概要

近年，ブログの急速な発達により，これまでは情報源として活用することが難しかった個人の意見や体験談などが，容易に収集できるようになった．観光地開発のために情報を収集，分析しようと考えた場合において，実際に観光地に出かけた個人の意見，体験談というものが記述されているブログ記事は非常に有益なリソースであると考えられる．観光地開発のためにブログ記事から観光情報分析を行うためには，ブログ記事に記述されている地名を正確に判定することが重要である．しかし，地名を扱う先行研究 [1] では，複数の都道府県に存在する地名を扱うことができていない．

そこで本研究では，複数の都道府県に存在する地名を扱うことが可能な，ブログ記事における地名解析を行うことを目的とする．本研究で提案する地名解析の手法は,「手がかり語検出」および「都道府県名の曖昧性軽減」という二つの手法から構成する．

手がかり語検出を行うために,「手がかり語辞書」を作成する．手がかり語辞書は，都道府県を判定する手がかりとなる語と，その都道府県名の組から構成する辞書である．手がかり語辞書には，市区町村名のほか，施設名，イベント名，地形名，路線名などを登録する．この手がかり語辞書を用いて，手がかり語検出を行う．手がかり語検出は，まずブログ文章の形態素解析を行って名詞を抽出する．次に，抽出した名詞が手がかり語辞書に登録されているならば，その都道府県をブログ文章中にタグ形式で挿入する．

都道府県名の曖昧性軽減を行うために,「都道府県コーパス」を作成する．都道府県コーパスは，Wikipedia における各都道府県について記述されたページに出現した固有名詞を抽出して作成する．各都道府県のページから抽出した固有名詞を，その都道府県の共起語とする．この都道府県コーパスを用いて，都道府県名の曖昧性軽減を行う．手がかり語検出にて検出された手がかり語には，複数の都道府県名が出力される場合がある．この余計に出力された都道府県名を抑制する処理が，都道府県名の曖昧性軽減である．都道府県名の曖昧性軽減は，まずブログ記事内の固有名詞を抽出し，都道府県コーパスと照合し，ブログ記事単位の有力な都道府県名を決定する．次に，一つの手がかり語に対して複数出力された都道府県名の中に決定した有力な都道府県名が存在すれば，有力な都道府県名のみを出力し，複数出力された都道府県名の中に決定した有力な都道府県名

が存在しない場合は，そのまま出力する．

以上の提案手法に対する評価実験において，手がかり語検出では，正解となる手がかり語の文字列のうち一部を手がかり語として検出できればよいという評価においてF値で0.662という評価結果であった．都道府県名の曖昧性軽減では，評価対象を上述の手がかり語検出において正しく検出できた手がかり語に限定した場合において，F値で0.566という評価結果であった．正解手がかり語のうち一部を検出できればよいという評価において手がかり語検出を行い，評価対象を限定せずに都道府県名の曖昧性軽減を行ったときの，本手法全体の評価は，F値で0.336という評価結果であった．一方，ブログ記事単位で有力な都道府県名を判定する性能の評価は，一致率で60%となった．よってこれら評価結果より，手がかり語辞書を用いて手がかり語検出を行い，その後都道府県コーパスを用いて都道府県名の曖昧性軽減を行うという本研究で提案した地名解析の手法の有用性を確認した．

目 次

表 目 次

図 目 次

第1章　はじめに

近年，ブログの急速な発達により，これまでは情報源として活用することが難しかった個人の意見や体験談などが，容易に収集できるようになった．観光地開発のために情報を収集，分析しようとした時，実際に観光地に出かけた個人の意見，体験談というものは非常に有益なリソースであると考えられる．観光地開発のためにブログ記事から観光情報分析を行うためには，ブログ記事に記述されている地名を正確に判定することが重要である．しかし，地名を扱う先行研究 [1] において，複数の都道府県に存在する地名を扱うことができていない．そこで本研究では，複数の都道府県に存在する地名を扱うことが可能な，ブログ記事における地名解析を行うことを目的とする．地名解析は,「手がかり語検出」および「都道府県名の曖昧性軽減」という二つの手法から構成する．以下に，各手法と，その手法で用いるデータベースについて説明する．

手がかり語検出を行うために,「手がかり語辞書」を作成する．手がかり語辞書は，都道府県を判定する手がかりとなる語と，その都道府県名の組からなる辞書である．手がかり語辞書には，市区町村名のほか，施設名，イベント名，地形名，路線名などを登録する．この手がかり語辞書を用いて，手がかり語検出を行う．手がかり語検出は，まずブログ文章の形態素解析を行って名詞を抽出する．次に，抽出した名詞が手がかり語辞書に登録されているならば，その都道府県をブログ文章中にタグ形式で挿入する．

都道府県名の曖昧性軽減を行うために,「都道府県コーパス」を作成する．都道府県コーパスは，Wikipedia における各都道府県について記述されたページに出現した固有名詞を抽出して作成する．各都道府県のページから抽出した固有名詞を，その都道府県の共起語とする．この都道府県コーパスを用いて，都道府県名の曖昧性軽減を行う．手がかり語検出にて検出された手がかり語には，複数の都道府県名が出力される場合がある．余計に出力された都道府県名を抑制する処理が，都道府県名の曖昧性軽減である．都道府県名の曖昧性軽減は，まずブログ記事内の固有名詞を抽出し，都道府県コーパスと照合し，ブログ記事単位の有力な都道府県名を決定する．次に，一つの手がかり語に対して複数出力された都道府県名の中に決定した有力な都道府県名が存在すれば，有力な都道府県名のみを出力するようにする．

本論文の構成は以下の通りである．第 2 章で，地名解析を行う先行研究の概要と問題点について述べる．第 3 章では，手がかり語辞書の作成方法について説明する．第 4 章では，都道府県コーパスの作成方法について説明する．第 5 章では，地名の解析手法について説明する．第 6 章では，評価実験とその結果を示す．第 7 章では，本手法における考察と今後の課題について述べる．第 8 章では，本研究のまとめについて述べる．

第2章　先行研究

本章では，地名解析を行う先行研究の手法とその問題点について説明する．

2.1　先行研究の概要

安田ら [1] は，ブログ記事からブログ作者の居住域の推定を行った．まず，地名辞書を作成し，地名を含む文をブログ記事から抽出した．地名辞書は，goo 地域情報サイトの各都道府県の主要エリア名，国内観光情報サイトの分類に基づく地名（ランドマーク名や施設名を含む），および郵便番号データを用いて作成した．次に，抽出した文に対し，その地名がブロガーの居住域にあるかどうかを二値分類器を用いて分類を行った．ここで，二値分類器の学習には，地名の周囲の文脈を用い，地名そのものは用いない．よって，地名辞書に変更があっても，分類器の訓練をやり直す必要がないと期待できる．評価実験として，比較用の素朴な手法であるブログ記事中に出現した地名が所属する都道府県の中で最も出現回数が多かった都道府県に決定する手法と，上記の提案手法を比較した．ブログ記事中に出てきた地名が所属する都道府県の中で最も出現回数が多かった都道府県に決定する手法は精度 48.2%，二値分類器を用いる提案手法は精度 50.7%となった．提案手法によって，素朴な手法を用いるより精度が 2.5%向上することを示した．

2.2　先行研究の問題点

先行研究では，地名とその都道府県名との関係は先行研究全体で依存している重要な情報となっているので，地名と都道府県名が一対一でない関係は望ましくない．よって，地名辞書の作成の際に，複数の都道府県に存在する地名は排除している．

このように，地名を扱う研究において，複数の都道府県に存在する地名というものは，非常に大きな問題となっている．そこで本研究では，複数の都道府県を扱うことができる地名解析を行う．

