多言語 Wikipedia と BERT を用いた歴史的建造物の観光特性抽出の改良を目指して

谷口明都

広島市立大学 言語音声メディアグループ

概要 本研究では， 3 つの言語版の Wikipedia のダンプデータから 9 か所の歴史的建造物の記事のテキストデータを取得し，BERT を用いてそれらの類似度を算出することで日本と海外の歴史的建造物に関する紹介の仕方の共通点や相違点を明確化することを目指す．また，本報告ではリンク情報や画像情報を扱った追加実験の結果と分析，および今後の展望についても述べる．

1 はじめに

新型コロナウイルス感染症の世界的な蔓延の収束に伴い，再び海外への渡航が活発になっている中，日本においても 2025 年に開催される日本国際博覧会では多くの外国人観光客が訪日することが予想されている．観光地に従事する人々にとっては現地の良さを世界にアピールするチャンスだが，その際，多くの旅行者に来てもらうために重要となってくるのが観光案内である．

一方で，国土交通省観光庁が公表した 2023 年の年次報告書[1]では，訪日外国人を対象に「出発前に得た旅行情報源で役立ったもの」を調査したところ，SNS や個人ブログと回答した割合は 30%前後であるのに対し，宿泊施設ホームページは約 15%，地方観光協会ホームページは約 5%と低い割合を示している．

旅行者の母国の文化や国民性，目的地との関係などにより，旅行者が観光で何を求めるかも国ごとに違ってくる．SNS や個人ブログが旅行情報源として採用されやすいのは，旅行希望者がインターネットを通して自分の需要にマッチした情報を得ているからというのも一つの要因だろう．観光地側からの発信においても，こうした旅行者の国柄を考慮した観光案内が必要となる．

本研究では，その第一歩として，多言語 Wikipedia を用いて，日本と海外の歴史的建造物に関する紹介の仕方の共通点や相違点を明確化する方法を提案する．また，改良として，厳島神社に焦点を当て，Wikipedia の記事内のリンク情報や画像情報も扱った追加実験を行う．

2 提案手法

本研究では，まず海外観光客から人気の観光地上位 20 位から歴史的建造物を含む 9 か所を選択し，日本語版，英語版，フランス語版の Wikipedia のダンプデータからそれぞれがタイトルの記事をデータとして取得する．その後，各記事をセクションごとに分割し，BERT を用いて日本語と英語，日本語とフランス語で比較したときの類似度を算出する．ここで選択したのは，伏見稲荷大社，金閣寺，浅草寺，新宿御苑，明治神宮，奈良公園，姫路城，宮島，兼六園である．

3 実験

3.1 データ作成

まず，実験データとして，先に挙げた 9 つの歴史的建造物について，Wikipedia のダンプデータから日本語版，英語版，フランス語版の記事のテキストデータを取得する．このダンプデータは 2023 年 11 月 1 日に取得した．

次に，各記事をセクションで分割し，日本語と英語，日本語とフランス語でセクションごとに比較する．このとき，2 つの間に関連があれば 1 を，関連がなければ 0 を正解データとして入力する．

その後，日本語と英語，日本語とフランス語でセクションごとに BERT を用いて類似度を算出する．このとき，類似度が閾値以上であれば 1 を，閾値未満であれば 0 を類似度データとして出力する．

最後に，正解データと類似度データから精度，再現率，および F 値を評価値として算出する．

3.2 予備分析

類似度データを計算する際の閾値を 0.70 から 0.95 まで 0.05 ずつ変化させた．このときの日本語と英語，日本語とフランス語で比較した各データの評価値ごとの平均の変化をそれぞれ図 1, 2 に示す．

図 1：日英で比較したときの評価値の変化

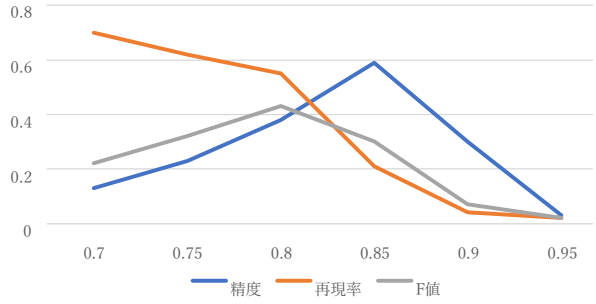
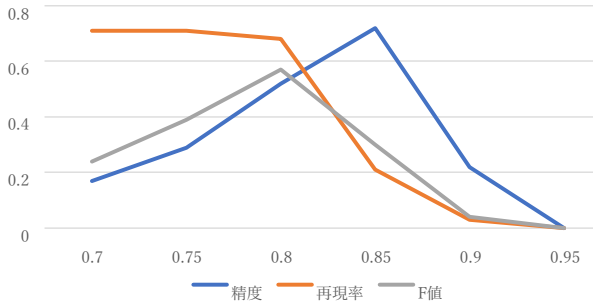


