地理的距離と有名度を用いた地名の曖昧性解消*

平野 徹† 松尾 義博† 菊井 玄一郎†
†日本電信電話株式会社 NTT サイバースペース研究所
{hirano.tohru,matsuo.yoshihiro,kikui.genichiro}@lab.ntt.co.jp

1 はじめに

近年，固有表現抽出技術の発展により文書中の地名を高精度で抽出できるようになってきた．一方，文書中の地名は「日本橋」のように都道府県や市区町村名が省略され，その実世界での位置を一意に決定できないことが多い．

文書中の地名の実世界での位置が特定できれば，文書検索結果において「大阪の日本橋」と「東京の日本橋」の文書を分けて提示することが可能になるだけでなく，ユーザの位置情報が利用できるモバイル機器と連携した文書検索サービスも可能になる．

2 従来研究

文書中の地名の実世界での位置を特定する従来手法の多くは次の 3 ステップから成る [1, 2, 3, 4].

1. 入力文書中の各地名 $W_h(1 \leq h \leq m)$ の実世界での位置候補 $C_{h,i}(1 \leq i \leq n)$ を住所 DB から取得する．ここで，$C_{h,i}$ は地名 W_h の i 番目の候補を意味する．
2. 入力文書中で地名 W_h と共起する他の地名 W_j の候補 $C_{j,k}$ との "距離" が最短となる候補 $C_{h,i}$ を地名 W_h の実世界での位置 G_h と特定する．"距離" として，緯度・経度に基づく地理的距離や住所の階層距離などが用いられている．なお，最短距離が一定の距離以上の時，本ステップはスキップされる．

$$G_h = \underset{C_{h,i}}{\operatorname{argmin}}\, Distance(C_{h,i}, C_{j,k}) \quad (j \neq h)$$

3. 実世界での位置 G_h が特定されていない地名 W_h の候補の中で，"有名度" （各候補がどの程度多くの人に想起されうるかを示したもの）が最大となる候補 $C_{h,i}$ を地名 W_h の実世界での位置 G_h と特定する．"有名度" として，住所階層や人口数などの情報から算出されたスコアが用いられている．

$$G_h = \underset{C_{h,i}}{\operatorname{argmax}}\, Popularity(C_{h,i})$$

*Location Disambiguation using Geographic Distance and Popularity
Toru Hirano†, Yoshihiro Matsuo†, Genichiro Kikui†
†NTT Cyber Space Laboratories, NTT Corporation

3 提案手法

本稿では，1. 有名度の算出方法と 2. 距離と有名度の組み合わせ方法において，新たな手法を提案する．

3.1 有名度の算出方法

従来手法の住所階層や人口数を用いた有名度は，地名 W_h の全ての候補が同スコアになり，機能しないことがある．例えば，住所階層を用いた有名度は，上位階層の候補の方が高スコアだが，階層ごとにスコアが定まるため，同階層の候補は同スコアになる．また，人口数を用いた有名度は，人口数の多い候補の方が高スコアだが，一般に利用可能な人口数情報は都道府県と市区町村に限られており，人口数情報の不明な候補は同スコアになる．例えば，「日本橋」の候補である「大阪府大阪市中央区日本橋」と「東京都中央区日本橋」は，人口数情報がわからないため，同スコアになる．

本稿では，全ての住所階層に対して有名度が算出可能な手法を提案する．基本アイデアとして「有名な場所=店の多い場所」と考え，店舗 DB を用いて，各候補にある店の件数を有名度のスコアとする．

$$Popularity(C_{h,i}) = count(shop\ \ located\ \ in\ \ C_{h,i})$$

例えば，4 節の評価実験で用いた店舗 DB では，「大阪府大阪市中央区日本橋」には 60 件の店が，「東京都中央区日本橋」には 215 件の店があり，従来手法では同スコアだった候補に対しても有名度に基づいた特定ができるようになる．

3.2 距離と有名度の組み合わせ方法

従来手法では，距離に基づく処理の後に有名度に基づく処理を行なうため，常に距離に基づく処理を優先している．距離に基づく処理を優先することで，有名でない候補 $C_{h,i}$ を地名 W_h の実世界での位置 G_h と特定できるようになるが，多くの人に想起されうる有名な候補が実世界での位置として正しい場合でも，誤って有名でない候補を地名 W_h の実世界での位置 G_h と特定することがある．

本稿では，常に距離に基づく処理を優先するのではなく，有名度が他の候補に比べ突出している場合には，有名度に基づく処理を優先する手法を提案する．提案手法は，従来手法の候補取得処理（2節のステップ1）の後に，以下のステップを加えた5ステップから成る．

