旅行口コミサイトからの旅行ノウハウ情報の自動抽出

石川 綾美† 難波 英嗣†† 石野 亜耶††† 竹澤 寿幸††

†広島市立大学 情報科学部 〒731-3194 広島県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 〒731-3194 広島県広島市安佐南区大塚東 3-4-1
†††広島経済大学ビジネス情報学科 〒731-0138 広島県広島市安佐南区祇園 5-37-1

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

あらまし 旅行ノウハウ情報とは，ある場所を訪れた者が経験的に得た情報であり，こうした情報は，その場所を訪れる旅行者にとって，旅先のトラブルを避け，また，充実した旅行にするために有益である．しかし，この旅行ノウハウ情報を人手で整備するには，多大なコストがかかる．そこで，本研究では，旅行口コミサイトから，旅行ノウハウ情報を自動的に抽出・整備する手法を提案する．旅行者によって書かれた口コミから旅行ノウハウ情報を抽出することで，その場所をこれから訪れる旅行者は，経験的にしか得ることのできない情報を事前に得ることができ，旅行の手助けができると考えられる．

キーワード ノウハウ，文書分類，旅行

1. はじめに

近年，情報網の発達により，個人の意見を発信する機会が増えるに伴い，旅行口コミサイトを通して，旅行先での個人的な体験を発信する人が，国を問わず増えている．そのような口コミの中には，観光スポットに関する感想などの他に，おすすめ情報，注意すべきこと，アドバイスなど，その観光スポットを訪れた者が経験的に得た情報が含まれている．本研究ではこのような，観光スポットを訪れた者が経験的に得た情報を，旅行ノウハウ情報と呼ぶ．旅行ノウハウ情報には以下のような例がある．

- 「宮島で食べ歩きをしていると鹿に食べ物を取られるので気を付けてください」
- 「広島の代表的な料理のひとつであるお好み焼きには，材料として豚肉が使われているのでイスラム教徒の旅行者は気を付けてください」

一つ目の例については，食べ歩きをしている時は鹿に注意した方がいいという情報をあらかじめ知っていれば，鹿に注意しながら宮島の散策を楽しむことができる．二つ目の例については，お好み焼きには豚肉が使われているという情報であり，豚肉を禁忌するイスラム教徒にとって非常に重要な情報である．

上記の例のように，旅行ノウハウ情報は，これから同じ観光スポットを訪れる旅行者が，事前に旅行ノウハウ情報を得ることにより，旅行先でのトラブルを避け，快適で充実した旅行をするために非常に有益な情報である．しかし，旅行ノウハウ情報を人手で抽出するには，多大なコストを要する．

そこで本研究では，旅行口コミサイトから，旅行ノウハウ情報を自動的に抽出する手法を提案する．本研究では，日本語と英語で記述された口コミを対象に旅行ノウハウ情報の抽出を行う．旅行ノウハウ情報を抽出することで，その場所をこれから訪れる旅行者は，経験的にしか得ることのできない情報を事前に得ることができ，旅行の手助けができると考えられる．

本論文の構成は，以下に示す通りである．2 節では，本研究に関連する研究を述べる．3 節では，旅行ノウハウ情報の抽出方法を述べる．4 節では，実験について述べる．5 節では，実験結果に基づいた考察を述べる．6 節で本研究をまとめ，7 節では，今後の課題について述べる．

