社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

旅行ブログからの観光情報の自動抽出

石野　亜耶† 　難波　英嗣† 　田熊　遥‡
尾崎　貴紘‡ 　小林　大祐† 　竹澤　寿幸†

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

‡広島市立大学　情報科学部　〒731-3194 広島市安佐南区大塚東 3 丁目 4 番 1 号

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

あらまし　本研究では、自動的に観光情報を収集するための手法を提案する。我々は観光情報を収集するため、ブロガーが日記形式で綴った旅行記である旅行ブログに焦点を当てた。多くのブロガーが旅行記をこの形で記述するため、旅行ブログは観光情報を得るための有益な情報源であると考えられる。よって本研究では、ブログデータベースから旅行ブログを検出し、その中から観光情報を抽出する手法を提案した。また実験により提案手法の有効性を示した。旅行ブログの検出に関しては、再現率 38.1%、精度 86.7%を得た。また、旅行ブログからの観光情報の抽出に関しては、抽出された上位 100 種類の土産物において精度 74.0%を得ることができたため、旅行ブログは観光情報の有益な情報源であるといえる。

キーワード　ブログ，　情報抽出，　観光情報

Automatic Compilation of Travel Information from Automatically Identified Travel Blogs

Aya ISHINO† 　Hidetsugu NANBA† 　Haruka TAGUMA‡
Takahiro OZAKI‡ 　Daisuke KOBAYASHI† 　and　Toshiyuki TAKEZAWA†

† Graduate School of Information Sciences, Hiroshima City University 3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194, Japan

‡ School of Information Sciences, Hiroshima City University 3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194, Japan

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

Abstract　In this paper, we propose a method for compiling travel information automatically. For the compilation, we focus on travel blogs, which are defined as travel journals written by bloggers in diary form. We consider that travel blogs are a useful information source for obtaining travel information, because many bloggers' travel experiences are written in this form. Therefore, we identified travel blogs in a blog database and extracted travel information from them. We have confirmed the effectiveness of our method by experiment. For the identification of travel blogs, we obtained scores of 38.1% for Recall and 86.7% for Precision. In the extraction of travel information from travel blogs, we obtained 74.0% for Precision at the top 100 extracted local products, thereby confirming that travel blogs are a useful source of travel information.

Keyword　Blog,　Information Extraction,　Travel Information

1. はじめに

2007 年 1 月に「観光立国推進基本法」が施行され、2008 年 10 月には国土交通省の外局として観光庁が設置されるなど、日本では今、「観光」を 21 世紀の基幹産業と位置付けた多様な取り組みが、国や地方公共団体、民間で積極的に推進されている。観光を支援する媒体としては、地方公共団体や旅行会社などが運営する観光ポータルサイトや、旅行情報雑誌など、既に観光情報データベースがいくつか作成されており、Web 上で公開されているものも少なくない。しかし、これ

らの旅行情報誌や観光ポータルサイトは人手で構築されたものであり、作成に多大なコストを要する。

そこで本研究では、ブログから自動的に観光情報を抽出することで、低コストでのデータベース作成を目指す。同時に、網羅性の高さや最新の観光情報を素早く獲得できる点などで、既存のデータベースよりも有用なデータベースになることが期待される。また、ブログ著者の属性(性別、年齢、居住域など)を文体や記載内容から自動的に推定する研究が進んでいるが[1, 2, 3]、このような技術を利用することで、例えば「女性に人気のスポット」や「若い人に人気のスポット」などユーザに適した観光情報も自動的に抽出できるようになると期待できる。

本論文の構成は以下の通りである。2 章では関連研究、3 章では提案手法、4 章では実験結果について述べる。また結論については 5 章で述べる。

2. 関連研究

'www.travelblog.org'と'travel.blogmura.com'は旅行ブログのポータルサイトである。これらのサイトでは、旅行ブログはブロガー自身により人手で登録され、目的により分類される。しかし、ブログ空間にはたくさんの旅行ブログが存在するため、これらのポータルサイトに登録されていない旅行ブログも多数存在する。よって本研究では、旅行ブログの完全なデータベースの構成を目的とし、旅行ブログの自動検出について研究を行った。

