Blogの表層的特徴と格フレームを利用したBlog文書からの訪問場所名詞の獲得

小林　卓弥†　　吉川　正俊††

† 京都大学工学部情報学科
〒 606-8501 京都市左京区吉田本町
†† 京都大学大学院情報学研究科社会情報学専攻
〒 606-8501 京都市左京区吉田本町
E-mail: †tkoba@db.soc.i.kyoto-u.ac.jp, ††yoshikawa@i.kyoto-u.ac.jp

あらまし　近年，ブログコンテンツなどから評判情報を抽出する研究が注目を浴びている．そのような評判の対象が商品でなく店舗や施設であった場合，実際の訪問に基づかない文書は評判情報の抽出対象として相応しくないと考えられる．そこで本論文では，ブログ文書における表層的特徴ならびに格フレームを利用し，与えられたブログ文書から訪問場所名詞を獲得する手法を提案する．評判情報を抽出する前処理として本手法を適用することで，一定の情報価値のあるブログから情報抽出が行えることが期待できる．

キーワード　ブログ, 自然言語処理, データマイニング

Acquisition of Visited Locative Nouns from Blogs by Using Surface Features of Blogs and Case Frames

Takuya KOBAYASHI† and Masatoshi YOSHIKAWA††

† Department of Informatics, Faculty of Engineering, Kyoto University
Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8501 Japan
†† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †tkoba@db.soc.i.kyoto-u.ac.jp, ††yoshikawa@i.kyoto-u.ac.jp

Abstract　Recently, research on extracting reputation information from Blog contents is emerging. In case of planimetrical feature reputation, documents not based on an actual visit are unsuitable as the extraction object of reputation information. This paper proposes a method of acquiring visit locative nouns from Blog documents by using surface features of Blogs and case frames. We expect that information extraction from Blogs with higher certainty will be possible by applying this technique before extracting reputation information.

Key words　Blog, Natural Language Processing, Data Mining

1. はじめに

近年，Web のネットワーク環境が整備され，Web2.0 [1] と呼ばれる形に進化したと言われている．Web2.0 とは，2004 年に Tim O'Reilly 氏が提唱した概念であり，従来とは異なる Web 関連の技術や，Web サイト・サービスなどの総称である．SNS やソーシャルブックマークなど，そのような Web2.0 的サービスとして注目を浴びているものの一つに「ブログ（Weblog）」がある．ブログは従来の Web ページに比べ作成・更新が容易であるため急速に広まっており，総務省の発表では平成 18 年 3 月末の時点で，ブログ登録者数は 868 万人となった [2]．この数が示すように，ブログに代表される個人が自由に情報を発信するメディアは CGM（Consumer Generated Media）と呼ばれる新しいメディアとして大きく成長しており，個人の実体験や生の声が Web 上でリアルタイムに投稿されている．

現在，対象をこのブログに限定して情報を探し出そうと，テクノラティ [3] に代表される様々なブログ検索サービスが登場している．一般的な検索エンジンでなくブログ検索エンジンを利用する状況として，例えば「あるキーワードに対して人々がどのように語っているかが知りたい」ということが考えられる．

ブログ記事は一般の Web ページに比べ，個人が発信する評判や感想情報が期待されるという性質を持っているためである．あるキーワードを含むブログ全てを一個人が読み切ることは現実的に不可能であることが多いため，この膨大なブログの集より評判情報を抽出しようと試みる研究が活発である．それらの研究の多くは，ブログ文書中に現れる形容詞をいかに評価対象の名詞と対応付けるか，またその評価の好悪を判断するものが主である [4], [5]．しかし，評価の対象が商品ではなく，店舗や施設といった地物であった場合，実際の訪問に基づかない文書は評判情報の抽出対象として相応しくないノイズになると考えられる．本論文では，ブログ記事の内容はすべて真実であるという前提の下で，ブログ記事から実際の訪問に基づく地物情報を抽出することを目的とする．

ブログ記事とは，各個人が自由な表現で文章や画像などを日記のような形式で投稿したものであり，その記述の様式は個人によって違いが大きい．その違いを吸収するために，我々は訪問場所名詞に関して表層的な特徴を利用することを考える．また，ブログ記事における文章は口語的な日本語であることが多いが，言語解析ツールなどの進化により，ある程度の水準で格フレーム辞書を利用することができる．

本論文では，表層的な特徴と場所名詞を格要素に持つ格フレームを利用することで，与えられたブログ文書から訪問場所名詞を獲得する手法を示す．地物に対する評判情報を抽出する前処理として本手法を適用することで，一定の情報価値のあるブログから情報抽出が行えることが期待できる．

本論文の以下では，2 節で関連研究について述べ，3 節で提案手法の詳細を記述する．4 節では実際のブログ記事に提案手法を適用し実験を試み，5 節でその実験結果を踏まえて提案手法について検証する．そして 6 節で提案手法の応用について考え，最後に 7 節でまとめと今後の課題を記述する．

2. 関 連 研 究

ブログからの場所名詞抽出を利用したサービスとして，maplog [6] が挙げられる．maplog は，表示された地図範囲について書かれているブログ記事を検索することができるサービスである．地図インタフェース上にブログをマッピングするために，自社で作成した「位置キーワード」と呼ばれる単語セットにヒットするブログを予め検索用データベースに格納している．地図インタフェースでブログ検索ができるため操作性に優れたシステムであるが,「位置キーワード」として挙げられているのは地名や駅名，ランドマーク名などであるため，単語セットに無い場所名詞はヒットしない．提案手法ではそのような名詞辞書を持たないため，辞書に無い未知語にも対応することができる．また，このように名詞辞書を人手で作成するのは多大なコストがかかる．ここが本研究との大きな違いの一つであり，さらに実際の訪問に基づく記事であるかどうかを判定する点がもう一つの違いである．

