ブログ作者の居住域の推定

安田宜仁, 平尾努, 鈴木潤, 磯崎秀樹
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町光台 2-4
{n-yasuda, hirao, jun, isozaki}@cslab.kecl.ntt.co.jp

概要

ブログ作者が居住している地域を推定する方法を提案する. 居住域の推定はブログを対象とした情報検索を行う際の制約として使えるだけでなく, 評判分析の結果との組み合わせや, 曖昧な語の解消に役立てることができる. 居住地の近くにある地名とそうでない地名とでは, 地名の周囲の局所的な文脈に違いがある場合が多いことに着目し, 地名の出現箇所が作者の居住地域であるかどうかを判定する分類器を用いる. さらに, ブログは継続して書かれることに注目し, 多数の地名の出現に対する分類器の出力を用いた投票を行いブログ作者の居住地域の推定を行う. 収集したブログを対象に都道府県を推定する実験を行い有効性を検証する.

1 はじめに

本稿では, ブログ作者 (ブロガー) の居住している地域の推定する方法を提案する. 居住している地域の推定とは, ブログから得られた情報を元に, その作者が現在住んでいる地域を県や市といった固定された単位で推定することである. このようなブロガーの居住域の推定の直接的な利用方法として, 情報検索時のひとつの制約に用いることが考えられる. つまり特定の地域に住んでいる人が書いたブログに限定して検索を行うことが可能となる. このような直接的な利用方法だけでなく居住域の推定は間接的にも有用である. Herring ら [1] は, 70% のブログは個人的ジャーナルとして分類されると報告している. つまり, 相当量のブログが, 個人的な出来事や意見を述べる日記のような形で提供されている. このような個人的な出来事や意見は近年注目を浴びている評判や意見の分析 (例えば [2, 3] など) にとって重要な対象である. 評判や意見のの分析結果に, 居住域の情報を加わることによって, 特定の地域の人がどのような意見を持っているかの分析が可能となるだろう. 別の間接的な利用方法としては, テキスト情報からだけでは解決が困難な照応の解決や, 地理情報の観点において曖昧な語の解決の補助に用いることが考えられる. たとえば, 「今日は新しい図書館に行ってきました. デザインがシンプルで機能的でかっこ良かったです」といった表現があったとする. ここでの「図書館」はこの表現を書いた人にとっては特定の図書館を差しているはずであるが, この図書館を特定することは文の外の知識がない限りほぼ不可能である. このような場合に, 作者の居住域が分かればその図書館を特定するヒントとして用いることができると考えらえる.

居住域の素朴な推定のために, ブログ中に出現した地名が居住域を示していると仮定することが考えられる. しかし, 旅行先についてのブログや, どこか遠くで起こった事件に関心を持った場合のブログなどは多数の地名を含むが, 出現した地名とブロガーの居住地とは関係がない.

すべての地名を同様に扱うのではなく, 本稿ではブログ中に地名が出現した場合, 地名の周囲の局所的な文脈に注目する. 地名がブロガーの居住域にある場合とそうでない場合とでは, 周囲の言葉の使い方や言い回しに違いが出ることが多いと考えられるからである. たとえば以下の文例ではどちらも「博多駅」という地名が出現していて, しかも作者の居住地については言及されていない. しかし, 注意深く読んだ場合, 一方はおそらく博多駅の近く (福岡県) に居住している人で, 他方はそうでないことが推測できる.

1.「新型の通勤車両があるらしいという情報を聞きつけて, 自転車で博多駅まで行ってきました」
2.「博多駅には行くのは初めてでした. けっこう大きくてホームの数も多いんですねー」

もし, ある場所の近くに居住していることが分かれば, 対象のブロガーの居住域は場所の所属する地域であると推測できる. そこで本稿では, その地名の近くに作者が居住しているかどうかを, 地名の周囲の文脈を特徴として用いる二値分類器で判定し, 「博多駅 - 福岡県」のような地名と地域名の対応が得られるような地名辞書と組み合わせることで居住域の推定を行う. さらに, ブログは多くの場合同一の個人によって継続的に書かれてあるため, 特定のブログエントリだけを用いて居住域を推定する必要はない. 居住地に変更がないことを前提とすれば, 推定のためには対象ブロガーが記述したブログエントリをすべて用いることができる. そこで, 二値分類器の結果の投票を行うことによって精度の高い居住域の推定を行う方法を提案する.

