

旅行ブログ記事からの名所とその付随情報の抽出

中嶋 勇人[†] 太田 学[†]

[†] 岡山大学大学院自然科学研究科 〒700-8530 岡山県岡山市北区津島中 3-1-1

E-mail: †{nakajima,ohta}@de.cs.okayama-u.ac.jp

あらまし 旅行の体験が綴られたブログ記事には、どのような場所で、どのような体験をしたかや、その場の状況、様子から得た感想等が書かれていることが多い。これらの情報は、その地を訪れる者にとって、非常に参考になる情報である。本研究では、場所の移動を表す記述にみられる移動表現を用いて、観光地の旅行ブログ記事を収集し、その地における名所を抽出した。さらに、書き手の体験情報、評価表現、書き手がその場を訪れることで得られる状態情報、名所の由来や歴史などの記述箇所の四つを名所付随情報として抽出する手法を提案する。これらの付随情報は、それぞれの記述文の品詞の特徴から抽出ルールを作成し、抽出する。評価実験では、名所の抽出精度で適合率 0.84、付随情報の抽出で適合率 0.65 を確認した。

キーワード 情報抽出, 観光情報

Extraction of Torist Spots and Their Related Information from Travel Blog Entries

Yuto NAKAJIMA[†] and Manabu OHTA[†]

[†] Graduate School of Natural Science and Technology, Okayama University

3-1-1 Tsushima-naka, Kita-ku, Okayama-shi, Okayama 700-8530 Japan

E-mail: †{nakajima,ohta}@de.cs.okayama-u.ac.jp

Key words Information Extraction, Tourist Information

1. はじめに

近年、ブログや SNS などのサービスを用いて、多くの人が日々の出来事や体験を日記の形で発信している。その中には旅行体験に関して書かれたブログ記事も多数存在する。観光情報を主に扱っているブログとしては「TravelBlog」, 「フォートラベル」, 「旅行・観光ブログ村」などが挙げられる。旅行の体験を綴ったブログ記事では、いつ、どこへ行き、どのような名所を観光し、そこはどのような場所で、その場でどのような体験をしたのかなどの情報が書かれていることが多い。これらの情報は、その地を訪れる予定の人にとって、非常に有用な情報となり得る。Web で名所やお土産などの観光情報を探す人にとっては、旅行会社の観光ポータルサイトや各自治体の観光情報サイトなどと同様に、ブログは観光情報に関する重要な情報源の一つである。しかし、Web 上には無数のブログ記事が存在するため、旅行ブログに特化した効率的な情報収集手段が必要であると考えられる。

本研究では、指定された観光地に関する旅行ブログ記事を自動収集し、そこからその地における名所とその名所の付随情

報を抽出する手法を提案する。本研究で扱う付随情報とは、「食べた」、「見た」などの書き手の行動を表す体験情報、「良かった」、「綺麗だった」などの評価表現に基づく書き手が得た感想、「～されていた」や「～があった」など書き手が名所を訪れたことで得られた状態情報、また「〇〇は江戸時代に建設された」などの名所に関する歴史や補足情報を解説した説明記述とした。これらの情報は、名所を理解する上で有用な情報である。

本研究では、旅行ブログ記事の収集と名所の抽出には、場所を表す格助詞と動詞の係り受けにより定義した移動表現を用いた。さらに、付随情報それぞれの記述にみられる品詞の特徴から、抽出ルールを作成し、名所ごとの付随情報を抽出した。

本稿の構成は以下のとおりである。2. で観光情報や体験情報の抽出に関する研究について、3. で本稿で扱う移動表現の定義と旅行ブログ記事の収集について述べる。4. で収集した記事からの名所の抽出と各名所の説明の記述箇所を特定する手法について、5. で名所ごとの付随情報の抽出について述べる。6. では旅行ブログ記事の収集と名所の抽出、付随情報の抽出の評価実験について述べ、7. でまとめる。

2. 関連研究

本節では、ブログなどの Web コンテンツから観光情報や体験情報を抽出する研究を紹介する。

2.1 観光情報の抽出に関する研究

近年、Web における観光情報提供、およびその分析に関する様々な研究やシステム開発などの取り組みが行われている [1]。郡ら [2] は、ブログからユーザの行動時の代表的な経路とその文脈を抽出し、それらを地図上にマッピングして提示するシステムを提案している。小田原ら [3] は、機械学習を用いて、旅行ブログ記事から旅行者の行動経路、またマイクロブログ Twitter から災害時の被災者の避難経路を抽出している。行動経路は、実際にその地を訪れた人々がどのような順序で、名所や観光スポットなどを周ったのかがわかり、旅行の計画時には非常に参考になる情報である。一方、本研究では、移動や観光地での行動に関する語を手がかりとして用いることで、旅行ブログ記事の本文中で、それぞれの名所に関する記述箇所を特定している。そこから体験情報などの名所付随情報を抽出することで、効率的に名所に対する情報収集を行う。