また，郡ら [7] はブロガーの行動経路を抽出する過程でブロガーが実際に訪れた地名を抽出する地名フィルタを作成している．その手法は，場所に関する深層格を表しうる格助詞と同じ文節に現れ，なおかつ動作動詞かサ変名詞に係っている名詞を全て列挙し，それから予め保持した地名辞書によりフィルタリングを行うというものである．そのため，地名辞書でフィルタリングを行う前は，例えば「ケーキを食べる」という文章から抽出された「ケーキ」という名詞も地名候補として含まれる．辞書を保持することでそのような名詞が抽出されることを防ぐが，このような場所名詞辞書を保持することは多大なコストである．

地名フィルタでは格表現の深層格とその係り受け関係までしか考慮していないが，我々はさらに格フレームを考慮する．なお，西田ら [8] はこの格フレームを利用し，料理教示発話の省略語獲得を試みている．我々はさらにブログに見られる表層的特徴も合わせて利用することで，地名辞書を保持せずに「訪問場所名詞」を抽出することを目指す．そのため，本手法では辞書にない未知語も抽出することができる．本手法により，地名フィルタにおける失敗例をある程度克服するフィルタを作成できると考えられる．

3. 提 案 手 法

本提案手法で対象とする「訪問場所名詞」は，日本語語彙大系 [9] の一般名詞意味体系で [/1 名詞/2 具体/388 場所] に分類されうる名詞とし,“ 日本 ”や“ 京都 ”など範囲の広すぎる名詞や，伏せ字が用いられているなど具体名でない名詞，また記事のみから具体名が得られない場合は抽出の対象外とする．

ここで，本手法は次の三つのステップから成る．

Step1: 「訪問場所名詞」候補を抽出し，それぞれに初期確信度を付与

Step2: Step1 で抽出したそれぞれの候補について条件を満足する場合，確信度を増大

Step3: Step2 で決定した確信度を判定し，出力する「訪問場所名詞」を決定

この概要図を図 1 に示す．次節以降でそれぞれのステップの詳細について述べる．

3.1 候補の抽出と初期確信度の付与

Step1 では「訪問場所名詞」候補の抽出ならびに初期確信度の付与を行う．その際，ブログの表層的特徴ならびに格フレームを考慮した属性を考える．なお，格フレームを考慮する際に使用する係り受け解析器は CaboCha [10] を使用する．まず，Step1 で使用する atr1〜atr5 の属性を以下に示す．これらの属

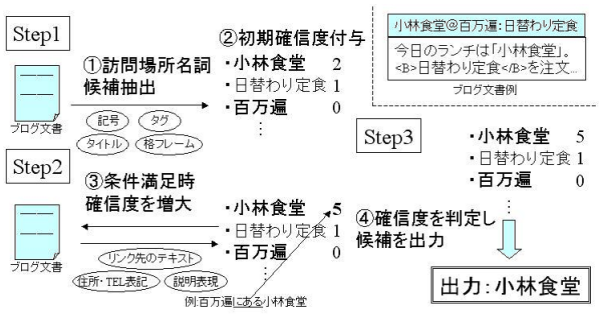

図 1 提案手法の概要図

性により,「訪問場所名詞」候補を抽出することが可能となる.

atr1: 強調表現

<B>タグやタグまたは<font>タグで囲われた語

atr2: 外部リンク

<A>タグで外部リンクが張られている語（<A href= ”http://xxx”>ＸＸＸ</A>からＸＸＸが抽出される．ただし，そのブログの他の記事などへの内部リンクであるものは対象外とし，href 属性が示すリンク先 URL (http://xxx) はそのブログ外の HP でなければならない）

atr3: 記号で囲われている

「」,『』などの記号で囲われている語

atr4: ブログ文書のタイトル

そのブログ文書のタイトル（ただし空白や@などの記号が表れる場合はそれで区切った語も候補とし，また連体助詞「の」が表れる場合はその直前にくる名詞も候補とする．その際丸括弧で囲われている語については，3.1.2 節で説明する初期確信度を-1 して付与するものとする．これは，丸括弧は特にその直前の語に対する注記を加えるために用いられることが多く，直前の語の読みなど内容の本質の理解には不要な語であることが多いため，他の候補より確信度の低い候補として扱うための処置である）

atr5: 格フレーム

格フレームを利用した自然言語処理により抽出される（3.1.1 節で説明する）

3.1.1 格フレーム

格フレームとは，動詞を基準として，取り得る格とその値に関する制約を記述したものである．例えば「日本語用言の結合価」（朝倉日本語新講座 3 『文法と意味Ｉ』付録 2) [11] によると，用言「行く」に対する文型は以下が挙げられている．

- N[具象物] が + N[場所] から + N[場所] に + V
- N[人間] が + N[場所] から + N[場所] に + V
- N[具象物] が + N[場所] から + N[場所] へ + V
- N[人間] が + N[場所] から + N[場所] へ + V

このような格フレームで場所を内容とする［動詞，格助詞］の組を列挙し，ブログ文書に対し自然言語処理を行う．日本語語彙大系では用言意味体系を 36 属性にカテゴリ分けしており，ここで抽出の対象とする動詞は“ 18 物理的移動 ”のカテゴリに属する動詞とする.“ 訪問した ”という動作を明らかにするために，特にこのカテゴリに限定した．さらに，その中から主体が人であるものに限定するため自動詞に限定した．自動詞に限定したのは，地名フィルタにおける“ 主語が Blog の書き手ではない ”という誤りを回避するためである．そして対象とする格助詞は，一般名詞意味体系で [/1 名詞/2 具体/388 場所] を内容とするものだけでなく，[/1 名詞/1000 抽象/2422 抽象的関係/2610 場] を内容とするものも含めた．これは，コーパスの構築の際，形態素解析により具体名詞が抽象名詞に分類されることがあるためである．例えば,“ 出町柳駅 ”や“ 丸太町駅 ”といった名詞は“ 駅 ”という単語にまとめられて分類されている．

