知能と情報（日本知能情報ファジィ学会誌）Vol.22, No.6, pp.000－000（2010）

原著論文

旅行ブログエントリからの観光情報の自動抽出†

石野 亜耶＊・難波 英嗣＊・竹澤 寿幸＊

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エントリに焦点を当てた．多くのブロガーが旅行記をこの形で記述するため,旅行ブログエントリ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．まず本研究では，ブログデータベースから旅行ブログエントリを検出した．その中から観光情報として土産物情報と観光名所情報を抽出する手法を提案した．更に，旅行ブログエントリからリンクを抽出することで，観光情報リンク集の構築を行った．また実験により提案手法の有効性を示した．旅行ブログエントリの検出に関しては，再現率38.1%，精度86.7%を得た．また，旅行ブログエントリからの観光情報の抽出においては，抽出された上位100件の土産物において精度74.0%，観光名所において精度71.0%を得ることができたため，旅行ブログエントリは観光情報の有益な情報源であるといえる．旅行ブログエントリからの観光情報リンク集の自動構築においても，高い精度・再現率を得られており，提案手法の有効性を示すことができたと言える．

キーワード：ブログ，情報抽出，観光情報

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「観光」を21世紀の基幹産業と位置付け，観光を支援する多様な取り組みが積極的に推進されている．Web上で利用可能な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が挙げられる．観光ポータルサイトでは，土産物や観光名所などの情報，ホテルやレストランへのリンクが観光情報として紹介されている．しかし観光情報は，土産物の商品開発や，テーマパークなどの施設の建設などにより日々新しくなり，ホテルやレストランを紹介するWebページも新しく作成される．そのため観光情報を新たに獲得し，古くなった情報は削除するといった更新作業が不可欠である．しかし，既存のデータベースは人手で観光情報を抽出し，整理，保守するため，非常に時間とコストがかかる．

そこで，本研究では，旅行者が気軽に観光情報を発信する場としてよく用いられるブログに注目した．本研究では，旅行記が記述されたブログエントリを旅行ブログエントリと呼ぶこととする．旅行ブログエントリには，土産物，観光名所，旅行の際に参考にしたWebページへのリンクなど様々な観光情報を含んでいる．このような旅行ブログエントリから自動的に観光情報を抽出することで，低コストで観光情報データベースを作成することが可能になると考えられる．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，既存の観光ポータルサイトよりも有用なものになることが期待される．

本研究では，旅行ブログエントリから，土産物情報・観光名所情報を抽出し，さらに，旅行ブログエントリ中に存在するリンクを収集，分類することで観光情報リンク集の構築を行う．

また，近年ブログ著者の属性(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[1, 2, 3]．このような技術を利用し，ブログ著者の属性と，観光情報の利用者の属性を照らし合わせることで，例えば「女性に人気の土産物」や「若い人に人気の観光名所」など，利用者に適した観光情報を推薦することができると考えられる．

本論文の構成は以下の通りである．２節では関連研究，３節では提案手法，４節では実験結果と考察について述べ，５節で本稿をまとめる．

2．関連研究

本章では，「地理情報検索」と「リンクの分類」に関連する研究について述べる．

---

† Automatic Compilation of Travel Information from Automatically Identified Travel Blog Entries
Aya ISHINO, Hidetsugu NANBA and Toshiyuki TAKEZAWA

＊ 広島市立大学大学院 情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

2.1 地理情報検索

近年，地理情報検索に関する様々な研究が行われている．CLEF（Cross Language Evaluation Forum）という評価ワークショップ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うブログに焦点をあてた．

次に，Webを用いた地理情報検索に関連する研究について述べる．本研究と同様に，Webから地域情報を自動収集しようとする研究がある．大槻ら[5]は，地域情報ウェブディレクトリを自動編集するシステムを提案している．地域情報ウェブディレクトリは地域情報検索に利用される．本研究では，地域情報を収集するにあたって，旅行ブログエントリを対象としているが，大槻らは，自治体が提供する地域情報サイトと，そのリンク先の地域サイトを対象としている点で異なる．

佐藤[6]は，Webを利用した住所探索を提案している．この提案手法では，まず検索エンジンを利用して住所情報が記載されている可能性が高いWebページを収集し，そのWebページから住所データを抽出する．このようにして得られた住所データを整理・統合して，目的の住所情報を出力する．

村山ら[7]は，位置情報が明記されていない文書に位置情報をメタデータとして付与するため，お店の名前等の固有名と対応する位置情報のデータベースをWeb上の文書から自動的に作成する手法を提案している．

安田ら[8]は，携帯端末等を持ちながら歩いている状況で，できるだけ近くの店舗に関する情報を検索するといった，小さな領域における距離を考慮した検索を可能とする手法を提案している．この手法は，例えば「東京都新宿区歌舞伎町」周辺の情報を知りたいという検索が行われた場合，「東京都新宿区歌舞伎町」に言及する文書を優先的に検索することができる．

