多言語旅行ブログエントリを用いた観光情報提示システム

藤井 一輝† 石野 亜耶† 藤原 泰士† 前田 剛†

難波 英嗣† 竹澤 寿幸†

†広島市立大学大学院情報科学研究科 〒731-3194 広島県広島市安佐南区大塚東 3-4-1

E-mail: †{fujii, ishino, fujiwara, maeda, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

あらまし 旅行ブログエントリは，ブロガーの実体験に基づいて記載されているため，旅行情報として有益な情報源と考えられる．旅行ブログエントリを地図上にマッピングすることにより，観光地周辺の知識を持たない旅行者でも容易に情報を取得できる．しかし，旅行ブログエントリは数多く存在するため，旅行者が必要とする情報にアクセスするのが容易ではない．そこで，本研究では，日本語と英語で記述された旅行ブログエントリを対象に，これらを「買う」,「食べる」,「体験する」,「泊まる」,「見る」,「その他」の 6 種類のタイプへ自動的に分類する手法を提案する．エントリをタイプごとに地図上に提示することにより，旅行者の目的に沿った旅行ブログエントリへ容易にアクセスすることができる. 本研究で提案した手法により, 日本語の旅行ブログエントリの場合, 精度 0.637, 再現率 0.462，英語で記述の旅行ブログエントリの場合，精度 0.597，再現率 0.337 の分類精度が得られた.

キーワード 旅行ブログエントリ，文書分類，地図，多言語

1. はじめに

インターネット上には，多くの観光情報が存在しており，その 1 つとして旅行ブログエントリがある．旅行ブログエントリは，ブロガーが旅行先で実際に体験した体験談や感想などが記載されており，国を問わず多数存在している．そのため，その様な情報は旅行計画を行う際，有益な情報源と考えられる．

そこで本研究では，日本語または英語で記述された旅行ブログエントリを対象に，観光情報提示システムの構築を目指す．このシステムでは，旅行ブログエントリを地図上に提示する．視覚的にそれらの情報を提示することで，観光地の知識を持たない旅行者でも，観光スポットやその周辺の情報を入手することが可能となる．しかし，地図上の旅行ブログエントリを閲覧すると，宿泊施設や見るだけで楽しめるような観光スポット，美味しいランチのお店についての旅行ブログエントリが全て混在して提示されてしまう. このため，旅行者がランチのお店について情報を得たい場合，多くの旅行ブログエントリの中から情報を見つけ出す必要がある．そこで，旅行者の目的に沿って旅行ブログエントリを提示する．旅行者が観光スポットについての情報を得たい場合はタイプ「見る」, 飲食店についての情報を得たい場合はタイプ「食べる」というカテゴリに分類された旅行ブログエントリを提示することにより，必要とされる情報を効率的に提示するシステムを構築する．

本研究では，旅行ブログエントリを観光の主な目的となる「買う」,「食べる」,「体験する」,「泊まる」,「見る」の 5 種類のタイプへ自動的に分類する手法を提案する．旅行ブログエントリをタイプごとに地図上へ提示することにより，旅行者は目的に沿った旅行ブログエントリへ容易にアクセスすることができる．

本研究の構成は以下の通りである．2 節ではシステムの概要・動作例について述べ，3 節では関連研究を紹介する. 4 節では旅行ブログエントリのタイプ分類, 5 節では実験，6 節では考察について述べる．7 節では本稿のまとめについて述べる．

2. システムの概要・動作例

本節では，本研究で構築するシステムの概要と，その動作例について説明する．はじめに，システムの概要について述べる．地図上へ旅行ブログエントリをマッピングすることにより，観光スポットの場所を明確にする．そして，旅行者が必要とする情報へのアクセスを容易にするため，6 種類のタイプごとに旅行ブログエントリを提示するシステムである．

本研究で構築するシステムは，スマートフォンなどのタブレット端末での閲覧を想定している．図 1 は，本研究の提案手法を用いて構築した「ぶらり広島電停散歩 MAP[1]」である. 現時点における本システムでは，広島電鉄の各電停に焦点を当てており，日本語で記載された旅行ブログエントリのみを対象にしている．図 1 は，広島県の宮島周辺の地図が表示され，右側に観光の主な目的となる 6 種類のタイプが表示されている. 旅行者が牡蠣の美味しいお店について知りたい場合は「グルメ」ボタンをクリックする．また，鳥居や花火大会についての情報が得たい場合は「見る」ボタンを

[1] http://p2walker.jp/peace/ja/blog/ は広島 P2 ウォーカー運営協議会と共同で開発したシステムである.

