

LLM による海外の観光地紹介文生成のための インストラクションデータ自動構築

出口 湧也、酒井 浩之、仁科 慧

成蹊大学 理工学部 理工学科 コンピュータ科学専攻

us212004@cc.seikei.ac.jp, h-sakai@st.seikei.ac.jp, kei-nishina@st.seikei.ac.jp

概要

本研究では、観光地に関する WEB サイトから観光情報を抽出し、抽出された観光情報からインストラクションデータを自動生成する手法を提案する。そのデータで LLM をインストラクションチューニングし、その LLM でユーザーの嗜好に合った観光地紹介文を生成し、提示するシステムの構築を目指す。観光情報の抽出のため、本研究では観光地の6つの観点（観光、街並み、食べ物、歴史、文化、お土産）に関する情報を収集する。評価の結果、本手法によって構築されたインストラクションデータによってチューニングされた LLM は、より良い観光地紹介文を生成することが可能であることを示した。

1 はじめに

旅行先を選ぶ際、景観の美しさや食事の魅力、歴史的価値、文化的背景など、多様な観点から評価できる一方で、それらの情報は Web 上に点在しており、旅行者が一つひとつ比較・検討するのは容易ではない。また有名ではないが、魅力的な観光地を Web 検索で探すのは困難である。そこで、本研究では、ユーザーが望む海外旅行の条件をプロンプトとして入力すると、その条件に応じた観光地と観光地の魅力を紹介する文章（紹介文と呼ぶ）が出力される、観光情報検索システムの開発を目指す。このシステムの紹介文を自動生成する機能を実装するため、独自の手法で観光情報に関するインストラクションデータを自動構築し、それをを用いて LLM のファインチューニングを行った。このインストラクションデータは、観光地に関する情報が記載された Web サイトをクロールして得られたテキストデータの中から自動的に推定された「景観、食事、歴史、文化」などの観点に関わる、より詳細な観光地の魅力を表す固有表現に基づいて自動構築される。

関連研究として、文献[1] は対話から話者情報を要約し観光地推薦に活用しているが、嗜好が明示されない場合は精度が低下しやすく、対話データを前提とする点で本研究とは異なる。文献[2] は既存の構造化データを活用して LLM に推薦知識を効率的に学習させる手法を提案しているが、非構造的な Web 文章を対象とする本研究とは対象が異なる。文献[3] はレビューから文化的含意を抽出し推薦に活用するが、既存レビューが前提であり、文収集から分類まで自動で行う本研究とはアプローチが異なる。さらに、LLM を用いてインストラクションデータを自動構築する手法として、Wang ら[4]が提案した手法が挙げられる。この手法は、LLM の生成能力を活用して少数のタスクから多様な学習データを生成するものである。本研究の手法は、LLM の生成能力に加えて、Web サイトからクロールした客観的な観光情報に基づいた根拠のある学習データを生成できるという点で異なる。また Li ら[5]は、観光に関する質問に対して適切な POI を提示するため、質問文と候補 POI を別々に埋め込み、類似度に基づいて検索する手法を提案している。一方、本研究は検索やランキングではなく、Web 上の非構造化文から観点別情報を抽出し、紹介文およびインストラクションデータを自動生成して LLM をインストラクションチューニングする点でアプローチが異なる。文献[6]では、LLM に対してユーザー情報やアイテム情報をプロンプトとして与え、推薦結果を直接生成させる手法を提案している。一方、本研究は LLM に観光地の紹介文を生成させるため、その学習用のインストラクションデータを自動構築する手法を取っており、LLM を直接推薦器として用いる点でアプローチが異なる。文献[7]は、Web 上に存在する質問回答形式に近い文書構造を手がかりとして、instruction-response ペアを自動抽出・精錬し、インストラクションデータを構築する手法を提案している。一方、本研究は観光地に特化し、Web 文を観点

別に抽出した上で質問と回答セットを生成するドメイン特化型のインストラクションデータ自動構築を目的としており、対象領域と構築プロセスが異なる。文献[8]は、Web 文書をもとに指示文と応答文の対応関係を再構成することで、インストラクションデータを自動生成する手法を提案している。Web 文書そのまま用いるのではなく、指示となるように書き換えや応答文の再構築を行っている。一方、本研究は観光地に特化し、Web の文を観点別に抽出することで作成した固有表現と紹介文を用いてインストラクションデータセットを構築しており、対象領域とデータ構築プロセスが異なる。