2. 関連研究

本節では，本研究に関連する研究を紹介する．本研究では，旅行口コミサイトから，旅行ノウハウ情報を自動的に抽出する手法を提案している．本研究と同様に，Web テキストから情報を抽出する研究がある．阿部ら[1]の研究に，経験マイニングと呼ばれる，商品やサービスなど，様々な事物（トピック）の利用に関するセンチメントの情報やセンチメントを暗に含むイベントの情報を広く Web 文書集合から抽出し，意味的な索引付けを行う技術がある．経験マイニングは，経験を分類する基準として特定の利用シーンに特化した基準を仮定するのではなく，事態タイプや事実性のような一般性の高い意味的なタグによって経験情報を索引付けする．次に，倉島ら[2]は，ブログに記述された人間の経験を構造化する試みとして，経験を構成する 5 要素（時間，空間，動作，対象，感情）をブログから抽出し，さらに「成功/失敗は主に動作主の感情に因る」として，それぞれの経験情報に成功/失敗要素を付与する手法を提案している．また，人間の感情を８カテゴリに分類することで，それぞれの経験情報が，動作主にとって成功だったか，失敗だったかを導き出している．さらに，得られた大量の経験情報集合から，相関ルール抽出技術を用いて状況と行動と主観との関係をルール形式で抽出している．本研究においても，Web

からの情報抽出を行なっているが，抽出する情報を旅行ノウハウ情報としている点で異なる．

本研究と同様，Web からノウハウを獲得する研究がある．小澤ら[3]は，ノウハウには少なくとも 1 つはモノが含まれていることが多く，それらの利用がノウハウにおいて重要な役割を果たしている点に着目し，モノとその使われ方によりノウハウを獲得している．Web からモノを含むパッセージを獲得し，モノの用途表現と手がかり表現パターンを用いて，ノウハウを獲得する手法を提案している．ここで，小澤らは，物事の手順やアドバイスをノウハウとしている．次に，守谷ら[5]は，質問回答サイトおよびウェブからノウハウ知識を相補的に収集する手法を提案している．この研究では，ある検索対象についてのノウハウ知識を網羅的に収集し，集約・俯瞰する手法を確立している．質問回答サイトには多くのノウハウ知識が含まれているが，質問回答サイトだけでは十分でないことが想定されるため，守谷らは質問回答サイトおよび一般のウェブページという二種類の情報源を併用することにより，ノウハウ知識を相補的に収集し，集約・俯瞰する手法を提案している．次に，服部ら[6]は，耳より情報の抽出を行い,その耳より情報のタイプ分類を行っている．まず，SNS からコメントを取得し，LDA を用いて単語数の少ない（データが疎な）多数のコメントからトピックの推定を行う．そしてトピック毎にクラスタリングされたコメントから経験情報を述べているコメントを取得する．次に予備実験により抽出した耳より情報の要因となるキーワードの辞書をあらかじめ作成し，その辞書を用いることにより経験情報から耳より情報の抽出を行っている．さらに，この耳より情報を「提案・推薦」,「制止・抑制」,「現状・状況説明」,「可能・不可能」の 4 つのタイプに分類している．本研究においても，Web からノウハウを獲得しているが，本研究では，手掛かり語を使用し，また，ある場所を訪れた者が経験的に得た情報をノウハウと定義し，ノウハウの抽出を行う点で異なる．

本研究と同様に，Web から観光情報を抽出する研究がある．石野ら[7]は，日本語で記述された旅行ブログエントリを使用し，自動的に観光情報を収集するための手法を提案している．観光情報の抽出手法は，ブログデータベースから旅行ブログエントリを検出し，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している．さらに，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ている．石野らの研究のような Web から観光情報を抽出する研究は様々あるが，本研究ではロコミに含まれている旅行ノウハウ情報に着目する．

ノウハウに着目した研究として Nanba ら[8]の研究がある．Nanba らは，英語旅行ブログから自動的にノウハウ，イベント情報を抽出することで，低コストでのデータベース生成を目指している．英語旅行ブログサイトに投稿されている英語旅行ブログの，“Japan”というカテゴリに分類されている旅行ブログを対象とし，ノウハウ，イベントブログを抽出する手法を提案している．Nanba らは，着物の着方などの手順を示す内容が含まれる部分をノウハウとしている．また,“酒祭り”のような行事の開催地，開催日，内容などの情報が記載されている部分をイベントとしている．英語旅行ブログからノウハウ，イベントブログを自動抽出する手法として，手掛かり語の有無を素性として機械学習を行っている．本研究では，旅行ロコミサイトを対象としている点,また,旅行ノウハウ情報の定義を，観光スポットを訪れた者が経験的に得た情報としている点で Nanba らの研究とは異なる．