図 1: ぶらり広島電停散歩 MAP の動作例

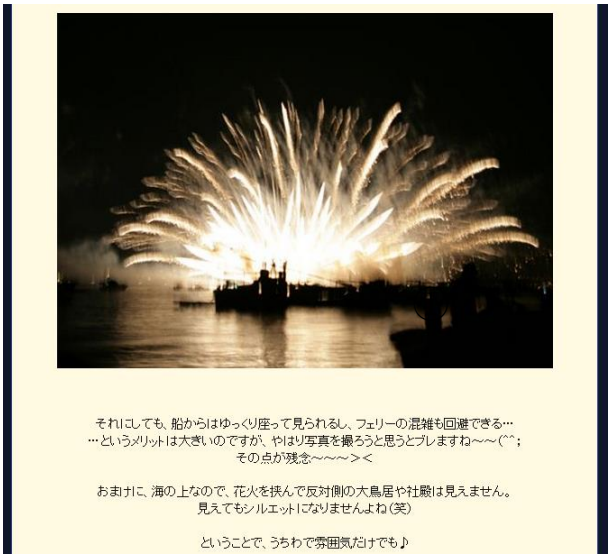

図 2: 提示された旅行ブログエントリの例

クリックする．図 1 は，「見る」ボタン(図中①)をクリックした場合の例であり，見るだけで楽しめるような場所にピンが表示される．ピン(図中②)をクリックするとそのピンの場所について記載された「見る」に関する旅行ブログエントリのタイトルが図中③に表示される．そして，そのタイトルをクリックすると，旅行ブログエントリを閲覧することができる．図 2 は，提示された旅行ブログエントリであり，宮島水中花火大会について書かれている．記事の内容は，船から花火を観覧すると混雑に巻き込まれないが，大鳥居は見ることができないと書かれている．以上のように，本研究で構築したシステムでは，旅行者の目的に沿って，ブロガーの実体験に基づいた多様な情報が提示される．

現在は，日本語で記載されている旅行ブログエントリのみを広島市に限定して提示している．また，広島で全国菓子大会博覧会が行われたため，「菓子・スイーツ」に関するブログ記事も閲覧できる仕様となっているが，本論文では，「菓子・スイーツ」の自動分類は行わない．今後は，英語の旅行ブログエントリを世界各国の地図上に提示するシステムの構築を想定している．

3. 関連研究

本研究では，旅行ブログエントリを自動的にタイプ分類し，地図上にマッピングし提示するシステムを提案する．本節では，本研究に関連する研究やサービスを紹介する．

本研究では，旅行のための有益な情報源として，ブロガーが日記形式で綴った旅行記である旅行ブログエントリに焦点を当てた．本研究では，地図上に旅行ブログエントリをマッピングするシステムの構築を行うが，まず，ブログデータベースから旅行ブログエントリをどのように収集するのかといった問題がある．日本語の旅行ブログやそのエントリを登録したポータルサイトとして，旅行・観光ブログ村[2]，フォートラベル[3], 英語の旅行ブログエントリを登録したポータルサ