2 提案手法

本手法の概要を以下に示す。

- Step1: 観光地ごとに WEB サイトをクロールして HTML ファイルを収集し、HTML ファイルから文を抽出
- Step2: Step1 で抽出した文の中でそれぞれの観光地の観点（観光、街並み、食べ物、歴史、文化、お土産）に関する文を手で抽出し、LLM の Few-shot prompting のための学習データを作成
- Step3: LLM の Few-shot prompting により、それぞれの観光地の観点に関する文を観光情報として抽出
- Step4: LLM の Few-shot prompting により、Step3 で抽出した文から観光地の魅力を表現する固有表現を抽出
- Step5: Step3 で抽出した文からそれぞれの観光地の紹介文を LLM で生成
- Step6: Step4 の固有表現と Step5 の紹介文を組み合わせるインストラクションデータセットを構築
- Step7: Step6 のデータセットで LLM をインストラクションチューニングし、その LLM から観光地紹介文を自動生成

2.1 観点ごとの観光地に関する文の抽出

本節では、Step2 の観点ごとの観光地に関する文を抽出する処理について説明する。まず、我々は、観光情報の重要な項目として、観光地の紹介で頻出する「観光」「街並み」「食べ物」「歴史」「文化」「お土産」の6つの項目を選び、観光地を選ぶときに重視される観点として設定した。この観点に関連した観光地に関連する文を Web 上から収集するため、Google Custom Search API を使用し、観光地地名リストから「国名 都市名 観点」で検索した検索結

果の上位 10 件の WEB サイトの HTML ファイルを取得した。なお、観光地名として 92 か国 3285 の観光地名を用意した。

2.2 観光情報文の抽出

本節では、Step3 の観光地の観点に関する文、観光情報文を抽出する処理について説明する。HTML ファイルから取り出した文の中には観点（例:街並み）とは無関係な文が多く含まれているため、LLM の Few-shot prompting により、観光地の観点に関する文のみを抽出する。

Few-shot prompting を行うために、観点ごとに人手で少数の学習データを作成した。2.1 節で抽出した観点ごとの文の中からランダムに 25 都市選び、それらの都市の観点に関する文を手で 25 文選出した。2.1 節で抽出した文を作成した学習データを用いて LLM の Few-shot prompting により観光地の各観点に関する文かそうでない文かに分類する。分類ラベルは観光地における各観点に関する文を「True」、そうでない文を「False」と出力させる。ここで、LLM には Llama-3-ELYZA-JP-8Bⁱ を用いた。

LLM の Few-shot prompting によってラベル付けされた結果と、人手でテストデータにラベル付けした結果で正解率を求めた。ここで、各観点 50 文ずつランダムに選出しテストデータとした。結果を表 1 に示す。

表 1 各観点の分類結果の正解率

観点	正解率
観光	81.3%
文化	60.8%
街並み	89.8%
歴史	74.0%
お土産	72.0%
食べ物	78.0%

正解率は最高で「街並み」の 89.8%、最低で「文化」の 60.8%となった。これは、「文化」の定義が広範囲かつ抽象的であり、判断基準を明確に学習できなかったことが要因と考えられる。正解率が低かった文化、歴史、お土産において GPT-OSS-20Bⁱⁱを使用し、再度、文の抽出を行った。その結果、正解率は文化 84%、歴史 78%、お土産 78%まで上昇した。

ⁱ <https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B>

ⁱⁱ <https://huggingface.co/openai/gpt-oss-20b>

2.3 観光情報文からの固有表現の抽出

本節では、Step4 の観光情報文から LLM の Few-shot prompting によって、観光地の魅力を表す固有表現を抽出する処理について説明する。これは、2.2 節で抽出した観光地の各観点に関する文から、LLM の Few-shot prompting により観光地の各観点に関する固有表現を抽出することによって行われる。ここで固有表現は、観光地の各観点の魅力や特徴を構成する要素を表す語と定義する。例えば「街の真ん中にはエジンバラのシンボル・エジンバラ城が聳え立ち、その両側に中世の旧市街と 18 世紀に計画的に建設された都市計画の最高傑作と言われるジョージアン様式の建物とその庭園が整然と配置された新市街が広がっています。」という文から固有表現として「エジンバラ城」、「中世の旧市街」、「ジョージアン様式の建物」を抽出する。

LLM の Few-shot prompting を行うために、観点ごとに人手で学習データを作成した。2.2 で抽出した文の中から観点ごとに 50 文選び、固有表現を人手で抽出し学習データとした。2.2 で抽出した文と作成した学習データを用いて、LLM の Few-shot prompting により固有表現を抽出する。ここで、LLM には GPT-OSS-20B を用いた。2.2 節で抽出した文の中から固有表現が含まれなかった文を削除し、最終的な固有表現を含む観光地に関する文とした。表 2 に抽出できた文の数を示す。