3. 旅行ロコミサイトからの旅行ノウハウ情報の自動抽出

本節では，旅行ロコミサイトから旅行ノウハウ情報を抽出する手法について述べる．3.1 節では，旅行ノウハウ情報であるかそうでないかの人手による判定基準について説明する．3.2 節では，旅行ノウハウ情報の自動抽出手法ついて説明する．

3.1. 旅行ノウハウ情報の判定基準

本節では，旅行ノウハウ情報かどうかの人手による判定基準について述べる．本研究では，ある場所を訪れた者が経験的に得た情報を旅行ノウハウ情報と定義する．人手により旅行ノウハウ情報を含むと判定された日本語で記述されたロコミの例を図 1，英語で記述されたロコミの例を図 2 に示す．

知る人ぞ知る観光名所。
奈良の東大寺と同様、鹿がそこらじゅうにいるので、食べ歩きの際はつつかれないように注意しましょう。

図 1：旅行ノウハウ情報を含む
日本語で記述されたロコミの例

Cool Island with temple and tori in the middle of the water.
When low tide - I would recommend to walk to the tori.

図 2：旅行ノウハウ情報を含む
英語で記述されたロコミの例

図 1 は宮島について日本語で記述されたロコミである．波線が旅行ノウハウ情報であると判定された部分である.この文には,「注意」という単語含まれており，注意すべきことの情報となっている．よって，旅行ノ

ウハウ情報であると判定できる．

図 2 は宮島について英語で記述された口コミである．波線が旅行ノウハウ情報であると判定された部分である．この文には，「recommend」という単語が含まれており，おすすめ情報となっている．よって，旅行ノウハウ情報であると判定できる．

このように，おすすめ情報や注意すべきこと，アドバイスなど，その場所を訪れた者にしかわからない，有益な情報部分を旅行ノウハウ情報と判定する．

3.2. 旅行ノウハウ情報の自動抽出手法

本研究では，機械学習により旅行口コミサイトから旅行ノウハウ情報を自動抽出する手法を提案する．学習には，「口コミに出現する各単語」と「手掛かり語の有無」を素性として与える．

口コミには，「注意」，「おすすめ」など，旅行ノウハウ情報に頻繁に出現する単語がある．このような語を，日本語で記述された口コミ，英語で記述された口コミから収集し，手掛かり語とした．日本語の手掛かり語を表 1，英語の手掛かり語を表 2 に示す．

表 1：日本語の手掛かり語

• よ！	• でしょう	• ことができ
• よ。	• 可能	• しょう
• 注意	• OK	• くださいね
• 気を付け	• ＯＫ	• どうぞ
• 気をつけ	• しましょう	• からです
• きをつけ	• みたい	• 難しい
• 必要	• 時	• 厳し
• 残念	• ベスト	• 苦労
• 便利	• コツ	• ただし
• 方がいい	• ですので	• ようです
• ほうがいい	• いいかも	• しれません
• 方が良い	• 良いかも	• 難点
• 方がよい	• よいかも	• 断念
• ほうがよい	• チェック	• 十分
• ほうが良い	• 得	• ポイント
• 不満	• 欠点	• 利用
• ください	• 必ず	• 無料
• 方が	• 時間	• 大変
• ほうが	• 必須	• しまい
• 思います	• れる	• なる
• 思いました	• 危険	• 調べて
• (が、|ので、|為|ため) かつ (おすすめ|おススメ|お勧め|オススメ|お薦め)		

表 2：英語の手掛かり語

• recommend	• remember	• least
• recommended	• early	• beware
• careful	• want	• possible
• carefully	• stay	• section
• check	• watch	• holding
• holiday	• best	

4. 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と結果について述べる．4.1 節で実験方法，4.2 節で実験結果について述べる．