[2] http://travel.blogmura.com/
[3] http://4travel.jp/

イトとして Travel Blog[4]などがある. これらのポータルサイトでは，ブロガーが自身のブログを旅行ブログとして登録することで，旅行ブログの集積を行う．しかし,ブログ空間にはたくさんのブログが存在するため，このようなポータルサイトに登録されていない一般ブログの中にも，旅行ブログエントリが多数存在する．ブログデータベースから旅行ブログエントリを自動的に収集する手法としては，Nanba ら[1]や Ishino ら[2]の手法がある．Nanba ら[1]は，一般ブログから，機械学習を使用して旅行ブログエントリを自動的に検出する手法を提案している. 機械学習の手法には CRF を採用し，精度 0.867 と高い精度で旅行ブログエントリを検出することに成功している．Ishino ら[2]は，広島県の特徴のひとつである，広島電鉄の電車（広電）を使用した観光を支援するための枠組みの一つとして，広電の電停に関する旅行ブログエントリを収集する手法を提案している．石野らは，0.824 と高い精度で広電の電停に関する旅行ブログエントリの検出に成功している．本研究では，Nanba らの手法により収集された日本語の旅行ブログエントリを利用する．英語の旅行ブログエントリは，十分な件数が得られたため，Travel Blog に登録されている旅行ブログエントリを利用する．

次に，文書分類に関する研究を紹介する．福田ら[3]は，ドメイン適応を用いて論文や特許のデータを対象に要素技術を抽出し，論文の解析を行っている．論文ドメインの素性を用いてモデルを獲得し解析を行った後，特許ドメインの素性を用いてモデルを獲得し，さらに解析を行うというドメイン適応手法を提案している．この手法によって，精度・再現率の向上に成功している．本研究では，ドメイン適応に日本語と英語を用いて，旅行ブログエントリを分類する．観光情報を分類する研究して，中嶋ら[4]の研究がある．中嶋らは Twitter を用いて観光ルートを推薦している．その際，ツイートを「食事」,「景観」,「行動」の 3 タイプに分類している．しかし，本研究では，旅行ブログエントリをタイプ分類する手法を提案している．

本研究と同様に，旅行ブログエントリを自動分類する研究がある．石野ら[5]は，旅行ブログエントリから収集したリンクのタイプ判定を行うことで，観光情報リンク集の構築を行っている．石野らは，リンクのタイプ分類を行うが，本研究では，旅行ブログエントリのタイプ分類を行う点で異なる．徳久ら[6]は，ブログエントリから，観光開発のためのヒントを抽出するために，ブログエントリ中の文に対し，ヒント文であるか，ヒント文でないのかを，自動で分類する手法を提案している．本研究では，旅行ブログエントリの 1 つの記事を，4.1 節で定義する観光に特化したタイプに分類する点で異なる．また，遠藤ら[7]は，ブログエントリを「見る・遊ぶ」,「イベント・祭り」「食べる・泊まる」,「お土産・特産品」「自然・文化」の 5 つタイプに自動的に分類する手法を提案している. 本研究では，英語で記述された旅行ブログエントリを用いている点で異なる.

[4] http://www.travelblog.org/

4. 旅行ブログエントリのタイプ分類

本節では，観光に関する情報が記載されている旅行ブログエントリを自動的に 6 種類のタイプへ分類する手法を説明する．4.1 節ではタイプ分類の判定基準，4.2 節では情報利得を用いた手掛かり語の収集について，4.3 節ではドメイン適応について説明する.

4.1. タイプ分類の判定基準

本節では，人手による旅行ブログエントリのタイプ判定について述べる．分類項目は,「買う」,「食べる」,「体験する」,「泊まる」,「見る」,「その他」の 6 種類である．なお，1 件の旅行ブログエントリに対して複数のタイプが付与される場合もある．各タイプの判定基準を表 1 に示す．また，図 3, 4 では，それぞれ人手によりタイプ「見る」, タイプ「食べる」と「体験する」に判定された旅行ブログエントリの例を示す.

表 1: タイプとその判定基準

タイプ	判定基準
買う	お土産に関する情報が記載されている.
食べる	飲食に関する情報が記載されている.
体験する	○○体験やスキューバダイビングなど，自分の体を使って楽しめる物についての情報が記載されている
泊まる	宿泊に関する情報が記載されている.
見る	観光名所などの見て楽しめる物やイベントについての情報が記載されている.
その他	上記のタイプに該当しない情報が記載されている.