CLEF(Cross Language Evaluation Forum)という会議のタスクのひとつとして、地理系に特化した情報を検索する Geo CLEF[1]が 2005 年から開催されている[4]。このタスクの目的は、新聞記事集合から「ヨーロッパにある川の周りはワイン作りが盛んな地域だ」のような地理情報の関連記事を探すというものである。本研究では、新聞記事の代わりに、一般の旅行者が気軽に観光情報を発信する場としてよく使われる旅行ブログに焦点をあてた。

3. 観光情報の自動抽出

観光情報の自動抽出の段階は、以下の 2 つのステップに分けられる。

Step1. 旅行ブログの検出
Step2. 旅行ブログからの観光情報の抽出

この 2 つのステップについては 3.1、3.2 節で、それぞれ説明する。

3.1. 旅行ブログの検出

旅行ブログには‘旅行’、‘観光’、‘ツアー’などの旅行に関する手掛かり語を含む可能性が高いと言える。しかし、すべての旅行ブログに、このような手掛かり語は含まれているわけではない。例えば、あるブロガーがノルウェー旅行について複数のブログエントリにわたって日記を書いていた場合、最初のエントリには‘私たちはノルウェーに旅行に行った’と書いてあっても、2 ページ目のエントリには‘野生の羊にあったんだ！’としか書かれていないこともある。この場合、2 ページ目のエントリには旅行に関連した表現が含まれていないため、2 ページ目のエントリを旅行ブログであると判定することは困難である。そこで本研究では、それぞれのターゲットとなるエントリについてのみ見るのではなく、前後のエントリにも注目した。

そこで本研究では、旅行ブログの検出を系列ラベリング問題として解き、機械学習を用いて解決する手法を考案した。機械学習の手法には、近年自然言語処理の分野において、実験に用いられ高い精度を示している CRF を使用した。CRF に与える素性とタグは以下のとおりである。

(1) ターゲットとなるエントリより前の k 個のエントリに付与されたタグ
(2) ターゲットとなるエントリの前に存在する、ターゲットからの距離が k 以内のエントリに存在する手掛かり語の有無
(3) ターゲットとなるエントリの後に存在する、ターゲットからの距離が k 以内のエントリに存在する手掛かり語の有無(図 1)

我々は予備実験の結果から k=4 と定めた。ここで、‘旅行’、‘ツアー’、‘出発’や地名[2]など 416 個の素性が各エントリに含まれるかどうかを機械学習に与えた。

3.2. ブログからの観光情報の抽出

Step1 で検出した旅行ブログから、観光情報として地域名と土産物の対を効率的に抽出する。

まず、482 の地域名と土産物の対を用意した。これらの対は Google から提供されている‘Web 日本語 N グラム’データベース[3]から自動で抽出した。このデータベースは、Web 上にある日本語で書かれた 20 億文から抽出された N グラム(N=1～7)で構成されている。本研究では‘[地名]　名物　「[名物]」’という表層パターンをデータベースにあてはめ、それぞれ一致した一区切りから、地域名と土産物の対を 482 対抽出した。

