

大規模言語モデルを用いた観光地レビューに基づく 観光地推薦手法に関する研究

20A3009
指導教員

飯塚 柊
二宮 洋

1. はじめに

近年、情報技術の発展により観光施設や公共団体から公開されている情報や、Q&A サイトでのお薦め観光地の質問・回答、ウェブマッピングサービスのレビューなど、多種多様な観光情報が蓄積されている。しかし、これらから得られる情報は膨大であり、利用者が自分の好みや嗜好に合った観光地を迅速に見つけることは容易ではない。

この問題に対し[1]では、インターネットから取得した観光地周辺の施設情報、地図画像等から観光地をベクトルで表現し、得られた観光地ベクトルと他の観光地ベクトルのコサイン類似度を用いて類似した観光地を推薦する手法を提案している。しかし、この手法では生成された観光地ベクトルが定量的な情報源から形成されており、観光地の情緒的な特徴を考慮していない問題がある。

この問題に対して[2]では、観光地の情緒的な特徴を考慮するため、観光情報サイトじゃらん[3]から観光地のレビュー文章を収集し、Distributed Bag of Word version of Paragraph Vector(PV-DBOW)を用いて観光地レビューをベクトルに変換することで観光地ベクトルを表す手法を提案している。この手法では、PV-DBOW の学習に膨大なレビュー文章が必要であるという問題がある。

本研究では、[2]の問題を解決するため、事前学習済み大規模言語モデル(LLM)を用いて観光地のレビュー文章をベクトルに変換することで観光地をベクトルで表現する方法を提案する。事前学習済み LLM は膨大な文章を大規模なネットワークで学習することで、高い文章理解能力を有している。提案手法は、[2]と比べ観光地をベクトルに変換するための PV-DBOW の学習が不要であるという利点がある。また、事前学習済みの LLM は高い文章理解能力を有しているため、PV-DBOW より観光地をベクトルで表現する能力が高いことが考えられる。

実験として、提案手法と従来手法を用いて新宿御苑に似た大阪の観光地推薦を行い、その結果を比較する。

2. 事前学習済み大規模言語モデル

事前学習済み LLM は、インターネットから取得した膨大な文章を Transformer に基づくアーキテクチャに学習させたモデルであり、高い文章理解能力を有している。LLM は文章生成や文章分類等多岐にわたり応用されている。

Sentence-LUKE[4]は BERT の学習時に単語のエンティティを考慮したモデルであり、Sentence-

LUKE は文章を入力とし、ベクトルを出力するモデルである。従来の LLM と比べ日本語言語理解ベンチマーク等のタスクで最高スコアを有している。

3. 提案手法

本研究では、事前学習済み LLM を用いて観光地のレビュー文章をベクトルに変換することで観光地をベクトルで表現する手法を提案する。提案手法では、[2]と同様に観光情報サイトじゃらん[3]の観光地のレビュー文章をスクレイピングにより取得する。じゃらんの観光地のレビュー文章の概要を図 1 に示す。

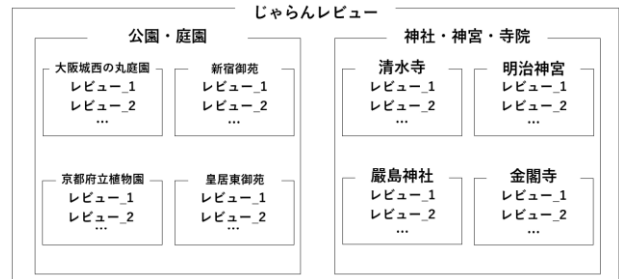


図 1 観光地レビューの概要

図 1 に示すように、じゃらんのレビュー文章は観光地名、観光地の位置する都道府県、じゃらんが設定した観光地のカテゴリが公園・庭園、神社・神宮・寺院のように紐付けられている。得られた観光地ベクトルはレビュー文章毎に事前学習済みの LLM に入力することでベクトルに変換する。本研究では、同じ観光地のレビュー文章を平均したベクトルを観光地ベクトルとする。ある観光地 S を表す観光地ベクトル \mathbf{v}_S を(1)に示す。

$$\mathbf{v}_S = \frac{1}{N_S} \sum_{i=1}^{N_S} \mathbf{v}_{S,i} \quad (1)$$

ここで、 N_S は観光地 S のレビューの数、 $\mathbf{v}_{S,i}$ は観光地 S の i 番目のレビュー文章を事前学習済み LLM に入力し得られたベクトルである。また、都道府県を表す都道府県ベクトルは同じ都道府県に属する観光地レビュー文章を LLM に入力し得られたベクトルの平均したベクトル、カテゴリを表すカテゴリベクトルは同じカテゴリに属する観光地レビュー文章を LLM に入力し得られたベクトルを平均したベクトルとする。

本研究では、[2]と同様の方法で観光地を推薦する。はじめに、利用者は所望する観光地と、都道府県またはカテゴリを定める。次に、所望する観光地を表す観光地ベクトルからその観光地が属