しかしこれだけでは,“ ＸＸＸで食事した ”といった深層格と動作動詞の関係から抽出できる語が抽出できなくなってしまう．ここで，このような形で場所を値に持ちうる格助詞「に」と「で」について注目する．どちらも動作・作用の行われる空間的な場所を表す格助詞であるが，用法に違いがあり,「に」＋状態動詞,「で」＋動作動詞となる．そこで，デ格を持つ名詞も格フレームによる「訪問場所名詞」候補として扱うことする．以上を考慮して，日本語語彙大系を基本の格フレームとして参照する．日本語語彙大系は日本語 30 万語を 3000 種類の意味属性で分類した日本最大のシソーラスであり，基本として参照する対象としては十分な規模であると考えられる．

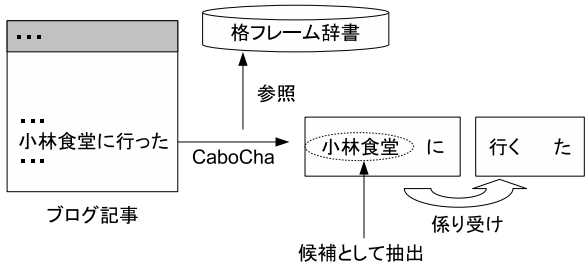

図 2 格フレーム処理の概要図

さらに，Web から自動構築した大規模格フレーム [12] でその拡張を行う．これは Web 上の 5 億文の文章を利用して構築したコーパスであり，これを利用して拡張することでブログ文書のような自由な文章への対応が期待できる．その拡張例を動詞「戻る」で示す．日本語語彙大系では，物理的移動を示す「戻る」の構文は

- N1 が N2 から／より N3 に／へ 戻る　N1 return from N2 to N3

と表されており，場所を内容とし，組を作る格助詞は「から」,「より」,「に」,「へ」となる．ここで，Web から自動構築した大規模格フレームを参照すると,「まで」,「を」も場所を内容とする格助詞となることが分かる．このようにして基本の格フレームを拡張する．この拡張の結果，300 強であった［動詞，格助詞］の組数は 1 割弱増え，およそ 350 組となった．

ブログ記事の本文を CaboCha で係り受け解析する際，こうして作成された格フレーム辞書を参照し「訪問場所名詞」候補を抽出する．例えば,“ Xへ行く ”からXが抽出される．その概要図を図 2 に示す.

3.1.2 初期確信度の付与

atr1〜atr5 までの五つの属性で抽出された語を「訪問場所名詞」候補として以下扱う．Step1 の最後に，この候補に対し確信度の初期値を与える．初期値が与えられる組み合わせを表 1 に示すようにヒューリスティックに与える．表 1 に見られる通り，単一の属性しか見られない候補の確信度は 0 である．三つ以上の属性が見られた場合は，最も高い数値を与える組み合わせを初期確信度とし，さらに三つ目の属性が見られた際に確信度を 2 増やし，それ以上見られた場合はその都度 1 増やすものとする．例えばある語について atr2〜4 の三つの属性が見られた場合，最も高い初期確信度を与える組は [atr2,atr4] であるので，3+2 で確信度は 5 となる．

	atr1	atr2	atr3	atr4	atr5
atr1	0	-	-	-	-
atr2	1	0	-	-	-
atr3	0	2	0	-	-
atr4	1	3	2	0	-
atr5	2	3	2	2	0

表 1　属性の組合せの確信度 1

	atr6	atr7	atr8	atr9
atr1	1	2	2	2
atr2	3	3	3	3
atr3	2	3	3	2
atr4	0	2	2	2
atr5	2	3	3	2

表 2　属性の組合せの確信度 2

3.2　条件付確信度増大処置

Step2 では，Step1 で抽出したそれぞれの候補について，ある条件を満足する場合に確信度を置換し増大させる処置を行う．その際，以下に挙げる atr6〜atr9 の属性を考慮する．

atr6:　タイトルに候補が含まれる

そのブログ文書のタイトルに前述の属性で抽出した候補が含まれる（タイトルが“ ＸＸＸでラーメン ”である場合など，タイトル中に場所名詞である“ ＸＸＸ ”を区切る文字が表れない場合のための属性であり，atr6 は atr4 を包含する(注1)）

atr7:　外部リンク先

外部リンク先に候補で抽出された単語が存在する．ただしリンクが張られている語はその候補から除くものとする（リンクが張られている語がリンク先のページを示す語であるのは当然であり，この属性は例えば http://〜といった語に対しリンクが張られている場合に対応するための処置である．<A href=”http://xxx”>http://xxx</A>という記述があっても atr2 では http://xxx しか抽出できない）

atr8:　説明表現が用いられている語

前述の属性で抽出した候補の直前に説明表現が用いられている．例えば“ 河原町にあるＸＸＸ ”といった表現で，日本語語彙大系の用言意味体系における“ 4 存在 ”のカテゴリに表れる表現を利用した

atr9:　住所や電話番号を利用した Google との関連付け

3.2.1 節で説明する

3.2.1　住所や電話番号を利用した Google との関連付け

ブログ文書中に住所や電話番号のラベルを発見すると，そのワードで Google に対しフレーズ検索を行う．その結果上位 5 件の文書を走査し，atr1〜atr5 で抽出した候補が文書中に含まれているかどうかを調べ，含まれていればその候補はこの属性を持つものとする．そのような検索結果の文書は電話帳のように［名称，住所，電話番号］がリスト表示されたページ，また検索ワードが場所名詞であった場合はその場所へのアクセス方法を示したページなどがヒットすることが予想され，この手法は理にかなっていると思われる．この属性により，特にグルメなどのレビューを行い，対象店舗の情報（住所，電話番号，営業時間，定休日など）の提示を行っている記事における確信精度の向上が見込まれる．