1. 地名 W_h の候補の最大の有名度を，全候補の有名度の和で割った値を尤度とし，尤度が閾値（th）より大きいならば，有名度が最大となる候補 $C_{h,i}$ を地名 W_h の実世界での位置 G_h と特定する．

$$G_h = \underset{C_{h,i}}{\operatorname{argmax}}\, Popularity(C_{h,i}) \;\; if \;\; Likelihood > th$$

$$Likelihood = \frac{\max Popularity(C_{h,i})}{\sum_{i=1}^{n} Popularity(C_{h,i})}$$

2. 実世界での位置 G_j が特定されていない地名 W_j と入力文書中で共起する他の地名 W_h の実世界での位置 G_h（上記ステップで特定されたもの）との距離が最短となる候補 $C_{j,k}$ を地名 W_j の実世界での位置 G_j と特定する．なお，最短距離が一定の距離以上の時，本ステップはスキップされる．

$$G_j = \underset{C_{j,k}}{\operatorname{argmin}}\, Distance(C_{j,k}, G_h)$$

4 評価実験

評価実験には，gooブログ[1]でカテゴリに各都道府県名もしくは「食べ歩き」と設定された文書に，人手で地名とその実世界での位置を付与した1,908文書を用いた．この中には3,872個の地名が存在し，その実世界での位置候補は平均17.34個である．なお，候補が1つのみの地名は1,284個あり，最も候補が多かったのは「上野」の355個である．

地名の実世界での位置候補を取得するために，国土交通省の「街区レベル位置参照情報」（13,045,497件）を住所DBとして，「国土数値情報（鉄道データ）」（8,918件）を駅DBとして用いた．さらに，提案した有名度算出手法で用いる店舗DBにはgoo地域[2]のデータ（約250,000件）を利用した．

評価実験では，地名の実世界位置の特定において，提案手法の有効性を示すため，次の3つの手法を比較した．

従来手法 地理的距離と有名度（住所階層）を3ステップで組み合わせた手法

提案手法1 地理的距離と有名度（店舗数）を3ステップで組み合わせた手法

提案手法2 地理的距離と有名度（店舗数）を5ステップで組み合わせた手法

[1] http://blog.goo.ne.jp/

[2] http://local.goo.ne.jp/

表1: 実世界位置の特定精度

	精度
従来手法	67.5% (2615/3872)
提案手法1	82.2% (3184/3872)
提案手法2	92.7% (3589/3872)

なお，全手法で候補取得処理と地理的距離の計算式（次式）は共通である．次式で，lat_x は位置 x の緯度を，lng_x は位置 x の経度を意味している．

$$Distance(x, y) = \sqrt{(lat_x - lat_y)^2 + (lng_x - lng_y)^2}$$

また，各ステップでの閾値として，有名度では0.9，地理的距離（十進経緯度）では0.2を与えた．閾値はデベロップメントセットを用いてチューニングした．

実験結果（表1）から，提案手法は従来手法よりも特定精度が25.2%向上したことが確認でき，提案手法が文書中の地名の実世界位置を特定するのに有効であることがわかる．残りの約7%の誤りは，(a) 文書中に地名が1つしか出現せず，共起する地名を利用できなかったものと，(b) 文書中で共起するが，文脈上関係ない地名を利用して誤ったものに大別できる．(a) には，共起する地名だけでなく，周辺の単語も考慮した手法が，(b) には，文脈における関連度などを考慮した手法が必要であると考えられる．

5 おわりに

本稿では，文書中の地名の実世界での位置を特定する手法として，店舗数を用いた有名度の算出方法と，距離と有名度の組み合わせた方法を提案した．評価実験では，従来手法に比べ，特定精度が25.2%向上したことがわかり，提案手法の有効性が確認できた．今後は，更なる特定精度の向上を目指し，上記の考察で述べた問題に取り組む予定である．

参考文献

[1] Li, H., Srihari, R. K., Niu, C. and Li, W.: InfoXtract location normalization: a hybrid approach to geographic references in information extraction, *In Proceedings of the HLT-NAACL 2003 workshop Analysis of geographic references*, pp. 39–44 (2003).

[2] Li, Y., Moffat, A., Stokes, N. and Cavedon, L.: Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval, *Workshop on Geographic Information Retrieval* (2006).

[3] Rauch, E., Bukatin, M. and Baker, K.: A confidence-based framework for disambiguating geographic terms, *In Proceedings of the HLT-NAACL 2003 workshop Analysis of geographic references*, pp. 50–54 (2003).

[4] Smith, D. A. and Crane, G.: Disambiguating Geographic Names in a Historical Digital Library, *In Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pp. 127–136 (2001).