4.1. 実験手法

実験に用いるデータ

本実験では，旅行口コミサイト TripAdvisor[1]を利用した．TripAdvisor では，地域や観光スポットごとに口コミが投稿されている．実験には，しまなみ海道，縮景園，鞆の浦，マツダスタジアム，宮島の 5 件の日本国内の観光スポットに登録されている日本語の口コミ 1,353 件，英語の口コミ 1,114 件に含まれる文に対して，人手により旅行ノウハウ情報であるかを判定した結果を使用した．人手により旅行ノウハウ情報かどうかを判定した結果を表 3，表 4 に示す．

表 3：日本語の口コミから
人手により旅行ノウハウ情報を判定した結果

観光スポット	旅行ノウハウ情報でないと判定した文の数	旅行ノウハウ情報であると判定した文の数
しまなみ海道	568	45
縮景園	526	43
鞆の浦	756	55
マツダスタジアム	1,203	102
宮島	3,424	492
合計	6,477	737

[1] https://www.tripadvisor.jp/

表 4：英語の口コミから
人手により旅行ノウハウ情報を判定した結果

観光スポット	旅行ノウハウ情報でないと判定した文の数	旅行ノウハウ情報であると判定した文の数
しまなみ海道	470	104
縮景園	556	97
鞆の浦	37	13
マツダスタジアム	298	73
宮島	5,074	904
合計	6,435	1,191

機械学習

機械学習には，TinySVM を用いた．2 次の多項式カーネルを使用し，2 分割交差検定を行った．

評価尺度

評価尺度は，以下の示す精度と再現率を用いた．

$$\text{精度} = \frac{\text{システムが抽出した正解数}}{\text{システムが抽出した数}}$$

$$\text{再現率} = \frac{\text{システムが抽出した正解数}}{\text{人手で抽出した正解の数}}$$

比較手法

口コミに含まれる単語の出現回数を素性として機械学習を行った結果を比較手法とした．

提案手法

口コミに含まれる単語の出現回数と，手掛かり語の有無を素性として機械学習を行った結果を提案手法とした．

4.2. 実験結果

日本語の口コミから旅行ノウハウ情報を自動抽出した結果を表 5，英語の口コミから旅行ノウハウ情報を自動抽出した結果を表 6 に示す．

表 5：日本語の口コミからの
旅行ノウハウ情報の自動抽出結果

	精度	再現率
比較手法	0.18	0.28
提案手法	0.39	0.29

表 6：英語の口コミからの
旅行ノウハウ情報の自動抽出結果

	精度	再現率
比較手法	0.52	0.32
提案手法	0.55	0.35

5. 考察

4.2 節の実験結果について，考察を行う．まずは日本語の旅行ノウハウ情報抽出の実験結果について考察を行う．各文における人手による判定とシステムによる判定の結果を表 7 に示す．

表 7：日本語の旅行ノウハウ情報抽出における
人手による判定とシステムによる判定の結果

		人手による判定		合計
		旅行ノウハウ情報である	旅行ノウハウ情報でない	
システムによる判定	旅行ノウハウ情報である	213	327	540
	旅行ノウハウ情報でない	524	6,150	6,674
合計		737	6,477	7,214

実験結果により，人手によって旅行ノウハウ情報であると判定された 737 文のうち，213 文がシステムで正しく判定された．正しく判定された文の例を図 5，図 6 に示す．下線部が素性である．

島には鹿が多くいますが、すぐに寄ってくるので、食べ物をもっていたら注意した方がいいですよ。

図 3：正しく判定された例（宮島）

図 3 において，「注意」，「方がいい」，「よ。」の 3 つの素性が含まれている．このように，おすすめ情報や注意すべきこと，アドバイスなどに関する素性が一文の中に複数含まれていたため，システムが正しく旅行ノウハウ情報であると判断できたと考えられる．

実験結果により，システムによって旅行ノウハウ情報であると判定された 540 文のうち，213 文が，人手で判定した結果と一致した．しかし，人手により旅行ノウハウ情報でないと判定された 6,477 文のうち，327 文がシステムにより誤って検出された．以下に，検出誤りと再現誤りについて述べる．