石野ら [4] は、機械学習を用いて、旅行ブログ記事から自動的に、土産情報と観光名所情報を抽出する手法を提案している。さらに、旅行ブログ記事からリンクを抽出することで、観光情報リンク集を構築している。しかし、彼女らの研究では「日本の草津温泉」という文から「日本」という観光地、「草津温泉」という名所を抽出し、このように観光地とその名所が存在する地域名の範囲が適切でないことがあるという問題があった。本研究では、観光地を指定し、その地名と名所の共起度を、Web の検索結果を用いて計算することで、観光地における適切な名所を抽出する。また、各名所における付随情報も抽出する。

寺西ら [5] は、観光ガイドブックの各ページをカテゴリ分類する手法と、観光ガイドブックと観光コンテンツを対応付ける手法を提案している。各ページをそれぞれ「見る」、「体験する」、「買う」、「食べる」、「泊まる」、「その他」に分けられる。実験により、ページのタイプ分類において精度 74.1 %、再現率 42.2 % を示している。

2.2 体験情報の抽出に関する研究

池田ら [6] は、ユーザの体験記述に現れる特徴を体験表現として定義し、体験が記述された体験情報を提供する体験情報抽出システムを提案している。倉島 [7] は、ある場所について書かれた個々のテキストから、人々の体験を、時間・空間・動作・対象属性間の相関ルールマイニングによって抽出している。これは「京都駅でお土産と切符を買う」という文から、空間は「京都駅」、対象は「おみやげ」と「切符」、動作は「買う」を抽出するものである。しかし、ブログ記事にはこの例文のように場所と体験が同時に出現する文ばかりではない。本研究では、旅行ブログ記事の本文から、名所について記述された箇所を特定し、名所が出現しない文からも体験情報を含む付随情報を抽出する。

3. 旅行ブログ記事の収集

3.1 概要

観光地の観光情報を収集するためには、その旅行体験が綴られたブログ記事が有用である。ここではそのような旅行ブログ記事の収集について述べる。

本研究では、Yahoo! ブログ検索 API^(注1) を利用して、旅行ブログ記事を収集する。具体的には、観光地名（以下、「検索地名」と、「旅行」または「観光」の2語の AND 検索を行い、検索結果から旅行ブログ記事を収集する。「札幌」に関する旅行ブログ記事を得たい場合、検索クエリは「札幌 旅行」と「札幌 観光」となる。

しかし、検索結果の中には旅行の予定を綴ったブログ記事や、「検索地名」とは別の観光地の記事なども混在している。そのため、検索されたブログ記事が「検索地名」に訪れて書かれた旅行ブログ記事かどうかを判定して収集する。

3.2 検索地名を訪れて書かれた記事かどうかの判定

判定には、ブログ記事検索の結果から得たタイトルとサマリを用いる。その際、サマリでの判定には、3.2.1 節で定義する移動表現を用いる。

本研究では、地名が現れる文を係り受け解析することにより、その地名を訪れたのか、あるいはその地名から離れたのかどうかを判定した。なお、係り受け解析器については、CaboCha^(注2) を利用した。

郡ら [2] は、地名が現れる文に対して係り受け解析を行うことで、書き手がその場に訪れたかどうかの判定を行っている。場所を表しうる格助詞が地名と同じ文節に現れ、なおかつその地名が動作動詞かサ変名詞に係っている場合、書き手がその場を訪れたと判定している。場所を表しうる格助詞とは、「から」、「へ」、「まで」、「を」、「に」、「で」、「より」、「において」、「に対して」、「にたいして」である。本研究では、郡らの手法を参考にし、移動表現を定義した。

3.2.1 移動表現

本研究では、場所を表しうる「格助詞」と移動や名所での行動を表す際に用いられる動詞の組み合わせを移動表現として定義する。

場所を表しうる格助詞は、それが用いられる際の文脈上の意味から、表 1 に示す 4 種類に分類した。

また、動詞は、表 2 に示す動作動詞、サ変名詞とした。サ変名詞は、名詞+「する」の形で構成され、動詞化する名詞の総称である。これらの動作動詞とサ変名詞に関しては、「到着を表す語」、「通過を表す語」、「出発を表す語」、「名所での行動に関する語」に分類した。表 2 の各動作動詞や各サ変名詞については、旅行ブログ記事の本文から人手で収集した。

