Webにおける実世界の位置情報類推に関する研究

Identifying real world locations on the Web

長岡 諒[*1] 松本 光弘[*1] 沼尾 正行[*2] 栗原 聡[*2]
Ryo Nagaoka Mitsuhiro Matsumoto Masayuki Numao Satoshi Kurihara

[*1]大阪大学大学院情報科学研究科情報数理学専攻
Department of Information and Physical Sciences, Graduate School of Information Science and Technology, Osaka University

[*2]大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

Recently, there are abundant documents on the Internet, in web diary systems such as weblogs and social networking servicies, which do not have location information specified. In order to utilize location information for mining real world, most of the current solution use gazetteers and location names. However, the terms that are related to a location can afford a clue to specify the location it-self. In this paper, we propose an effective approach to utilize related terms, which automatically derives the location that are mentioned in the documents.

1. はじめに

インターネットの発展とともに，既存のメディアに頼らず，個人が情報を発信できる環境が整ってきている．中でもブログや SNS(Social Networking Service) は，手軽な情報発信の手段として近年とくに注目を集めており，世界中の人が日々の実体験の情報を言葉にして Web 上に公開している．インターネットを利用できる環境さえあれば，その瞬間に見たことや感じたことを気軽に Web 上に公開して友達と共有することができることから，世の中の関心をリアルタイムに反映する新たなメディアとしての注目を集めており，ブログから有用な情報を抽出したいという需要から，現在様々なアプローチから研究が進められている [8, 9]

しかし，実世界の情報を知るための有用な情報が多いにも関わらず，個人が書く文章は言及している地域が不明確なことが多く，現在の検索システムでは地理情報が効果的に活用されていない．現在の検索システムは，人手による登録が必要であったり，クエリーとしての地名が文中に含まれているか，詳細な住所が記載されているもののみを抽出しているのがほとんどである．そこで本研究では，地名とコンテンツを機械的にマッチさせるために，抽象的な記述の多い個人によって記述されているブログ記事において最も言及している地名を類推することを目的とする．

2. 関連研究

従来の言及地名類推に関する研究は，Amity ら [1] のように地名のみを用いて類推しているものや，Ding ら [2] のように文章内のメタ情報を用いているものがほとんどである．日本語文章に対して利用できるサービスとして，LocoSticker[12] もその一つで，位置情報のみを用いて言及地名を類推している．メタ情報を用いることで正確な地名と文章のマッチングが可能となるが，人手によりメタ情報を付与することはコストが高く，適切なメタ情報が付与されたブログ記事は少ないのが現状である．本研究では地名のみに頼らず，その地名を特定づける関連語を用いることで精度を向上させている点において異なるといえる．

Li ら [4] は地名辞書を用いて Named Entity Recognition(NER) の分野で用いられいる地名の特定を行っており，地名の曖昧性の除去のために，地名階層の上下関係を用いている．また，Chuang Wang ら [5] は Provider location（文章内の地名）と Content location（地名間の関連性）と Serving location（リンク関係やユーザログ）に分けて場所の分類を行っている．しかし，Chong Wang ら [6] によれば，Li らと Chuang Wang らの手法では詳細な地名が記述されていない短い文章では利用できないとして，全ての文章は多かれ少なかれ地名と関係があるという仮定に基づき，Latent Dirichlet Allocation(LDA,[7]) を拡張した Location Aware Topic Model（LATM）モデルを提案している．Chuan Wang らの LATM モデルによって語と関連度の高い地名が抽出されているが，Bush と US，Jintao Hu と China であるなど明白な語しか利用できていない．そこで本研究では，学習に基づいた地名抽出モデルを用意し，従来研究を踏まえつつ，関連語情報を用いた言及地名の類推を試みている．

3. 提案手法

本研究は，特定の地名に関する文章を作成する際に，主題と考える地名に関しては，文章内の他の地名よりも文章内で説明する割合が高いというヒューリスティックに基づいている．例えば，京都での出来事に関して記述する際，南禅寺に行った場合は湯豆腐を食べた事や琵琶湖疎水を見たこと，嵐山ならトロッコ電車を使ったことや渡月橋を渡った事が記述されている可能性が高いといったことが挙げられる．そのため，従来手法で用いられている位置情報のみによる言及地名の類推を行うよりも，それら手法と併せて特定の地名の関連語を用いて，関連語が多く含まれている地名を言及地名とするアプローチは効果的であると考えた．そこで，本章ではその有効性を確認するために，関連語を抽出して，関連語と地名との関連度に応じて言及地域を類推する図１に示すようなシステムの提案を行う．