表 2 抽出できた観光地に関する文の数

1 観光地当たりの最大文数	1526
1 観光地当たりの最小文数	13
1 観光地当たりの平均文数	368.7
全観光地の総文数	1,211,150

2.4 観光地の紹介文の生成

本節では、Step5 の LLM による観光地の紹介文の生成の処理について説明する。これは、2.2 節で抽出した観光地の各観点に関する文を使用し、各観光地の紹介文を生成することによって行われる。観光地ごとに各観点で抽出した文を統合し、LLM に対する入力データを作成した。ここで、LLM には GPT-OSS-20B を用いた。付録[1]に生成された観光地紹介文の例を示す。本手法により全 3285 都市の観光地紹介文を作成した。

2.5 インストラクションデータの自動構築

本節では、Step6 のインストラクションデータの自動構築について説明する。これは、2.3 節で抽出した各観点の固有表現と 2.4 節で作成した観光地の紹介文を使用して、インストラクションデータセットを自動的に作成することによって行う。まず観点ごとに抽出された固有表現とそれに合う文を組み合わせて質問文を作成する。例えば観点「文化」では「～といった文化を感じられる観光地を教えてください。」が質問文となる。ここで「～」には抽出した固有表現が割り当てられる。その質問文に対する回答として、「おすすめは[都市名]です。」という文と 2.4 節で作成した観光地紹介文を組み合わせた回答文を作成する。都市ごとに 6 つの観点で作成したインストラクションデータを統合し、1 都市当たり最大 2734 セット、最小 37 セット、平均 719 セット、合計 2,361,694 セットのデータセットを構築した。作成したインストラクションデータの例を付録[2]に示す。

2.6 インストラクションチューニング

本節では、Step7 の LLM のインストラクションチューニングについて説明する。これは、2.5 節で作成したインストラクションデータセットを用いて LLM をチューニングすることによって実行される。構築した全てのインストラクションデータを使用した場合、学習時間が膨大になることが予想された。そのため、各都市につきランダムに抽出した 100 件のデータを選択して使用した。LLM には LLM-JP-3.1-13B-instructⁱⁱⁱを使用した。

2.7 観光情報検索システム

2.6 節でインストラクションチューニングをした LLM と、2.2 節で抽出した観光情報文を使って観光情報検索システムを構築した^{iv}。図 1 に構築した観光情報検索システムのスクリーンショットを示す。このシステムでは入力したプロンプトからキーワードを抽出し、そのキーワードを含む観光情報文をチューニングした LLM にプロンプトとして与えて、観光地ごとに観光地紹介文を生成する。そして、検索された観光地と生成した観光地紹介文をユーザーに提示する。

ⁱⁱⁱ <https://huggingface.co/llm-jp/llm-jp-3.1-13b-instruct4>

^{iv} <http://133.220.249.13:8000/>

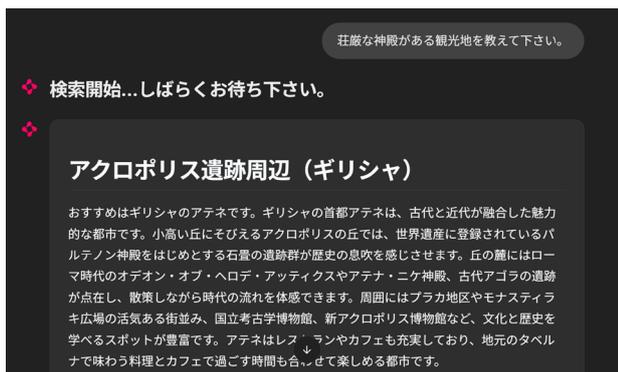


図1 観光情報検索システム

3 評価

本手法によって構築したインストラクションデータセットでチューニングをした LLM モデルの性能を評価するために、キーワードに特化した観光地紹介文の生成を行い、出力された紹介文を LLM によって比較評価した。

評価データを作成するために、まず 2 章の 2.3 節で抽出した固有表現の中から出現頻度が高い上位 1000 個を用意した。次にその固有表現を 2.2 節で抽出した観光地に関する文と 2.4 節で生成した観光地紹介文に含む観光地を検索して求める。これらをインプットとして LLM に入力し、固有表現を重要視した観光地紹介文の生成を行う。付録[3]にフライング(以降は FT と略記する)前後の LLM で生成した観光地紹介文の出力例を示す。FT 前後の LLM による生成結果を、GPT-5 を用いて評価した。GPT-5 には以下の 4 つの評価基準それぞれで、どちらの生成結果が優れているかとその理由を出力させ、最後に総合的な評価も出力させた。