移動表現は、場所を表しうる格助詞が出現する文節の係り先が、表 2 の「到着を表す語」または「観光地での行動を表す語」

(注1) : Yahoo!デベロッパーネットワーク

<http://developer.yahoo.co.jp/>

(注2) : CaboCha

<http://code.google.com/p/cabocha/>

表 1 場所を表しうる格助詞

意味	格助詞
移動の目標や到達点を表す	に, へ, まで
動作の直接的な対象や移動時の経路を表す	を
移動の起点や経由点を表す	から, より
動作の手段や起こる場所を表す	で, において, にたいして, に対して

表 2 移動を表す動作動詞とサ変名詞の一覧

種類	例
1 到着を表す語	行く, 着く, 到着, 至る, 渡る, やって来る, 戻る, 寄り道, 移動 直行, 立ち寄る, 来る, 訪れる, 帰る, 持ち帰る, 向かう
2 通過を表す語	通る, 通過, 過ぎる, 抜ける
3 出発を表す語	出発, 出る, 出掛ける, 離れる, スタート
4 観光地での行動を表す語	旅行, 観光, 散策, 散歩, 見学, 過ごす, 参拝, 拝観, 入館, 入場, 参内, 入城, 訪問

のいずれかである文を「到着表現」, 「通過を表す語」のいずれかである文を「通過表現」, 「出発を表す語」のいずれかである文を「出発表現」とした。

3.2.2 判定手法

検索地名を訪れて書かれた記事かどうかの判定には, ブログ記事検索の結果から得たタイトルとサマリを用いた。まず, タイトルに基づく判定を行い, そこで判定がつかなかった場合はサマリで判定する。

(1) タイトルに基づく判定

まず, ブログ記事のタイトルに「検索地名」と, 「旅」, 「観光」, 「巡り」, 「めぐり」, 「日目」といった旅行ブログ記事のタイトルに頻繁に現れる語が同時に出現した場合, その地を訪れた旅行ブログ記事と判定し, 収集する。例えば, 「検索地名」を京都とした場合は, 「夏の京都の旅行」や「京都一日目」といったタイトルのブログ記事を収集する。

(2) サマリに基づく判定

サマリの係り受け解析結果と「到着表現」を用いて, 「検索地名」の旅行ブログ記事かどうかを判定する。

具体的には, サマリの文の中に, 「検索地名」が現れ, 同じ文節に表 1 の「移動の目標や到達点を表す格助詞」または「動作の直接的な対象や移動時の経路を表す格助詞」のいずれかが出現する文を係り受け解析する。「検索地名」の現れた文節の係り先が, 表 2 の「到着を表す語」または「観光地での行動を表す語」に係っている場合, その地を訪れたと判定する。例えば, 「先日, 奈良に行った。」という文では, 「奈良」の後に「移動の目標や到達点を表す格助詞」の「に」が続き, この文節の係り先は「行った」となっている。「行っ」は, 「到着を表す語」の「行く」の連用形であり, これは「奈良」に移動した文であると判定される。

また, 「移動の目標や到達点を表す格助詞」は, 「東大寺

へ。」や「正倉院に。」のように動詞が省略されて移動を表す場合がある。そのため, この三語の格助詞に関しては, 「検索地名」にこの格助詞がつき, なおかつその格助詞で文が終了した場合は, 「検索地名」を訪れたと判定する。

なお以下の場合はその地を訪れていないと判定する。

- 動詞の活用形が未然形である場合
(例: 行かずに, 寄らずに, …)
- 動詞の後に希望を表す助動詞が続く場合
(例: 行きたかった, 寄りたい, …)
- 動詞の後に伝聞, 推定, 当然を表す助動詞が続く場合
(例: 行ったらしい, 寄るべきだ, …)

ここでの動詞とは, 動作動詞はその語そのものであり, サ変名詞に関してはその後続く「する」という動詞を指す。「金閣寺には行かずに」などの否定が含まれる文や「京都に行きたい」などの希望を表す文に対応した。

4. 名所の抽出と記述箇所の特定

3. 節で収集した「検索地名」の旅行ブログ記事から, その地における名所を抽出する。3. 節と同様に, 「到着表現」を用い, 場所を表す名詞を収集し, それらを名所候補とした。