I went to the Money Museum. It had a bit of history of money and how it evolved in Scotland and the world. It showed some examples of early paper money and how enterprising It shows a bit about how coins are struck, some ways that people attempted, and succeeded in making forgeries. ***** (省略) *****

図 3: タイプ「見る」に判定された英語の旅行ブログエントリの例
(http://www.travelblog.org/Europe/United-Kingdom/Scotland/Midlothian/Edinburgh/blog-657415.html)

レンタカーを借りて市場→山代温泉へと向かいました。ホテルのロビー・・・ツリーがきれいでした。
山代温泉では日帰り風呂に入りたかったのですが・・気乗りしない様子のだんな様。。。仕方なく、足湯にだけつかり、街を散策し軽めの昼食を。初めて来る山代温泉、こんな感じなんだ～。

******(省略)*****

やはり、昼は蕎麦ですよね～
路地でに見つけました『蕎麦 山背』
落ち着いた雰囲気で、蕎麦も美味しかったです。
デザートにと果物のサービスも。

図 4: タイプ「食べる」と「体験する」に判定された日本語の旅行ブログエントリの例
(http://blogs.yahoo.co.jp/ami33364/56891928.html)

4.2. 情報利得を用いた手掛かり語の収集

タイプ「見る」に判定された旅行ブログエントリには，日本語の場合「展示」,「見学」などが，英語の場合「museum」,「castle」などの単語が頻繁に出現する．この様に，各タイプに判定された旅行ブログエントリには，そのタイプを示す特有の単語が頻繁に出現する傾向がある．そのため，本研究では，各タイプに特有の単語を手掛かり語として収集し，手掛かり語の有無を機械学習の素性として与える.手掛かり語の収集は，情報利得により収集する．

情報利得とは，「ある単語の有無」の情報がクラスに関するエントロピーをどのくらい減少させるかを示す数値である．エントロピーを減少させる単語を選択することにより，クラスの特定が容易になる．そのため，情報利得を用いて収集した単語は，クラス分類において，有効な手掛かり語であると言える．

本研究では，「買う」,「食べる」,「体験する」,「泊まる」,「見る」,「その他」の 6 種類のタイプごとに，旅行ブログエントリ内で出現する単語に対して，情報利得を求める.情報利得を求める際に使用する単語は，形態素ごとに単語の分割を行い,品詞が動詞,名詞句,形容詞であるものとする.また,これらの単語のうち，出現回数が 1 回未満，単語の長さが 15 文字以上，単語の長さが半角 1 文字未満のいずれかに当てはまる単語は不要語として削除する.なお,形態素解析において，日本語で記述された旅行ブログエントリでは MeCab[5]，英語で記述された旅行ブログエントリは TreeTagger[6]を用いた．

上記の条件に当てはまる単語に対し，情報利得を求め，その値が閾値より高い単語を手掛かり語として収集する．閾値は，予備実験により設定した．旅行ブログエントリから収集した手掛かり語の例を表 2 に示す．

表 2: 情報利得により収集した手掛かり語の例

タイプ	日本語	英語
買う	お土産，購入	buy，shop
食べる	食べる，美味しい	sushi，dinner
体験する	温泉，浸かる	climb，hike
泊まる	部屋，ホテル	room，guest
見る	公園，紅葉	museum，visit
その他	機材，朝一番	go，see

4.3. ドメイン適応を用いたタイプ分類

本研究では，日本語で記述された旅行ブログエントリを日本語ドメイン，英語で記述された旅行ブログエントリを英語ドメインとして，ドメイン適応を行う．まず,日本語ドメインの素性を用いてモデルを獲得し，獲得したモデルを英語ドメインに適応する．その際，英語で記述された旅行ブログエントリを日本語へ翻訳する．翻訳は，Microsoft Translator[7]の API を用いた．

5. 実験

本節では，提案手法の有効性を確認するために行った実験について述べる．5.1 節では実験条件について説明し，5.2 節では提案手法，5.3 節では実験結果について述べる．

5.1. 実験条件

5.1.1. 実験に用いるデータ

実験に用いるデータとして，日本語または英語で記載された旅行ブログエントリを対象とした．日本語で記載された旅行ブログエントリは，Nanba らの手法により収集した旅行ブログエントリ 1,000 件，英語で記載された旅行ブログエントリでは，Travel Blog から収集した 660 件を対象に実験を行った．収集したデータに対して，表 1 の判定基準に従い人手でタイプ分類を行った結果を実験に用いた．人手によりタイプ分類を行った正解データの件数を表 3 に示す．