評価基準

1. キーワードに関する体験、雰囲気、魅力を十分に説明できているか
2. 他の情報で観光地の魅力を補足することができるか
3. 観光地の魅力を正しく・魅力的に伝えているか
4. 内容が正確かどうか

付録[4]に GPT-5 による評価例を示す。このような評価を 100 件行った結果、総合 59 件で FT 後の LLM で生成された紹介文の方が優れているという結果となった。表 3 に評価基準ごとの結果を示す。ここで表の Rate は FT 後 LLM の勝率を示す。

表 3 LLM による評価結果

評価基準	FT 後	FT 前	Draw	Rate
1	39	61	0	39 %
2	50	49	1	50 %
3	57	43	0	57 %
4	55	40	5	57 %
総合	59	41	0	59 %

4 考察

総合評価において、FT 後の LLM が 100 件中 59 件で優位となった最大の要因は、情報の正確性と対象都市の一貫性の向上が考えられる。評価基準 4 (正確性) において勝率 57% を記録した通り、FT 前の LLM では他都市の情報や誤情報が多く混在する事例がより多く見られた。これに対し FT 後の LLM は都市名とその都市の紹介文を紐づけたデータセットで学習しているため、正確な情報を出力する傾向が強まったと考えられる。評価基準 3 (魅力的な伝達) においても、FT 後のモデルが 57% の勝率を収めた。これは、単にキーワードに言及するだけでなく、紹介文を学習したことで、インプットとして与えられた情報以外の観光地の魅力を含める能力が向上した結果と言える。一方で、評価基準 1 (キーワードに関する説明) では勝率 39% に留まり FT 前のモデルを下回った。この理由として、FT 前のモデルは特定のキーワードを強調して冗長に説明する傾向があるのに対し FT 後のモデルは、キーワード以外の魅力も含むよう学習しているためキーワードへの記述の割合が低くなってしまい、相対的に低く評価された可能性がある。

5 まとめ

本研究では、観光地についての Web ページを活用し、LLM を用いて海外の観光地紹介文を生成するためのインストラクションデータ自動構築手法を提案した。本手法によって世界 92 か国 3285 都市に及ぶ観光地に対し、約 236 万件のインストラクションデータを自動的に構築することができた。評価においては、LLM の FT に使用した学習データを各都市 100 件に限定したが、さらに多くのデータの活用によって精度向上の余地がある。今後は、より多くの観点でのデータセットの収集と、構築するインストラクションデータの改良を進め、より有益な観光情報の生成を目指していく予定である。

参考文献

- [1] Ryutaro Asahara, Masaki Takahashi, Chiho Iwahashi, Michimasa Inaba, “SumRec:A Framework for Recommendation using Open-Domain Dialogue”, Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, 2023.
- [2] Yuwei Cao, Nikhil Mehta, Xinyang Yi, Raghunandan Keshavan, Lukasz Heldt, Lichan Hong, Ed H. Chi, Maheswaran Sathiamoorthy, “Aligning Large Language Models with Recommendation Knowledge”, proceeding of NAACL 2024, pp.1051–1066, 2024.
- [3] Hao Cheng, Shuo Wang, Wensheng Lu, Wei Zhang, Mingyang Zhou, Kezhong Lu, Hao Liao, “Explainable Recommendation with Personalized Review Retrieval and Aspect Learning”, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pp.51–64, 2023.
- [4] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi, “Self-Instruct: Aligning Language Models with Self-Generated Instructions”, proceeding of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pp.13484–13508, 2023.
- [5] Haonan Li, Martin Tomko, Timothy Baldwin, “Location Aware Modular Biencoder for Tourism Question Answering”, Findings of the Association for Computational Linguistics: AACL-IJCNLP 2023, pp. 101–113, 2023.
- [6] Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Christopher Leung, Jiajie Tang, Jiebo Luo, “LLM-Rec: Personalized Recommendation via Prompting Large Language Models”, Findings of the Association for Computational Linguistics: NAACL 2024, pp.583–612, 2024.
- [7] Xiang Yue, Tuney Zheng, Ge Zhang, Wenhui Chen, “MAMMO-TH2: Scaling Instructions from the Web”, arXiv preprint arXiv:2405.03548, 2024.
- [8] Yuxin Jiang, Yufei Wang, Chuhan Wu, Xinyi Dai, Yan Xu, Weinan Gan, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Wei Wang, “Instruction-Tuning Data Synthesis from Scratch via Web Reconstruction”, Findings of the Association for Computational Linguistics: ACL 2025, pp.6603–6618, 2025.