✧　システムが旅行ノウハウ情報であると誤って判定した場合

以下に，人手では旅行ノウハウ情報でないと判定したが，システムでは旅行ノウハウ情報であると判定した 327 文の検出誤りを分析し，主要な原因をいくつか示す．

しかし、近年観光化によって自家用車が押し寄せて岬の細い道が混雑するというので、岬をまたぐような橋の計画が立てられて、賛否両論です。

図 4：検出誤り例 1（鞆の浦）

図 4 において，「られ」という素性が含まれている．「られ」は，旅行ノウハウ情報に頻出する単語である．しかし，「られ」という単語のみであると，旅行ノウハウ情報でない文にも多く含まれている可能性がある．そのため，単語ひとつを素性とするだけではなく，単語と単語を組み合わせた素性の追加が必要である．

干潮の時には鳥居の下までくぐりに行っていた方が相当数いましたが、足元は大丈夫なのかな？

図 5：検出誤り例 2（宮島）

図 5 において，「方が」という素性が含まれている．「方が」を素性としているが，読み方としては，「ほうが」を想定していた．しかし，図 5 においては，「かたが」という読み方で利用されている．このように，漢字の読み方によって素性の意味が変わってしまったために，システムが誤って検出してしまったと考えられる．

外国人観光客がけっこう多いですよ。

図 6：検出誤り例 3（宮島）

図 6 において，「よ。」という素性が含まれている．しかし，旅行ノウハウ情報であるとは言えない．この文は，旅行ノウハウ情報であると判定したものに比べて，一文の長さが短いということが言える．このことから，旅行ノウハウ情報を抽出する際に，文の長さも旅行ノウハウ情報であるか判定する手がかりとする必要がある．

✧　システムが旅行ノウハウ情報でないと誤って判定した場合

以下に，人手では旅行ノウハウ情報であると判定したが，システムでは旅行ノウハウ情報でないと誤って判定した 524 文の再現誤りを分析する．

歩き疲れたら、鞆シーサイドホテルの日帰り入浴がお勧めです。

図 7：再現誤り例 1（鞆の浦）

図 7 おいて，「おすすめ」という素性が含まれており，旅行ノウハウ情報であると言える．おすすめのみを素性とすると，「宮島は本当におすすめです。」のような，スポット自体をおすすめしている場合も素性が含まれてしまう．そのため，本研究では，「(が、|ので、|為|ため)かつ(おすすめ|おススメ|お勧め|オススメ|お薦め)」を素性とし，スポット自体をおすすめしている場合を除けるようにしている．それにより，図 7 は旅行ノウハウ情報でないと判定されていると考えられる．これは，旅行ノウハウ情報である文で，おすすめと一緒によく出てくる単語などを再考する必要がある．

ボランティアがいて、時間内におさめて案内してくれます！

図 8：再現誤り例 2（縮景園）

図 8 において，素性が一つも含まれていない．このような例が多く存在していると考えられる．そのため，素性の追加，再考が必要であると考えられる．

次に，英語の旅行ノウハウ情報抽出の実験結果について考察を行う．各文における人手による判定とシステムによる判定の結果を表 8 に示す．

表 8：英語の旅行ノウハウ情報抽出における
人手による判定とシステムによる判定の結果

		人手による判定		合計
		旅行ノウハウ情報である	旅行ノウハウ情報でない	
システムによる判定	旅行ノウハウ情報である	413	341	754
	旅行ノウハウ情報でない	778	6,094	6,872
合計		1,191	6,435	7,626

実験結果により，人手によって旅行ノウハウ情報であると判定された 1,191 文のうち，413 文がシステムで正しく判定された．正しく判定された文の例を図 15 に示す．下線部が素性である．

I recommend going to Taichoro which is a building originally used as a guesthouse for visitors from Korea.