する都道府県またはカテゴリを表すベクトルを引き、所望する都道府県またはカテゴリを表すベクトルを足すことで得られる合成ベクトルを取得する。得られた合成ベクトルと全ての観光地を表すベクトルでコサイン類似度を計算し、類似度が高いものを推薦する観光地とする。所望する観光地が新宿御苑に似た、都道府県が大阪の観光地である場合の合成ベクトルの計算方法を(2)に示す。

$$\tilde{\mathbf{v}} = \mathbf{v}_{\text{新宿御苑}} - \alpha \mathbf{v}_{\text{東京}} + \alpha \mathbf{v}_{\text{大阪}} \quad (2)$$

新宿御苑は東京に位置する観光地であるため、新宿御苑を表す観光地ベクトルである $\mathbf{v}_{\text{新宿御苑}}$ から、東京を表す都道府県ベクトルである $\mathbf{v}_{\text{東京}}$ を引き、大阪を表す都道府県ベクトルである $\mathbf{v}_{\text{大阪}}$ を足している。ここで、 α は都道府県またはカテゴリの影響度を表すパラメータであり、 α の値を変更することで推薦される観光地が変わる。提案手法では事前学習済みの LLM として日本語を事前学習済みの Sentence-LUKE である[5]を用いる。

4 実験

本実験では、提案手法を用いて[2]と同様の条件で観光地を推薦し、その結果を比較する。[2]では、所望する観光地と似た観光地を新宿御苑、所望する都道府県を大阪としている。新宿御苑は東京に位置する公園・庭園カテゴリの観光地であり、このことから大阪に位置する公園・庭園カテゴリの観光地が推薦されることが望ましいと言える。[2]の推薦結果を表1に示す。表1では[2]の推薦結果の上位5つを抜粋しており、カテゴリ名である動物園・植物園をA、公園・庭園をBとした。また、大阪に位置する公園・庭園カテゴリの観光地を太字で表記している。表1から、[2]の推薦結果では所望する観光地は上位5つの内1つだけであることが分かる。

表1 [2]の推薦結果

観光地	類似度	都道府県	カテゴリ
新宿御苑の桜	0.6754	東京都	A
京都府立植物園	0.4791	京都府	A
皇居東御苑	0.4564	東京都	B
大阪城西の丸庭園	0.4370	大阪府	B
大谷山自然公園	0.4123	奈良県	B

4.1 実験 1

影響度パラメータ $\alpha = 1$ の際の提案手法の推薦結果を表2に示す。表2では影響度パラメータ $\alpha = 1$ の際の提案手法の推薦結果の上位5つを抜粋している。表2から、 $\alpha = 1$ の際の提案手法では、所望する観光地は上位5つの内1つだけであることが分かる。表1と比べ、提案手法はカテゴリ表現が優れていることが分かる。

4.2 等倍と定数倍での比較

次に、影響度パラメータ α を2から10の整数に変更しその結果を確認した。結果から、所望する観光地が上位5つの内に最も多く占めるのが影響度パラメータ $\alpha = 5$ の際であった。影響度パラメータ $\alpha = 5$ の際の提案手法の推薦結果を表3に示す。表3では影響度パラメータ $\alpha = 5$ の際の提案手法の推薦結果の上位5つを抜粋している。表3から、所望する観光地が上位5つの内4つを占めていることがわかる。このことから、影響度パラメータ α を適切に設定することで、都道府県ベクトルの影響を調整でき、推薦結果が改善することが確認できた。

表2 $\alpha = 1$ の際の提案手法の推薦結果

観光地	類似度	都道府県	カテゴリ
新宿御苑	0.9740	東京都	B
青葉の森公園	0.9476	千葉県	B
国営昭和記念公園	0.9466	東京都	B
大仙公園	0.9451	大阪府	B
都立水元公園	0.9397	東京都	B

表3 $\alpha = 5$ の際の提案手法の推薦結果

観光地	類似度	都道府県	カテゴリ
浜寺公園	0.8172	大阪府	B
久宝寺緑地	0.8122	大阪府	B
服部緑地	0.8114	大阪府	B
千葉県立柏の葉公園	0.8101	千葉県	B
山田池公園	0.8073	大阪府	B

5 まとめと今後の課題

本研究では、事前学習済み LLM を用いて観光地のレビュー文章をベクトルに変換することで観光地をベクトルで表現する手法を提案する。提案手法は、従来手法と比べ、レビュー文章をベクトルに変換するモデルを学習する必要が無いという利点がある。実験では、従来手法と同様の条件で観光地を推薦した結果、影響度パラメータ α を適切に設定することで従来手法よりも推薦結果が改善することを確認した。

今後の課題として、定量的に観光地を推薦する結果を評価すること、[1]で使用されている定量的な情報源も考慮して観光地をベクトルで表すことが挙げられる。

参考文献

- [1] 上原 尚, 他, 信学技報, NLC2012-35, Dec. 2012.
- [2] 吉田 朋史, 他, 観光情報学会, 第14回研究発表会講演論文集, pp.78-81, Nov. 2016.
- [3] Recruit Co., "じゃらん", <https://www.jalan.net/> (最終閲覧日:2024/1/21)
- [4] Yamada, Ikuya, et al., *arXiv preprint arXiv:2010.01057*, Oct. 2020.
- [5] Hugging Face, <https://huggingface.co/sonoisa/sentence-luke-japanese-base-lite> (最終閲覧日: 2024/1/21)