Webから地理情報を抽出する手法の１つに，あらかじめ検索対象のリストを作成し，クローリングによって得られた情報を各検索対象に関連づける登録型検索手法がある．しかし，この手法はリストに登録されていない対象に関する情報を抽出できないという欠点がある．そこで相良ら[9]は，実世界に存在する店舗を対象に新規店舗を検索し，店舗データを新たに登録する手法を提案している．この手法は，既に登録されている店舗情報を収集する際に得られたWebページ群から検索することで，新規店舗候補の検索を効率良く行う．上記で述べた研究はWebを対象としたものであるが，本研究では旅行ブログエントリを情報源とすることで，観光に特化した情報の抽出を目指す．

本研究と同様に，ブログを情報源とし，地域情報を自動抽出する研究がある．岡本ら[10]は，一般のブログ検索エンジンを利用することで，地名を含むブログエントリを収集し，それらのブログエントリから，地域イベント情報を抽出する手法を提案している．また，藤坂ら[11]は，Twitter[2]に代表されるマイクロブログから地域イベントを発見し，その特性を検証するためのシステムを提案している．しかし，ブログの中には観光と関係ないものも存在するため，ブログエントリを全て使うと，十分な精度で観光情報が抽出できない可能性がある．本研究では，まず，ブログ集合から旅行ブログエントリを自動検出し，次に，そこから観光情報の抽出をすることにより，高い精度での抽出を目指す．また，ブログ著者の属性を利用することで，利用者に適した観光情報を推薦できるようになると考えられる．

Webやブログで自らの行動を日記として発信することが盛んになってきている．郡ら[12]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することにより，集約して提示するシステムを提案している．また，Davidov[13]は，Webから交通手段や経路の地理的なネットワークを見つける手法を提案している．これらの研究と，本研究で提案した旅行ブログエントリから観光名所情報を抽出する手法を組み合わせることで，旅行者に最適な行動経路を推薦することができると考えられる．

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は，‘Travel Blog’[3]，‘旅行・観光ブログ村’[4]，‘フォートラベル’[5]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．一般ブログに焦点を当てることで，様々な層のより多くの旅行ブログエントリを収集できると考えられる．

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

2 http://twitter.com/
3 http://www.travelblog.org/
4 http://travel.blogmura.com/
5 http://4travel.jp/

2.2 リンクの分類

次に，リンクの分類に関する研究について述べる．リンクの分類の手法は主に，Kaleら[14]の評価表現の比率に基づく手法と，Martineauら[15]の機械学習に基づく手法に分けられる．本研究では，Martineauらの手法と同様に，機械学習を用いてリンクの分類を行う．Martineauらはブログ中のリンクについて，評価極性など複数の観点から分類を行っているが，本研究では，観光に特化した分類を行うため，3.3.3節で説明するタイプに分類する．このように，観光情報に特化した分類を行うことで，旅行者にとって有用なリンクを推薦することが可能になると考えられる．

本研究では，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た．観光情報リンク集を利用者に提示する場合に，スニペットをどのように生成するかという問題点がある．本研究では，リンクの分類を行う際に利用した引用個所をスニペットとして表示するようにした．このスニペットを読むことで，リンク先サイトに関する感想などの情報も得ることができるようになっている．戸田ら[16]は，地理情報検索における新しいスニペットの生成手法を提案している．この手法の主な特徴は，地理情報検索の検索クエリに含まれる地理制約と関連性の高い地理表現を利用するという点である．

3．旅行ブログエントリからの観光情報の自動抽出

本研究では，観光情報を抽出する際の情報源として，旅行ブログエントリを使用する．3.1節では，一般ブログから旅行ブログエントリを検出する手法について，3.2節では，旅行ブログエントリから土産物情報・観光名所情報を抽出する手法，3.3節では，旅行ブログエントリから観光情報リンク集を自動構築する手法について説明を行う．

3.1 旅行ブログエントリの自動検出

旅行ブログエントリには‘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エントリに，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エントリ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，旅行ブログエントリ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[6]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．なお，図1では，説明のためk＝2の場合を例として示している．

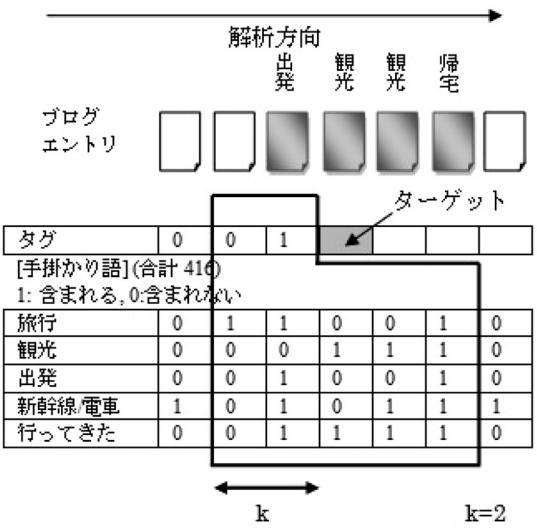

図1 CRFに与えた素性とタグ

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

3.2 土産物情報・観光名所情報の自動抽出

本節では，旅行ブログエントリから土産物情報，観光名所情報を自動抽出する手法について説明を行う．ここで，土産物情報は地域名と土産物の対，観光名所情報とは地域名と観光名所の対である．