図 2：日仏で比較したときの評価値の変化



どちらのグラフからも閾値が 0.85 のときに精度が最大となったことが読み取れる．よって，閾値が 0.85 のときに着目して歴史的建造物ごとの評価値を分析する．

3.3 歴史的建造物ごとの比較分析

表 1, 2, 3 に閾値が 0.85 のときの宮島，伏見稲荷大社，明治神宮の各言語の評価値をそれぞれ示す．

表 1：宮島に関する評価値

	精度	再現率	F 値
英語	0.20	0.03	0.06
フランス語	0.75	0.27	0.39

表 2：伏見稲荷大社に関する評価値

	精度	再現率	F 値
英語	0.28	0.20	0.23
フランス語	1.00	0.10	0.18

表 3：明治神宮に関する評価値

	精度	再現率	F 値
英語	0.64	0.37	0.47
フランス語	0.37	0.37	0.37

3 つの表から，伏見稲荷大社と明治神宮は英語版の方がフランス語版よりも評価値が高く，宮島はフランス語版の方が英語版よりも評価値が高いことが読み取れる．

また，他の 6 か所の歴史的建造物においては，英語版の方が高い値となったのは金閣寺，浅草寺，兼六園で，フランス語版の方が高い値となったのは新宿御苑，奈良公園，姫路城だった．このとき，金閣寺においてはフランス語版でも僅差の値だったのに対し，新宿御苑，姫路城においては英語版と大きな差があった．このことから，全体的に見れば，英語よりもフランス語の説明の方が日本語に類似していると考えられる．

4 追加実験

追加実験では，厳島神社の日本語版，英語版，フランス語版の記事について，3 章の実験で扱わなかったリンク情報や画像情報をもとに言語ごとの特徴を分析する．このとき，3 章の実験のようにテキストをセクションで分割はせず，記事内の本文全体を対象として比較を行う．

4.1 データ取得

リンク情報，画像情報ともに URL から Wikipedia の厳島神社のページへアクセスし，ページの HTML 構造を解析することで取得する．このとき，リンク情報として<a>タグからリンク先の URL と記事上の文字列，画像情報として<img>タグから画像の URL とタイトルを取得する．

4.2 差異情報抽出

取得したリンク付き文字列のうち，日本語版の記事にはないが英語版あるいはフランス語版の記事にはあるものは，その言語の特徴を表していると言える．よって，これを差異情報として抽出し，言語ごとの特徴を調べる．ただし，同じ記事内の注釈へのリンクや Wikipedia 以外のサイトへのリンクなどが付いた文字列は含まないものとする．

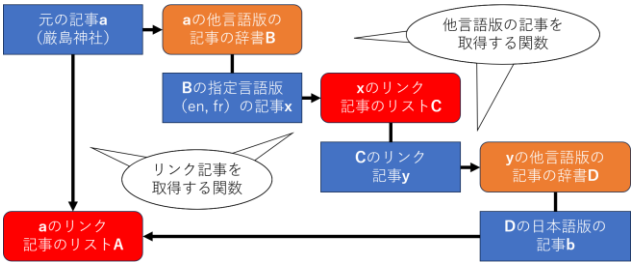
差異情報抽出のためのプログラムでは，以下の 2 つの関数を使用する．

I. 記事内のリンク付き文字列が持つ URL を取得し，URL のリストを作成する関数
II. 対応する他の言語版の記事の URL を取得し，その言語の言語コードをキー，URL を値とする辞書を作成する関数

以下にプログラムの処理の流れを示す．また，この流れを図で表したものを図 3 に示す．

① 差異情報を格納するための空の辞書 X を作成する．
② 厳島神社の記事の URL から，関数 I を用いてリスト A を作成する．
③ 厳島神社の記事の URL から，関数 II を用いて辞書 B を作成する．
④ 辞書 B 内の英語版とフランス語版の記事の URL から，関数 I を用いてリスト C を作成する．
⑤ リスト C 内のすべての記事の URL から，関数 II を用いて辞書 D を作成する．
⑥ 辞書 D 内の日本版の記事の URL がリスト A 内に存在しなければ，その記事のタイトルをキー，URL を値として辞書 X に格納する．
⑦ 辞書 X を出力する．

図 3：差異情報抽出の流れ



4.3 結果と分析

英語版とフランス語版の記事に含まれるリンク付き文字列の数，差異情報の数，およびリンク付き文字列の数に対する差異情報の数の割合を表 4 に示す．

表 4

	文字列	差異情報	割合（%）
英語	324	189	58.3
フランス語	139	53	38.1

この結果から，英語版ではおよそ 6 割，フランス語版ではおよそ 4 割の内容が日本語版と異なっていることになり，やはり言語によって説明の仕方に違いがあることが分かる．

次に，言語ごとに差異情報を見てみると，英語版では「神社神道」や「日本の仏教」など宗教に関する単語が特に多く，フランス語版では歴史のほか，「世界遺産の一覧」や「フランス国立図書館」など日本以外に関する単語が特に多かった．

また，抽出した画像の数を比較すると，日本語版は 43 枚，英語版は 32 枚，フランス語版は 6 枚であった．画像の被写体を比較すると，日本語版は境内の建造物が多いのに対し，

英語版では日本画，フランス語版では大鳥居の画像が多く，言語ごとに特徴が見られた．

5　まとめ

本研究では，多言語 Wikipedia からデータを取得し，BERT を用いてテキストデータから観光特性を抽出する手法について提案した．リンク情報と画像情報については Wikipedia のページから取得したが，ダンプデータから取得すれば実験の再現性を高めることができる．また，画像については目視による分析を行ったが，今後は生成 AI を用いて画像にキャプションを付け，文章を比較すれば，さらに特性を読み取ることができるだろう．

6　参考文献

[1] 国土交通省観光庁, “訪日外国人の消費動向 2023 年 年次報告書”, https://www.mlit.go.jp/kankocho/tokei_hakusyo/gaikokujinshohidoko.html（参照 2024-06-27）