第3章　手がかり語辞書

本章では，手がかり語辞書の作成方法と構成について説明する．手がかり語辞書は，場所を判定する手がかりとなる語と，その都道府県名の組の辞書である．手がかり語辞書は，日本郵便・郵便番号データ [2] から地名を抽出したもの，国内観光情報サイト「大好き日本」[3] から施設名やイベント名を抽出したもの，および日本語語彙大系 [4] から地形名や路線名を抽出したものから構成する．

3.1　郵便番号データからの抽出

本節では，日本郵便・郵便番号データから地名を抽出する方法とその結果について説明する．

3.1.1　郵便番号データとは

郵便番号データとは，郵便番号と住所が対応して登録されているデータベースである．郵便番号データには，住所が「”都道府県名”,”市区町村名”,”町域名”」という階層構造で記述されている．郵便番号データは，日本郵便のホームページにて公開されている．図 3.1 に，郵便番号データの記述形式の例を示す．

```
… ”北海道”,”札幌市中央区”,”以下に掲載がない場合” …
… ”北海道”,”札幌市中央区”,”旭ケ丘” …
… ”北海道”,”札幌市中央区”,”大通東” …
… ”北海道”,”札幌市中央区”,”大通西（１～１９丁目）” …
… ”北海道”,”札幌市中央区”,”大通西（２０～２８丁目）” …
… ”北海道”,”札幌市中央区”,”北一条東” …
```

図 3.1: 郵便番号データの記述形式の例

3.1.2　抽出方法

本節では，郵便番号データから地名を抽出する方法について説明する．

本研究で用いる郵便番号データは，2009 年 6 月 30 日更新分のものである．まず，郵便番号データより一行ずつ郵便番号と住所の組を取得する．取得した一行から，都道府県名，市区町村名，町域名が漢字で記述されている部分のみを抽出する．ここで，もし町域名が地名でなければその行は読み飛ばし，町域名に “（）”が書かれていればその部分を削除する．この時，“（）”が複数行に渡って書かれている場合は，その部分は読み飛ばす．このように処理を行い，階層構造を保ったまま都道府県名，市区町村名，町域名を抽出する．図 3.2 に，抽出した都道府県名，市区町村名，町域名の例を示す．

北海道, 札幌市中央区, 旭ケ丘
北海道, 札幌市中央区, 大通東
北海道, 札幌市中央区, 大通西
北海道, 札幌市中央区, 北一条東

図 3.2: 抽出した都道府県名，市区町村名，町域名の例

3.1.3　表記のゆれ対応

都道府県名，市区町村名，町域名それぞれにおいて，ブログ記事に記述される際に表記のゆれがある．例えば，都道府県名では「兵庫県」と記述される場合と「兵庫」と記述される場合がある．市区町村名では「神戸市」と記述される場合と「神戸」と記述される場合がある．このような表記のゆれに対応するために，抽出した都道府県名，市区町村名，町域名それぞれに対し，一行ずつ図 3.3 のルールを適用して，単語の分割処理を行う．

○都道府県名
・□□県　→　□□，□□県（都，府も同じルールを適用）

○市区町村名
・□□市　→　□□，□□市
・□□郡△△町　→　□□郡△△町，□□郡，□□，△△町，△△
・□□市△△区　→　□□市△△区，□□市，□□，△△区，△△
・□□郡△△村　→　□□郡△△村，□□郡，□□，△△村，△△

○町域名
・□□町　→　□□，□□町
・□□町△△　→　□□町△△，□□町，□□，△△
・□□区△△　→　□□区△△，□□区，□□，△△
・□□町△△町　→　□□町△△町，□□町，□□，△△町，△△

図 3.3: 表記ゆれ対応ルール

例として「愛知郡長久手町」という市区町村名の場合，
・□□郡△△町　→　□□郡△△町，□□郡，□□，△△町，△△
というルールに当てはめ分割処理を行い,「愛知郡長久手町」,「愛知郡」,「愛知」,「長久手町」,「長久手」という単語を得る．ここで,「愛知郡長久手町」は愛知県に存在することが階層構造より分かるので，分割処理を行って得た単語それぞれを，“愛知”という都道府県名と組にして手がかり語辞書に登録する．

3.1.4 抽出結果

郵便番号データから抽出した都道府県名と手がかり語の組の例を，図 3.4 に示す．図 3.4 に示す通り，都道府県名と手がかり語が組になって記述される形式になっている．

愛知, あし原
愛知, あし原町
愛知, あずら
愛知, あま
愛知, あま市
愛知, いろは
愛知, いろは町

図 3.4: 郵便番号データから抽出した都道府県名と手がかり語の組の例

表 3.2 に，郵便番号データから抽出した手がかり語の件数を示す．ここで，都道府県名ありとは，手がかり語と都道府県名が組になっている場合の抽出件数である．都道府県なしとは，手がかり語のみを抽出しユニークをとった場合の抽出件数である．手がかり語のみを抽出しユニークをとると件数が減少するのは，異なる都道府県に同一の手がかり語が存在するためである．

表 3.1: 郵便番号データから抽出した手がかり語の件数

状態	件数
都道府県名あり	152,903 件
都道府県名なし	108,992 件

3.2　国内観光情報サイトからの抽出

本節では，国内観光情報サイト「大好き日本」から施設名やイベント名を抽出する方法とその結果について説明する．

3.2.1　国内観光情報サイト「大好き日本」とは

国内観光情報サイト「大好き日本」には，各都道府県ごとに宿泊施設や交通機関，観光施設やアウトドア，レジャーといったカテゴリに分かれて観光情報が掲載されている．観光情報には，ホテル名や交通機関名など，本研究で有用な手がかり語となるものが記述されている．各都道府県ごとに観光情報がまとめられているため，都道府県名と手がかり語を組にして抽出することが可能である．

3.2.2　HTML ファイルのダウンロード

国内観光情報サイト「大好き日本」より，各都道府県ごとの観光情報が記述されている HTML ファイルをダウンロードした．本研究で用いる HTML ファイルは，2009 年 9 月 3 日にダウンロードを行ったものである．

国内観光情報サイトより HTML ファイルのダウンロードを行ったカテゴリを，図 3.5 に示す．

観光協会・協会組合	農業共同組合	漁業共同組合
旅行会社・ガイド	宿泊施設・温泉・ホテル	交通機関
博物館・美術館	動物園・水族館	歴史
道の駅	自然	アウトドア
農業観光	海・川・湖	遊園地
スキー場	バリアフリー・子供	お酒
街歩き	お祭り	日本百選

図 3.5: HTML ファイルのダウンロードを行ったカテゴリ

図 3.5 に示されているカテゴリに属する観光情報から，施設名やランドマーク名，イベント名などを抽出する．

3.2.3 抽出方法

国内観光情報サイトよりダウンロードを行った HTML ファイルから一行ずつ文字列を取得し，手がかり語が書かれている部分を抽出する．HTML ファイルの記述形式の解析を行った結果，図 3.5 のカテゴリによって 3 パターンの記述形式が存在した．記述形式ごとに，正規表現を用いて手がかり語の抽出を行う．図 3.6 に，パターンごとの正規表現を示す．

・カテゴリ「日本百選」
```
/\A(.*)<SPAN.*/
```
・カテゴリ「ホテル」
```
/<TD.*><DIV class=.*?>(.*)<\/DIV><DIV class=.*><\/TD>/
```
・それ以外
```
/\A ■<.*?>(.*?)<\/A>.*/
```