本研究では，自動抽出の際，表層パターンと機械学習を用いる．表層パターンを用いた手法とは，あらかじめ決められたパターンを埋める形でテキスト中の情報を抽出する手法である．そのため，高い精度で情報を抽出できるが，網羅性の点で十分でないという問題がある．そこで，表層パターン手法で抽出されたデータから機械学習用データを自動的に作成し，機械学習を用いることにより，低コストで網羅性の高い観光情報を抽出する．詳しい手順について，以下で説明を行う．

まず，土産物リスト(地域名と土産物の対)，観光名所リスト(地域名と観光名所の対)を作成する．これらのリストはGoogleから提供されている‘Web日本語Nグラム’データベース[7]に，表層パターンをあてはめ，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．土産物リストの作成には，‘[地域名]　名物　「[土産物]」’，観光名所リストの作成には，‘[地域名]にある観光名所「[観光名所]」’という表層パターンを使用した．その結果，土産物リストには地域名と土産物の対が482対，観光名所リストには地域名と観光名所の対が35,827対登録された．

次に，旅行ブログエントリに，情報抽出技術に基づいた機械学習を用いることで，新しい地域名と土産物の対，地域名と観光名所の対を得る．土産物情報の抽出のための機械学習の訓練用データは，以下の方法で準備した．

1．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’(地域名)と‘product’(土産物)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．[8]
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．

タグを付与した例を以下に示す．

> ＜location＞広島＜/location＞名物＜product＞もみじ饅頭＜/product＞をどうぞ．
>
> ＜location＞福岡＜/location＞では銘菓＜product＞ひよ子＜/product ＞．

同様にして，‘location’(地域名)と‘spot’(観光名所)タグを付与することで，観光名所情報の抽出のための訓練用データを作成した．タグを付与した例を以下に示す．

> この週末，＜location＞新潟県＜/location＞＜spot＞瀬波温泉＜/spot＞へ行ってきました．
>
> PTAのお友達4人で＜location＞京都＜/location＞の＜spot＞嵐山＜/spot＞に行きました．

本研究では機械学習としてCRFを使用した．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，土産物情報の抽出のときk＝2，観光名所情報の抽出のときk＝4と定めた．また，機械学習には以下の素性を使用した．

土産物情報の抽出の際に使用した素性

・単語
・品詞[9]
・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

7 http://www.gsk.or.jp/catalog/GSK2007-C/catalog.html
8 予備実験において，負例(地域名を含み土産物を含まない文)を含めずに機械学習を行った結果，得られた抽出器が全ての文から土産物を抽出しようとし，低い精度しか得られなかったため，本実験では，手順1の正例(地域名と土産物を含んだ文)と手順2の負例(地域名のみを含んだ文)の両方を訓練用データとして用いて学習を行った．
9 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．

観光名所情報の抽出の際に使用した素性

・単語
・品詞[10]
・単語に引用記号がついているかどうか
・単語が動詞かどうか
・単語が表層格かどうか
・‘博物館’や‘ランド’のような観光名所の名前によく使われる単語が含まれているかどうか
・‘市’や‘出張’などの不用語が含まれているかどうか

3.3 観光情報リンク集の自動構築

本節では，旅行ブログエントリからの観光情報リンク集を自動構築する手法について説明を行う．

3.3.1 リンク集構築の手順

リンク集構築の手順を以下に示す．

1．旅行ブログエントリのテキストを入力する．
2．入力テキストからリンク部分を見つけて，そのリンクに関する情報が記述されている文(引用個所)を抽出する．
3．引用個所を用いてリンクタイプの判定を行う．
4．システムが判定したリンクタイプの結果と，人手で判定したリンクタイプの結果の比較を行う．
5．システムによるリンクタイプ判定の精度，再現率を出力する．

3.3.2 引用個所の抽出

引用個所の抽出について説明を行う．リンクに関する情報は，リンクの周辺に記述される傾向があるが，リンクから離れた場所にも記述される場合もある．よって本研究では，手掛かり語により，引用個所を自動で抽出する．サイトを紹介する際には，リンク先サイトのタイトルが‘「」’や‘『』’などの記号で囲まれている場合がある．また，‘紹介’，‘のHP’などの語が使われるため，これらを手掛かり語として使用した．以下に，手掛かり語と，手掛かり語を用いた引用個所の抽出ルールを示す．