3.2.2　確信度置換条件とその実行

atr1〜atr5 により抽出された候補にこれらの属性が見られた場合，表 2 を参照し確信度が増大する時に限り確信度の置換を行う．ただし，atr4 がすでに見られているものについては atr6 を考慮しない．また，三つ以上の属性が見られた場合の処置は Step1 に示した通りである．

3.3　出力する「訪問場所名詞」の決定

Step3 では，Step2 で決定した確信度を判定し，出力する「訪問場所名詞」を決定する．Step1,2 により，抽出された候補はそれぞれ 0 以上の値の確信度を持っている．これに対し，次のルール 1〜3 で出力する「訪問場所名詞」を決定する．ただしルール 1 から見ていき，いずれかを適用するものとする．

ルール 1:　3 以上の値を持つ候補が存在する場合はそれを全て出力し終了．

ルール 2:　2 の値を持つ候補が一つだけ見つかればそれを出力し終了，複数あれば出力せず終了．

ルール 3:　1 の値を持つ候補が一つだけ見つかればそれを出力し終了，それ以外は出力せず終了．

ルール 2,3 のような処置を行うのは，イベント名やメニュー名など，場所名詞以外のものが出力されることを防ぐための処置である．

3.4　適　用　例

本章で示した提案手法について，いくつか例をとって適用を試みる．簡潔のため本文は一文とする．

> タイトル：小林食堂
> 本文：今日は小林食堂に行った．

この場合，“ 小林食堂 ”という語が atr4 と atr5 で抽出されたため，表 1 を参照し，確信度は 2 となる．ここでは他の候補が見られないためルール 2 が適用され，“ 小林食堂 ”が出力される．

> タイトル：小林食堂@百万遍
> 本文：百万遍へ行って「小林食堂」の中華そばを食べた．

この場合は“ 小林食堂 ”という語が atr3 と atr4 を持って確信度 2，“ 百万遍 ”という語が atr4 と atr5 を持って確信度 2 となる．3 以上の確信度を持つ候補が無いためルール 2 が適用され，値 2 を持つ候補が複数あるため，何も出力しない．

> タイトル：中華そば@小林食堂
> 本文：百万遍にある「小林食堂」の「中華そば」を食べた．

まず，Step1 で“ 中華そば ”と“ 小林食堂 ”が atr3 と atr4 を持ってそれぞれ初期確信度 2 を持つ．それから Step2 を適用すると，“ 小林食堂 ”に対し atr8 が見られる．表 2 を参照すると atr3 と atr8 による組み合わせの確信度が最も高いため，3

(注1)：atr4 はタイトルから候補を列挙するための属性であるのに対し，atr6 はタイトル中に atr4 以外の属性で抽出された候補が含まれることを示す属性と捉えてほしい

ブログホスト	Yahoo!	goo1	goo2	平均
無回答再現率	15/16	13/14	7/7	35/37
抽出率	18/34	23/36	23/43	64/113
正解率	100%	100%	100%	100%
全体の精度	66%	72%	60%	66%

表 3　グルメカテゴリに対する実験結果

を“ 小林食堂 ”の初期確信度とし，さらに三つ目の属性が見られたため 2 増やし 5 となった．候補は確信度 2 の“ 中華そば ”と確信度 5 の“ 小林食堂 ”であるため，ルール 1 が適用されて“ 小林食堂 ”が出力される．

次節で実際のブログ記事に対し本手法を適用した実験結果を示す．

4. 実　　験

4.1 節，4.2 節における評価実験の際に使用するデータは，Yahoo!ブログ [13] と goo ブログ [14] のカテゴリを指定してヒットする中から無作為に抽出したものとした．特に，まず表層的な特徴がよく見られると筆者が感じた，グルメレビューを記したブログに対して行った．それから一般的なブログ文書ではどの程度抽出できるかを評価するため，［地域 - 京都府］カテゴリの記事について実験を行った．京都府のデータとしたのは，正解データの作成は人手で行うため，京都府に在住している筆者が「訪問場所名詞」の判定を行いやすくするための処置である．さらに 4.3 節では，それらの結果を踏まえて，テクノラティ [3] でタグ名を指定して行うタグ検索の結果から無作為に抽出したものを実験データとして行った．また，それぞれの実験において改良手法を考え適用した．

4.1　グルメカテゴリに対する実験

まず，Yahoo!ブログにおける［生活と文化 > グルメ，ドリンク > 飲食店］カテゴリから 50 件，goo ブログにおける［お出かけ - グルメ］カテゴリと［グルメ・クッキング - 食べ歩き］カテゴリからそれぞれ 50 件無作為に抽出し実験を行った．goo ブログで二つのカテゴリから抽出しているのは，同様のカテゴリが 2 件見つかったためである．

