移動軌跡に関する質問応答データセット

浅野 輝 [1,2]　大内 啓樹 [3,4,2]　春日 瑛 [4]　米谷 竜 [4]
[1] 東京大学大学院　[2] 理化学研究所
[3] 奈良先端科学技術大学院大学　[4] サイバーエージェント
asano-hikaru19@g.ecc.u-tokyo.ac.jp, hiroki.ouchi@is.naist.jp
{kasuga_akira, yonetani_ryo}@cyberagent.co.jp

概要

本研究では，移動軌跡データに関する質問応答（QA）データセットを構築した．より具体的には，事実照会問題，選択式問題，自由記述式問題という 3 種類の QA タスクを定義し，それぞれに対して 500 件の問題・回答ペアを作成した．構築したデータを使用して大規模言語モデルの性能を評価し，一日における移動軌跡データに関する QA よりも，複数日にまたがる移動軌跡データに関する QA の方が難易度が高いことがわかった．

1　はじめに

携帯電話や車両に搭載された GPS による位置測位技術が発展し，人々がいつ・どこを・どのように移動したかを詳細に記録した**移動軌跡データ**を取得できるようになった．移動軌跡データは都市計画・公共交通・感染症対策・ターゲット広告など多様な分野で活用されている．人々が頻繁に立ち寄る地物（POI; Point-of-Interest）や典型的な移動パターンを分析することで，交通混雑の緩和施策の実施や位置に応じた観光・商品情報の提供といった応用につながることが期待される．

一般に，移動軌跡データは不特定多数の人物による複数日にわたる位置情報（緯度・経度，あるいは環境地図における座標値）の系列として与えられており，その解釈は必ずしも容易ではない．既存研究では移動軌跡データの可視化手法やクラスタリング手法が提案されているが，移動の背後にある意図や文脈を深く理解するためには，人手によるさらなる分析が必要となる．

この課題を自然言語処理によって解決する一つのアプローチとして，我々は移動軌跡を言語と結びつけ，言語を介して移動を解釈・説明あるいは探索する **Mobility Question Answering** [1] に取り組んでいる．特に本研究では，移動軌跡データに関する機械学習モデルの質問応答（QA）能力を測定するベンチマークデータセット **Mobility QA Dataset** を新たに導入する．本データセットでは，移動軌跡データに対して事実照会問題，選択式問題，自由記述式問題という 3 種類の QA タスクを定義し，それぞれに対して 500 件の問題・回答ペアが与えられている．その例を図 1 に示す．本稿では，同データセットを用いて複数の LLM を対象とした基礎実験の結果を報告する．

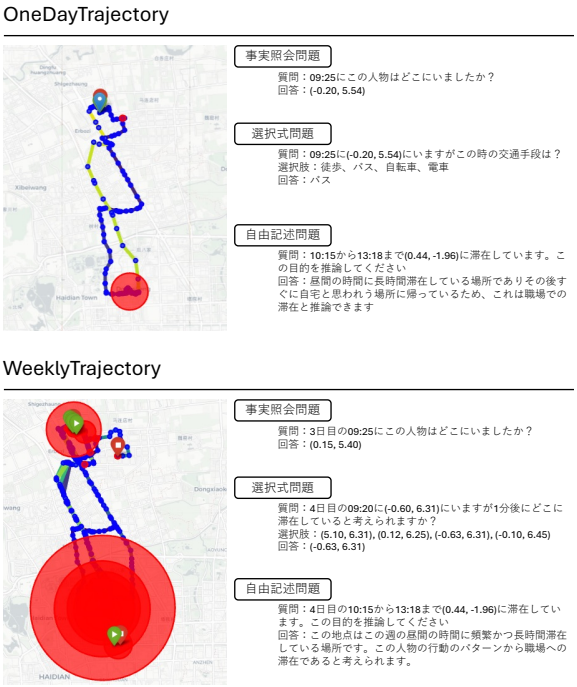

図 1　Mobility QA の例

2　関連研究

本研究はマルチモーダル QA 研究 [2, 3, 4] に位置付けられ、特に移動に関する QA [5, 6, 7] と関係が深い．これらの研究では移動中の一人称視点の画像