手掛かり語 (26個)

・‘「」’，‘『』’などリンク先サイトのタイトル周辺に使用される記号 (6個)
・‘紹介’，‘HP’，‘公式サイト’，‘こちら’などリンク先サイトを紹介する際に使用される単語(20個)

引用個所の抽出ルール

1．リンクが含まれている文を抽出する．
2．リンクが含まれている文の前後X文を抽出する．(予備実験よりX＝2とする．)
3．リンク先サイトを指し示す語(Keyword)を，リンクが含まれている文，リンク前後X文から抽出する．
   ・手掛かり語(記号)が含まれていれば，記号で囲まれている文字列をKeywordとする．
   ・手掛かり語(単語)が含まれていれば，手掛かり語の周辺の文字列をKeywordとする．
4．Keywordが含まれている文を抽出する．

次のエントリを用いて，引用個所抽出ルールを説明する．

1 チェックアウト後，いつものようにパパ&ママの寄り道が始まります!!
2 ということで，まずは河津の【バガテル公園】に行ってきました☆
3 四季の蔵から，車で数分圏内にあります．
4 ワンコもお散歩OKなので，犬連れには嬉しい場所です
5 メッチャ，綺麗でしたよ〜♪
6 ※バガテル公園のHPは，こちら→
7 http://www.bagatelle.co.jp/index.html
8 ↑いうまでもなく，美しいバラの数々(写真)
9 四季の蔵の朝ごはんがボリューム満点だから，これくらいで充分です!!
10 初めて来たバガテル公園ですが，ワンコOKだし，
11 季節によってはお花が綺麗なのでいいかも〜♪
12 ランチメニューも充実しているし，また今度も来ようっと(ノ∇≦*)キャハッッッ♪

ルール1により，リンクが記述されている7文目を引用個所として抽出する．次に，ルール2により，リンクが含まれている文の前後2文(5，6，8，9文目)を引用個所として抽出する．ルール3により，6文目に‘のHP’という手掛かり語が含まれているため，‘のHP’の直前の単語である‘バガテル公園’をKeywordとする．ルール4により，Keywordである‘バガテル公園’という単語が含まれている2，10文目を引用個所として抽出する．よって，上記のエントリから抽出される引用個所は，2, 5, 6, 7, 8 , 9, 10文目である．

10 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．

3.3.3 リンクタイプ

リンクタイプは以下のように判定する．

（1）S（Spot）
旅行者が訪れた名所，施設に関する情報（歴史，生息する動物など）かどうか．

（2）H（Hotel）
旅行者が宿泊したホテルや宿に関する情報かどうか．

（3）R（Restaurant）
旅行者が食事をとったレストラン，食べ物，食べ物を販売するお店に関する情報かどうか．

餃子スタジアムやたこせんべいの里などは，食を売りにした観光スポットであるため，リンクタイプはSとR両方に判定される．このように各リンクは複数のタイプに判定される場合もある．

S,H,Rのいずれにも判定されないものをOとする．Oに判定されたリンクには以下のようなものがある．

・旅行に持っていくために購入したデジタルカメラのサイトへのリンク
・車を運転する際のモラルを掲載したサイトへのリンク

3.3.4 リンクタイプの判定

本研究では，機械学習によりリンクタイプの判定を行う．学習には，「引用個所に出現する各単語」，「手掛かり語の有無」を素性として与える．

リンクタイプSのリンク周辺には，観光名所の名前や，‘観光’，‘見学’，‘訪れる’など，旅行者が観光名所に訪れた際によく使う単語が頻繁に出現すると考えられる．このような手掛かり語をWikipediaなどのWebページから収集しリストを作成した．R，Hについても同様の観点から手掛かり語の収集を行った．

(1) Sの手掛かり語　(17,812個)
・Wikipediaから収集した観光名所の名前（17,371個）
・‘動物園’や‘博物館’など観光名所の名前に使用される単語（138個）
・‘見学’や‘散策’など観光の際に使用される単語（172個）
・その他（131語）

(2) Hの手掛かり語　(73個)
・‘ホテル’や‘旅館’など宿泊施設の名前に使用される単語（9個）
・‘フロント’，‘客室’などの宿泊施設の構成要素（29個）
・‘泊る’や‘チェックイン’など宿泊する際に使用される単語（14個）
・その他（21個）

(3) Rの手掛かり語　(3,028個)
・Wikipediaから収集した料理名（2,779個）
・Wikipediaから収集した料理の種類（114個）
・‘レストラン’や‘食堂’など食事をとる施設の名前に使用される単語（21個）
・‘食べる’や‘おいしい’など食事をとる際に使用される単語（52個）
・‘ご飯’や‘料理’など，食べ物を指す単語（31個）
・その他（31個）