その実験結果を表 3 に示す．goo におけるグルメカテゴリ，食べ歩きカテゴリはそれぞれ goo1，goo2 とした．無回答再現率とは，訪問場所が無いブログ記事に対して本手法が地物を何も返さなかった率である．抽出率は「訪問場所名詞」が存在するブログ記事に対してどれだけ抽出できたかであり，正解率はその出力が実際に「訪問場所名詞」であるかの判定を行ったものである．出力したが実際は「訪問場所名詞」の無い記事であったという誤りは無回答再現率に表されている．全体の精度とは，無回答時の正解率と抽出時の正解率を合わせたものである．また，抽出率は同時に「訪問場所名詞」の存在率も表しており，例えば表 3 における Yahoo!での存在率は，50 件のデータに対し行ったため 68%であったことがわかる．実験で実際に抽出された「訪問場所名詞」の具体例は，{ふぐ 福治，ふぐ八丁，三ます，szechwan restaurant 陳，大森軒} などが挙げられる．

表 3 に見られるように，goo2 の食べ歩きカテゴリの方が「訪問場所名詞」の存在率が高かった．直感通り，グルメカテゴリより食べ歩きカテゴリの方が実際に訪問した記事が投稿されやすいことが伺える．抽出した「訪問場所名詞」の正解率は 100%と，今回のデータでは非常に良い結果が得られた．無回答再現率に見られるように，「訪問場所名詞」が存在しない記事に対し，2 件回答してしまった．この原因を見ると，Yahoo!における誤りは晩御飯の記事であり，その晩御飯の食べ物を「訪問場所名詞」として回答した．しかし Yahoo!ブログから抽出した実験データのカテゴリは［生活と文化 > グルメ，ドリンク > 飲食店］であり，これは自宅における記事であるから，ブロガーが記事に相応しくないカテゴリに登録したためである．goo1 における誤りも食べ物を「訪問場所名詞」として回答していた．この食べ物はどこかに買いに行ったものではなく，頂き物であるとの記述があった．この実験データのカテゴリは［お出かけ - グルメ］であったから，どこにも出かけていないこの記事は同様に相応しくないカテゴリに登録されていると考えられる．また，全て合計した精度の平均は 66%であった．

ここからは提案手法の改良を考える．本手法で抽出できなかった記事の原因の多くは，記事中で一度しか場所名詞が出現しない等，語を特徴づける属性が一つしか見られなかったためであった．例えば文書中に記号で囲われる形で 1 度だけ出現するなどである．このような場合は，これまで示した手法で入力記事のみから，候補を確信するのは不可能であると思われる．そこで次のような手法を考える．タイトルに表れる語はその記事の内容を要約した語であると考えられ，例えば訪問場所名詞だけでなく，食べたメニューや地域名などが挙げられている．テクノラティ [3] によると一日に 120 万個のブログ記事が書かれているということであるので，同じ要約した語を使用している記事が存在することが見込まれる．表 3 の通り抽出した語の正解率は高いため，他のブログ記事を利用してそこから抽出した候補とマージし，元のブログ記事における候補のどれが場所名詞であるかを得ることを考える．そこでこれまで示した手法で何も返さず，また atr4 においてタイトル中の区切り文字で区切られた語が 2〜3 語である場合，それをクエリとしてブログ検索を行い，返ってきたブログ記事に対してこれまでの手法を適用し，クエリ中の 1 語だけを返すブログが検索結果の上位 5 件までに存在した場合，それを「訪問場所名詞」として確信するものとする．その際，ブログ検索にはテクノラティを利用する．ただし，タイトルから得られる区切り語が 3 語である場合は 2 語のペアで作られるクエリで行う（すなわち 3 回行う）．この改良手法を適用することで，全体の精度におよそ 7%の向上が見られ，平均 73%となった．

4.2　京都府カテゴリに対する実験

次に，Yahoo!ブログにおける［地域 > 日本 > 京都府］カテゴリ，goo ブログにおける［地域 - 京都府］カテゴリからそれぞれ 50 件無作為に抽出し実験を行った．その実験結果を表 4 に示す．実験で実際に抽出された「訪問場所名詞」の具体例は，{遊形，東寺，東北院，おはりばこ，六曜社珈琲店} などが挙げられる．

ブログホスト	Yahoo!	goo	平均
無回答再現率	20/20	17/17	37/37
抽出率	11/30	13/33	24/63
正解率	73%	92%	83%
全体の精度	56%	58%	57%

表 4　京都府に対する実験結果

	食べ歩きタグ
無回答再現率	7/8
抽出率	30/42
正解率	90%
全体の精度	68%

表 5　食べ歩きタグに対する実験結果

表 4 の無回答再現率の分母に表れるように，実際に訪問した場所名詞が存在しない記事がグルメカテゴリに比べて多く見られた．これは京都府カテゴリの分類は地域カテゴリに属しているだけであり，お出かけなど外出に関する内容を示すことに限っていないためである．京都府に在住している人の記事ではあるが，地域に関する内容とは言いがたい日記記事も多く見られた．そして京都府カテゴリの特徴と思われるが,「訪問場所名詞」には寺社・仏閣が多く見られた．また，全て合計した精度の平均は 57%であった．

