6A-04

GPT-2 を用いた人々の一日の移動軌跡の生成

藤本　祥二†　　石川　温†　　水野　貴之‡
金沢学院大学†　　国立情報学研究所‡

1. はじめに

個人の日々の移動に関するビッグデータは、災害、テロ、治安、感染症、空間隔離、マーケティング、交通渋滞などの問題に取り組む上で重要である。個人の移動軌跡の統計的性質を満たすモデルを開発することによって、新しいインフラの導入や疫病の蔓延、テロ攻撃、万博のような国際的イベントによる都市移動の変化をシミュレーションすることが可能となる。さらに、生成モデルは軌跡データのジオプライバシーを保護するためにも有用である。人間の移動のモデリングはいくつかの種類があるが、本研究では個人の移動の軌跡を生成するモデルを構築する[1]。

自然言語生成において Recurrent Neural Network の代替となりつつある Transformer モデルの一つに GPT-2[2]モデルがある。本研究ではこの GPT-2 を用い、自然言語の代わりに個々の日々の移動軌跡を生成する。学習済みのモデルに、朝の初期位置（例：自宅周辺）を入力し、個々の一日の軌跡（例：公共交通機関で観光地へ行き、観光して食事をし、帰宅する経路の座標）を出力する。このモデルによって生成された軌跡は、1）軌跡の 1 時間毎の移動距離の分布は対数関数に従うファットテイルを持つ、2）移動距離の自己相関関数は短時間記憶を示す、3）長距離移動では 1 時間の移動方向に正の自己相関が存在する、4）個々の軌跡において最終位置は初期位置の近くになることが多い、5）人の拡散は移動の時間スケールによって変化する、等の 5 つの現実的な特性を再現している。

2. データ

本研究では株式会社 Agoop 提供による、2021 年 11 月と 2022 年 1 月に京都駅周辺（京都市下京区）を通過した計 170 万台のスマートフォン（1 日あたり約 2 万 8 千台）の位置情報データ（2 億 8 千万ログ）を用いた。位置情報は主に GPS による緯度・経度の情報で、精度は通常 20m 以内である。各軌跡を 250m グリッド、30 分オーダーで 1 分のスライディングウィンドウを用いて調整し、このスライディングウィンドウにより、1 分の時間分解能を持つ 170 万個の軌跡は、30 分の時間分解能を持つ 51（=170×30）万個の軌跡に変換される。また、各ユーザのジオプライバシーを保護するために、自宅グリッドを削除しユーザが外出した際の軌跡のみに着目する。10 時間以上外出した個人の 1 日の軌跡の総数は 840 万時系列である。これらの時系列を独立に 8:2 の割合でモデルの学習用データと検証用データに分け、検証データを用いて移動軌跡の生成を行った。

3. グリッド位置情報のトークン化

データの位置情報は緯度・経度で記録されているが、これを日本地域グリッドコード「JIS X0410」の 5 次メッシュを用いてグリッドコードに変換することで「5235/36/80/2/3」のように、5 つのサブコードの組み合わせで表現される。第 1 レベルのサブコード（例:5235）は、緯度 40 分差、経度 1 度差の正方形で囲まれた固有の領域を表している 4 桁の数字、第 2 レベルのサブコード（例:36）は、第 1 レベルのグリッドを緯度方向と経度方向に 8 等分した領域を示す 2 桁の数字、第 3 レベルのサブコード（例:80）は第 2 レベルのグリッドを緯度および経度方向に 10 等分して得られる領域を示す 2 桁の数字、第 4 レベルのサブコード（例:2）は第 3 レベルのグリッドを緯度と経度で 2 等分した 4 つの 1 桁の数字、第 5 レベルのサブコード（例:3）は第 4 レベルのグリッドを緯度・経度で 2 等分した 4 つの 1 桁の数字である。このコード化によって 1 辺の長さ約 250m の解像度で、日本の陸地に存在する約 1800 万のユニークなグリッドコードを第 1 レベルから第 5 レベルまでのたった 348 個のサブコードの組み合わせで表現できる。この結果、
(34.716, 135.586)　(34.696, 135.548)
(34.697, 135.535)　(34.695, 135.529)
(34.788, 135.689)
のように緯度経度で表されている軌跡は
5235045644　5235043342　5235043242
5235043214　5235154531
のようなコードに変換される。

次に、言語モデルを適用するための技術的な処理として「グリッドサブコードからバイト文字(UTF-8)への変換マップ」を作成した。ただし、第 2 レベルのサブコードの「36」と第 3 レベルのサブコードの「36」は意味が異なるので、異

なるバイト文字を割り当てている。第1レベルのサブコードを176文字、第2レベルを64文字、第3レベルを100文字、第4レベルを4文字、第5レベルを4文字に対応する変換マップを作成し、頻出文字の組み合わせ（＝頻出サブワード）を、合計5万トークンになるまで日本の陸地に関するトークンとして設定した。この変換によって上記の軌跡は