連絡先: 長岡諒，大阪大学産業科学研究所沼尾研究室
567-0047 大阪府茨木市美穂ヶ丘 8-1
Tel: 06-6879-8426 Fax: 06-6879-8428
E-mail:nagaoka@ai.sanken.osaka-u.ac.jp

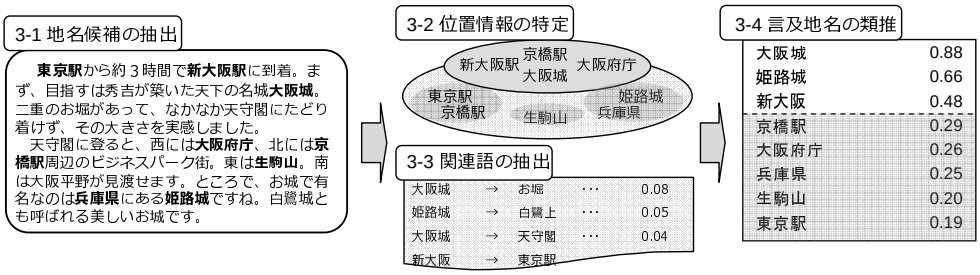


図 1: システム概要

以下 3.1 節で文章からの地名候補の抽出，3.2 節で位置情報の特定，3.3 節で文章中の地名に関する関連語句と関連度の抽出，3.4 節で 3.1 節〜3.3 節の結果を踏まえた言及地名の類推，と 4 つに分けて説明する．

3.1 地名候補の抽出

地名辞書を用いて文章中から地名を抽出するだけでは「山口さん宅を訪問する」といった文章の「山口」が地名なのか人名なのか判断できない．そこで本研究では，自然言語処理の固有表現抽出技術を用いて地名候補を抽出した．まず人手によって地名のラベル付けがされたブログを用いて CRF(Conditional Random Field)[10] による識別モデルを作成し，文脈からの地名の抽出を行った．

地名の定義に関しては基本的に IREX[*1] よる地名の定義に従ったが，本研究では「特定の場所を示すもの」を地名と定義し,「ORGANIZATION」に属する「京大」や「沖電気関西研究所」などの場所を伴う組織,「ARTIFACT」に属する「エンパイアステートビル」や「敦賀原発１号機」などの建築物など，その１単語で場所を識別できるものを全て地名とした．また，IREX の定義によれば「大阪弁」や「日本人」には LOCATION タグが付与されるが，それらは場所そのものを意味しているわけではないため，地名からは除外した．

CRF の精度を検証するためにアメブロ[*2]のブログ記事 1,000 件に関して地名のラベル付けを行い，1,000 記事を 22,803 文に展開して 10 分割交差検定を行ったところ，F 値で 82.75%(precision=87.55%,recall=78.46%) の結果となった．斉藤ら [11] の結果では，ブログ記事で作成したモデルを用いてブログ記事の固有表現を識別した結果，F 値で 76%であったため想定内の結果であると考えられる．

3.2 位置情報の特定

日本全国において「日本橋」など同名の地名が多数存在するため，地名だけを用いて位置を特定することは難しい．そこで，経験的に文章内の地名は地域ごとに固まっていることが考えられるため，文章内地名の地理的な近接関係を用いて特定を行った．まず，2.1 節で抽出したそれぞれの地名 ℓ_i に対して緯度・経度情報を返す Yahoo!Local Search[*3]を用いてそれぞれの同名地名の位置情報 p_{ij} を得た．次に，全 p_{ij} に対して 2 点間の直線距離を用いた階層的クラスタリングを行い，10km 以内にクラスタリングされたクラスタ内における位置情報 p_{ij} の地名 ℓ_i 数の対数に比例した得点を与えた $(clust(p_{ij}))$．図 2 では，クラスタ $c1$，$c3$，$c4$ に対して近接位置情報の多いクラスタ $c2$ の得点が高くなっていることを示している．

*1 http://nlp.cs.nyu.edu/irex/NE/df990214.txt
*2 http://ameblo.jp/
*3 http://developer.yahoo.co.jp/webapi/map/

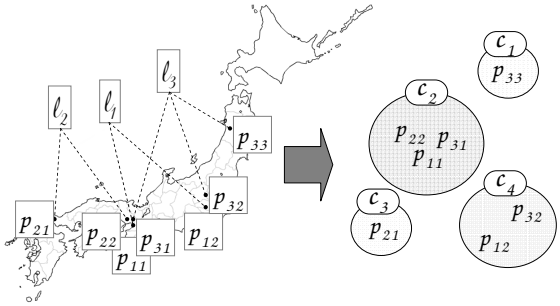


図 2: クラスタリング概念図

また，それぞれの位置情報 p_{ij} に対して各 ℓ_i ごとに一番近い p_{ij} との地理的な距離の逆数に応じた得点を与えた $(near(p_{ij}))$．図 3 では，それぞれの一番近い位置情報（$p11$ は $p22$ と $p31$，$p12$ は $p22$ と $p32$）に関して，距離の近い位置情報が多い $p11$ の得点が高くなっていることを示している．