図 3.6: パターンごとの正規表現

図 3.6 に示されている正規表現にマッチした行から，() で囲まれている部分の文字列を手がかり語として抽出する．どこの都道府県に属する HTML ファイルから手がかり語を抽出したかを記録しておくことで，手がかり語と都道府県名が階層構造を持った形で抽出できる．例として，北海道の観光情報から抽出した手がかり語を，図 3.7 に示す．

円山動物園
のぼりべつクマ牧場
ノーザンホースパーク
ノースサファリサッポロ
おたる水族館

図 3.7: 抽出した手がかり語の例（北海道の HTML ファイルより）

3.2.4 表記のゆれ対応

ブログ記事に記述される際の表記のゆれに対応するため，抽出した手がかり語に対し，以下のルールを適用する．

- アルファベットおよび記号が含まれている手がかり語は，アルファベットおよび記号が全角のものと半角のものの両方を手がかり語辞書に登録

- スペースが含まれる施設名は，スペースで区切ったそれぞれと，スペースを削除して詰めたものをそれぞれ手がかり語辞書に登録

例として「わかさ氷ノ山　自然ふれあいの里」という手がかり語の場合,「わかさ氷ノ山」,「自然ふれあいの里」,「わかさ氷ノ山自然ふれあいの里」という 3 つの単語に加工して，それぞれを手がかり語に登録する．

3.2.5　抽出結果

国内観光情報サイト「大好き日本」から抽出した都道府県名と手がかり語の組の例を，図 3.8 に示す．図 3.8 に示す通り，都道府県名と手がかり語が組になって記述される形式となっている．

```
鳥取, 鳥取しゃんしゃん祭
鳥取, 鳥取グリーンホテルモーリス
鳥取, 鳥取シティホテル
鳥取, 鳥取ワシントンホテルプラザ
鳥取, 鳥取空港
```

図 3.8: 国内観光情報サイトから抽出した都道府県名と手がかり語の組の例

表 3.2 に，国内観光情報サイトから抽出した手がかり語の件数を示す．ここで，都道府県名ありとは，手がかり語と都道府県名が組になっている場合の抽出件数である．都道府県なしとは，手がかり語のみを抽出しユニークをとった場合の抽出件数である．手がかり語のみを抽出しユニークをとると件数が減少するのは，異なる都道府県に同一の手がかり語が存在するためである．

表 3.2: 国内観光情報サイトから抽出した手がかり語の件数

状態	件数
都道府県名あり	59,479 件
都道府県名なし	55,702 件

3.3　日本語語彙大系からの抽出

本節では，日本語語彙大系から地形名や路線名を抽出する方法とその結果について説明する．

3.3.1　日本語語彙大系とは

日本語語彙大系は,「意味体系」,「単語体系」,「構文体系」によって構成されている.「意味体系」は，日本語の一般名詞，固有名詞，用言の意味的用法を意味属性体系で体系づけている.「単語体系」は，一般名詞や固有名詞などの意味的用法を約 3,000 の意味属性体系を用いて定義している.「構文体系」は，日本語の用言約 6,000 語の表現構造を結合価パターン約 14,000 件にまとめたものである．

本研究では,「単語体系」にまとめられている固有名詞意味属性体系から，場所を特定する手がかりとなるような単語を抽出し，手がかり語辞書に登録する．固有名詞意味属性体系に登録されている各単語には，意味属性が付与されている．この意味属性を考慮して，手がかり語となりうる単語を抽出する．なお，固有名詞意味属性体系から抽出した手がかり語の都道府県名は不明である．図 3.9 に，日本語語彙大系の固有名詞意味属性体系の例を示す．図 3.9 に示す通り，一行には「単語，読みかた，単語の属性（一般名詞か固有名詞か），意味属性番号と意味属性名」が記述されている．

峰島 (みねじま)[固]　67 姓
峯須川 (みねすかわ)[固]　50 河川湖沼名
岑介 (みねすけ)[固]　69 名 (男)
峰巣鼻 (みねすばな)[固]　49 陸上地形名
美祢線 (みねせん)[固]　63 路線名

図 3.9: 日本語語彙大系の固有名詞意味属性体系の例

3.3.2 抽出方法

日本語語彙大系の固有名詞意味属性体系の意味属性を解析し，場所を特定する手がかりとなるような単語に付与されている意味属性を調査した．調査結果を図 3.10 にまとめる．図 3.10 には，手がかり語の抽出対象となる意味属性名と，その意味属性番号を示している．

44 地方名	58 公園名等
45 地区名	59 農牧場名
49 陸上地形名	60 黄泉・泉等名
50 河川湖沼名	63 路線名
52 海洋名	64 交通施設名
53 海底地形名	65 駅名等
57 建物名	98 大学・高専

図 3.10: 手がかり語の抽出対象となる固有名詞意味属性

固有名詞意味属性体系から一行ずつ文字列を取得し，図 3.10 に示した固有名詞意味属性が付与されている単語を，手がかり語として抽出する．その際，図 3.9 に示したように,「単語，読みかた，単語の属性（一般名詞か固有名詞か），意味属性番号と意味属性名」という並びで記述されているので，一行の先頭にある単語の部分のみを正規表現を用いて抽出する．抽出した単語が，郵便番号データおよび国内観光情報サイトから抽出した手がかり語と一致していた場合，その単語は新たに手がかり語辞書に登録しない．

3.3.3 抽出結果

日本語語彙大系から抽出した都道府県名（不明）と手がかり語の例を，図 3.11 に示す．日本語語彙大系から抽出した手がかり語は，その都道府県名が不明である．しかし，都道府県名と手がかり語を組にして登録するという手がかり語辞書の記述の仕方の整合性を保つために，不明な都道府県名を”*”で表現して，手がかり語と組にして手がかり語辞書に登録する．

*, 伯耆大山
*, 剥岳
*, 博士山
*, 博士峠
*, 博多浦

図 3.11: 日本語語彙大系から抽出した都道府県名（不明）と手がかり語の組の例

表 3.3 に，日本語語彙大系から抽出した手がかり語の件数を示す．

表 3.3: 日本語語彙大系から抽出した手がかり語の件数

件数
34,495 件

3.4 合計件数

手がかり語辞書の合計件数を表 3.4 に示す．手がかり語辞書には，郵便番号データ，国内観光情報サイト，および日本語語彙大系から抽出した手がかり語が登録されている．手がかり語辞書の合計件数には，手がかり語のみを取り出してユニークをかけた場合の件数を示している．

表 3.4: 手がかり語辞書の合計件数

件数
224,144 件

手がかり語辞書の登録件数の内訳を表 3.5 に示す．表 3.5 には，手がかり語辞書に登録されている手がかり語に対して，その都道府県名に関する分類によりそれぞれの登録件数を示している．

表 3.5: 手がかり語辞書の登録件数の内訳

都道府県名	件数
不明	34,495 件
1 県	148,216 件
2 県以上	41,433 件

第4章　都道府県コーパス

本章では，都道府県コーパスの作成方法と構成について説明する．

4.1　作成方法

都道府県コーパスを作成するために，まずWikipedia[5]における各都道府県について記述されたWebページを収集する．次に，収集したWebページからMeCab[6]を用いて形態素解析を行い，固有名詞を抽出して作成する．具体的には，形態素解析において「名詞, 固有名詞, 地域」および「名詞, 固有名詞, 一般」と分類された単語のみを抽出する．

ここで,「名詞, 固有名詞,組織」および「名詞, 固有名詞,人名」に分類された固有名詞は，各都道府県ごとの共起語にふさわしくないと判断し，抽出対象としない．理由は,「名詞, 固有名詞, 組織」はアルファベットの羅列がこれに分類されることが多く,「名詞, 固有名詞, 人名」は人名に用いられる固有名詞がこれに分類されるからである．よって，アルファベットの羅列や人名に用いられる固有名詞を共起語として都道府県コーパスに登録しないために,「名詞, 固有名詞, 組織」および「名詞, 固有名詞, 人名」は抽出対象としない．