また, 旅行ブログ記事で, 名所についての記述がなされている箇所を特定する。

4.1 名所の抽出

まず, 旅行ブログ記事本文中で「移動の目標や到達点を表す格助詞」または「動作の直接的な対象や移動時の経路を表す格助詞」のいずれかが出現する文を係り受け解析し, この格助詞が含まれる文節の係り先が, 表 2 の「到着を表す語」または「観光地での行動を表す語」かどうかを調べる。係っていた場合は, 「移動の目標や到達点を表す格助詞」または「動作の直接的な対象や移動時の経路を表す格助詞」の前の名詞を抽出する。そのような名詞の前にさらに名詞が連続する場合は, 連結して抽出し, それを「検索地名」の名所候補とする。

CaboCha では形態素解析器として MeCab を用いており, 「金閣寺」や「清水寺」などの有名な名所に関しては「B-ORGANIZATION」というタグを付ける。このタグがついた名詞は, その前に名詞が連続していた場合も名詞の連結を行わず, その一語で抽出する。

また以下に挙げる名詞は, 名所候補として抽出しない。

- 各都道府県, 国名 (例: 北海道, 東京都…)
- 交通施設, 宿泊施設 (例: 東京駅, ホテル, …)
- 日付と時刻 (例: 十日, 三時, …)
- 代名詞 (例: そこ, ここ, …)
- 一単語のみの一般名詞 (例: 頂上, 入口, …)
- 副詞可能名詞 (例: 午前中, 昨日, …)

このようにして抽出した名所候補が全て「検索地名」の名所であるとは限らない。そこで, 「検索地名」と各名所候補の共起の強さを Simpson 係数を用いて求めた。「検索地名」x を含む

Web ページ集合を X , 名所候補 y を含む Web ページ集合を Y として, x, y の単独の検索ヒット数を $|X|, |Y|$ とし, x と y の AND 検索の検索ヒット数を $|X \cap Y|$ とする. このとき, Simpson 係数 $S(x, y)$ は次式で計算される.

$$S(x, y) = \frac{|X \cap Y|}{\min(|X|, |Y|)} \tag{1}$$

一般的には, Simpson 係数が 1 に近いほど, 2 語の相関が強いとされる. しかし, Simpson 係数は, 比較される一方の語の検索結果が, もう一方の語と比べて極体に少ない場合には, 関係性がさほど強くない語同士でも高い値が出てしまう. そのため, x, y それぞれの単独の検索ヒット数に 1000 件以上という制限を設けた. 検索結果数が 1000 件未満の名所候補は, 名所としない. 予備実験において, 「小豆島」, 「倉敷」, 「沖縄」をそれぞれ「検索地名」とし, 旅行ブログ記事を収集した. 収集した旅行ブログ記事から名所を抽出した結果, いずれも Simpson 係数の値が 0 から 0.45 未満の間では「検索地名」の名所ではない名詞が多く含まれていた. この結果より, Simpson 係数の閾値を 0.45 に設定し, 閾値以上の名所候補を「検索地名」の名所とした.

4.2 名所の説明記述箇所の特定

旅行ブログ記事の多くは, 本文中で旅行時に訪れた複数の名所についての記述がなされている. 各名所の付随情報を抽出するためには, それらの名所について記述された箇所の特定が必要である.

旅行ブログ記事では, 時系列に書き手が訪れた名所について記述されることが多いため, 本研究では, 「名所 A への到着表現」または「名所 A の通過表現」が現れる文から「別の場所 B への到着表現または出発表現」または「名所 A からの出発表現」が現れるまでの文を名所 A に関する説明記述として特定する. ここでの場所 B は, 名所だけでなく, 都道府県名などの地名, 駅や空港などの交通施設, ホテルなどの宿泊施設も含む.

以下の例を用いて説明する.

- 1

まずは, 清水寺へ行きました.
- 2

やはり京都といえば, ここですよ.
- 3

修学旅行生が多いのでかなり賑やかでした.
- 4

清水の舞台からこの時期鮮やかな紅葉を見て,
四季の素晴らしさを感じました.
- 5

次は, 銀閣寺へと向かいました.

この例では, 文 1 で「清水寺」への「到着表現」があり, 文 2-4 では「清水寺」での体験や感想等が記述されている. 文 5 では「銀閣寺」への「出発表現」の記述があるため, 文 2-4 を「清水寺」に関する説明記述とした.