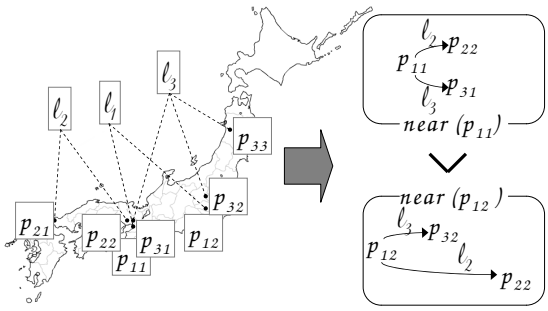


図 3: 近接度スコアリング概念図

最終的に各位置情報のスコアを 0〜1 の定数 α を用いて $c(p_{ij}) = (1-\alpha)clust(p_{ij}) + \alpha near(p_{ij})$ で更新し，各 ℓ_i の中で一番高いスコアを持つ位置情報 p_{ij} を文章中での ℓ_i の位置と推定した．

3.3 関連語の抽出

経験的に，地名と関連の高い語は多くの文章において，その地名が出現した場合に，他の語と比べて近い位置に出現する傾向があると考えられる．そのため，地名 ℓ_i と単語 w との関連度は，地名と，地名と最も近い単語のペアの距離の合計値から算出される．

そこでまず，候補地名 ℓ_i，単語 w の関連度を，最も近い単

語との距離の逆数の合計を抽出対象の全ブログ記事に対する ℓ_i の出現数 $count(\ell_i)$ で割ったもの

$$rel_{avg}(\ell_i, w) = \sum_{\ell_i, w \text{ を含む文章}} \frac{1}{\min_w dist(\ell_i, w)} / count(\ell_i) \quad (1)$$

で関連度を計算してみた．ここで $dist(w_1, w_2)$ は w_1,w_2 の文単位での距離であり，同一文に出現すれば 1，前後の文なら 2,3... となる．また，関連語対象とする語 w は，その出現数が全対象記事中に 10 回以上出現していて，$dist(\ell_i, w) \leq 3$ のもののみを対象とした．これは，語の出現数が極端に少ない場合，地名 ℓ_i との同時発生数も少なくなってしまい，関連度に偏りが生じることが考えられるためである．

対象の地名と全ての語との距離の逆数の合計をしただけでは表 1(嵐山（count(嵐山)=720）に関する関連度の抽出結果，語の出現数は対象語の全ブログ記事に対する出現数で，同時出現数は地名と対象単語が同時に記述された回数，距離の合計は $\sum_{\ell_i, w \text{ を含む文章}} \frac{1}{\min_w dist(\ell_i, w)}$) のようになってしまい，「こと」「人」「お店」といったその地名を特徴付ける語でない一般的な語が高い関連度で抽出されてしまった．

表 1: 嵐山の関連度（距離の合計/同時出現数）

順位	関連語	語の出現数	同時出現数	距離の合計	関連度
1	京都	20477	394	247.9142	0.344325
2	紅葉	7270	148	91.3322	0.126850
3	こと	200576	134	71.8323	0.099767
4	渡月橋	226	75	45.1663	0.062731
5	人	80165	81	39.4990	0.054860
6	一	41668	57	30.7495	0.042708
7	私	106559	64	30.0827	0.041782
8	方	60597	47	27.9164	0.038773
9	時	50181	53	27.0829	0.037615
10	お店	27674	41	24.8331	0.034490
11	竹林	236	43	24.3331	0.033796
12	清水寺	1287	32	24.2498	0.033680
13	ここ	39858	56	24.2495	0.033680
14	ところ	35278	47	23.8328	0.033101
15	これ	63852	53	23.0829	0.032060

そこで，地名に対する対象単語の条件付き確率を導入することで，多くの地名と共起している一般的なキーワードの重みを下げ，相対的にその地名を特徴付ける語の関連度が向上するように修正し，

$$rel(\ell_i, w) = \frac{P(\ell_i|w)}{count(\ell_i)} \times \sum_{\ell_i, w \text{ を含む文章}} \frac{1}{\min_w dist(\ell_i, w)} \quad (2)$$