各都道府県ごとのWebページから抽出した固有名詞を，それぞれ都道府県ごとにコーパスとしてテキストファイルに登録した．つまり，都道府県コーパスは，47個のコーパスから構成される．

4.2 作成結果

都道府県コーパスの登録例（鳥取コーパス）を図 4.1 に示す．都道府県コーパスには，図 4.1 に示すように，共起語とその点数を組にして登録している．本研究では，点数はすべて 1 点としている．なお，本研究では具体的な手法は提案できなかったが，共起語ごとに重み付けを行う場合には，点数を調整する方法が考えられる．

```
青谷,1
千代川,1
倉吉,1
倉吉線,1
倉吉平野,1
```

図 4.1: 都道府県コーパス（鳥取コーパス）の登録例

都道府県コーパスの共起語の件数を表 4.1 に示す．表 4.1 の共起語の件数は，各都道府県ごとのコーパスに登録されている共起語の件数を合算したものである．

表 4.1: 都道府県コーパスの共起語の件数

件数
12,548 件

第5章　地名の解析手法

本章では，地名の解析手法について説明する．地名の解析は,「手がかり語検出」と「都道府県名の曖昧性軽減」で構成する．

5.1　地名の解析手法の概要

本節では，本研究において行う地名解析の全体像について説明する．本研究における地名の解析は,「手がかり語検出」と「都道府県名の曖昧性軽減」という 2 つの手法から構成する．地名解析における入力データはブログ記事である．

まずブログ記事中の，場所を判定する手がかりとなる語にその都道府県名をタグ形式で文章中に挿入する．この処理が手がかり語検出である．手がかり語検出を行うために，場所を判定する手がかりとなる語とその都道府県名が組になって登録されている「手がかり語辞書」を用いる．

次に，1 つの手がかり語に対して複数出力された都道府県名に対し，ブログ記事単位の有力都道府県名を判定してマスク処理を行い，余計な都道府県名の出力を抑制する．この処理が都道府県名の曖昧性軽減である．都道府県名の曖昧性軽減を行うために，各都道府県名と共起する単語が登録されている「都道府県コーパス」を用いる．

以上の 2 つの提案手法を用いて，場所を判定する手がかりとなる語に曖昧性が軽減された都道府県名がタグ形式で付与されているブログ記事が出力される．

5.2　手がかり語検出

本節では，手がかり語検出の方法とその動作例について説明する．手がかり語検出は，形態素解析器（MeCab）および手掛かり語辞書を用いて行う．

5.2.1　手がかり語検出のアルゴリズム

以下に，手がかり語検出のアルゴリズムを示す．

手順 1 手がかり語検出を行う対象のブログ記事から，1 文取得する．

手順 2 取得した 1 文に対して，形態素解析を行い，名詞を抽出する．

手順 3 抽出した名詞が手がかり語辞書に登録されていれば，文章中のその名詞の箇所に，都道府県名と共にタグを挿入する．

次に，手がかり語検出の際に指定できる動作条件を説明する．手順 2 において名詞を抽出する際に，抽出対象に含める品詞情報を以下の 4 種類に指定することができる．

- 「名詞」すべて
- 「名詞, 一般」および「名詞, 固有名詞」
- 「名詞, 一般」および「名詞, 固有名詞」から人名を排除したもの
- 「名詞, 固有名詞, 地域」および「名詞, 固有名詞, 一般」

これは，手がかり語辞書には地名であるが人名でもある単語や，地名であるが一般名詞でもある単語が存在するため，抽出対象を変化させることによる手がかり語検出の性能の違いをみるためである．

最後に，検出結果の出力条件を説明する．手順 3 において挿入するタグの形式を以下に示す．

- <pl name="都道府県名 1, 都道府県名 2,…">手がかり語</pl>

上記に示すように，手がかり語の前後にタグが挿入される形式になっている．タグの中の name 欄に，手がかり語辞書に登録されている都道府県名が記述される．この name 欄には，単一の都道府県名が記述される場合，複数の都道府県名が記述される場合，および都道府県名不明の”*”が記述される場合がある．

5.2.2　手がかり語検出の動作例

図 5.1 に，ブログ記事の例を示す．図 5.2 に，そのブログ記事に手がかり語検出を行った動作例を示す．なお動作例では，名詞抽出の際に「名詞, 固有名詞, 地域」および「名詞, 固有名詞, 一般」を抽出対象とした場合の例を示している．

以前，携帯で移した分は投稿しましたが …
再度，デジカメで写した分を投稿しますね
今日の夕陽は綺麗であって欲しいなァ～
東名自動車道の由比を過ぎた辺りだったと思います
浜名湖のＳＡから対岸を写しました
恋人の聖地のシンボルです
夕陽が沈むまでには，少し時間がありましたが …
緊急付き添いで借り出された奈々ｃｈａｎ
ここはお気に入りなのでご機嫌です
教え子の … 生きるの死ぬので呼び出された
大変なドライブでしたが …
景色が救ってくれたかな
明日は，この帰り道の …
ドライブインで寝ちゃった後の明け方の
富士山を …

図 5.1: ブログ記事の例

以前，携帯で移した分は投稿しましたが …
再度，デジカメで写した分を投稿しますね
今日の夕陽は綺麗であって欲しいなァ～
<pl name="愛知, 静岡">東名</pl>自動車道の由比を過ぎた辺りだったと思います
<pl name="静岡">浜名湖</pl>のＳＡから対岸を写しました
恋人の聖地のシンボルです
夕陽が沈むまでには，少し時間がありましたが …
緊急付き添いで借り出された奈々ｃｈａｎ
ここはお気に入りなのでご機嫌です
教え子の … 生きるの死ぬので呼び出された
大変なドライブでしたが …
景色が救ってくれたかな
明日は，この帰り道の …
ドライブインで寝ちゃった後の明け方の
<pl name="山梨, 愛知, 長野, 静岡">富士山</pl>を …

図 5.2: 手がかり語検出の動作例

手がかり語検出は，図 5.2 に示すような形式で動作する.「東名」,「富士山」は該当する都道府県名が複数あるため，複数の都道府県名が付与されている.「浜名湖」は該当する都道府県名が静岡のみであったため，静岡のみが付与されている．また,「由比」は検出すべき手がかり語であると考えられるが，形態素解析において「名詞, 固有名詞, 人名」と判定されたため，この例では検出することはできなかった.