データに基づく QA を扱っている．一方で我々の研究では，画像データは使用せず，緯度・経度・時刻の三つ組を基本単位とする系列から構成される移動軌跡データを使用している点が異なる．また，緯度・経度等の位置・空間情報に基づく QA 研究 [8, 9] とも関連する．これらの研究では地物や経路等の地理情報に関する内容に主眼があり，人の移動を扱っているわけではない点が我々の研究と異なる．我々の知る限り，移動軌跡データに関する QA タスクの提案，および，そのデータセットの構築を行った研究は，本研究が初である．

3 Mobility QA Dataset

3.1 タスク設定

一般的に QA タスクは，質問 Q に対し，情報源 C が与えられた上で答え A を生成することを求められる．Mobility QA においては，この「情報源」として**移動軌跡データ** T が用いられる．すなわち，本タスクは $(Q,T) \mapsto A$ という形で，ユーザの移動軌跡を分析し，自然言語による適切な回答を出力することを目的とする．移動軌跡 T は，時刻 t_i と 2 次元座標 (x_i, y_i) を持つイベントの列 e_i から構成される．1 日もしくは複数日にわたる一連のイベントが積み重なり，移動軌跡 $T = \{e_0, \ldots, e_N\}$ を形成する．本研究ではこの移動軌跡 T をテキスト形式に変換し，LLM による解決を目指す．

3.2 移動軌跡データ

移動軌跡が長くなるほどモデルが処理すべき時系列トークンも膨大となり，ロングコンテキストの推論が必要となる．一方で，長期間の時系列データを適切に処理できれば，行動パターンや反復的に訪問する場所といった傾向をとらえることが可能となる．本研究では，短期間と長期間の移動軌跡における LLM の推論能力を比較するため，以下の 2 種類のデータセットを構築する．

- **OneDayTrajectory:** 特定の個人の 24 時間（1 日）にわたる移動データ．
- **WeeklyTrajectory:** 同一個人の連続 7 日間にわたる移動データ．

特に本稿では，著名な軌跡データセットである Geolife[1] [10] から N 名のユーザ・M 日分の軌跡を抽出し，それぞれのユーザに対して OneDayTrajectory と WeeklyTrajectory を構築した．

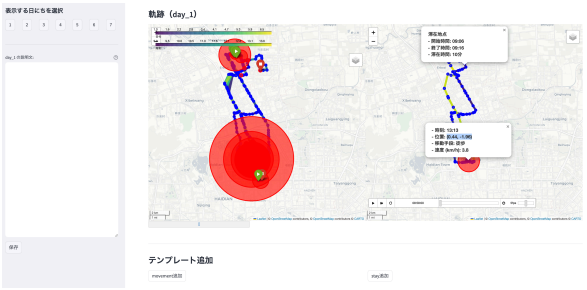

図 2 アノテーションツール

3.3 質問の種類と QA ペアの構築

Mobility QA における LLM の能力を包括的に評価するため，以下の 3 種類の質問形式を定義する．

1. **事実照会問題** 軌跡から特定の事実情報を直接取得する質問（例：「ユーザは 10 時 15 分にどの座標付近にいたか？」）．
2. **選択式問題** 用意された選択肢から正答を選ぶ質問（例：「ユーザが次に移動する可能性が最も高い目的地はどれか？」）．
3. **自由記述式問題** 解釈や説明を含む文章回答を求める質問（例：「3 日目の行動パターンを要約し，移動の目的を推論せよ．」）．

具体的には，まず 3 種類の質問形式それぞれに対して複数の質問テンプレートを作成した．そして移動軌跡 T の中からランダムにイベント e を抽出し，イベント e の情報を使用しつつテンプレートに沿って質問を作成した．

一方で，回答の生成には以下の 2 つの方法を本研究では採用した．

ルールベース生成　事実照会問題および選択式問題に対して，軌跡データから質問 Q に対応するイベント列 $\{e_{i^1}, \ldots, e_{i^n}\}$ を抽出し，これに基づいて回答 **A** を作成した（例：最も訪問回数が多い座標を特定する，ある時刻の位置を取得する）．