[1] http://ir.shef.ac.uk/geoclef/

[2] 地名の判定には CaboCha を用いた。
http://chasen.org/~taku/software/cabocha/

[3] http://www.gsk.or.jp/catalog/GSK2007-C/catalog.html

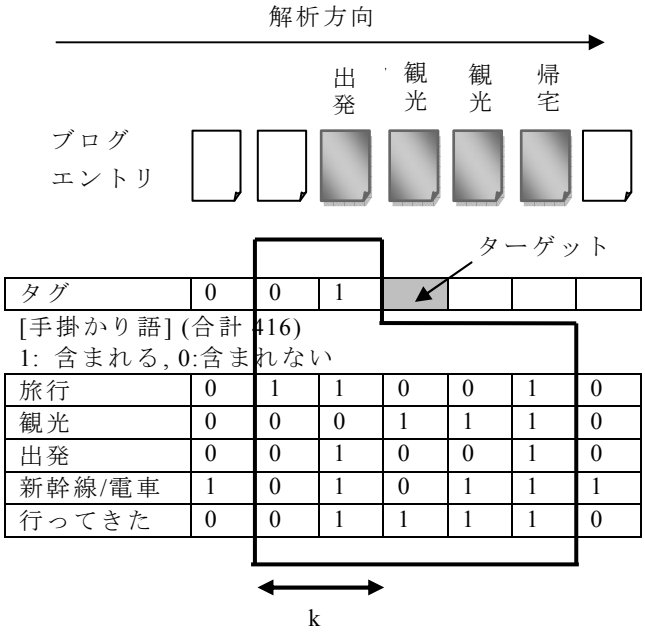

図 1　CRF に与えた素性とタグ

次に、Step1 で検出した旅行ブログに、情報抽出技術に基づいた機械学習を用いることで、新しい対地域名と土産物の対を得た。機械学習の訓練用データは以下の方法で準備した。

1. 482 対から地域名と土産物両方を含む 200 文を選ぶ。ここで自動的に‘location’と‘product’タグを付与したタグ付きの 200 文を生成する。
2. 地域名だけを含む 200 文を準備する。[4]
   またこれらの文に‘location’タグを付与したタグ付きの 200 文を生成する。
3. タグ付きの 400 文を機械学習に与え、これらの文に自動的に‘location’と‘product’タグを付与する。

本研究では機械学習として CRF を使用した。Step1 と同じように、CRF 基本手法は与えられた文に含まれる語を分類するのに使用した。素性とタグは以下のように CRF に与える。

(1)ターゲットとなる単語から、CRF に与える前後の単語数 k
(2)ターゲットとなる単語の前に存在する、ターゲットからの距離が k 以内に現れる単語
(3)ターゲットとなる単語の後に存在する、ターゲットからの距離が k 以内に現れる単語

我々は予備実験の結果から k=2 と定めた。また、以下の 6 つの素性を機械学習に使用した。

- 単語
- 品詞[5]
- 単語に引用記号がついているかどうか
- 単語が‘名物’、‘名産’、‘特産’、‘銘菓’、‘土産’のような手掛かり語であるかどうか
- 単語が表層格かどうか
- ‘ケーキ’や‘ラーメン’のような土産物や名産物の名前によく使われる単語が含まれているかどうか

4. 実験

本研究では 2 種類の実験を行った：
(1)旅行ブログの検出
(2)旅行ブログから観光情報の抽出
これらについては、それぞれ、4.1、4.2 節で述べる

4.1. 旅行ブログの検出

日本語で書かれた約 1,100,000 エントリから 317 人のブロガーによって書かれた 4,914 エントリをランダムに選んだ。我々はこの 4,914 エントリを人手で旅行ブログかどうかを判定した。その結果、“旅行ブログ”と判定されたのは 420 エントリと少数であったため、4 分割交差検定を行い評価することとした。機械学習器には CRF++[6]を使用した。また、精度と再現率を用いて評価を行った。

比較手法

提案手法の有効性を確かめるため、比較手法として、前後のエントリの素性を使用せず、注目しているブログエントリのみの素性を使用した旅行ブログの検出を行った。

実験結果と考察

実験結果を表 1 に示す。表 1 より、我々の提案手法は、精度は 26.2%上がったが、再現率は 13.3%下がった。現在のステップで精度が低いと、次のステップでも精度が低くなってしまうため、本研究では再現率よりも精度を重要視した。

