The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

新聞記事中の地名に対する地理的位置推定における有効な素性の調査

An Investigation of Effective Features for Toponym Resolution of Words in Newspaper Articles

関 龍 *1　乾 孝司 *1
Ryo Seki　Takashi Inui

*1筑波大学大学院システム情報工学研究科
Graduate School of System and Information Engineering,University of Tsukuba

In this paper, we investigate which features are efficient on toponym resolution of words in newspaper articles. Speriosu's TRIPDL algorithm are used in the investigation, and an extension version of TRIPLD is also used to diminish data sparseness problem. We focus on nouns and four types of named entity classes (ORGANIZATION, PERSON, LOCATION, and ARTIFACT), which are used in standard named entity recognition task. Through the experiments, it turns out that using only LOCATION words achieves better performance in terms of both accuracy and computational complexity.

1. はじめに

近年，SNS などの様々なサービスが普及し，人々が気軽に情報を発信できる時代になってきている．これら発信された文書内にある，場所を示す表現と実世界上での地理的位置を結びつけたいというニーズが存在する．SNS の代表とも言える Twitter では投稿時の位置情報をつぶやきに付与することが可能であるが，Middleton ら [1] によればこの機能を利用しているものは全体の 1%にも及ばない．文書内の表現に対して実世界上でのその地理的位置が同定できれば，特定区画に対象を限定したテキストベースの社会分析などがこれまで以上に高精度・高被覆で実現できるようになる．

文書内の場所を示す単語と実世界での地理的位置を結びつけるには様々な問題があるが，その中でも場所を示す単語の曖昧性の問題がある [2]．曖昧性とは例えば,「中央区」という単語があった場合に，日本に数ある中央区のうちのどの位置を実際に参照しているのかわからない状況のことである．実際に日本には数多くの中央区が存在するが，簡単のため「大阪市中央区」と「東京都中央区」に絞った例を図 1 に示す．この地理的位置推定の研究は以前から行われており，どのように場所を示す単語の曖昧性を解消するかの方法には様々な手法が提案されてきた．文書レベルでの地理的位置推定や単語レベルでの地理的位置推定，教師データの有無などによって手法は異なるが，本研究では単語レベルかつ教師データを用いる地理的位置推定について検討する．

一般にこの地理的位置推定のタスクではデータに SNS データが用いられることが多い．Lianhua ら [3] はデータに Twitter を用い，地理的位置推定には地名やハッシュタグ，メンションの情報が有効であることを示した．本研究ではデータに新聞記事を利用するときに，どのような素性，特にどの固有表現が地理的位置推定に有効なのか調査する．

本論文では文書中に表れる実世界上の地理的位置を指す単語を「地名」と呼び，その地名に対応する可能性のある実世界上の住所を「候補地」あるいは単に「候補」と呼ぶ．

2. 関連研究

文書中の地名と実世界上の位置を関連付ける研究は数多く行われてきたが，問題設定の違いにより場合分けを行うことができる．まず，文書中に出現する「地名」つまり単語を位置と結びつけるタスクを toponym resolution と呼ぶ．次に「文書」の記述内容を位置と結びつけるタスクを document geolocation と呼ぶ．また，ある「ユーザ」と関連付けられた文書集合と位置を結びつけるタスクを user geolocation と呼ぶ．この分類に従うと，本研究は toponym resolution に該当する．

図 1: 中央区に対する曖昧性

Speriosu は TRIPDL[4] を提案した．この手法は実世界をセルで区分けし，それらと文書との対応付けを行っている．入力文書の単語確率分布と学習済みのセルの単語確率分布間の KL ダイバージェンスを測り，より近いセルを出力とする方法である．本研究ではベース手法として TRIPDL を採用する．手法の詳細については 5. 章で述べる．

オリジナルの TRIPDL の問題設定は document geolocation であり，実世界上の位置にはセル（地理的位置を格子状に区切った領域）を採用している．セルを採用している理由としては入手が簡単，扱いもしやすいという点が挙げられる．まず，場所に対する経度緯度情報さえあれば場所とそれに対応するセルがすぐに関連付けできるから入手が簡単であるといえる．また，セルとは地図を正規化したものであるとも言え，複雑な地形などもすべて格子状に落とし込むことができるため扱いやすいといえる．しかし，本研究ではセルは使用せず，行政地区を使用する．行政地区を使用するメリットとしては非常に出力が直感的で他の用途に使いやすい点が挙げられる．例えばラベル付けを考えた場合，ある曖昧性のある地名に対して出力がセルだとなんの情報かわかりにくいが，出力が行政地区の場合はラベル付けもしやすく，見る側にとってもわかりやすい．