また, 移動表現の他に, 旅行ブログ記事では名所を箇条書きにして, 名所に関して記述するものも見られる. そのため, 文が名所の表現止めで終わっており, なおかつ名所と句読点を除く文字数が 1 文字の文は, 箇条書きと判定した. 例は, 「・太秦映画村」や「□銀閣寺」といった文である. 名所 A の箇条書きから場所 B の箇条書きまでの間の文を名所 A に関する記述とした.

表 3 体験情報の抽出パターン

抽出ルール	例
動詞 (連用形) + 「た/ました」	「登った」, 「食べました」
動詞 (基本形) + 文末文字	「登る。」, 「食べる！」
動詞 (未然形) + 動詞「す」 (未然形) + 接尾動詞「れる」 (連用形) + 「た/ました」	「並ばされた」, 「濡らされた」
サ変名詞+動詞「する」 (連用形) + 「た/ました」	「散策した」, 「体験した」
サ変名詞+動詞「する」 (基本形) + 文末文字	「散策する。」, 「体験する。」
動詞 (連用形) + 接続助詞「て」 + 「みた/みました」	「してみた」, 「やってみた」
サ変名詞+動詞「できる」 (連用形) + 「た/ました」	「拝観できた」, 「休憩できた」

表 4 評価情報の抽出パターン

抽出ルール	例
形容詞 (連用形) + 「た」	「良かった」, 「楽しかった」
形容詞 (基本形) + 文末文字	「美味しい。」, 「良い。」
形容動詞+ 「だ/だった/です/でした」	「綺麗だった」, 「変でした」

5. 名所付随情報の抽出

この節では, 4.2 節で特定した名所の記述から, 名所の付随情報を抽出する手法について述べる.

5.1 抽出する文

本研究では, 付随情報を以下の四つのいずれかの表現が出現する文であると定義している.

- 書き手の体験情報
- 書き手の評価表現に基づく感想
- 書き手が名所で得た客観的情報
- 名所に関する説明記述

「書き手の体験情報」とは, 「～した」や「～できた」といった自発的な体験情報と, 「～させられた」のような強制的な体験情報の二つとした.

「書き手の評価表現に基づく感想」とは, 「楽しかった」, 「おもしろかった」のように形容詞, 形容動詞などの評価表現を用いられたものとした.

「書き手が名所で得た状態情報」とは, 「～していた」, 「～されていた」などの書き手が実際にその名所を訪ねることで得た名所の状態情報を示す文とした.

名所に関する説明記述とは, 「○○は世界遺産に登録されている」や「○○は江戸時代に建設された最古の～である」といった, 名所に関する歴史や補足情報を記述した文とした.

5.2 付随情報の抽出

各付随情報として抽出する文の抽出パターンを表 3-6 に示す. それぞれ文における品詞の特徴から抽出パターンを作成した. 表中の丸括弧で囲まれた部分は品詞の活用形であり, 連用形は「美味しかった」の「美味しかっ」のような連用タ接続も含む. 文末文字とは, 句点の「。」と「.」, その他に「！」と「♪」と「(」などの文の最後に頻繁に現れる記号を指す.

旅行ブログ記事の中には, 今その場で書き手が体験しているかのように, 動詞の基本形で書かれたブログ記事も見られるため, 文末に動詞の基本形が出現する文は, 付随情報と判定する.

表 3-6 の抽出パターンに当てはまる文を, 名所の付随情報として抽出した.

表 5 状態情報の抽出パターン

抽出ルール	例
{動詞 (連用形)/動詞「する」(連用形) /動詞「する」(未然形)+動詞「れる」(連用形)} +接続助詞「て」(+非自立動詞「い」) +「る/た/ます/ました」	「やっていた」, 「開催されていた」
動詞「ある」(基本形)+文末文字 /動詞「ある」(連用形)+「た/ます/ました」	「案内板があった」 「絵馬が飾ってあった」

表 6 説明記述の抽出パターン

抽出ルール	例
名所+係助詞「は」+…+動詞「する」(未然形) +「た/ました」	「金閣寺は 1955 年に再建された」
名所+係助詞「は」+…+名詞+「です/でした /である/だった」	「金閣寺は世界 遺産です」

表 7 旅行ブログ記事の収集の精度

検索地名	適合率	再現率
京都	0.750	0.450
奈良	0.785	0.579
ローマ	0.746	0.545
パリ	0.750	0.571
計	0.746	0.537