LLM を用いた生成　自由記述式の回答では，移動の意図や日常パターンといったセマンティック情報の把握が不可欠となる．しかし，LLM に生のユーザーの移動軌跡データをそのままテキストとして与えた場合，時系列が膨大なトークンとなり，移動速度や移動方向などの情報を十分に解釈できない可能性がある．本研究では図 2 の可視化ツールを用いて，移動軌跡 T を以下の 2 種類のセマンティックな情報によってラベル付けをした：

1） https://github.com/jeffmur/geoLife

- **Movement Phase（移動区間）**：出発地・到着地，出発時刻・到着時刻，近似的な POI（Point of Interest），主要交通手段，移動目的など
- **Stay Phase（滞在区間）**：滞在の開始時刻・終了時刻，合計滞在時間，座標，滞在目的

このようなセマンティックな情報を移動軌跡 **T** に付与することによって，LLM が移動軌跡からより効率的に質問に対する回答が生成できると期待される．そこで，本研究では，自由記述問題に対する回答を $\mathbf{A} = F(\mathbf{Q}, \mathbf{T}, \mathbf{L})$ のように生成した．

3.4 評価指標

本研究では，質問タイプに応じて異なる評価指標を用いる．事実照会問題と選択式問題に対しては，正解（Ground Truth）となる回答が存在する．したがって，評価指標として**正解率（accuracy）**を用いてモデル性能を評価する．具体的には，モデルが出力した回答が正解ラベルと一致した割合を算出する．

4 実験

本節では，新たに提案した Mobility QA データセットを用いて，コストとパフォーマンスの効率からよく利用されるモデルである **gpt-4o-mini** での性能を評価する．ここでの評価では，LLM に対して自己修正型の解答生成を行う仕組みを導入する．具体的には，以下の 2 つのエージェントを用いたフレームワークを構築した．

- **回答用エージェント**（Agent-A）：与えられた「質問」と「軌跡データ」から初回の回答を生成し，その後，評価用エージェントからのフィードバックを受け取った上で再度回答を修正・更新する．
- **評価用エージェント**（Agent-B）：回答用エージェントが生成した回答を参照し，その回答内容を「質問」「軌跡データ」と照らし合わせて評価（フィードバック）を行う．

この枠組みでは，まず Agent-A が（質問，軌跡）を入力とし回答を生成する．続いて Agent-B が（質問，軌跡，回答）を入力とし，回答内容に対する評価・フィードバックを与える．最後に，再び Agent-A が（質問，軌跡，前回の回答，フィードバック）を入力として，新しい回答を作成する．最終的に得られた修正後の回答をモデルの出力として評価する．

Prompt

You are tasked with extracting accurate and detailed facts from an individual's trajectory data.

Key Information:

- The trajectory data is normalized relative to the coordinates of the initial event on Day 1.

Objective:

- Utilize the provided trajectory data to precisely answer the fact retrieval question below.

Guidelines for Fact Retrieval:

1. **Event Extraction:**
   - Identify and extract relevant event details such as date, time, and coordinates that pertain to the question.
2. **Location and Time Analysis:**
   - Determine the individual's exact location at specified times.
   - Count the number of visits to specific coordinates.
3. **Frequency and Pattern Detection:**
   - Assess the frequency of visits to designated coordinates.
   - Identify and analyze any patterns or irregularities in the timing of visits.

Answer Construction:

- Base all responses solely on the data provided.
- Provide a clear and concise answer that directly addresses the fact retrieval question.
- If multiple answers are possible, select and present only one (either a single time or location).
- Use precise, logical, and unambiguous language.
- Apply critical thinking and sound judgment to interpret the data effectively.

 # Trajectory Data
`{trajectory}`

 # Question
`{question}`

図 3 事実照会問題用の回答用プロンプト

4.1 テストセット

gpt-4o-mini に対して，以下の 4 種類のテストセット（合計 2000 件）を用意し，それぞれについて上記の自己修正型フレームワークを適用して解答を生成させた．

1. 事実照会問題＋ OneDayTrajectory（500 件）
2. 選択式問題＋ OneDayTrajectory（500 件）
3. 事実照会問題＋ WeeklyTrajectory（500 件）
4. 選択式問題＋ WeeklyTrajectory（500 件）

Feedback Prompt

You are a helpful "feedback assistant" that provides constructive feedback to improve the user's answer.