4．実験

3節で述べた提案手法の有効性を確かめるため，実験を行った．4.1節では，一般ブログから旅行ブログエントリを検出する手法，4.2節では，旅行ブログエントリから土産物情報・観光名所情報を抽出する手法，4.3節では，旅行ブログエントリから観光情報リンク集を自動構築する手法の実験について説明を行う．

4.1 旅行ブログエントリの自動検出

実験に用いるデータ

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エントリかどうかを判定した．その結果，‘旅行ブログエントリ’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．

機械学習と評価尺度

機械学習器にはCRF++[11]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

旅行ブログエントリの検出を系列ラベリング問題として解く手法を提案した．この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エントリの検出を行った．

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．旅行ブログエントリの検出の精度が低いと，観光情報を抽出する際の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．

11 http://www.chasen.org/~taku/software/CRF++/

表1 旅行ブログエントリの検出

	精度	再現率
提案手法	86.7	38.1
比較手法	60.5	51.1

人手では‘旅行ブログエントリ’と判定したが，提案手法では‘旅行ブログエントリでない’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‘旅行ブログエントリ’と判定できなければ，残りのエントリも‘旅行ブログエントリ’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エントリにあてはめ，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‘旅行ブログエントリではない’と判定したが，提案手法では‘旅行ブログエントリ’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エントリが存在(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エントリが存在(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エントリが存在していた．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．これはブログ著者の性別，年齢，居住区などの属性を文体から自動推定する研究の成果を利用することで，解決できるのではないかと考えられる．

(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‘旅行ブログエントリでない’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‘旅行ブログエントリ’と判定されてしまった．

4.2 土産物情報・観光名所情報の自動抽出

実験に用いるデータ

旅行ブログエントリ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つの情報源を用いて観光情報を抽出する．

旅行ブログ(提案手法)：3.1節の手法により，日本語で書かれた約1,100,000エントリから，旅行ブログエントリとして検出した17,266旅行ブログエントリ中の全ての文(80,000文)
一般ブログ：約1,100,000ブログエントリから選択した任意の80,000文
Web文書：ウェブ5億文データベース[17]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土産物情報・観光名所情報を抽出し，出現頻度によりランク付けを行った．

評価尺度

評価尺度としては，上位にランク付けられた土産物情報，観光名所情報に対して，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．

$$\text{精度}=\frac{\text{正しく抽出された地域名と(土産物/観光名所)の対}}{\text{抽出された地域名と(土産物/観光名所)の対}}$$

実験結果と考察

土産物情報・観光名所情報の上位100種類の抽出結

果を図２，図３にそれぞれ示す．土産物情報の抽出において，旅行ブログ手法(提案手法)では74.0%，Web文書手法では7.0%，一般ブログ手法では20.0%の精度を得た．また，観光名所の抽出において，旅行ブログ手法(提案手法)では71.0%，Web文書手法では37.0%，一般ブログ手法では31.0%の精度を得た．土産物情報・観光名所情報の抽出において，旅行ブログ手法(提案手法)は，Web文書手法や一般ブログ手法に比べ，高い精度を得ることができた．よって旅行ブログエントリは，観光情報の抽出のための有益な情報源であるといえる．

Google N-gramデータベースから作成した土産物，観光名所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新しく抽出された土産物，観光名所の種類を表２に示す．

土産物情報の抽出において，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，Web文書手法では７種類，一般ブログ手法では15種類であった．また，観光情報の抽出において，旅行ブログ手法では32種類の観光名所を抽出することができた．一方で，Web文書手法では24種類，一般ブログ手法では16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エントリの有益性を示せたといえる．

次に，(1)土産物情報の抽出，(2)観光情報の抽出における抽出誤りについて考察を行う．

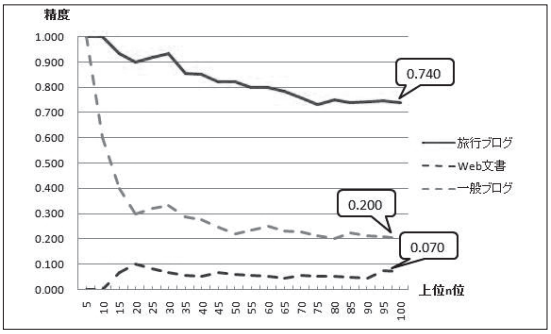

図２　上位ｎ位の土産物情報の抽出精度

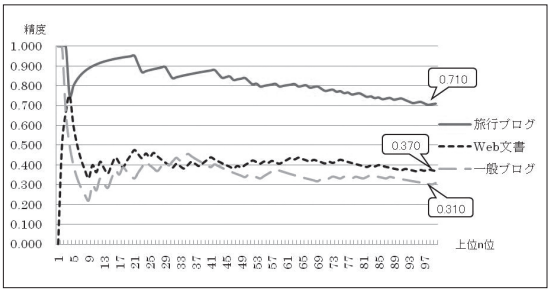

図３　上位ｎ位の観光名所情報の抽出精度

表２　各手法で新しく抽出された土産物と観光名所

	土産物	観光名所
旅行ブログ(提案手法)	41	32
Web 文書	7	24
一般ブログ	15	16

(1) 土産物情報の抽出

土産物情報の抽出における提案手法の上位100位間の典型的な抽出誤りは，土産物として土産物の販売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．また，土産物の販売店と土産物は，地域名と土産物と似たパターンで記述されることがあるため，誤って抽出されたと考えられる．この問題は，土産物とその土産物の販売店の対を抽出することで解決できると考えられる．