6. 評価実験

旅行ブログ記事の収集の精度、名所の抽出と各名所の付随情報の抽出精度を評価した。使用した観光地名は、国内では「京都」と「奈良」で、国外では世界的に有名な観光地であるイタリアの「ローマ」とフランスの「パリ」とした。

6.1 旅行ブログ記事の収集実験

3. 節で述べた「検索地名」に関する旅行ブログ記事の収集の精度について評価する。それぞれの観光地について、Yahoo!ブログ検索 API で検索した 2012 年 12 月初頭時点の最新 100 件のブログ記事を実験データとした。検索結果を手手で「検索地名」に関する旅行ブログ記事かどうか判定したものを正解データとして用いて、収集した旅行ブログ記事の適合率、再現率を求めた。適合率は、収集した旅行ブログ記事の内、人手で「検索地名」の旅行ブログ記事であると判定したブログ記事の割合であり、再現率は、人手で「検索地名」の旅行ブログ記事であると判定した内、収集した旅行ブログ記事の割合である。

実験結果を表 7 に示す。適合率に比べ再現率の低さが目立つが、ここでの適合率の精度が、6.2 節以降の実験にも影響を与えるため、適合率が高いことは重要である。収集する旅行ブログ記事数の確保は、検索件数を増やすことで対応する。

以下に提案手法で誤って正解判定した例と、「検索地名」の旅行ブログ記事であるが、そう判定できなかった例を示す。

- ・ サマリに「検索地名」を訪れたという記述がない
- ・ 「検索地名」とは別の観光地の旅行ブログ記事
- ・ その他

ブログ記事本文には「検索地名」を訪れたという記述が現れるが、サマリには現れないブログ記事を「検索地名」の旅行ブログ記事として収集できなかった。

複数にわたる旅行ブログ記事で、「検索地名」と別の観光地をはしごして観光しているものを誤って、「検索地名」の旅行ブログ記事として収集した。例は、「京都」に関する旅行ブログ記事を収集時に、タイトルで「京都・奈良観光日記一日目」から「京都・奈良観光日記三日目」といったタイトルに複数の地名が現れる一連のブログ記事を全て「京都」に関するブログ記事として収集してしまった。前半の記事は、「京都」に関するものであったが、後半は「奈良」に関するものであった。これは、ブログ記事本文で改めて訪れた場所に関して判定することで解決できると考える。

その他の例としては、サマリで「検索地名」に訪れた旨の記述はされているが、「詳細は後日書きます」などの「検索地名」に訪れた報告だけを記述されたブログ記事を収集していた。これはブログ記事本文のテキスト量に制限を設けることで改善できると考える。また、旅行からの帰宅後、現地で買ったお土産に関するブログ記事を誤って収集した。また、旅行時の写真だけを掲載しているブログ記事なども情報を取得できないため不正解とした。

6.2 名所の抽出実験

収集した「京都」、「奈良」、「ローマ」、「パリ」の旅行ブログ記事から名所を抽出し、評価した。抽出された名所を手手で、「検索地名」の名所かどうか判断した。実験データは各検索地名について 500 件収集したブログ記事の中から文書量の多いものを各 50 件ずつ使用した。この 200 件のブログ記事からはあらかじめ人手により、各検索地名の名所と著者が判断した名詞を収集し、それを正解データとして使用して再現率、適合率、F 値を求めた。再現率は、人手で収集した名所に含まれる抽出した名所の割合であり、適合率は、抽出した名所に含まれる人手で収集した名所の割合である。

実験結果を表 8 に示す。どの検索地名においても適合率は 0.83 以上となった。「京都」と「パリ」について抽出した名所の一覧を表 9 に示す。

誤って抽出した名詞としては、一語では特定の名所を示さない語があった。例えば、「京都」では「パワースポット」や「美術館」、「パリ」では「ビストロ」などである。また、「検索地名」と関連の深い観光地の名所は、Simpson 係数が高くなり、誤って「検索地名」の名所として抽出した。例えば、「京都」の名所として「東大寺」を抽出したが、「東大寺」は「奈良」の観光地である。これは、「京都」と「奈良」が地理的に近く、二つが密接に関係した観光地であるからと考える。

また名所を抽出できなかった原因として、使用した係り受け解析器 Cabocha の仕様がかった。Cabocha では形態素解析器として MeCab を使用しているが、MeCab では「京都」の観光地である「嵐山」のような一部の語を形態素解析すると、「名詞、一般」とする。これは、4.1 節で説明したように、一語のみの「名詞、一般」は名所候補として抽出しないため、名所として抽出できなかった。また、定義した移動表現とともに書かれていない名所は抽出できなかった。これに対応するには、移動表現の拡充等が必要である。