ここからは提案手法の改良を考える．グルメカテゴリにおける実験と同様，抽出できなかった理由に特徴付ける属性が一つしか無かったことが挙げられるが，多くの記事に次のような特徴が見られた．多くのブログ記事には写真が掲載されており,「訪問場所名詞」が写真の解説文中に出現した（例：ＸＸＸの紅葉です）．そこで，画像が存在するブログ文書においては,「〜です」で表現される，助動詞「です」の直前にくる名詞を追加属性として扱うこととした．ただし連体助詞「の」がその名詞に係っている場合は，その直前にくる名詞を追加属性として扱うものとし，また助詞「は」がその文節に存在する場合は,「ＸＸＸは〜です」というようにＸＸＸの説明表現をしている文章であることが考えられるので，その場合は係助詞「は」の直前にくる語を追加属性として扱う．同様に写真の解説文としてしばしば見られる表現に,“ ＸＸＸにて ”が挙げられる．そこで，画像が存在するブログ文書において格助詞「にて」に注目する．「にて」は，格助詞「に」に接続助詞「て」がくっついた言葉であり,「に」と大体同じ用法である．ニ格が表す内容は「基礎日本語文法」[15] によると，人やものの存在場所，所有者，移動の着点，動作の相手，動作の対象，状態の対象，原因，移動動作の目的，事態の時である．ただし，現代語においてニテ格が表す内容は主に場所や時間に限られる．そこで，ニテ格を持つ候補も追加属性として扱うこととする．語の内容が時間であるものを抽出しないために，数字が混ざった語は除くものとする．atr1〜5 で抽出された候補にこの追加属性が見られた場合，その候補を「訪問場所名詞」として出力するものとする．また，単一の属性しか持たない候補はこれまで確信度 0 として扱ってきたが，1 以上の確信度を持つ候補が一つも見られなかった場合，このニテ格を持つ語を出力するものとする．

この改良手法をグルメカテゴリに適用すると，食べたメニュー等が抽出されることが予想されるので，適用するのは京都府カテゴリのみとする．この改良手法を適用することで全体の精度で 5%の向上が見られ，平均 62%となった．4.1 節で適用した改良手法は適用しなかった．これはタイトル中に区切り文字があまり現れず，語の羅列から「訪問場所名詞」候補を取得するのが困難であったためである．

4.3　食べ歩きタグに対する実験

これまで Yahoo!ブログと goo ブログの特定のカテゴリに属するブログ文書に対し実験を行ってきた．しかし，このカテゴリはユーザが自分で作成した名称ではなく，ブログホスト側が作成した名称であるため，そのカテゴリに相応しくないブログ記事が含まれることが多々見られた．そのため，他のカテゴリに対してもいくつか実験を行ったが,「訪問場所名詞」の存在率が低く実験対象としてふさわしくなかった．そこで，ブロガーが自分で作成した記事のカテゴリ（ここではタグと呼ぶ）を指定して得られるデータに対し実験を行うものとした．テクラノティ [3] ではタグ検索を提供しており，これを利用する．タグ検索とは，ブログ記事に登録されている「タグ」を利用して記事を検索する仕組みであり，タグクリエイターによって生成されるコードを記事に記述することで結果に反映される．またそのようなコードを記述せずとも，ブログツールの機能で独自にカテゴリを作成し，その名称を結果として反映させたいタグ名にするという方法もある．おそらくこちらの方が容易であり，一般的であろう．PING サーバに PING を送信すれば確実に反映される．検索ワードとするタグ名であるが，これまでの実験で最も「訪問場所名詞」の存在率が高かったのは goo ブログにおける［グルメ・クッキング - 食べ歩き］であったので“ 食べ歩き ”とした．それでタグ検索を行い，ヒットする中から無作為に 50 件抽出したものに対し同様に実験を行った．その実験結果を表 5 に示す．実験で実際に抽出された「訪問場所名詞」の具体例は, {九十九ラーメン，マジックスパイス，鬼はそと福はうち，ラフェ・クレール，Bonheur} などが挙げられる．

表 5 に見られるように,「訪問場所名詞」の存在率は 8 割強であり，これまでで最も高かった．また，抽出した「訪問場所名詞」の正解率も 90%と，グルメカテゴリ同様非常に高い数値が得られた．無回答再現率に見られるように,「訪問場所名詞」が存在しない記事に対し 1 件回答してしまった．この原因を見ると，通信販売の店名が抽出されたものであり，これはそのブログ記事が記事の内容に相応しくないタグ名を付けていたためである．全体の精度は 68%であった．

ここからは提案手法の改良を考える．抽出したのは 30 件であったが，そのうち 3 件が誤りであった．原因を見ると，抽出された語は場所名詞ではなく，そのブロガーが食べた物であった．これは，食べ歩きといったグルメカテゴリ特有の現象であるが，このようなグルメカテゴリにおいては次のような改良策が考えられる．日本語語彙大系の一般名詞意味体系で，[/1 名詞/2 具体/533 具体物/706 無生物/760 人工物/838 食料] に分類される名詞がどのような構文で使用されるか，日本語語彙大

	atr1	atr2	atr3	atr5	atr6	atr7	atr8	atr9
グルメカテゴリ	3.4	6.0	27.5	8.1	36.2	3.4	3.4	12.1
京都府カテゴリ	3.7	11.1	24.1	20.4	35.2	0	3.7	1.9
食べ歩きタグ	7.3	13.4	19.5	12.2	34.1	4.9	4.9	3.7

表 6　属性の出現割合

系から取得した．例えば，動詞「食べる」の構文は次のように表されている．

- N1（人）が N2（仕事）で 食べる　N1 earn N1's living with N2
- N1（人，動物）が N2（食料，生物）を 食べる　N1 eat N2

これにより，食料を内容とする［動詞，格助詞］,［形容詞，格助詞］の組が得られた．その件数は 131 件であった．例えば上の例からは［食べる，を］が取得できる．ここで得られた［動詞，格助詞］の組に 3.1.1 節で取得した組と重複が無いかどうか調べたが，そのような組は無かった．そのため，この格フレームに該当した「訪問場所名詞」候補は候補から除外するものとする．この改良手法により 1 件改善され，全体の精度は 70%となった．4.1 節におけるグルメカテゴリの実験データで同様の改善が見られないか調べたが，今回のデータでは見られなかった．しかしこの手法は一般的なグルメカテゴリにおいて有効な手法であると考えられる．また，4.1 節におけるグルメカテゴリで用いた改良手法も適用したが，今回のデータでは改善結果が得られなかった．これはグルメカテゴリにおける抽出率が平均 56%であったのに対し食べ歩きタグでは 71%と，元々の提案手法での抽出率が高かったためであると思われる．改良手法による上乗せを加えてもグルメカテゴリの抽出率は平均 65%と，食べ歩きタグより低かった．