(2) 観光名所情報の抽出

観光名所情報の抽出における提案手法の上位100位間の典型的な抽出誤りについて考察を行う．抽出誤りの例として，イタリアの観光名所である‘ピサの斜塔’のように，観光名所に地域名が含まれている場合に，地域名として‘ピサ’，観光名所として‘斜塔’が抽出されるという誤りがあった．また，地域名‘日本’，観光名所‘草津温泉’という対のように，誤りではないが地域名の範囲が適切でないものもあった．これは，旅行記を記述したブロガーの居住地域により，地域名の表現が異なるために起こったと考えられる．これは観光情報の利用者によって，地域名の範囲を限定することで解決できると考えられる．

4.3　観光情報リンク集の自動構築

実験に用いるデータ

3.1節の手法により，日本語で書かれた約1,100,000エントリから，旅行ブログエントリとして17,266件のエントリを検出した．これらの旅行ブログエントリには7,421件のリンクが含まれていた．リンクの中には，Wikipediaやブログ，ニュースサイトへのリンクなど，リンク先 URL からリンク先サイトを判定することができるものも含まれている．よって本研究では，そのようなリンクを除外した4,155件のリンクから，1,000件のリンクを抽出し，人手でリンクタイプの判定を行った結果を機械学習に用いる．人手でリンクタイプの判定を行った結果を表３に示す．提案手法

表 3 1,000 件のリンクに含まれる各タイプの件数

リンクタイプ	S	H	R	O
リンク件数	353	98	343	250

の有効性を確かめるため，リンクが含まれている文の前後X文を引用個所として比較実験を行う．

提案手法：引用個所の抽出ルールにより，抽出された文を引用個所として使用．
比較手法：リンクの前後 X 文を，引用個所として使用．

機械学習と評価尺度

リンクタイプの判定の学習にはTinySVMを用いた．2次の多項式カーネルを使用し，4 分割交差検定を行った．また，精度と再現率を用いて評価を行った．

実験結果と考察

提案手法による実験結果を表 4 に示す．比較手法では引用個所として，リンクの前後 X 文を使用した．ここでは，最も実験結果の良い X＝ 2 のときの結果を示す．

表 4 実験結果

リンクタイプ	提案手法		比較手法	
	精度	再現率	精度	再現率
S	72.7	62.5	64.7	54.5
H	81.3	64.9	79.8	63.3
R	76.7	71.9	76.0	72.3
O	48.6	71.6	42.2	59.2

上記の実験結果より，比較手法に比べ，提案手法のリンクタイプ R の再現率が若干低下したが，その他では，提案手法が精度・再現率ともに，高い数値を記録することができた．特に，リンクタイプ S において，精度8.0ポイント(72.7-64.7)，再現率8.0ポイント(62.5-54.5)の改善を行うことができた．よって提案手法の有効性を示せたといえる．

提案手法，比較手法ともに，他のリンクタイプに比べ O の精度が低くなってしまったのは，S, H, R のいずれにも判定されないものを O としたためである．S, H, R の更なる精度の向上により，O の精度も改善できると考えられる．

次の各段階に分けて，提案手法を用いた際の，リンクタイプの判定誤りの原因ついて考察を行う．

(1) 引用個所の抽出
(2) リンクタイプの判定

以下に，それぞれの段階について説明する．

(1) 引用個所の抽出

本研究では，まず旅行ブログエントリから引用個所を抽出し，リンクタイプの判定を行っている．そのため，引用個所の抽出に誤りがあった場合に，リンクタイプを正しく判定することができないものがあった．引用個所の抽出には，以下の 2 つの問題がある．

(1-1) 引用個所の抽出不足

本研究では人手で収集した手掛かり語を用いて，引用個所の抽出を行った．しかし，リンクに関する情報が記述されている文を，引用個所として抽出できていない場合があった．

1 つ目の原因として，リンクの紹介方法がブロガーにより大きく異なるため，人手で収集した手掛かり語では対応できなかったことが挙げられる．これは，リンク周辺に出現する単語を収集し，手掛かり語を網羅的に集めることで解決できると考えられる．