5.3 都道府県名の曖昧性軽減

本節では，都道府県名の曖昧性軽減の方法とその動作例について説明する．都道府県名の曖昧性軽減は，手掛かり語検出後のブログ記事および都道府県コーパスを用いて行う．

5.3.1 都道府県名の曖昧性軽減のアルゴリズム

以下に，都道府県名の曖昧性軽減のアルゴリズムを示す．

まず，1 つのブログ記事内の有力な都道府県名 $\tilde{r}$ を，以下の式で判定する．このとき都道府県コーパスを用いる．

$$\tilde{r} = \arg\max_{r \in C} \sum_{n \in N} c(r, n)$$

$$c(r, n) = \begin{cases} 1 & if\ r \text{ のコーパスに } n \text{ が存在} \\ 0 & otherwise \end{cases}$$

ここで C は都道府県名の集合，N は 1 つのブログ記事内の固有名詞の集合である．ここで固有名詞とは，MeCab の形態素解析において,「名詞, 固有名詞, 地域」および「名詞, 固有名詞, 一般」と判定された単語を表す．$c(r, n)$ は固有名詞 n が都道府県名 r のコーパスに存在する場合に 1 を返し，それ以外は 0 を返す関数である．

次に，以下の式で，$\tilde{r}$ を用いて都道府県名の曖昧性軽減を行う．

$$S' = \bigcup_{(w,P) \in S} (w, m(P, \tilde{r}))$$

$$m(P, r) = \begin{cases} \{r\} & if\ r \in P \\ P & otherwise \end{cases}$$

ここで，S は手がかり語 w と，都道府県名集合 P の対を 1 つのブログ記事から集めた集合である．$m(P, r)$ は都道府県名集合 P の中に都道府県名 r が存在すれば $\{r\}$ を，無ければ P を返す，すなわち r によるマスク関数である．S' は，S に対して都道府県名の曖昧性を軽減したものである．都道府県名の曖昧性軽減は，上記に示したアルゴリズムを用いて行う．まず，1 つのブログ記事内の有力な都道府県名を決定することにより，ブログ記事内で主に話題となっている都道府県名が判定できる．次に，判定した有力な都道府県名を用いて，1 つの手がかり語に対して複数出力された都道府県名の曖昧性軽減を行う．複数出力された都道府県名の中に有力な都道府県名が存在しなかった場合には，手がかり語辞書に登録されている都道府県名の方を信用し，曖昧性軽減を行わず，複数出力されたままにしておく．

5.3.2 都道府県名の曖昧性軽減の動作例

図 5.3 に，都道府県名の曖昧性軽減を行う前のブログ記事の例を示す．図 5.4 に，都道府県名の曖昧性軽減を行った後の動作例を示す．

```
今年で第 38 回目をむかえるパレードみたいです
中央通りを<pl name="三重, 京都, 岡山, 広島, 東京">京橋</pl>から
<pl name="大阪, 東京">日本橋</pl>の三越前まで
各地区の祭り集団がパレードしてました
<pl name="東京">東京</pl>音頭を踊りながら
<pl name="大阪, 東京">日本橋</pl>へ
三越<pl name="大阪, 東京">日本橋</pl>をバックにエイサーのパレード
```

図 5.3: 都道府県名の曖昧性軽減前

```
今年で第 38 回目をむかえるパレードみたいです
中央通りを<pl name="東京">京橋</pl>から
<pl name="東京">日本橋</pl>の三越前まで
各地区の祭り集団がパレードしてました
<pl name="東京">東京</pl>音頭を踊りながら
<pl name="東京">日本橋</pl>へ
三越<pl name="東京">日本橋</pl>をバックにエイサーのパレード
```

図 5.4: 都道府県名の曖昧性軽減後

図 5.4 に示すように，都道府県名の曖昧性を軽減することができた．

以下に，軽減に至る過程を説明する．このブログ記事に出現する手がかり語は，全て正解都道府県名は東京である．ブログ記事単位の有力な都道府県名判定において,「日本橋」および「東京」が都道府県コーパスの東京コーパスにヒットした．よって,「東京」が 1 回と「日本橋」が 3 回出現しているので，“東京”に 4 ポイントが計算される．都道府県コーパスの他の各都道府県ごとのコーパスにはヒットしなかった．よって，ブログ記事単位の有力な都道府県名は “東京” に決定される．都道府県名の曖昧性軽減前では「京橋」および「日本橋」に対して東京以外の余分な都道府県名を付与してしまっているが，都道府県名の曖昧性軽減後では余分な都道府県名の出力を抑え，正解の東京のみが出力されていることが分かる．

第6章　評価実験

本章では，評価実験の方法とその結果について説明する．評価実験の際に用いる正解データの作成方法についても本章で説明する．

6.1　評価実験の概要

本節では，本研究において行う評価実験の概要について説明する．評価実験では，まずそれぞれの提案手法ごとの性能を確認するために,「手がかり語検出の性能評価」および「正しく手がかり語検出が行えた範囲に評価対象を限定した場合の都道府県名の曖昧性軽減の性能評価」を行う．次に本手法の総合性能を確認するために,「評価対象を限定しない場合の本手法全体の性能評価」を行う．さらに，都道府県名の曖昧性軽減においてブログ記事単位で有力な都道府県名を判定する性能を確認するため,「ブログ記事単位で有力な都道府県名を判定する性能評価」を行う．

6.2　単語単位での地名解析

本節では，手がかり語検出および都道府県名の曖昧性軽減について，単語単位での評価を示す．

6.2.1　正解データ

本節では，単語単位での地名解析の評価実験を行う際に用いる正解データの作成方法について説明する．

本研究で用いる正解データの元となるブログ記事は，2010 年 11 月 1 日 14 時 22 分の時点でヤフーブログの旅行カテゴリに登録されていた最新のブログ記事 200 件である．ただし，以下に示す例に該当したブログ記事は除外して，最新のものから順に 200 件を選択する．

- 日本語で書かれていないブログ記事
- 海外旅行について書かれているブログ記事

正解データの元となるブログ記事 200 件について，場所を判定する手がかりとなる単語に対し，人手でタグを挿入して正解データを作成する．このとき挿入するタグの形式を，以下に示す．

- <pl name="都道府県名">手がかり語</pl>

タグの name 欄に記述される都道府県名は，ユニークに判定できればその都道府県名を記述する．ユニークに判定できず，複数の都道府県に存在する可能性がある場合は，”*” を記述する．

人手で手がかり語を判定する際に，タグを付与する対象を定めた．タグの付与対象を以下に示す．

- 地名 (島や山，峠，平野，半島，海，川，滝などの名称も含む)
- 施設名 (ランドマークや道の駅，ホテル，温泉，飲食店，寺院，城などの名称も含む)
- 交通機関名 (駅名や空港，港，電車，船，道路や路線の名称も含む)
- 構造物名 (橋やトンネル，門や像など)
- 地域固有のイベント名 (イベント，祭りなど)

タグを付与する際の条件を以下に示す．

- 付与対象の所在が明確な場合は，その所在に該当する都道府県名のタグを付与する．
- 付与対象の所在を判断するにあたり，Web 検索などの地図情報および，ブログ記事内の文脈情報を考慮してよい．
- 複数都道府県にまたがる山や道，および複数都道府県にまたがる地域の名称 (ex“九州”,“瀬戸内”)，または文脈上所在が判断できないチェーン店の店名など，付与対象が地名や施設名であることは読み取れるが，その所在が一都道府県に断定できない場合は，“*” を付与する．
  しかし文脈上所在が断定できるときは，その都道府県名のタグを付与する．
- タグは，地名および施設名であると判断できる文字列の最長に対して付与する．
  (例：“鳥取県” ならば “<pl>鳥取</pl>県” ではなく “<pl>鳥取県</pl>”)

表 6.1 に作成した評価実験用正解データの統計情報を示す．表 6.1 に示すように，ブログ記事ごとに文字数や手がかり語の出現数など，大きくばらつく結果となった．ここで，総手がかり語数と総都道府県名数に差があるのは，ユニークに都道府県名を付与できず“*”を付与した手がかり語が存在するためである．

表 6.1: 正解データの統計情報

評価実験用データの統計情報	数値
総記事数	200 件
総文字数	106,635 文字
平均文字数	533.18 文字
文字数の分散	252,341.62
総手がかり語数	1,497 語
総都道府県名数	1,386 件
最も文字数の多い記事の文字数	3,852 文字
最も文字数の少ない記事の文字数	13 文字
最も手がかり語の多い記事の手がかり語数	63 個
最も手がかり語の少ない記事の手がかり語数	0 個