5. 提案手法の検証

提案手法で八つの属性を挙げ（atr6 は atr4 を包含するので八つ），それを利用して「訪問場所名詞」を抽出してきた．ここで,「訪問場所名詞」を抽出できた場合の属性の出現数を見て提案手法を検証する．4 節においてグルメカテゴリ，京都府カテゴリ，食べ歩きタグで実験を行ったが，その際，改良手法を適用した上で出現したそれぞれの属性の数を全て合計したグラフを図 6 に示す．格フレームを利用して抽出できるものは，場所名詞が地名辞書に登録されていれば基本的に地名フィルタにおいても抽出できる（もちろん地名辞書が巨大になればなるほど「訪問場所名詞」でない場所名詞を抽出しやすくなり，精度は落ちる）．そこで，地名辞書に今回抽出した「訪問場所名詞」が全て登録されていると仮定した場合に地名フィルタで抽出できるものを数え上げた．atr5 の色の違う棒グラフがそれを表す．また，4 節でのそれぞれの実験における八つの属性の出現割合を表 6 に示す．

図 3 の atr6 に見られるように,「訪問場所名詞」はタイトル中に最も多く見られた．次に多く見られたのが「」など記号で囲われているという属性 atr3 であり，格フレームによる自然言語処理で抽出されるという属性 atr5 はその次であった．地

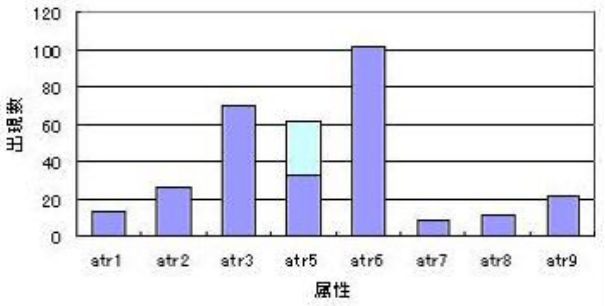

図 3　属性の出現数

名フィルタによって抽出可能な「訪問場所名詞」の数は 60 程度で,「訪問場所名詞」がタイトル中に表れた場合に限っても，提案手法の方が atr6 との差である 40 程度多く抽出できる．この結果は，今回のような実験対象において「訪問場所名詞」と場所に関する深層格を表しうる格助詞が同時に表れることはそれほど多くなく，そのような特徴を利用するだけで抽出できる「訪問場所名詞」はごく限られることを示す．よって，地名フィルタでは抽出できないブログ記事からある程度抽出が可能になっており，本提案手法の有効性が確認できる．

また表 6 からカテゴリ・タグ間での属性出現数を比較すると，atr5 の値からグルメ系のカテゴリでは京都府カテゴリに比べ「訪問場所名詞」が格助詞と共に出現する割合が低いことが分かる．そのため，特にグルメ系のカテゴリにおいて地名フィルタとの差異が大きく表れ，本手法が有効であると考えられる．

ブログ文書では日記のような記述表現がよく見られるため，核フレームによって抽出される語が増えると考え本論文では対象データをブログ文書としたが，図 3 が示すように本手法の抽出率を最も高めている要素はタイトル属性であるから，一般の Web 文書でも，タイトルに場所名詞が出現している文書であればある程度本手法による抽出が可能であると考えられる．しかし，通常の Web 文書全体に対し本手法を適用することは現実的では無く，適用データを絞り込むことが必要であろう．また口語的な文章に関しては，本来大文字である日本語が小文字で記述されているなどといった場合は正しく抽出できたが，スペースで区切る必要の無い「訪問場所名詞」がスペースで区切られているといった場合は抽出できなかった．

本手法は地名辞書の代わりに格フレーム辞書を保持するが，そのサイズに大きな違いがある．本手法の格フレーム辞書が保持する［動詞，格助詞］の組は 350 件程度で，また 4.3 節における改良手法適用時に保持することになる［動詞，格助詞］,［形容詞，格助詞］の組は 131 件と，合計しても 500 件に満たない．それに対し，maplog における「位置キーワード」セットや地名フィルタにおける地名辞書のような場所名詞を全国レベルで保持すればどれほどのサイズになるか，NTT 情報開発株式会社の HP [17] で検索したところ，旅館・ホテルなど［生活関連サービスすべて］で 1,101,825 件，公園・動物園など［施設・機関］で 118,263 件，飲食業に当たる［和風飲食店］,［洋風・中華飲食店］,［スナック・バー・酒場・喫茶店］で 822,456 件であった．全て合計すると 200 万件を超え，保持にかかるコス

トだけでなく，フィルタリングの際に必要な処理コストも膨大になる．また，そのデータの性質についても，特に飲食店においては閉店によるデータの陳腐化，新規出店による更新の必要性も問題となる．提案手法ではそのような大規模な場所名詞辞書を保持することなく，小サイズの格フレーム辞書のみで「訪問場所名詞」の抽出が可能となる．

6. 応　　用

性別や年齢といった，ブロガーの属性を推定する研究が試みられている [16]．本手法の応用として，そのようなブロガー属性の一つである，居住地域を推定することが考えられる．本論文では対象とするブログ記事を無作為に抽出したが，一つのブログにおける全てのブログ記事を対象とすれば，そのブロガーの訪問した場所名詞集合が抽出できる．それら場所名詞から住所集合を得れば，居住地域が推定できるであろう．同様の推定を行うのに場所名詞の辞書を持つことは，対象とするブロガーがどこの地域に住んでるか分からないため，全国レベルの場所名詞辞書を持つことは 5 節に示したようにあまりに多大なコストであり，実用的ではない．さらに本手法による居住地域推定では，都道府県に止まらず市区町村まである程度推定することが期待できると思われる．