地域を考慮する技術として, ウェブページと地理情報を対応付けるための研究が多数なされている ([4] など). 本手法においては特定のページがどの地域と関連があるかではなく, その作者がどの地域に居住して

いるかに注目しているので, これらの手法と相補的に用いることも可能であると考えられる.

以降の節では提案する居住域の推定方法について述べ, 実際のブログサイトより収集したブログを対象にして提案法の検証を行う.

2 居住域の推定

居住域の特定は基本的には複数クラスの分類問題と捉えて, 複数クラスの分類器を用いることも考えられる. しかし, 一般的にクラス数が非常に多い場合には適用困難であるにもかかわらず, 地域名の数は都道府県を地域の単位とした場合でさえ 47 となるなど, 大きくなりがちである. 居住域の特定をこの問題を複数クラスの問題と扱うのではなく, 地域と関連付けられた地名辞書と, その地名がブロガーの居住域にあるかどうを判定する二値分類器を組合せることによって解決することを試みる. ここで, 二値分類器の学習には地名の周囲の文脈を用い, 地名そのもは用いる必要はないので, 判定対象となる地域が追加されたり, 地名辞書に変更があった場合でも, 分類器の訓練をやり直す必要がないと期待できる.

辞書は地名を現す語の網羅的なリストである必要はない. 本稿の目的は, ブログ中に現れた地名の場所や地域を特定することではなく, ブロガーの居住域を特定することであるので, 地名辞書として必要な特性は, 網羅性よりはむしろ, 近くに住んでいる人とそうでない人によって書き方が変わるような地名を多く含んでいることである. また, 地名とその地域名との関係は本手法全体で依存している重要な情報となるので, 不正確な関係や, 地名と地域名が一対一でないような関係は望ましくない. 地名辞書の作成については 4 節で述べる.

2.1 地名出現箇所の分類

対象のブロガーが書いた一連のブログを入力として, 地名辞書の見出し語が出現した箇所を特定する. 本手法は地名の周囲の言葉の使い方や言い回しを含めた局所的な情報に注目しているので, 地名を含む一文を用いる. 訓練および分類には, 地名を含む文から地名を除いたもの bag-of-words 表現による AdaBoost [5] を用いる. 新しい推定対象の地域や, 新しい地名を取り扱うことになった場合でも分類器を再び訓練せずに済むよう地名そのものは分類器の入力には加えない.

2.2 地名の重み付け

分類器は出現した地名とは独立に動作するのが望ましいが, 一方で, 地名自身でも居住域を表す文脈で出現しやすいかそうでないかという傾向がある. たとえば, 主に地元の人に利用される地方の小さな駅名は, その駅の近くに居住する人々によって記述されることが多いと考えられるし, 観光地として有名な寺の名前は, 旅行記のようなブログで記述されることが多く, 結果として居住者以外によって書かれることが多いと考えられる.

そこで, 分類器の出力と組合せるために地名自身の重みを考慮する. 地名 t の重みを以下のように定める:

$$w(t) = \log \frac{P(t)+1}{N(t)+1}$$

ここで $P(t)$ は地名 t が居住域に現れた回数, $N(t)$ は非居住域に現れた回数を示す. 訓練データ中で一度も出現しなかあった地名については, $w(t) = 0$ とする.

2.3 投票

2 つの種類の投票方法を考える. 1 つは投票について, 分類器の出力の符号を用いる方法であり, もう 1 つは分類器の出力のスコアを用いる方法である. どちらの方法についても 2.2 で述べた地名の重み付けは同様に行う. それぞれの方法における地域 a の投票のスコアは以下のように与えられる:

- 投票方法 1

$$V(a) = \sum_{loc} ((\text{classifier sign for } s(loc)) + \alpha\, w(loc))$$

- 投票方法 2

$$V(a) = \sum_{loc} ((\text{classifier score for } s(loc)) + \beta\, w(loc))$$

ここで, loc は対象のブロガーの記述したブログ中に出現した地名, $s(loc)$ は loc を含む文, $w(loc)$ は loc の地名重み, そして α と β は定数である.

もっとも高い正のスコアを得た地域名を最終的なブロガーの居住域の推定結果として出力する. もし, どの地域も正のスコアを得られなかった場合, 回答を断念する. 本手法は地名辞書に依存しているため, もし出現したすべての地名がブロガーの居住域と関係ない場合には正しい答を出力することはできない. また, このような状況は十分起こり得る. きっと間違っているような回答をするよりも, 断念する方が適切であると考える.