5.1.2. 機械学習と評価尺度

機械学習には，TinySVM を用いて，2 分割交差検定を行った．評価尺度として精度，再現率を使用した．旅行ブログエントリは日々作成され膨大に存在するため，本研究では再現率よりも精度を重視した．

[5] http://mecab.sourceforge.net/
[6] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[7] https://datamarket.azure.com/dataset/bing/microsofttranslator

表 3: 旅行ブログエントリのタイプ件数

タイプ	日本語の旅行 ブログエントリ	英語の旅行 ブログエントリ
買う	163	30
食べる	382	97
体験する	241	143
泊まる	134	61
見る	395	316
その他	193	155

5.2. 実験手法

以下に示す提案手法について実験を行った．また，提案手法の有効性を確認するため，比較手法(Baseline 手法)を用いて実験した．

＜日本語の旅行ブログエントリのタイプ分類＞

- Baseline 手法：日本語の旅行ブログエントリに出現する全単語を素性とした手法．
- IG 手法:日本語の旅行ブログエントリを対象に情報利得を利用する．情報利得により収集された手掛かり語を素性として用いる手法．

＜英語の旅行ブログエントリのタイプ分類＞

- Baseline 手法：英語の旅行ブログエントリに出現する全単語を素性とした手法．
- IG 手法:英語の旅行ブログエントリを対象に情報利得を利用する．情報利得により収集された手掛かり語を素性として用いる手法．
- DA 手法：ドメイン適応を用いる．日本語ドメインから IG 手法を用いて獲得したモデルを用いて，日本語へ翻訳した英語ドメインに適応する手法．
- IG+DA 手法：DA 手法の出力結果の値を IG 手法の素性として与える手法．

5.3. 実験結果

日本語で記述された旅行ブログエントリのタイプ分類の結果を表 4，英語で記述された旅行ブログエントリのタイプ分類の結果を表 5 に示す．

表 4，5 より，IG 手法は Baseline 手法を上回る結果を得た．DA 手法では，IG 手法のタイプ「買う」と「泊まる」と比較し，精度を向上することができた．IG+DA 手法では，IG 手法，DA 手法の平均精度を上回ることができ，提案手法の有効性を示すことができた．

6. 考察

日本語で記述された旅行ブログエントリのタイプ分類において，最も精度の低かったタイプ「買う」について考察を行う．本研究では，人手によりタイプ判定を行う際に，1 件の旅行ブログエントリに対して，複数のタイプに判定することを許している．そこで，タイプ「買う」と判定された旅行ブログエントリが，ほかのタイプにも判定された割合を求めた．その結果を表 6 に示す．

表 4: 日本語で記述された旅行ブログエントリの実験結果

	評価尺度	買う	食べる	体験	泊まる	見る	平均
Baseline	精度	0.317	0.769	0.278	0.567	0.462	0.479
	再現率	0.050	0.579	0.021	0.037	0.297	0.197
IG	精度	**0.549**	**0.777**	**0.602**	**0.589**	**0.667**	**0.637**
	再現率	0.318	0.671	0.337	0.343	0.640	0.462

表 5: 英語で記述された旅行ブログエントリの実験結果

	評価尺度	買う	食べる	体験	泊まる	見る	平均
Baseline	精度	0.011	0.122	0.140	0.087	0.381	0.148
	再現率	0.375	0.676	0.970	0.527	0.961	0.702
IG	精度	0.222	0.728	0.726	0.452	0.744	0.574
	再現率	0.125	0.342	0.311	0.083	0.617	0.296
DA	精度	**0.250**	0.389	0.535	**0.538**	0.580	0.458
	再現率	0.200	0.577	0.266	0.115	0.873	0.406
IG+DA	精度	**0.250**	**0.810**	**0.741**	0.410	**0.773**	**0.597**
	再現率	0.094	0.473	0.295	0.149	0.672	0.337