と定義した．また，ここでも対象語は出現数が 10 以上のものに限定した．嵐山の結果に条件付き確率を導入した結果を表 2 に示す．

この結果，語の距離を考慮すると同時に条件付き確率 $P(\ell_i|w)$ を用いることで，「お店」のように多くの文章に出現する一般的な語の重みを下げ，その地名を表している語の関連度が相対的に高くなっていることが確認できた．

表 2: 嵐山の関連度（条件付き確率導入後）

順位	関連語	語の出現数	同時出現数	距離の合計	関連度
1	渡月橋	226	75	45.1663	0.020818
2	京都	20477	394	247.9142	0.006625
3	嵯峨野	111	25	20.4999	0.006413
4	竹林	236	43	24.3331	0.006158
5	天龍寺	102	27	16.2498	0.005974
6	野宮神社	54	14	7.9999	0.002881
7	嵐山	156	28	11.4165	0.002846
8	天龍寺	17	7	4.6666	0.002669
9	トロッコ列車	81	16	9.6665	0.002652
10	トロッコ電車	35	9	7.3333	0.002619
11	紅葉	7270	148	91.3322	0.002582
12	鈴虫寺	159	21	13.8333	0.002538
13	人力車	200	21	14.1665	0.002066
14	清涼寺	11	6	2.5832	0.001957
15	桂川	25	7	4.5833	0.001782

最終的に本研究では，2008 年 10 月 11 日〜11 月 23 日までのアメブロのブログから同一文を排除した 3,348,113 件を対象に，各地名と語の関連度を計算した．表 3 に「平安神宮」「南禅寺」「嵐山」の関連語の一部を関連度の高い順に示す．

表 3: 関連語抽出例

	平安神宮		南禅寺		嵐山	
順位	関連語	関連度	関連語	関連度	関連語	関連度
1	平安遷都	0.026309	永観堂	0.029412	渡月橋	0.020818
2	例祭	0.015000	水路閣	0.010983	京都	0.006625
3	平安京	0.010000	インクライン	0.005435	嵯峨野	0.006413
4	時代祭	0.009555	湯豆腐	0.005356	竹林	0.006158
5	桓武天皇	0.008470	三門	0.005088	天龍寺	0.005974
6	長岡京	0.008465	順正	0.004280	野宮神社	0.002881
7	創建	0.008102	平安神宮	0.003296	嵐山	0.002846
8	宇太村	0.007168	山門	0.002165	天龍寺	0.002669
9	新京	0.007168	銀閣寺	0.001941	トロッコ列車	0.002652
10	葛城郡	0.006547	哲学	0.001899	トロッコ電車	0.002619

3.4 言及地名の類推

ブログ筆者が文章中で地名 ℓ_i を言及している度合いを示す言及度は，地理的に密集している地名は経験的に言及している可能性が高いことから，3.2 節の位置情報のスコア（$cl(\ell_i)$）と 3.3 節の関連度（$rel_{sum}(\ell_i)$）を用いて以下のように定義した．

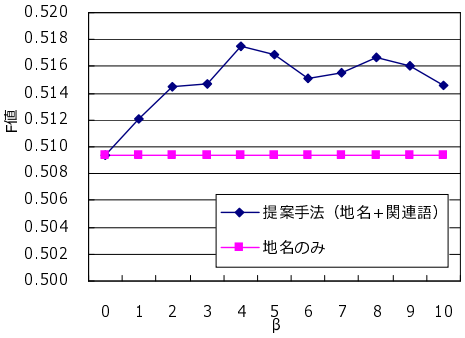
$$score(\ell_i) = \frac{(1+\beta^2) \times cl(\ell_i) \times rel_{sum}(\ell_i)}{\beta^2 \times cl(\ell_i) + rel_{sum}(\ell_i)}$$

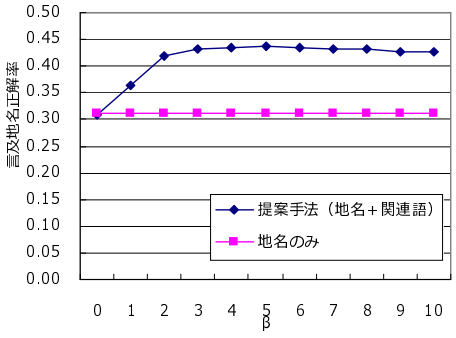
各 ℓ_i に対して $cl(\ell_i) = max(c(p_{ij}))$ であり，文章内の各地名同士の近接度を表す．$rel_{sum}(\ell_i)$ は文章内での各地名 ℓ_i とその関連語との関連度の合計で，$rel_{sum}(\ell_i) \propto \sum_w rel(\ell_i, w) \times \frac{1}{dist(\ell_i, w)}$ となり，$rel_{sum}(\ell_i)$ の標準偏差を用いて 0.5 の周囲に変換した．パラメータ β は 0 以上の正の整数をとるパラメータであり，近接度に対して関連度を何倍重視するかを表している．そして，$score$ が閾値 θ 以上であれば文章中で言及されている地名と判断した．