表 1：旅行ブログの検出

	再現率	精度
提案手法	38.1	86.7
比較手法	51.1	60.5

[4] 試験的な実験で、最初は機械学習に否定的なケースは使用しなかったため、低い精度しか得ることができなかった。これは、我々のシステムが旅行ブログにおいて地域名を含んでいる全ての文から、土産物を抽出しようと試みているためである。

[5] このステップでは、CaboCha を用いて自動的に地域名を判定した。

[6] http://www.chasen.org/~taku/software/CRF++/

人手では“旅行ブログ”と判定したが、提案手法では“旅行ブログでない”と誤って判定したエントリが266件存在した。このエントリの中から50エントリを任意に選び、検出誤りについて分析を行った。以下に検出誤りの主要な原因を示す。

(1) 複数エントリにわたる旅行記の一部(50%)
(2) 記載内容が3行以下のエントリ(10%)
(3) その他(40%)

以下に、それぞれの検出誤りについて説明する。

(1) 複数エントリにわたる旅行記の一部(50%)

50件のうち25件(50%)が複数エントリにわたる旅行記の一部であった。複数エントリにわたる旅行記の場合、最初のエントリが“旅行ブログ”と判定できなければ、残りのエントリも“旅行ブログ”と判定することはできない。この検出誤りの原因は、手掛かり語の不足であった。提案手法では、人手で選択した手掛かり語を使用しているが、手掛かり語を増やす一つの手法として、Nグラムを自動的に検出した旅行ブログにあてはめ、手掛かり語を網羅的に集めることで問題を解決することができると考えられる。

(2) 記載内容が3行以下のエントリ(10%)

50件のうち5件(10%)が、記載内容が3行以下のエントリであった。この検出誤りの原因は、提案手法で判定するためには短すぎるからだと考えられる。

また、人手では“旅行ブログではない”と判定したが、提案手法では“旅行ブログ”と判定したエントリが26件存在した。この26件の検出誤りは、大きく次の4種類に分類することができる。

(1) エントリの前後に旅行ブログが存在(38.5%)
(2) 地元紹介のエントリ(34.7%)
(3) 他人の旅行を紹介しているエントリ(11.6%)
(4) その他(15.2%)

以下に、それぞれの検出誤りについて説明する。

(1) エントリの前後に旅行ブログが存在(38.5%)

26件のうち、10件(38.5%)がエントリの前後に旅行ブログが存在していた。あるブロガーがA・B・C・Dというエントリを記述したとする。エントリA・B・Dに旅行記を記述し、Cには旅行とは全く関係のない内容を記述しているとき、提案手法ではエントリCも旅行記の一部であると判断してしまっていた。

(2) 地元紹介のエントリ(34.7%)

26件のうち9件(34.7%)が地元住民による地元の紹介エントリであった。この検出誤りの原因は、ブロガーの居住区情報が反映されていないため、旅行で訪れた場所か、日常生活圏で訪れた場所なのか判定できないためである。

(3) 他人の旅行を紹介しているエントリ(11.6%)

26件のうち3件(11.6%)が他人の旅行を紹介しているエントリであった。自らが体験した旅行についての記事ではないため、人手では“旅行ブログでない”と判定される。しかし、他人の旅行について記事を書いているため、旅行に関する単語が頻繁に出現し、提案手法では“旅行ブログ”と判定されてしまった。

4.2. 旅行ブログからの観光情報の抽出

データセットと実験セット

旅行ブログは観光情報の抽出のための有用な情報源であることを確かめるため、我々は以下の3つの情報源を用いた観光情報を抽出する。

- 旅行ブログ(提案手法):3.1節で述べた手法を用いて判定した1,100,000エントリから自動的に判定した17,268旅行ブログ中の全ての文(80,000文)
- 一般ブログ:1,100,000ブログエントリから選択した任意の80,000文
- 一般ウェブ:ウェブ5億文データベース[5]から選択した任意の80,000文