表 6: タイプ「買う」と判定された旅行ブログエントリが他のタイプに判定された割合

タイプ	他のタイプに判定された割合(%)
買う	28.2
食べる	49.7
体験する	14.7
泊まる	9.2
見る	39.3

表 6 より，人手によりタイプ「買う」と判定されている旅行ブログエントリは，タイプ「食べる」やタイプ「見る」にも判定されている割合が高い．そのため，情報利得により手掛かり語を収集する際に，「チョコ」や「展望」などのタイプ「食べる」やタイプ「見る」に関する手掛かり語も収集されてしまう可能性が高いと考えられる．そのため，タイプ「買う」においては，タイプ分類の精度向上があまりみられなかったと考えられる．

次に，英語で記述された旅行ブログエントリのタイプ分類について考察を行う．情報利得を用いた IG 手法により，タイプ「食べる」，「体験する」，「見る」では，それぞれ精度 0.728，0.726，0.744 を得ることができ，大幅な精度向上が見られた．これらの結果に比べ，タイプ「買う」と「泊まる」では，それぞれ精度 0.222，0.452 であった．Baseline 手法を上回ることができたが，タイプ「食べる」，「体験する」，「見る」と比べ，精度の向上は見られなかった．この理由として，実験に使用したデータ件数が少ないためだと考えられる．英語により記述された旅行ブログエントリのデータ件数を補うため，ドメイン適応を用いた DA 手法を行った．その結果，タイプ全体での平均精度において，DA 手法は IG 手法に比べ低下したが，データ件数の少なかったタイプ「買う」と「泊まる」において，それぞれ，精度 0.028，0.086 向上させることができた． また，IG+DA 手法では，DA 手法から得られた結果を素性として加えることにより，IG 手法の平均精度と比べ，精度を 2.25 ポイント向上することができた．この結果より，ドメイン適応を用いることにより，データ数が少ない問題を解決でき，精度を向上させることができた．

7. おわりに

旅行ブログエントリを用いた観光情報提示システムを構築した．旅行者が必要とする情報へのアクセスを容易にするために，本研究では日本語または英語で記載された旅行ブログエントリを「買う」，「食べる」，「体験する」，「泊まる」，「見る」の 5 種類のタイプへ自動的に分類する手法を提案した．タイプの自動分類では，情報利得とドメイン適応を用いた手法により，精度 0.597，再現率 0.337 を得ることができ，提案手法の有効性を確認できた．

本研究では，日本語と英語の旅行ブログエントリを対象とし，単語をベースとしたアプローチをとっている．そのため，構文解析などの技術は用いておらず，今回実験で使用した言語以外の旅行ブログエントリに対しても同様の手法により，タイプ分類を行えると考えられる．

参 考 文 献

[1] Nanba, H., Taguma, H., Ozaki, T., Kobayashi, D., Ishino, A. and Takezawa, T., “Automatic Compilation of Travel Information from Automatically Identified Travel Blogs”, Proc. of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, Short Paper, pp. 205-208, 2009.

[2] Ishino, A., Nanba, H. and Takezawa, T., “Construction of a System for Providing Travel Information along Hiroden Streetcar Lines”, Proc. of the 3rd IIAI International Conference on e-Services and Knowledge Management (IIAI ESKM 2012), 2012.

[3] 福田 悟志，難波 英嗣，竹澤 寿幸，“論文と特許からの技術動向情報の抽出と可視化”，情報処理学会論文誌，データベース，Vol.6, No.2, pp.16-29, 2013.

[4] 中嶋 勇人，新妻 弘崇，太田 学，“位置情報ツイートを利用した観光ルート推薦”，情報処理学会研究報告，第 158 回データベース・システム研究報告, No.28, pp.1-6, 2013.

[5] 石野 亜耶，難波 英嗣，竹澤 寿幸，“ブログを中心とした観光情報の組織化”，第 3 回楽天研究開発シンポジウム, 2010.

[6] 徳久 雅人，村田 真樹，“観光開発のヒントをブログ記事から得るための支援技術～SVM を用いる場合～”，第 8 回観光情報学会全国大会発表概要集，pp.44-45, 2011.

[7] 遠藤 雅樹，大野 成義，石川 博，“地域サイト及びブログの観光情報融合のためのキーワード自動抽出手法の検討”，情報処理学会研究報告，第 158 回データベース・システム研究報告，No.2, pp.1-8, 2013.