4. 評価実験

関連語を用いた言及地名類推の有効性について，被験者実験を通して検証した．人手によって地名のラベル付けがされたブログのうち 500 件をランダムに抽出して，1 記事に対して 2 人によって (1) 言及している地名全て (2) あれば最も言及している地名 1 つ，を選んでもらい，両者が共に言及しているとした地名を言及地名とした．関連語を用いない場合（β=0,$score(\ell_i) = cl(\ell_i)$）をベースラインとして，$score(\ell_i) \geq \theta$ となる地名の類推結果の F 値と，一番高い score を持つ ℓ_i が最も言及している地名である場合を正解とした言及地名正解率を表 4 に示す．パラ

表 4: F 値（左）と言及地名正解率（右）





メータには $\theta = 0.3$, $\alpha = 0.5$ を用いた．両者とも関連語を用いることによって精度の向上が確認できた．

5. おわりに

本研究では，従来の地名のみを用いた言及地名類推に加えて，地名とその関連語を用いることによって，ブログ記事において言及している地名類推精度が向上することを確認した．地名を 1 つ以上含むブログ記事から抽出した関連語を用いることによって，抽出精度を示す値である F 値の結果には大きな影響はなかったが，最も言及している地名の類推精度では位置情報のみの場合の 0.3158 ポイントから 0.4375 ポイントと，0.1217 ポイントの精度の向上を確認し，関連語を用いることの有用性を示した．

しかし，本研究を通していくつかの課題点も見つかっている．現段階では，関連語を抽出するにあたり，全学習対象記事中で 10 回以上記述されている語のみを関連語の対象としたが，約 80 万件のブログを用いただけでは 10 回以上の出現を閾値とした場合，ノイズの影響を受けやすいことを確認している．

また，今回は地名と関連語の関連度の抽出においては前後関係なく，単純に文単位の距離を用いたが，一般的に日本語の文章では特定の地名よりも後にその地名と関連の深い語が多く出現する傾向が見られる．そのため，それらヒューリスティックを用いて，地名と関連語との位置の重みに変化を与えることで，より精度の高い類推も行えるのではと考えられる．

参考文献

[1] Eniat Amitay, Nadav Har'El, Ron Sivan, and Aya Soer. Web- a-where: Geotagging web content. In Proceedings of the 27th Special Interest Group on Information Retrieval(SIGIR), pages 273-280, 2004.

[2] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing geographical scopes of web resource. In Proceedings of The International Journal on Very Large Data Bases(VLDB), 2000.

[3] Huifeng Li, Rohini K. Srihari, Cheng Niu, and Wei Li. Location normalization for information extraction. In Proceedings of the 18th International Conference on Computational Linguistics(COLING), 2002.

[4] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and laveling sequence data. In Proceedings of the 18th International Conference on Machine Learning 2001.

[5] Chuang Wang, Xing Xie, Lee Wang, Yansheng Lu, Wei-Ying Ma. Detecting Geographic Locations from web Resources. In Proceedings of the 2005 workshop on Geographic information retrieval, 2005.

[6] Chong Wang, Jinggang Wang, Xing Xie, Wei-Ying Ma. Mining Geographic Knowledge Using Location Aware Topic Model. In Proceedings of the 4th ACM workshop on Geographical Information Retrieval(GIR), 2007.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993-1022, January, 2003.

[8] Kotaro Nakayama, Takahiro Hara and Shirakawa Nishio. Wikipedia Mining for an Association Web Theauras Construction. In Proceedings of Web Information Systems Engineering(WISE) 2007, pp. 322-334, 2007.

[9] 稲川雅之，大島裕明，小山聡，田中克己. Web からの語集合間の特定関係抽出とその可視化. 日本データベース学会論文誌，Vol.7，No.1，pp.175-180，2008. 7.

[10] John. Lafferty, Andrew. McCallum, and Fernando. Pereira. Conditional random fields: Probabilistic models for segmenting and laveling sequence data. In Proceedings of the 18th International Conference on Machine Learning(ICML), 2001.

[11] 齋藤邦子, 鈴木潤, 今村賢治. CRF を用いたブロ グからの固有表現抽出. 言語処理学会第 13 回年次 大会論文集, 2007.

[12] LocoSticker, http://locosticker.jp/.