6.2.2 手がかり語検出の評価

本節では，手がかり語検出の性能の評価を示す．正解データはヤフーブログから取得したブログ記事 200 件から作成したものである．正解データには検出すべき手がかり語が 1,497 件存在する．

適合率と再現率および F 値は以下のように定義する．

$$\text{適合率} = \frac{\text{正しい出力手がかり語数}}{\text{出力した手がかり語の総数}}$$

$$\text{再現率} = \frac{\text{正しい出力手がかり語数}}{\text{正解データの手がかり語の総数}}$$

$$F\text{ 値} = \frac{2 * (\text{適合率} * \text{再現率})}{\text{適合率} + \text{再現率}}$$

手がかり語検出において，正解手がかり語の一部のみ検出できる場合がある．例として,「東京タワー」という手がかり語を検出する場合を考える．形態素解析器が「東京タワー」を途中で区切らずに名詞と判定し，かつ「東京タワー」が手がかり語辞書に登録されている場合に，正しく「東京タワー」を手がかり語として検出できる．しかし，形態

素解析器が「東京」と「タワー」のようにそれぞれを品詞として区切ってしまい，かつ「東京」は手がかり語辞書に登録されていた場合，正解手がかり語の「東京タワー」の部分文字列である「東京」のみ検出できる．このように，正解手がかり語の文字列全体を正しく検出できた時のみを正解とする評価方法と，正解手がかり語の文字列の一部分を検出できたら正解とする評価方法の 2 つが考えられる．本研究の評価実験において，前者の評価方法を「部分マッチ」，後者の評価方法を「完全マッチ」と呼ぶ．

以下に，単語単位の評価における評価方法を示す．

- 部分マッチ
  正解手がかり語の文字列の一部分を検出できたら正解とする．
  適合率を計算する場合，一つの正解手がかり語に対して複数手がかり語検出を行ったとき，複数回正解とする．
  再現率を計算する場合，一つの正解手がかり語に対して一部分でも手がかり語検出ができれば正解とする．
  例として，正解データが「<pl name="東京">東京</pl>に行きました!!<pl name="東京">東京タワー</pl>に登りました !! 」というブログ記事があった場合を考える．
  このとき手がかり語検出において「<pl name="東京">東京</pl>に行きました !! <pl name="東京">東京</pl>タワーに<pl name="大阪">登</pl>りました !! 」という結果が出力されたとき，適合率は 2/3，再現率は 2/2 となる．

- 完全マッチ
  正解手がかり語の文字列全体を正しく検出できた時のみを正解とする．例として，正解データが「<pl name="東京">東京</pl>に行きました !! <pl name="東京">東京タワー</pl>に登りました !! 」というブログ記事があった場合を考える．
  このとき手がかり語検出において「<pl name="東京">東京</pl>に行きました !! <pl name="東京">東京</pl>タワーに<pl name="大阪">登</pl>りました !! 」という結果が出力されたとき，適合率は 1/3，再現率は 1/2 となる．

表 6.2 に，手がかり語検出を部分マッチにて評価した適合率，再現率および F 値を示す．表 6.2 には，手がかり語検出の際の品詞情報による抽出対象の区分ごとの評価を示している．ここで，適合率の分子と再現率の分子の値が異なるのは，一つの正解手がかり語に対して複数回手がかり語検出を行ったときに，複数回正解としているからである．

表 6.2: 手がかり語検出（部分マッチ）の評価

抽出対象	適合率	再現率	F 値
名詞	0.312(1824/5851)	0.855(1280/1497)	0.457
名詞,(一般 or 固有名詞)	0.401(1511/3770)	0.810(1212/1497)	0.536
名詞,(一般 or 固有名詞) 人名排除	0.399(1343/3367)	0.742(1111/1497)	0.519
名詞, 固有名詞,(一般 or 地域)	0.777(933/1200)	0.576(863/1497)	0.662

表 6.3 に，手がかり語検出を完全マッチにて評価した適合率，再現率および F 値を示す．表 6.3 には，手がかり語検出の際の品詞情報による抽出対象の区分ごとの評価を示している．

表 6.3: 手がかり語検出（完全マッチ）の評価

抽出対象	適合率	再現率	F 値
名詞	0.101(593/5851)	0.396(593/1497)	0.161
名詞,(一般 or 固有名詞)	0.157(593/3770)	0.396(593/1497)	0.225
名詞,(一般 or 固有名詞) 人名排除	0.158(532/3367)	0.355(532/1497)	0.219
名詞, 固有名詞,(一般 or 地域)	0.408(490/1200)	0.327(490/1497)	0.363

表 6.4 に，形態素解析を行わずに前方最長一致で手がかり語検出を行い，正解手がかり語を文字単位に評価した場合の適合率，再現率および F 値を示す．適合率と再現率および F 値は以下のように定義する．

$$\text{適合率} = \frac{\text{正しい出力文字数}}{\text{出力した手がかり語の総文字数}}$$

$$\text{再現率} = \frac{\text{正しい出力文字数}}{\text{正解データの手がかり語の総文字数}}$$

$$F\text{ 値} = \frac{2 * (\text{適合率} * \text{再現率})}{\text{適合率} + \text{再現率}}$$

表 6.4: 文字単位の評価

適合率	再現率	F 値
0.205(4844/23683)	0.829(4844/5845)	0.328

表 6.4 に示すように，形態素解析を用いない場合，正解手がかり語の文字列のうち約 83%の文字を手がかり語検出できていることが再現率より分かる．よって，作成した手がかり語辞書を用いたときに，その運用をうまく行えば，完全マッチで評価した場合でも再現率は最大約 83%の性能が見込めることが示された．

6.2.3 都道府県名の曖昧性軽減の評価

本節では，都道府県名の曖昧性軽減の性能の評価を示す．正解データはヤフーブログから取得したブログ記事 200 件から作成したものである．ここでは，都道府県名の曖昧性軽減の評価を純粋に行うために，手がかり語検出において正しく検出できた手がかり語のみを評価対象とする．さらに，正解データにおいて手がかり語として判定したが都道府県名を断定できずに “*” を付与したものと，手がかり語検出において都道府県名が不明で “*” が出力されたものは評価対象から除外する．このように評価対象を限定したとき，例えば抽出対象が「名詞, 固有名詞, 地域 or 名詞, 固有名詞, 一般」の場合に部分マッチで評価を行ったとき，正しく検出できた手がかり語は 6.2.2 節の表 6.2 より 863 件である．その中で正解データにおいて都道府県名に “*” が付与されているものを除くと，ここで検出すべき都道府県名は 701 件存在するということになる．

適合率と再現率および F 値は以下のように定義する．

$$\text{適合率} = \frac{\text{正しい出力都道府県名数}}{\text{出力した都道府県名の総数}}$$

$$\text{再現率} = \frac{\text{正しい出力都道府県名数}}{\text{正解データの都道府県名の総数}}$$

$$F\text{ 値} = \frac{2 * (\text{適合率} * \text{再現率})}{\text{適合率} + \text{再現率}}$$

表 6.5 に，評価方法に部分マッチを用いたときに正しく検出できたと判定された手がかり語について，都道府県名の曖昧性軽減を行ったときの適合率，再現率および F 値を示す．

表 6.5: 都道府県名の曖昧性軽減（評価範囲限定，部分マッチ）の評価

入力データの抽出対象	軽減	適合率	再現率	F 値
名詞	前	0.080(970/12082)	0.771(849/1101)	0.145
名詞	後	0.191(908/ 4764)	0.726(799/1101)	0.302
名詞,(一般 or 固有名詞)	前	0.099(897/ 9098)	0.793(818/1032)	0.175
名詞,(一般 or 固有名詞)	後	0.233(842/ 3611)	0.746(770/1032)	0.355
名詞,(一般 or 固有名詞) 人名排除	前	0.098(781/ 7979)	0.765(717/ 937)	0.174
名詞,(一般 or 固有名詞) 人名排除	後	0.234(732/ 3129)	0.719(674/ 937)	0.353
名詞, 固有名詞,(一般 or 地域)	前	0.164(646/ 3949)	0.864(606/ 701)	0.275
名詞, 固有名詞,(一般 or 地域)	後	0.431(615/ 1428)	0.825(578/ 701)	0.566