図 9：正しく判定された例（鞆の浦）

図 9 において,「recommend」という素性が含まれている．このように，おすすめ情報や注意すべきこと，アドバイスなどに関する素性が含まれていたため，システムが正しく旅行ノウハウ情報であると判断できたと考えられる．

実験結果により，システムによって旅行ノウハウ情報であると判定された 754 文のうち，413 文が，人手で判定した結果と一致した．しかし，人手により旅行ノウハウ情報でないと判定された 6,435 文のうち，341 文がシステムにより誤って検出された．以下に，検出誤りと再現誤りについて述べる．

✧ システムが旅行ノウハウ情報であると誤って判定した場合

以下に，人手では旅行ノウハウ情報でないと判定したが，システムでは旅行ノウハウ情報であると判定した 341 文の検出誤りを分析し，主要な原因を示す．

The stadium has fantastic food, decent beer, and the best fans!

図 10：検出誤り例 1（マツダスタジアム）

図 10 において,「best」という素性が含まれている．「best」は，旅行ノウハウ情報に頻出する単語である．しかし,「best」という単語のみであると，旅行ノウハウ情報でない文にも多く含まれている可能性がある．そのため，単語ひとつを素性とするだけではなく，単語と単語を組み合わせた素性の追加が必要である．

I recommend it!

図 11：検出誤り例 2（マツダスタジアム）

図 11 において,「recommend」という素性が含まれている．しかし，旅行ノウハウ情報であるとは言えない．これは，旅行ノウハウ情報であると判定したものに比べて,一文の長さが短いということが挙げられる．このことから，旅行ノウハウ情報を抽出する際に，文の長さも旅行ノウハウ情報であるか判定する手がかりとする必要がある．

✧ システムが旅行ノウハウ情報でないと誤って判定した場合

以下に，人手では旅行ノウハウ情報であると判定したが，システムでは旅行ノウハウ情報でないと誤って判定した 778 文の再現誤りを分析する．

All the info you need including maps you can get from the tourist office in Onomichi.

図 12：再現誤り例（しまなみ海道）

図 12 において，素性が一つも含まれていない．このような例が多く存在していると考えられる．そのため，素性の追加，再考が必要であると考えられる．

6. おわりに

本研究では，旅行口コミサイトから，旅行ノウハウ情報を自動的に抽出する手法を提案した．提案手法では，旅行ノウハウ情報に含まれやすい語を手掛かり語とし，機械学習を行った．その結果，日本語の旅行ノウハウ情報の自動抽出の場合は精度 0.39,再現率 0.29,英語の旅行ノウハウ情報の自動抽出の場合は精度 0.55,再現率 0.35 を得た．今後，精度，再現率を向上させるために，さらなる改善が必要である．

謝辞

本研究の一部は，総務省による戦略的情報通信研究開発推進制度（SCOPE）の支援を受けて行われた．

参 考 文 献

[1] 阿部修也，江口萌，隅田飛鳥，大崎梓，乾健太郎，“みんなの経験：ブログから抽出したイベントおよびセンチメントの DB 化”，言語処理学会第 15 回年次大会，pp.296-299，2009.

[2] 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 第 19 回データ工学ワークショップ(DEWS2008)論文集 A1-4，2008.

[3] 小澤俊介，内元清貴，松原茂樹，“モノの用途表現を手がかりとした Web からのノウハウの獲得”，情報処理学会研究報告，Vol.196，No.1，pp.1-7，2010.

[4] Marti A. Hearst，“TexTiling: Segmenting text into multi-paragraph subtopic passages”，Computational linguistics，Vol.23，No.1，pp. 33-64，1997.

[5] 守谷一朗，今田貴和，井上祐輔，聶添，宇津呂武仁，河田容英，神門典子，“質問回答事例および検索エンジン・サジェストを用いたノウハウ知識の相補的収集”，DEIM Forum 2015，2015.

[6] 服部祐基，灘本明代，“ソーシャルメディアからのタイプ別耳より情報の抽出手法の提案”，DEIM Forum 2013，2013.

[7] 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.

[8] H. Nanba，S. Douke and T. Takezawa，“Automatic Identification of Know-How Blog Entries from a Travel Blog Database”，Proceedings of the 25th International Conference on Advanced Tourism Informatics (ICATI2013)，2013.