付録

[1] 観光地紹介文の出力例 香港・中環

中環は香港の中心で、植民地時代の建物と近代的な高層ビルが混在し、ショッピングやグルメを楽しめるエリアです。中央大観覧車は 60 メートルの高さからビクトリアハーバーを一望でき、晴天ならビクトリアピークの凌曉閣まで見渡せます。また、セントラルマーケットは歴史的な建築を保存しつつ、カフェやショップが立ち並ぶ広場に生まれ変わりました。PMQ (元警察署) はクリエイターの拠点として人気です。ハリウッドロードやアートギャラリー、ボツラ通りは歴史とカルチャーを感じさせる通りで、世界最長の屋外エスカレーターとしても知られています。こうした多彩な風景は、初めての旅行者にとっても訪れる価値が高い観光地となっています。

[2] 生成したインストラクションデータの例

Instruction: 美しいビーチといった街並みを楽しめる観光地を教えてください。

Response: おすすめはトルコのチェシメです。トルコ・チェシメは、イズミル近郊に位置し、エーゲ海に面したリゾート町です。16 世紀に建てられた城塞からは青く輝く海と赤い屋根の街並みが一望でき、城内は博物館化されており、オスマン時代の噴水や各時代の大理石彫像・陶器が展示されています。美しいビーチや高級別荘が立ち並び、街歩きでは白壁の古民家や地元のレストランでトルコ風味のグルメを楽しめます。ギリシアの影響も残る雰囲気の中で、毎週土曜日に開催される青空市場で食材や土産を買えるのも人気です。フェリーでギリシャ・キオス島からもアクセスが楽で、観光客にとってはまさに海と歴史が調和した理想のリゾートです。

[3] スイス「トゥルン」の観光地紹介文 FT 前の LLM (Model B)

スイス トゥルンは都市を楽しむのに最適な場所です。歴史と自然が調和する美しい都市で、高台にそびえるトゥーン城を中心に、中世の雰囲気を残す街並、装飾された噴水、印象的な教会、ブティックやレストランが点在しています。ベルナーアルプスの湖の奥に位置し、12 世紀末にツェーリングン公が創設しました。ソロトゥルンは、16 世紀から 18 世紀の建物が立ち並ぶスイスで最も美しいバロック都市として知られ、先史時代に遡る歴史を誇ります。さまざまな建築様式が混ざり合い、特に貴族の邸宅など美しいバロック建築があり、隠れた宝物を発見し、町の豊かな歴史に触れることができるでしょう。

FT 後の LLM (Model A)

おすすめはスイス・トゥルンです。スイス・トゥルンはアルプスの湖畔に位置し、12 世紀末にツェーリングン公が創設した歴史ある都市です。古いトゥーン城を中心に広がる旧市街では、中世の雰囲気を残す噴水や豪華な教会、ブティック、レストランが並び、訪れる旅行者に歴史と自然の両方を満喫させます。街を散策しながら、丘の上からトゥーン湖を一望できる景色も楽しめます。

[4] GPT-5 による評価例 (Model A : FT 後の LLM、 Model B : FT 前の LLM)

1. 「都市」に関する体験・雰囲気・魅力の説明
 - 優れている方: Model_A
 - 理由: 旧市街の散策、ブティックやレストラン、噴水や教会、中世の雰囲気など、都市滞在中で体験できる具体的な情景がまとまっていて、歴史と日常の街歩きの魅力が伝わる。Model_B も「都市」を強調するが、後半でソロトゥルンの説明に逸れて一貫性を損ねている。
 2. 他の情報でトゥルンの魅力を補足できているか
 - 優れている方: Model_A
 - 理由: トゥーン城を中心とした旧市街、アルプスの湖畔、丘の上からの眺望など、都市以外の自然・景観の魅力を適切に補足している。Model_B はバロック都市としての記述などがソロトゥルンの情報であり、トゥルンの補足情報として不適切。
 3. 魅力を正しく・魅力的に伝えているか
 - 優れている方: Model_A
 - 理由: 歴史ある城と旧市街、湖と展望を組み合わせて、トゥルンの定番の魅力をバランスよく訴求している。Model_B は記述自体は魅力的だが、他都市の要素が混在し、トゥルンの魅力を正確に伝えきれていない。
 4. 内容の正確性
 - 優れている方: Model_A
 - 理由: トゥーン城、ツェーリングン家の関与、湖畔の立地、旧市街の雰囲気など概ね正確。Model_B はソロトゥルン (別都市) のバロック都市としての情報を混在させており、誤りが明確。
- 総合評価 「Model_A」