表 6.5 には，手がかり語検出の際の品詞情報による抽出対象の区分ごとに，それぞれ都道府県名の曖昧性軽減を行う前と行った後について示している．抽出対象が「名詞, 固有名詞, 一般 or 名詞, 固有名詞, 地域」の場合の評価において，都道府県名の曖昧性軽減を行う前の適合率が 0.164 と低く，1 つの手がかり語に対して大量の都道府県名を出力してしまっていることが分かる．しかし，都道府県名の曖昧性軽減を行う前から再現率は 0.864 と高いため，大量に出力してしまった都道府県名の中に，正解の都道府県名が含まれている率は高いことが分かる．都道府県名の曖昧性軽減を行う前と行った後で再現率は 0.864 から 0.825 へわずかに下がっているが，適合率が 0.164 から 0.431 へと大幅に上昇しているため，都道府県名の曖昧性軽減の有効性が確認できる．また，適合率の分子と再現率の分子の値が異なるのは，一つの正解手がかり語に対して複数回手がかり語検出を行ったときに，複数回検出した手がかり語において正解都道府県名が出力できていれば，複数回正解としているからである．

表 6.6 に，評価方法に完全マッチを用いたときに正しく検出できたと判定された手がかり語について，都道府県名の曖昧性軽減を行ったときの適合率，再現率および F 値を示す．表 6.6 には，手がかり語検出の際の品詞情報による抽出対象の区分ごとに，それぞれ都道府県名の曖昧性軽減を行う前と行った後について示している．

表 6.6: 都道府県名の曖昧性軽減（評価範囲限定，完全マッチ）の評価

入力データの抽出対象	軽減	適合率	再現率	F 値
名詞	前	0.234(442/1887)	0.940(442/470)	0.375
名詞	後	0.537(426/ 793)	0.906(426/470)	0.675
名詞,(一般 or 固有名詞)	前	0.234(442/1887)	0.940(442/470)	0.375
名詞,(一般 or 固有名詞)	後	0.537(426/ 793)	0.906(426/470)	0.675
名詞,(一般 or 固有名詞) 人名排除	前	0.235(393/1671)	0.952(393/413)	0.377
名詞,(一般 or 固有名詞) 人名排除	後	0.582(378/ 650)	0.915(378/413)	0.711
名詞, 固有名詞,(一般 or 地域)	前	0.247(356/1444)	0.954(356/373)	0.392
名詞, 固有名詞,(一般 or 地域)	後	0.570(345/ 605)	0.925(345/373)	0.706

6.2.4 本手法の総合評価

本節では，本手法の総合性能を評価する．正解データはヤフーブログから取得したブログ記事 200 件から作成したものである．正解データにおいて手がかり語として判定したが都道府県名を断定できずに “*” を付与したものと，手がかり語検出において都道府県名が不明で “*” が出力されたものは評価対象から除外する．このとき，正解データには検出すべき都道府県名が 1,386 件存在する．

適合率と再現率および F 値は以下のように定義する．

$$\text{適合率} = \frac{\text{正しい出力都道府県名数}}{\text{出力した都道府県名の総数}}$$

$$\text{再現率} = \frac{\text{正しい出力都道府県名数}}{\text{正解データの都道府県名の総数}}$$

$$F\text{ 値} = \frac{2 * (\text{適合率} * \text{再現率})}{\text{適合率} + \text{再現率}}$$

表 6.7 に，手掛かり語検出の評価方法に部分マッチを用いた時の都道府県名の曖昧性軽減を行ったときの適合率，再現率および F 値を示す．表 6.7 には，手がかり語検出の際の品詞情報による抽出対象の区分ごとに，それぞれ都道府県名の曖昧性軽減を行う前と行った後について示している．

表 6.7: 都道府県名の曖昧性軽減の評価（部分マッチ）

抽出対象	軽減	適合率	再現率	F 値
名詞	前	0.023(970/41409)	0.613(849/1386)	0.045
名詞	後	0.045(908/20194)	0.576(799/1386)	0.083
名詞,(一般 or 固有名詞)	前	0.034(897/26731)	0.590(818/1386)	0.064
名詞,(一般 or 固有名詞)	後	0.067(842/12557)	0.556(770/1386)	0.120
名詞,(一般 or 固有名詞) 人名排除	前	0.034(781/22889)	0.517(717/1386)	0.064
名詞,(一般 or 固有名詞) 人名排除	後	0.070(732/10423)	0.486(674/1386)	0.123
名詞, 固有名詞,(一般 or 地域)	前	0.114(646/ 5675)	0.437(606/1386)	0.181
名詞, 固有名詞,(一般 or 地域)	後	0.281(615/ 2185)	0.417(578/1386)	0.336

表 6.8 に，手掛かり語検出の評価方法に完全マッチを用いた時の都道府県名の曖昧性軽減を行ったときの適合率，再現率および F 値を示す．表 6.8 には，手がかり語検出の際の品詞情報による抽出対象の区分ごとに，それぞれ都道府県名の曖昧性軽減を行う前と行った後について示している．

表 6.8: 都道府県名の曖昧性軽減の評価（完全マッチ）

抽出対象	軽減	適合率	再現率	F 値
名詞	前	0.011(442/41409)	0.319(442/1386)	0.021
名詞	後	0.021(426/20194)	0.307(426/1386)	0.039
名詞,(一般 or 固有名詞)	前	0.017(442/26731)	0.319(442/1386)	0.031
名詞,(一般 or 固有名詞)	後	0.034(426/12557)	0.307(426/1386)	0.061
名詞,(一般 or 固有名詞)」人名排除	前	0.017(393/22889)	0.284(393/1386)	0.032
名詞,(一般 or 固有名詞)」人名排除	後	0.036(378/10423)	0.273(378/1386)	0.064
名詞, 固有名詞,(一般 or 地域)	前	0.063(356/ 5675)	0.257(356/1386)	0.101
名詞, 固有名詞,(一般 or 地域)	後	0.158(345/ 2185)	0.249(345/1386)	0.193

6.3　ブログ記事単位での地名解析

本節では，都道府県名の曖昧性軽減を行う際の，記事単位に有力な都道府県名を判定するタスクの評価を示す．これは，都道府県名の曖昧性軽減を行う際の，ブログ記事単位の有力な都道府県名判定の性能を評価するためである．

正解データはヤフーブログから取得したブログ記事 50 件に対して，人手で記事ごとに有力な都道府県名を付与して作成したものである．図 6.1 に，正解データの例を示す．例に示したブログ記事では，人手による有力な都道府県名判定において正解都道府県名を “北海道” と判定した．

北海道旅行の続編で，サッポロビール園についてもう少し詳しく紹介します．
色々なコースがあるのですが，やっぱりお勧めは
ジンギスカン食べ飲み放題コースです !!
生ラム，トラディショナルジンギスカン，野菜，ビール
他ソフトドリンクなど食べ飲み放題で 3600 円くらい
でした．
ビールは隣のビール工場で作られたばかりのもので，
非常に飲み易く，味が全然違います！
生ラムだけとか好きな野菜だけのおかわりもＯＫです．
今までに北海道旅行は２０回くらい行きましたが，
必ず毎回サッポロビール園には行ってま〜す !!