ßħγB2　ßħδB3　ßħλB3　ßħλC2　ßμξD4

と表される。

最後に、一時帰宅にカンマトークン「,」、最終帰宅にピリオドトークン「.」を追加して各ユーザの一日の移動軌跡を表す。

3. GPT-2 による軌跡の生成

前節のように変換し、トークン列に変換した各個人の移動軌跡の訓練データで GPT-2 モデルの学習を行い、学習済みモデルに検証データの軌跡の最初の 5 地点を入力し、残りの軌跡の生成をピリオドトークンが生成されるまで行った。図1は生成された軌跡の1例である。

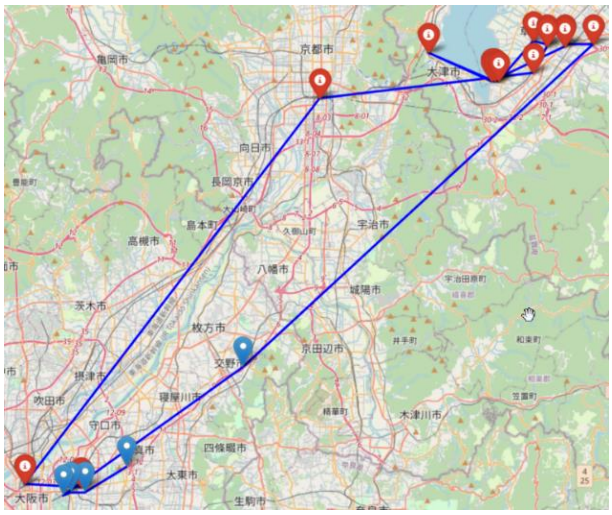

図 1　GPT-2 モデルを用いて生成した移動軌跡の例、青色のピンは入力に用いた最初の 5 地点、赤色ピンは生成された一日の移動軌跡

4. 他のモデルとの比較による精度検証

比較のため本研究では、4 種類のモデル、a) 2-gram モデル、b) 3-gram モデル、c) Catboost モデル、d) GPT-2 モデル、を用いて軌跡の生成を行い、検証データの軌跡と生成された軌跡の統計的性質の違いを確認した。確認したのは、1) 移動距離の分布、2) 移動距離分布の自己相関、3) 移動距離と次に向かう角度の関係、4) 初期位置に再帰する確率、5) 人々の拡散係数、の 5 つの統計的性質である。いずれも GPT-2 で生成した軌跡が検証データの性質を最も再現した。図 2 は一日の移動軌跡の最終時刻（帰宅時刻 Time=0）までの5時間について、初期座標から 3km 圏内にいる再帰確率を検証データと各モデルの生成データの比較である。

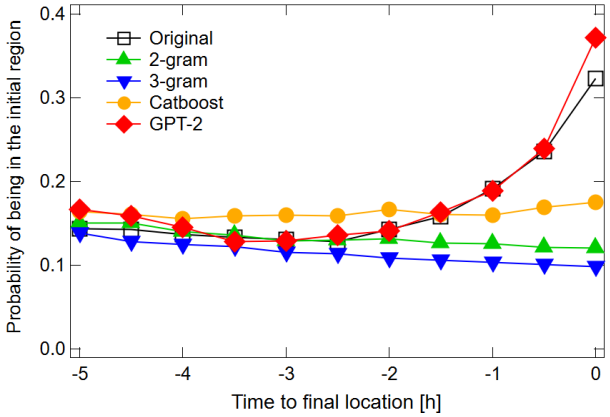

図 2　初期位置に再帰する確率

最後に予測精度について、表 1 に生成した軌跡の 30 分後、1 時間後、最終時刻において、予測位置が検証データの実際の位置座標から 1km（又は 10km）以内である確率を示す。GPT-2 モデルで生成した 1 日の最終時刻の位置は、検証データの最終時刻の位置を他のモデルと比較して高い確率で再現している。

表 1　予測精度の比較

	30 分後	1 時間後	最終時刻
a)	0.29(0.67)	0.16(0.54)	0.011(0.20)
b)	0.33(0.75)	0.20(0.61)	0.014(0.19)
c)	0.15(0.70)	0.070(0.54)	0.016(0.25)
d)	0.40(0.82)	0.26(0.70)	**0.12(0.40)**

謝辞

本研究は JST CREST JPMJCR20D3, JSPS 科研費 JP19K22852, JP21H01569, JP21K04557 の助成を受けています。

参考文献

[1] T. Mizuno, S. Fujimoto and A. Ishikawa. “Generation of Individual Daily Trajectories by GPT-2”. Frontiers in Physics 08 November 2022 Sec. Interdisciplinary Physics

[2] A Radford, J Wu, R Child, D Luan and D Amodei, “Language models are unsupervised multitask learners” OpenAI blog (2019) 1:9.

Generation of people's daily movement trajectory by using GPT-2

†FUJIMOTO Shouji, ISHIKAWA Atsushi, Kanazawa Gakuin University

‡MIZUNO Takayuki, National Institute of Informatics