我々はそれぞれの情報源から観光情報(地域名と土産物の対)を抽出し、出現頻度によりランク付けを行った。

評価方法

評価尺度としては、上位にランク付けられた観光情報に対して、次の式で精度を求めた。5間隔で上位5位から100位まで精度を計算した。

$$\text{精度}=\frac{\text{正しく抽出された地域名と土産物の対}}{\text{抽出された地域名と土産物の対}}$$

実験結果と考察

実験結果を図2に示す。この図で示したように、特に高い順位で、一般ブログ手法は一般ウェブ手法より高い精度を示した。また、旅行ブログ(提案手法)は一般ブログ手法よりも有益であり、観光情報の抽出のための有益な情報源であることがわかる。

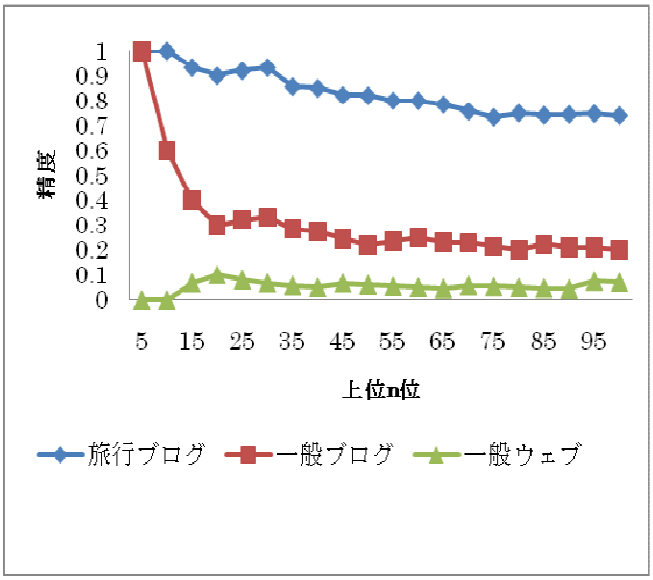

図 2：上位 n 位の観光情報の抽出の再現率

Google N-gram データベースから作成した土産物のリストに含まれていないが、本研究で行った各手法により抽出された土産物の種類を表 2 に示す。表 2 より、旅行ブログ手法では 41 種類の土産物を抽出することができた。一方で一般ブログ手法では 15 種類、一般ウェブ手法では 7 種類であった。これらの結果より、観光情報の情報源として旅行ブログの有益性を示せたといえる。

提案手法の上位 100 位間の典型的な抽出誤りは、店の名前を間違えて抽出したことである。これらの店の多くでは土産物を売っている。この問題を解決するためには、土産物とその土産物の販売店の対を抽出する必要がある。

表 2：各手法で新しく抽出された土産物の種類

旅行ブログ(提案手法)	41
一般ブログ	15
一般ウェブ	7

5. 結論

本研究では、ブログデータベースから旅行ブログを自動で検出する手法を提案し、それらから観光情報を抽出した。旅行ブログの検出では、再現率 38.%、精度 86.7%を得ることができた。また、旅行ブログからの観光情報の抽出では、抽出された上位 100 位の土産物で精度 74.0%を得ることができた。

文　献

[1] Norihito Yasuda, Tsutomu Hirao, Jun Suzuki, and Hideki Isozaki. 2006. Identifying bloggers' residential areas. *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp.231-236.

[2] Daisuke Ikeda, Hiroya Takamura, and Manabu Okumura. 2008. Semi-Supervised Learning for Blog Classification. *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp.1156-1161.

[3] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. *Proceedings of AAAI Symposium on Computational Approaches for Analyzing Weblogs*, pp.199-205.

[4] Fredric C. Gey, Ray R. Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. 2005. GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. *Lecture Notes in Computer Science*, LNCS4022, pp.908-919.

[5] Daisuke Kawahara and Sadao Kurohashi. 2006. A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp.176-183.