3 コーパス

収集したコーパスについて述べる. 多くのブログサービスサイトでは, ブログを容易に発行するためのツールを供えている. そのうちのひとつとして「プロフィール」を公開するためのツールがある. このようなツールを使って公開されたブログは, 特定のフォーマットを持っているためブログ中から自動的にブロガーのプロフィールを取り出すことが可能である場合が多い. 特に, goo ブログ (http://blog.goo.ne.jp/) におけるプロフィール公開用ツールにおいては, 居住地はユーザの自由記述ではなく, ツールから提供される地域名のメニューからの選択となるため, 地域の粒度に関する曖昧性がない. このような理由からコーパス収集の対象としてこのサイトを選択した. このサイトでは「新着記事一覧」として RSS フィードを配信

地名	都道府県
...	...
渋谷	東京
お台場	東京
みなとみらい	神奈川
関内	神奈川
北新地	大阪
...	...

表 1: 地域名辞書エントリの例

しており, この RSS フィードから辿ることができた 74,155 ブログの 2005 年 1 月から 2005 年 9 月までの 5,278,107 エントリを収集した. このうち, 40,354 ブログ (54.4%) で居住している都道府県がが公開されていた.

イメージファイルや HTML タグ, および日付や見出しといった情報を取り除いた本文を取り出したところ, 約 3GB のテキストを得ることができた. 得られたテキストは Chasen[6] を用いて形態素解析を行った. 地名辞書の見出し語を取り扱いやすくするために, 地名辞書の見出し語を標準の辞書に加えたものを形態素解析の辞書とした.

4 地名辞書の作成

今回収集したコーパスにおける居住域の粒度は都道府県単位であったため, 居住域の推定に用いる地域名辞書の地域の単位を都道府県単位と定めた. したがって地域名辞書は,「見出し地名- 都道府県」という組からなる辞書である. 地域名辞書の例を表 1 に示す. なお, ここで地名とは, 見出し語は場所を示すための狭い意味ではなく, 施設や建物の名前も含めた広い意味で取り扱っている.

提案手法では, 網羅的な地域名辞書を作成する必要はないが, 辞書の大きさによるの違いによる影響を見るために, 見出し語数に応じて 3 種類の辞書を作成した.

4.1 小規模辞書

小規模辞書は, 良く知られた地名によって構成した. このような良く知られた地名を得るために, goo 地域情報サイト (http://machi.goo.ne.jp/) の各都道府県の主要エリア名 147 を小規模辞書の見出し語とした. ひとつの都道府県あたりの見出し語は 1 から 21 の間である.

4.2 中規模辞書

中規模辞書の作成には, 国内観光情報サイト http://www.gojapan.jp/ の分類に基づく地名を用いた. このサイトでは目的別に 33 のカテゴリに分けて提供してある. このうち, ランドマーク, 施設名, お祭りなどの地域の特別なイベントなどを含むような 15 のカテゴリを選択し, そのカテゴリ中の地名を地名辞書のエントリとした. このようにして得られた地名のうちの一部は複数の地域に対して用いられるたり, 地名としても人名としても使われるような語であったため, そのような語は取り除いた. これらの語と先ほどの小規模辞書を組合せた結果, 7,531 の見出し語を含む地名辞書となった. 各都道府県あたりの見出し語は 59 から 665 である.

辞書	地名を含む文の数	ブロガー数
小規模	172,647	19,663
中規模	200,460	21,752
大規模	214,170	23,645

表 2: 各辞書によって抽出された文の数とブロガー数

4.3 大規模辞書

大規模辞書の作成には郵便番号辞書を用いた. 郵便番号辞書は都道府県名, 市町村名, 町域名の 3 階層で構成されており, 計 121,161 の町域を含んでいる. 市町村名や町域名を特定するためには, その上位の階層をすべて並べれば必ず特定できるが, 単一で用いたり, 上位の階層の一部だけを用いた場合でも特定できる場合もある. たとえば, ほとんどすべての市名はそれ自体で一意である. そこで, 3 階層をすべて並べた形式での地名に加えて, このような一意に特定できるような組合せも大規模辞書に加えた. 人名としても用いられる語を取り除き, 中規模と合わせた結果, 見出し語の数は 244,897 となった.

5 実験

収集したコーパスを用いて実験を行った. まず辞書の大きさが, 候補地名の出現に程度影響を与えるかどうかを見た. 表 2 はそれおぞれの辞書を用いた場合の, 抽出した地名を含む文の数およびブロガーの数である. 我々の予想に反して, 小規模辞書と大規模辞書では 1,600 倍の大きさの違いがあるにもかかわらず, 抽出された文の数はさほど変わらないことが分かった. このため, 以下の実験では中規模辞書のみを用いて, 少なくとも地名を一度は含んでいるような 21,752 ブロガーを対象に行った.