2 つ目の原因として，手掛かり語に頼った抽出手法では，文間の語彙的なつながりを見つけることが困難であることが挙げられる．これは，引用個所の抽出に，語彙的伝搬の情報を加えることで解決できると考えられる．

(1-2) 引用個所の過抽出

旅行ブログエントリにリンクが連続して出現している場合，他のリンクに関する情報を，ターゲットとしているリンクの引用個所として抽出する場合があった．また，リンクの直前や直後に，リンクに関係のない記述があるときに，その文を引用個所として抽出してしまう場合があった．このような場合は，他のリンクとの距離や，リンクの前後の文の語彙的なつながりを考慮に入れることで解決できると考えられる．

(2) リンクタイプの判定

次に，各リンクタイプにおける判定誤りについて考察を行う．

人手では‘リンクタイプ X’と判定したが，提案手法では，‘リンクタイプ X でない’と誤って判定したリンクについて考察を行う．(X＝S, H, R) 以下に，判定誤りの主要な原因を示す．

(2-1) リンク先サイトに関する記述内容の不足

（2-2）手掛かり語の不足

(2-1) リンク先サイトに関する記述の不足

判定誤りの主な原因として，リンク先サイトに関する記述が少ない場合に，判定を誤っていた．本研究では，手掛かり語を用いた手法を提案したが，リンク先サイトに関する記述が不足していると，手掛かりとなる語が含まれておらず，提案手法では正しく判定できなかったと考えられる．

(2-2) 手掛かり語の不足

本研究では，人手により収集した手掛かり語を用いて，リンクタイプの判定手法を提案した．リンクタイプの判定誤りの原因として，手掛かり語の不足が考えられる．例として，リンクタイプをRと判定する場合を挙げる．

リンクタイプをRと判定する際の手掛かり語として，‘おいしい’など食事をとる際に使用される単語を使用した．しかし本研究では，旅行ブログエントリを情報源として使用しているため，同じ‘おいしい’という意味でも‘おいしー’，‘おいし～’，‘美味しい’，‘オイシイ’など様々な記述が存在する．このため，人手により手掛かり語を網羅的に収集するのは困難である．この問題を解決する手法として，レストランの口コミサイトなどの口コミを利用することで，より多くの手掛かり語を収集することが考えられる．

人手では‘リンクタイプXでない’と判定したが，提案手法では，‘リンクタイプX’と誤って判定したリンクについて考察を行う．（X＝S, H, R）以下に，判定誤りの主要な原因を示す．

（2-3）周辺施設に関する記述が存在

（2-4）手掛かり語の重複

(2-3) 周辺施設に関する記述が存在

リンク先サイトを紹介する際に，タイプの異なる周辺施設を紹介する記述が存在した場合に，判定を誤っている場合があった．例えば，リンク先サイトがホテルに関するサイトであり，人手でリンクタイプはSでないと判定されていたとする．このとき，リンク周辺に，ホテルの部屋から眺める観光名所に関する情報が記述されていた場合に，リンクタイプSであると誤って判定されていた．

(2-4) 手掛かり語の重複

本研究では，各リンクタイプのリンク周辺に出現しやすい単語を，人手で収集し手掛かり語として使用した．リンクタイプSの手掛かり語として，‘訪れた’という語を登録している．しかし，‘訪れた’という語は，レストランに食事に行った際にもよく使われる単語であるため，リンクタイプRの手掛かり語としても登録している．そのためリンク周辺に‘訪れた’という記述があった場合に，誤って判定されてしまった．この問題は，リンクタイプを判定したリンク周辺に出現するNグラムを使用し，各リンクタイプに特化した手掛かり語を自動的に収集することで解決できると考えられる．

5．おわりに

本研究では，ブログから観光情報を自動抽出するための手法を提案した．旅行ブログエントリの検出に関しては，再現率38.1%，精度86.7%を得た．また，旅行ブログエントリからの観光情報の抽出においては，抽出された上位100件の土産物において精度74.0%，抽出された上位100件の観光名所において精度71.0%を得ることができたため，旅行ブログエントリは観光情報の有益な情報源であるといえる．旅行ブログエントリからの観光情報リンク集の自動構築においても，高い精度・再現率を得られており，提案手法の有効性を示すことができたと言える．また，これらの手法と，ブロガーの属性(性別，居住地域など)を抽出する技術を利用することで，利用者に適した観光情報を提供することができようになると考えられる．

参 考 文 献

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

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

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

[4] 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.