表 8 名所の抽出の結果

検索地名	抽出した名所数	正解数	適合率	再現率
京都	31	26	0.839	0.742
奈良	30	25	0.833	0.781
ローマ	19	16	0.842	0.727
パリ	24	20	0.837	0.833

表 9 抽出した名所

	抽出した名所
京都	鳥丸御池, 清水寺, 二条城, 京都水族館, 京都市美術館, 京都御所, 京都宇治, 京都タワー, 府立植物園, 伏見稲荷, 東福寺, 二年坂, 大河内山荘, 錦市場, 東大寺, 美術館, 京都競馬場, 下鴨神社, ゴールデンウィーク, 北野天満宮, 平安神宮, 龍安寺, 広龍寺, 詩仙寺, 三十三間堂, 仁和寺, 水路閣, 貴船神社, パワースポット, 金閣寺, 銀閣寺
パリ	リスボン, マレ地区, サクレクール寺院, セーヌ河沿い, コンコルド広場, シャンゼリゼ通り, オランジュリー美術館, モンマルトル, オルセー美術館, 展覧会, ヴェルサイユ宮殿, ノートルダム大聖堂, ルーヴル美術館, チュイルリー公園, サント・シャペル, コンシェルジュリー, パンテオン, バスティュー広場, マドレーヌ教会, 中心部, 旧パリ市庁舎 オペラ座, ビストロ, シテ島

6.3 付随情報の抽出実験

6.2 節と同じ実験データを用い,「検索地名」の抽出した各名所における付随情報を抽出し,評価した。ここでは,付随情報が過不足なく抽出できているかどうかと,正確に名所に対応した付随情報を抽出できているかどうかを評価した。実験データに用いた旅行ブログ記事の本文から著者が人手で付随情報とその付随情報の対応する名所や場所を調べた。名所 A についての付随情報として,人手で名所 B の付随情報とした文を抽出した場合,それは誤りとした。適合率については,抽出した名所 A に関する付随情報のうち,人手で名所 A の付随情報と判定した文の割合であり,再現率については,人手で名所 A の付随情報と判定した文のうち,抽出した名所 A に関する付随情報の割合である。

比較対象として,名所の記述部分の特定を,名所 A が現れた文から次に別の場所の名詞が現れるまでの間の文を名所 A の記述箇所として,そこから付随情報を抽出する方法を評価した。

全ての検索地名に関する評価結果を表 10 に,「検索地名」ごとの評価結果を表 11 に示す。全体の評価結果として適合率 0.654, 再現率 0.746 が得られ,どちらにおいても比較手法を上回ることができた。

誤って抽出された文の多くは,以下の 3 つに大別される。

- 書き手の体験ではない
- 体験が旅行した時点のものより過去の場合
- 別の場所の付随情報である

例えば「修学旅行のときにした」など,体験表現で書かれているが,実際に体験したのが,旅行時より以前のものである情報を誤って判定していた。

また,再現率,適合率をとともに低下させた原因として,別の場所の付随情報を抽出してしまう場合が多くあった。これは名

表 10 付随情報の抽出結果

	適合率	再現率	F 値
提案手法	0.654	0.746	0.697
比較手法	0.343	0.382	0.361

表 11 検索地名別の付随情報の抽出結果

検索地名	適合率	再現率	F 値
京都	0.575	0.813	0.674
奈良	0.625	0.743	0.770
ローマ	0.813	0.732	0.673
パリ	0.636	0.715	0.679

表 12 各付随情報の抽出結果

付随情報	適合率	再現率	F 値
体験情報	0.656	0.786	0.715
評価情報	0.673	0.810	0.731
状態情報	0.620	0.829	0.710
名所説明記述	0.780	0.294	0.429

所の記述箇所の特定制が正確にできていないためである。話が別の場所に移ったにも関わらず,そう判定できなかったため,別の場所の付随情報を抽出してしまい,適合率の低下につながった。その原因として,「次は金閣寺。」や「そのまま,ホテル。」のように,場所を表す格助詞や移動を表す動詞もなく,場所の転換を表す表現があった。しかし,これらの多くには「次に」や「続いて」のように,場所を表す格助詞や動詞以外に場面の転換を示す語が同時に出現していたため,これらの語を移動表現に加えることで,改善できると考える。また,これにより「次に来たときは,絶対に挑戦する。」などといった願望を表した体験ではない文も排除することができると考える。他には,「城を後にして」の「後にする」のように,今回使用した動詞以外の移動を表す表現があった。これらを考慮することで,適合率,再現率の向上,また 6.2 節で抽出できなかった名所も抽出可能になると考える。