2 つの素朴な手法と比較を行った. 1 つ目の素朴な手法は, 都道府県毎の分布が偏っていることを用いた手法である. 都道府県毎のブロガーの分布は非常に偏っている. このため, 常に「東京」と推定することによってある程度の精度を得ることができる. 2 つ目の素朴な手法は, 辞書の見出し語の出現に基づく投票である. 提案法と同じ辞書 (中規模辞書) の見出し語が出現毎に, その地名に対応する都道府県の投票を加算し, もっとも多い投票を得た都道府県を出力する.

既に述べた通り, 提案手法は, 地名辞書に依存しお

手法	精度
常に「東京」	26.8%
辞書に基づく投票	48.2%
提案手法 投票手法 1	50.7%
提案手法 投票手法 2	50.1%
提案手法 投票手法 1 (正しい断念)	57.6%
提案手法 投票手法 2 (正しい断念)	56.4%

表 3: 居住域の推定精度

	正の上位の特徴	負の上位の特徴
1	支部	北海道
2	自転車	温泉
3	南口	徳島
4	阿波踊り	ホテル
5	負担	戦
6	花火	旅行
7	シネマ	水戸
8	西口	青森
9	バー	群馬
10	F	京都
11	丁目	土産
12	沿線	県

表 4: 重みの大きかった上位の特徴

り, どの地域も正のスコアを得られない状況では回答を出すことはできない. 実際, 約 37%のブログにおいては, 正しく居住都道府県を示しているような地名が出現しなかった. このような状況では回答を断念する方適切であるとも考えられるため, 正しい断念 (どの地名も対象ブロガーの居住都道府県を表していない場合の断念) を正解とした場合の性能の評価も行った.

表 3 に実験結果を示す. 提案方は比較手法に比べて僅かながら良い結果であることが分かる.

6 議論

素朴な手法に比べて, 今回の提案手法における性能の向上は僅かであったが, 訓練によって得られた特徴は興味深い. 表 4 は居住域を示す特徴, 居住域でないことを示す特徴それぞれの上位である. 居住域を示す特徴の中には, 通常歩いていけるような場所が多く見て取れる. なお, 10 番目の 'F' は建物の中での階数を示すために多く用いられたものと推察される. 居住域でないことを示す特徴の中には, 県名が多く現れる. これは, あまり慣れてない地名に言及する場合に, より細かく記述景行があるためだと考えられる. 別の観点では, 居住域を示す特徴は小さな領域を示している場合が多いのに対し, 居住域でないことを示す特徴はより大きな領域を示している場合が多い.

7 まとめ

ブログからブログ作者の居住域を推定する方法を提案した. 居住域の推定は, ブログからの情報検索を行う際の新たな制約として使えるだけでなく, 評判分析の結果との組み合わせや, 曖昧な語の解消といった間接的な用途に役立てることができる. 地名の出現は必ずしもブロガーの居住地を意味しないため, 我々は, 地名の出現が居住地と関連しているかどうかを判定する二値分類器を導入し, 複数の分類器の結果の投票によって居住域の判定を行った.

実験の結果では素朴な手法に比べた性能の差は僅かであったが, 分類器の訓練によって得られた重みつきの特徴は, 地名の周囲に現れる語の興味深い特徴を示している.

提案法では, 地名辞書に含まれているような地名がまるで出現しないブログを取り扱うことができない. 今後の課題として, たとえば, 文書分類に基づく手法など, 辞書に頼らないカバレッジの高い手法と組合せることを検討したい.

参考文献

[1] Herring, S. C., Scheidt, L. A., Bonus, S. and Wright, E.: Bridging the Gap: A Genre Analysis of Weblogs., *HICSS* (2004).

[2] Pang, B. and Lee, L.: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, *Proceedings of the ACL*, pp. 271–278 (2004).

[3] Turney, P. D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, *CoRR*, Vol. cs.LG/0212032 (2002).

[4] McCurley, K. S.: Geospatial mapping and navigation of the web, *World Wide Web*, pp. 221–229 (2001).

[5] Freund, Y. and Schapire, R. E.: Experiments with a New Boosting Algorithm, *International Conference on Machine Learning*, pp. 148–156 (1996).

[6] Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K. and Asahara, M.: Japanese Morphological Analysis System ChaSen version 2.3.3 (2003).