当する記事 29 件をアノテーション対象の文章とした。KNB コーパスに含まれる各記事の 1 行目はテーマ名とタイトルを表しており、2 行目以降が本文を表している。そこで、旅行記 29 件の各記事の 2 行目以降をアノテーション対象とし、全体の文数は 455 となった。アノテーションツールとして doccano¹⁾ を用いた。なお、アノテーション作業中に場所参照表現かどうかわからない表現があった場合、検索エンジンを用いて調べてよいこととした。

1) <https://github.com/doccano/doccano>

表1 アノテーションされたラベルの数

	作業員 A	作業員 B
# VISIT	182	201
# OTHER	146	95

表2 作業員間アノテーション一致率 (F₁ 値)

VISIT	OTHER	VISIT+OTHER
78.33	69.71	75.00

4 結果と考察

表1は、2名の作業員によってラベル付けされたVISITとOTHERの数を表している。

4.1 作業員間アノテーション一致率

表2は作業員間のアノテーション一致率を表している。本実験では、一致率としてF₁値を採用する。F₁値を算出するにあたり、2名の作業員によるアノテーションのラベルおよびスパンの両方が一致していれば真陽性 (True Positive) とみなす。表2中のVISITは、アノテーション結果がVISITのラベルについて一致率を調べた結果であり、OTHERも同様である。また、表2中のVISIT+OTHERは、VISITとOTHERの両方のラベルの一致率について調べた結果である。

この結果は比較的高い数値であると解釈できる。本アノテーション作業では「実際に著者がその場所を訪れたか」の判断がポイントとなるが、これは事実性 (時制, アスペクト, モダリティ情報) の認識に深く関わる。この認識は個人の主観によって揺れる。例えば関連するものとして、日本語 TimeBankのアノテーション作業が挙げられ、概ね6割ほどの作業員間アノテーション一致率であったとの報告がある [12]。もちろん、本稿でのアノテーション作業と異なるため公平な比較はできないが、全体でも7割を超える一致率であり、最も重要なVISITの一致率は8割に迫るため悪くない数値であると言える。

4.2 固有表現の再現率

実際に本アノテーション (場所参照表現の抽出とラベル付与) を自動化する場合、シンプルな方法として、第一段階目で場所参照表現を抽出し、続いて第二段階目で抽出済み場所参照表現をVISITやOTHERに分類するといった二段階処理 (これは典型

表3 固有表現の再現率

	VISIT	OTHER	VISIT+OTHER
自動NER	62.45	76.62	68.06
人手NER	85.96	90.13	87.78

的な Entity Linking や Entity Typing と同じ方法) が考えられる。そこで本節では、第一段階目の場所参照表現抽出のために、一般的な固有表現抽出システムを用いるとどの程度場所参照表現を抽出できるか、その再現率の目安を付けるための分析を行う。さらに、固有表現抽出システムとの比較のため、人手による固有表現のアノテーションの再現率も確認する。

本稿では固有表現抽出システムとして spaCy²⁾ の ja_core_news_lg³⁾ を用い、モデルの予測結果のうち、FAC, LOC, GPE, ORG のいずれかのラベルを自動付与された言語表現のみをシステム出力 (以下、自動NER) とした。なお、FAC, LOC, GPE, ORG 以外のラベルとして予測されたアノテーションは全体の3%程度であり、システム出力に含まれない。また、人手によってラベル付けされたデータセットとして、KNB コーパスの固有表現アノテーションを用いた。固有表現アノテーションのうち、ラベルがLOCATIONもしくはORGANIZATIONの言語表現をシステム出力 (以下、人手NER) とした。作業員2名によるアノテーションを正解データとして各アノテーションで再現率を計算し、その平均を本節で報告する。

表3はラベルごとの再現率を表している。自動NERにおいて、VISIT+OTHERの再現率が7割程度という結果から、固有表現抽出システムが抽出できていない場所参照表現が3割程度存在することがわかる。言い換えると、一般的なサイズの固有表現データセットを用いたNERモデルの学習・構築では十分に場所参照表現を抽出できないことが示唆された。したがって、場所参照表現の追加アノテーションデータを用いたNERモデルのファインチューニングや、地名・施設名辞書を用いた新たな手法を考案することによって、場所参照表現の抽出精度を高めていくことが今後の技術的課題である。

4.3 事例の分析

本節では、アノテーション結果のより詳細な分析を行い、現状の到達点や改善点を洗い出す。2名の

2) <https://spacy.io>

3) https://spacy.io/models/ja#ja_core_news_lg

-
- on **Artificial Intelligence**, 2015.
- [5] 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.
- [6] Davy Weissenbacher, Arjun Magge, Karen O’Connor, Matthew Scotch, and Graciela Gonzalez-Hernandez. SemEval-2019 task 12: Toponym resolution in scientific papers. In **Proceedings of the 13th International Workshop on Semantic Evaluation**, pp. 907–916, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.
- [7] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. **IEEE Transactions on Knowledge and Data Engineering**, Vol. 27, No. 2, pp. 443–460, 2014.
- [8] Kentaro Inui, Shuya Abe, Kazuo Hara, Hiraku Morita, Chitose Sao, Megumi Eguchi, Asuka Sumida, Koji Murakami, and Suguru Matsuyoshi. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. In **2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology**, Vol. 1, pp. 314–321. IEEE, 2008.
- [9] Keun Chan Park, Yoonjae Jeong, and Sung Hyon Myaeng. Detecting experiences from weblogs. In **Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics**, pp. 1464–1472, Uppsala, Sweden, July 2010. Association for Computational Linguistics.
- [10] Yunzhong Liu, Yi Chen, Jiliang Tang, and Huan Liu. Context-aware experience extraction from online health forums. In **2015 International Conference on Healthcare Informatics**, pp. 42–47. IEEE, 2015.
- [11] 橋本力, 黒橋禎夫, 河原大輔, 新里圭司, 永田昌明. 構文・照応・評価情報つきプログコーパスの構築. 自然言語処理, Vol. 18, No. 2, pp. 175–201, 2011.
- [12] Masayuki Asahara, Sachi Yasuda, Hikari Konishi, Mizuho Imada, and Kikuo Maekawa. BCCWJ-TimeBank: Temporal and event information annotation on Japanese text. In **Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)**, pp. 206–214, Taipei, Taiwan, November 2013. Department of English, National Chengchi University.