次に,付随情報種類別の抽出精度を表 12 に示す。これを見ると,名所説明記述についての再現率が他のものに比べて低いことがわかる。これは,名所に関する説明記述として用意した抽出パターンより多くの表現があったためである。「ここは」など代名詞を用いて記述する場合も多くあり,抽出法のさらなる改善が必要である。

抽出された文を見てみると,「これもおいしかった。」のように,一文では対象が何を指しているのかわからないような文も見られた。そのため,今後の課題として,体験や評価の対象の有無の判定についても考える必要がある。

7. おわりに

本研究では,場所を表す格助詞と動詞の係り受けにより定義した移動表現を用い,旅行ブログ記事を収集し,そのブログ記事から名所と付随情報を抽出する手法を提案した。評価実験では,名所の抽出実験では,適合率 0.84, 再現率 0.77 を示し,付随情報の抽出では,適合率 0.65, 再現率 0.75 を確認した。

今後の課題としては、移動表現の拡充による名所と付随情報の抽出性能の向上がある。また、抽出した付随情報には、名所限定の体験や、季節限定や時間帯限定でのみ得られる情報も多く見られた。こういった情報は、旅行計画を立てる人にとって有用な情報である。そのため、抽出した付随情報から名所ごとの有用な情報の発見と、それを提示するインタフェースの実装に取り組む予定である。

文 献

- [1] 齊藤一：Web における観光情報提供と分析，人工知能学会誌 26(3)，234-239(2011).
- [2] 郡宏志，服部峻，手塚太郎，田島敬史，田中克己：ブログからのビジターの代表的な経路とそのコンテキスト抽出，情報処理学会研究報告データベースシステム研究会，Vol.2006，No.78，pp.35-42(2006).
- [3] 小田原周平，石野亜耶，難波英嗣，竹澤寿幸：ブログからのユーザの行動経路の自動抽出と可視化，電子情報通信学会第 20 回 Web インテリジェンスとインタラクション研究会 (2011).
- [4] 石野亜耶，難波英嗣，竹澤寿幸：旅行ブログエントリからの観光情報の自動抽出，日本知能情報ファジィ学会誌，Vol.22，No.6，pp.667-679(2010).
- [5] 寺西拓也，野村達二，平山智子，石野亜耶，難波英嗣，竹澤寿幸：観光ガイドブックへの旅行ブログエントリと質問応答コンテンツの対応付け，言語処理学会第 18 回年次大会，発表論文集，333-336(2012).
- [6] 池田佳代，田邊勝義，奥田英範，奥雅博：Blog からの体験情報抽出，情報処理学会論文誌，Vol.49，No.2(2008).
- [7] 倉島健，手塚太郎，田中克己：街 Blog からの体験抽出とその空間的提示手法の提案，情報処理学会研究報告．データベース・システム研究会報告，2005(67)，47-53(2005).
- [8] 西原陽子，佐藤圭太，砂山渉：出来事の画像表現によるブログからの体験談獲得，知能と情報 (日本知能情報ファジィ学会誌)Vol20，No.5(2008).
- [9] 奥村秀人，徳久雅人，村上仁一，村田真樹：観光地に対する長所情報の収集と分類の試み，電子情報通信学会技術研究報告．TL，思考と言語 110(244)，25-30(2010).
- [10] 倉島健，藤村考，奥田英範：大規模テキストからの経験マイニング，電子情報通信学会論文誌．D，情報・システム J92-D(3)，301-310(2009).
- [11] 佐々木健，小山聡，栗原正仁：実体験情報を含む Blog を抽出するシステム，情報処理学会創立 50 周年記念 (第 72 回) 全国大会，"2-497"- "2-498"(2010).
- [12] 高野太希，井上潮：文書構造に基づいた Blog からの体験情報抽出方法の提案，DEIM Forum 2011 A4-2(2011).
- [13] 岡田幸彦：現代日本語の移動動詞と場所名詞の格，日本アジア研究 6，39-61(2009).
- [14] Fredric C.Gey, Ray R.Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. GeoCLEF:The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview.Lecture Notes in Computer Science, LNCS4022, pp.908-919(2005)