Please analyze:

- The user's question
- The anonymized trajectory data
- The previous answer

Your tasks:

1. Identify any deficiencies or inaccuracies in the previous answer.
2. Provide clear, constructive suggestions to refine and improve the answer.

[Output format example] `[Feedback]`: (Your remarks about any issues in the previous answer and your suggestions for improvement)

Note: Ensure that your feedback follows the content policy and does not include any inappropriate or harmful content.

図 4 フィードバックエージェン用のプロンプト

表 1 事実照会問題・選択式問題における評価結果

Model	OneDayTrajectory		WeeklyTrajectory	
	事実照会	選択式	事実照会	選択式
GPT-4o-mini	0.906	0.290	0.432	0.296
+ 1 Iteration	0.974	**0.320**	0.534	0.342
+ 2 Iteration	**0.980**	0.284	**0.588**	0.342
+ 3 Iteration	0.970	0.292	0.558	**0.352**

4.2 評価結果

GPT-4o-mini を用いて性能を評価したところ, OneDayTrajectory に対しては事実照会で 98.0%の高い精度を達成した一方で, WeeklyTrajectory では精度が 58.8%まで低下することがわかった. 選択式では, 逆に WeeklyTrajectory の方が, 35.2%となり OneDayTrajectory の 32.0%よりわずかに高い精度となった. 自由記述式に関しては今回は評価を行っていない. この結果から, 事実照会のようなデータから自明に答えが決まるようなタスクに関しては, 短期間の移動軌跡であれば高い精度が達成できる一方で, 長期間の移動軌跡になると, 膨大なトークンを処理しきれずに顕著に精度が低下していると言える. 選択式問題は, いずれの期間でも低い精度を示した. 可能性の最も高い目的地を選ぶような推論タスクの場合, 大規模言語モデルが移動軌跡に関するデータを解釈して質問に回答する能力にはまだ限界があることが示された. この結果から, 本データセットを用いることで, 移動軌跡に対する推論能力に長けた大規模言語モデルの研究および開発が期待される.

5 おわりに

本研究では, Mobility QA Dataset を構築した. 実際に GPT-4o-mini を用いて評価することで, 移動軌跡データに関する機械学習モデルの質問応答（QA）のベンチマーク測定が可能になることが示された.

今後は、精緻なアノテーションによって自由記述式問題の評価も行うとともに、ベンチマークの評価指標の追加や、大規模言語モデルを用いた自動評価手法の開発を進める。また、本研究ではまず Geolife をベースとした屋外の移動軌跡を初期データセットとして用いたが、今後は屋内で計測された行動データも取り込み、さらなる拡張を図る。

さらに、軌跡モデルに対する推論能力を向上させるため、精密なアノテーションと回答エージェント・評価エージェントによる段階的な推論能力向上の仕組みを導入し、移動軌跡に対する Chain-of-Thought など、言語モデルの推論手法に関する研究も並行して進めていく予定である。

参考文献

[1] Hao Xue and Flora D. Salim. Human mobility question answering (vision paper), 2023.

[2] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images, 2021.

[3] Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. MIMOQA: Multimodal input multimodal output question answering. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 5317–5332, Online, June 2021. Association for Computational Linguistics.

[4] Darryl Hannan, Akshay Jain, and Mohit Bansal. Manymodalqa: Modality disambiguation and qa over diverse inputs, 2020.

[5] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, Elahe Arani, and Oleg Sinavski. Lingoqa: Visual question answering for autonomous driving, 2024.

[6] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario, 2024.

[7] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments, 2020.

[8] Gengchen Mai, Krzysztof Janowicz, Rui Zhu, Ling Cai, and Ni Lao. Geographic question answering: Challenges, uniqueness, classification, and future directions, 2021.

[9] Haonan Li, Martin Tomko, and Timothy Baldwin. Location aware modular biencoder for tourism question answering, 2024.

[10] Yu Zheng, Xing Xie, Wei-Ying Ma, et al. Geolife: A collaborative social networking service among user, location and trajectory. **IEEE Data Eng. Bull.**, Vol. 33, No. 2, pp. 32–39, 2010.