[5] 大槻洋輔，佐藤理史．地域情報ウェブディレクトリの自動編集．情報処理学会論文誌，Vol.42，No.9，pp.2310-2318，2001．

［6］ 佐藤理史. ワールドワイドウェブを利用した住所検索. 情報処理学会論文誌, Vol.42, No.1, pp.59-67, 2001.
［7］ 村山紀文, 南野朋之, 奥村学. メタデータ付与のための住所録自動生成. 言語処理学会, 第11回年次大会, pp.53-56, 2005.
［8］ 安田宜仁, 戸田浩之. 検索位置のごく周辺を対象とした地理情報検索. 人工知能学会論文誌, 23巻5号C, pp.364-373, 2008.
［9］ 相良毅, 喜連川優. Webからの効率的な新規店舗の発見・登録支援手法. 情報処理学会論文誌, Vol.48, No.SIG_11(TOD_34), pp.49-57, 2007.
［10］ 岡本昌之, 菊池匡晃. ブログからの地域イベント情報抽出. 情報処理, Vol.51, No. 1, pp.14-17, 2010.
［11］ 藤坂達也, 李龍, 角谷和俊. 地域イベント発見および特性検証のための実空間マイクロブログを用いたユーザ移動パターン分析システム. 情報処理学会創立50周年記念（第72回）全国大会, pp.845-846, 2010.
［12］ 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.
［13］ Dmitry Davidov. Geomining: Discovery of Road and Transport Networks Using Directional Patterns. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp.267-175, 2009.
［14］ Anubhav Kale, Amit Karandikar, Pranam Kolari, Akshay Java Tim Finin, and Anupam Joshi. Modeling Trust and Influence in the Blogosphere Using Link Polarity. International Conference on Weblogs and Social Media, 2007.
［15］ Justin Martineau, Matthew Hurst. Blog Link Classification. Proceedings of International Conference on Weblogs and Social Media, 2008.
［16］ 戸田浩之, 安田宜仁, 奥村学, 松浦由美子, 片岡良治. 地理情報検索のためのスニペット生成法. 人工知能学会論文誌, Vol. 24, No.6, pp.494-506, 2009.
［17］ Daisuke Kawahara, Sadao Kurohashi. 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, 2006.

（20　年　月　日　受付）
（20　年　月　日　採録）

［問い合わせ先］
〒731-3194　広島市安佐南区大塚東3-4-1
広島市立大学大学院 情報科学研究科 知能工学専攻
石野　亜耶
TEL：082-830-1584
FAX：082-830-1584
E-mail：ishino@Is.info.hiroshima-cu.ac.jp

著 者 紹 介

石野 亜耶［非会員］

2009年広島市立大学情報科学部知能情報システム工学科卒業．現在，同大学大学院同研究科博士前期課程在学中．

難波 英嗣［非会員］

1996年東京理科大学理工学部電気工学科卒業．1998年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．2001年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了．同年，日本学術振興会特別研究員．2002年東京工業大学精密工学研究所助手．同年，広島市立大学情報科学部講師．2010年広島市立大学大学院情報科学研究科准教授．現在に至る．博士(情報科学)．テキストマイニング，情報検索，自動要約，特許情報処理に関する研究に従事．言語処理学会，情報処理学会，人工知能学会，ACL，ACM各会員．

竹澤 寿幸［非会員］

1984年早稲田大学理工学部電気工学科卒業．1989年同大学大学院理工学研究科博士後期課程修了．工学博士．同年(株)国際電気通信基礎技術研究所入社．音声対話翻訳の研究開発に従事．2007年広島市立大学大学院情報科学研究科教授．知能工学専攻言語音声メディア工学研究室に所属．現在に至る．2006年電子情報通信学会ISS論文賞受賞．電子情報通信学会，情報処理学会，人工知能学会，日本音響学会，言語処理学会各会員．

Automatic Compilation of Travel Information from Automatically Identified Travel Blog Entries

by

Aya ISHINO, Hidetsugu NANBA and Toshiyuki TAKEZAWA

Abstract：

In this paper, we propose a method for compiling travel information automatically. For the compilation, we focus on travel blog entries, which are defined as travel journals written by bloggers in diary form. We consider that travel blog entries are a useful information source for obtaining travel information, because many bloggers' travel experiences are written in this form. First, we identified travel blog entries in a blog database. Next, we extracted souvenir information and tourist spots information as travel information from them. Furthermore, we extracted hyperlinks from travel blog entries and constructed the collection of travel information links. We have confirmed the effectiveness of our method by experiment. For the identification of travel blog entries, we obtained scores of 38.1% for Recall and 86.7% for Precision. In the extraction of travel information from travel blog entries, we obtained 74.0% and 71.0% for Precisions at the top 100 extracted local products and tourist spots, respectively, and thereby confirming that travel blog entries are a useful source of travel information. In the construction of the collection of travel information links, we obtained high precision and recall.

Keywords：Blog, Information Extraction, Travel Information

Contact Address： **Aya ISHINO**

Dept. of Intelligent Systems, Graduate School of Information Sciences, Hiroshima City University

3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima, JAPAN

TEL ：082-830-1584

FAX ：082-830-1584

E-mail ：ishino@Is.info.hiroshima-cu.ac.jp