また，プロモーション活動への応用が考えられる．人々が新たな施設や店舗に訪れようと考えそれらを探す際，施設のイベント内容や“ おいしい ”や“ 安い ”といった店舗の特徴による検索だけでなく,“ 自宅から 30 分以内で行ける ”や“ 特定の駅から徒歩で行ける ”といった場所に関する特徴も重要な要素であると考えられる．一つのブログから抽出した場所名詞を地図上にプロットした時，いくつか点が固まって見られる範囲は，そのブロガーが難なく訪れることができる場所であるということが予想できる．そのため，ある地物のプロモーションを行う際，その地物の場所をその範囲に含んでいるブロガーに限定して行うことにより，効果的なプロモーションが可能になると考えられる．

7. まとめと今後の課題

本論文では，ブログ文書における表層的特徴ならびに格フレームを利用し，与えられたブログ文書から訪問場所名詞を獲得する手法を示した．地名辞書なしでも，ある程度の精度を持った抽出が可能であることがわかった．対照実験ではないので参考程度であるが，地名辞書を使用した地名フィルタ [7] における精度を表 7 に示す．また地名フィルタでは，以下のような失敗例が挙げられていたが，これらはある程度克服できた．

- 主語が Blog の書き手ではない

（例:花子は銀閣寺へ行ったらしいが）

- 動作動詞も助詞「へ」「から」「まで」も存在しない

（例:次は銀閣寺です）

- タイトルがビジターの訪れた地名である

グルメカテゴリ，京都府カテゴリ，食べ歩きタグと三つの実験データに対し実験を行ったが，全体の精度はそれぞれ改良手法適用後 73%，62%，70%と，それほど大きな違いは見られなかった．無回答再現率についても 0〜2 件の誤りがあるだけで大きな違いは無く，その誤りの原因は適切なカテゴリにブログ記事が登録されていないためであった.「訪問場所名詞」が存在する記事からの改良手法適用後の抽出率は，それぞれ 65%，44%，74%と，京都府カテゴリに比べ他のグルメ系のカテゴリでは高い結果が得られた．そして,「訪問場所名詞」を出力した場合のその正解率はそれぞれ，97%，83%，90%といずれも高い結果が得られた．抽出率と合わせて見ても，グルメ系のカテゴリにおいては実用に十分な精度が得られたと考えられる．また，グルメ系のカテゴリにおけるブログ記事では「訪問場所名詞」が格助詞と共に出現する割合が低く，特に本手法はグルメ系のカテゴリにおいて有効性を増すことがわかった．

エントリ内の地名数	1	2	3	4	5	6	15
精度	74%	67%	33%	25%	75%	0%	100%

表 7　地名フィルタの精度

今後のさらなる改良案としては，カテゴリを持つというブログの構造を利用することが考えられる．あるブロガーの作成した同一カテゴリ内の記事は，そのブロガー個人によるテンプレートに従って記述しているケースが見られるため，あるブロガーに関して「訪問場所名詞」が見つかった場合はタイトル等におけるテンプレートを記録して，それに即した出力をするというものである．そのためには本手法による実験結果の蓄積が必要であるので行っていきたい．また，その過程でさらなる改良策も見つけたい．

謝辞　Web から自動構築した大規模格フレームの貴重なデータを提供して下さった京都大学情報学研究科の黒橋禎夫教授に深く感謝いたします．

文　　献

[1] http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html
[2] ブログ及びＳＮＳの登録者数（平成１８年３月末現在）, http://www.soumu.go.jp/s-news/2006/060413_2.html
[3] テクノラティ, http://www.technorati.jp/
[4] 高橋哲朗, 乾健太郎, 松本裕治,“ テキストから属性関係を抽出する ”, 情報処理学会自然言語処理研究会, 2004-NL-164, 2004.
[5] 鈴木泰裕, 高村大也, 奥村学,“ Weblog を対象とした評価表現抽出 ”, SIG-SWO-A401-02, 2004.
[6] maplog, http://maplog.jp/
[7] 郡宏志, 服部峻, 手塚太郎, 田島敬史, 田中克己,“ Blog からのビジターの代表的な行動経路とそのコンテキストの抽出 ”, DBWS2006-B3-35, 2006.
[8] 西田悠介, 柴田知秀, 河原大輔, 岡本雅史, 黒橋禎夫, 西田豊明,“ 料理教示発話の構造解析“ , NLP2003-B7-5, 2003.
[9] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦, 日本語語彙大系 CD-ROM 版, 岩波書店, 1999.
[10] CaboCha, http://chasen.org/~taku/software/cabocha/
[11] 水谷静夫他, 朝倉日本語新講座 3 文法と意味 I , 朝倉書店.
[12] Web から自動構築した大規模格フレーム, http://reed.kuee.kyoto-u.ac.jp/cf-search/
[13] Yahoo!ブログ, http://blogs.yahoo.co.jp/
[14] goo ブログ, http://blog.goo.ne.jp/
[15] 益岡隆志, 田窪行則, 基礎日本語文法, くろしお出版.
[16] 池田大介, 南野朋之, 奥村学,“ blog 著者の性別推定 ”, 言語処理学会 第 12 回年次大会, NLP2006-C2-3, 2006.
[17] NTT 情報開発株式会社 タウンページデータベース, http://www.nttbis.co.jp/townpage/townpage_db.shtml