図 6.1: 有力な都道府県名判定の正解データ（北海道）

5.3.1 節に示したアルゴリズムを用いて，ブログ記事単位の有力な都道府県名の出力を行い，正解データに付与した都道府県名との比較を行って，一致率を調査した．図 6.1 の例では，アルゴリズムを用いたブログ記事単位の有力な都道府県名の判定において，「北海道」が北海道コーパスにマッチし “北海道” に 2 ポイントが計算される．よって，アルゴリズムを用いたブログ記事単位の有力な都道府県名の判定において出力が “北海道” となり，正解データと同一の都道府県名を出力することができた．このように，正解データのブログ記事 50 件を 1 件ずつアルゴリズムを用いて都道府県名判定の出力を行い，正解都道府県名と同一の出力を得られれば一致したと評価する．

上記のように評価を行い，アルゴリズムを用いた記事単位の有力な都道府県名判定の出力と，正解データに付与した都道府県名との一致率を調査した結果，60%(30/50) となった．

第7章　考察

本章では，本研究で提案した手法について考察を行う．また，今後の課題について述べる．

7.1　手がかり語検出

手がかり語検出において，最もF値が高かったのは形態素解析において「名詞, 固有名詞, 地域」,「名詞, 固有名詞, 一般」のみを抽出対象とした場合であった．これは，他の3つの抽出対象による区分よりも適合率が高かったためである．抽出対象に強く制限をかけるほど，若干の誤差はあるものの適合率が上昇していき再現率は低下していくという形になった．これは場所を判定する手がかりとなる語には，一般に使われる名詞，人名にも用いられる名詞，固有名詞とさまざまなものが含まれており，抽出対象に制限をかければ手がかり語として信頼性の高いものが得られるが，同時に本当は手がかり語であるものを除外してしまうためだと考えられる．

上記のように形態素解析による検出誤りには，もうひとつの誤り原因が存在する．それは，名詞連続の過分割である．名詞連続から構成される単語を正解手がかり語とした場合，その名詞連続を形態素解析によって別の単語であると判定されてしまう場合がある．このとき，正解手がかり語が手がかり語辞書に登録されていても，正解手がかり語となる単語を分割してそれぞれにおいて手がかり語辞書を参照してしまったために，正しく手がかり語検出が行えない．このような，形態素解析が原因の手がかり語検出誤りがあると考えられる．この対策としては，本研究では実装できていないが，形態素解析結果において隣接する名詞は，それぞれにおいて手がかり語辞書を参照するとともに，隣接する名詞は一つの単語とみなして手がかり語辞書を参照するという処理がある．

さらに，正しく形態素解析が行えたが，もともとその単語が手がかり語辞書に登録されていなかったという誤り原因も存在する．この対策としては，手がかり語辞書の登録件数を増やすということが考えられる．しかし，手がかり語と都道府県名を信頼性高く同時に取得する手段は限られているという問題がある．

7.2　都道府県名の曖昧性軽減

都道府県名の曖昧性軽減を行うことにより，余計な都道府県名の出力を抑制することができた．しかし，曖昧性軽減を行っても出力を抑制できなかった場合もあった．それは，1 つの手がかり語に複数出力された都道府県名の中に，有力な都道府県名が存在しなかった場合である．このときは，複数出力されている都道府県名を抑制することができない．

この対策としては 2 つ考えられる．1 つ目の対策として，有力な都道府県名を判定するウィンドウ幅を，ブログ記事単位からもっと狭めて設定するということが考えられる．ブログ記事には，場所の移動が時系列に沿って記述されている場合が多い．ブログ記事単位に有力な都道府県名の推定を行うと，場所の移動が激しいブログ記事を扱うことは難しい．よって，有力な都道府県名を判定するウィンドウ幅を，3 文単位や 5 文単位といった文単位に設定することで，より精度よく有力な都道府県名を推定することができ，もっと効率よく都道府県名の曖昧性軽減を行うことができると考えられる．2 つ目の対策として，有力な都道府県名を判定する際に，その候補を複数出力するということが考えられる．例えば，あるウィンドウ幅において，有力な都道府県名を第一候補は “鳥取”，第二候補は “岡山” などと候補を複数出力すれば，第一候補の “鳥取” を用いて都道府県名の曖昧性軽減が行えなかった手がかり語に対して，第二候補の “岡山” を用いれば都道府県名の曖昧性軽減が行えるということが考えられる．

7.3　今後の課題

今後の課題として，本研究で提案した「手がかり語検出」および「都道府県名の曖昧性軽減」という 2 つの手法の改良を行っていく必要がある．手がかり語検出の改良方法としては，名詞連続の過分割を解決するために，形態素解析結果おいて隣接する名詞は一つの単語とみなして処理を行うことがあげられる．都道府県名の曖昧性軽減の改良方法としては，有力な都道府県名を推定するウインドウ幅を文単位に設定して処理を行うことと，有力な都道府県名の候補を複数出力して処理を行うことがあげられる．

これら手法の改良を行い，地名解析の精度を改善することができれば，ブログ記事からの観光情報分析に役立てることができると考えられる．

第8章　おわりに

本研究では，ブログ記事における地名解析を行うために,「手がかり語検出」と「都道府県名の曖昧性軽減」という 2 つの手法を提案した．

手がかり語検出では，場所を判定する手がかりとなる語とその都道府県の組を登録した手がかり語辞書を作成し，それを用いてブログ記事中の手がかり語にタグ形式で都道府県名を付与する処理を行った．

都道府県名の曖昧性軽減では，都道府県名と共起する語が登録された都道府県コーパスを作成し，それを用いてブログ記事単位に有力な都道府県名を推定し，1 つの手がかり語に複数出力された都道府県名を抑制する処理を行った．

さらに，提案した手法の評価実験を行った．手がかり語検出では，正解となる手がかり語の文字列のうち一部を手がかり語として検出できればよいという評価において F 値で 0.662 という評価結果であった．都道府県名の曖昧性解消では，評価対象を上述の手がかり語検出において正しく検出できた手がかり語に限定した場合において，F 値 0.566 という評価結果であった．正解手がかり語のうち一部を検出できればよいという評価において手がかり語検出を行い，評価対象を限定せずに都道府県名の曖昧性軽減を行ったときの，本手法全体の性能評価は，F 値で 0.336 という評価結果であった．一方，ブログ記事単位で有力な都道府県名を判定する性能の評価は，一致率で 60%という評価結果であった．

以上の評価結果により，手がかり語辞書を用いて手がかり語検出を行い，その後都道府県コーパスを用いて都道府県名の曖昧性軽減を行うという本研究で提案した地名解析の手法の有用性を確認した．

謝辞

本研究を進めるに当たり，種々の御助言を頂きました村田真樹教授，および，村上仁一准教授に心から御礼申し上げます．

また，徳久雅人講師には，終始に渡り研究の進め方や本論文の書き方など，細部に渡る御指導を頂きました．ここに深く感謝いたします．

その他様々な場面で御助力をいただいた計算機工学講座 C 村田研究室の皆様に感謝の意を表します．

参考文献

[1] 安田宜仁, 平尾努, 鈴木潤, 磯崎秀樹: “ブログ作者の居住域の推定”, 言語処理学会第12回年次大会, pp.512-515, 2006.

[2] 日本郵便: http://www.post.japanpost.jp/index.html

[3] 国内旅行観光情報・大好き日本: http://www.gojapan.jp/

[4] 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”, 岩波書店, 1997.

[5] Wikipedia: http://ja.wikipedia.org/wiki/

[6] MeCab: http://mecab.sourceforge.net/