表 1: 住所 DB の形式

地名	候補数	候補 1	候補 2	…
松屋町	8	岐阜県_岐阜市_松屋町	京都府_京都市_下京区_松屋町	京都府_京都市_上京区_松屋町
		京都府_京都市_中京区_松屋町	京都府_京都市_伏見区_松屋町	大阪府_寝屋川市_松屋町
		大阪府_大阪市_中央区_松屋町	香川県_丸亀市_松屋町	
本田	7	山形県_鶴岡市_本田	埼玉県_深谷市_本田	埼玉県_南埼玉郡_宮代町_本田
		新潟県_新発田市_本田	富山県_射水市_本田	岐阜県_瑞穂市_本田
		大阪府_大阪市_西区_本田		
亀田	3	福島県_郡山市_亀田	千葉県_富津市_亀田	新潟県_新潟市_江南区_亀田
旭区	2	神奈川県_横浜市_旭区	大阪府_大阪市_旭区	
パルプ町	1	北海道_旭川市_パルプ町		
芝浦	1	東京都_港区_芝浦		

3. コーパス

本研究で使用するコーパスは拡張固有表現タグ付きコーパス [5] の毎日新聞ジャンルである．このコーパスには固有表現として様々なタグが付与されているが，本論文では地名を扱うためタグは City タグのみを利用する．このコーパスに存在する事例数，つまり曖昧性を解決すべき地名の数はのべ 1723, 異なりで 373 である．また異なりで曖昧性を解決すべき地名の平均候補数は 10 である．曖昧性の定義は 4. 章で説明する．

4. 住所データベース

本研究では候補の情報を得るために住所データベース（以降，住所 DB）を参照する．住所 DB とは日本中全ての住所を含んだデータベースであり，実世界上の住所と一致していなくてはならない．しかし，住所は市区町村の合併などにより流動性をもったものであるから必ずしも住所 DB と実際の住所が合致しているとは限らない．本論文では住所 DB のおおもとのデータとして国土交通省の発行している「街区レベル位置参照情報」[*1] を利用した．今回使用するこのデータのサンプリングは 2016 年である．

本論文では一つの地名 t に対してどのような候補が存在しているのかを知りたいため，上記データに前処理を施し，表 1 の形式のデータを得た．例えば，地名「パルプ町」の場合，この名前をもつ地理的位置は北海道に 1 地点存在するだけであるので曖昧性がない．一方,「旭区」は 2 地点存在するため曖昧性の解消が必要である．なお，候補は都道府県名からの住所を含んでいるため候補間に曖昧性はない．住所 DB のエントリ数は 152,937 である．この内，候補数が 2 以上になる（曖昧性がある）エントリの割合は約 5%であり，これらエントリにおける平均候補数は 4.49 である．

5. TRIPDL

本研究のベース手法である TRIPDL について述べる．まず文書 d_k 中にあらわれる単語 w_j の生起確率 $\tilde{\theta}_{d_{kj}}$ を計算する．式 (1) が求める式である．

$$\tilde{\theta}_{d_{kj}} = \frac{\#(w_j, d_k)}{\sum_{w_l \in V} \#(w_l, d_k)} \tag{1}$$

$\tilde{\theta}_{d_{kj}}$ は実際に訓練データから計算する単語の生起確率だが訓練データが少ない場合に，この値だけでは多くの単語の生起確率が 0 となってしまうためスムージングを施した $\theta_{d_{kj}}$ を考える．このスムージングは good-turing 推定の考え方に基づく．まず，係数となる α_{d_k} を求めるために式 (2) を計算する．これは文書 d_k 中に一度だけ出現した単語 w_j の生起確率である．

$$\alpha_{d_k} = \frac{|w_j \in V s.t. \#(w_j, d_k) = 1|}{\sum_{w_j \in V} \#(w_j, d_k)} \tag{2}$$

次に単語の生起確率 $\tilde{\theta}_{d_{kj}}$ が 0 のときに使用する値 $\theta_{D_j}^{(-d_k)}$ である．その式が式 (3) である．なお，式 (3) で使用している θ_{D_j} は使用する全ての文書集合 D における単語 w_j の生起確率である．

$$\theta_{D_j}^{(-d_k)} = \frac{\theta_{D_j}}{1 - \sum_{w_l \in d_k} \theta_{D_l}} \tag{3}$$

最後に式 (2) と式 (3) を用いて求めた $\theta_{d_{kj}}$ である．

$$\theta_{d_{kj}} = \begin{cases} \alpha_{d_k} \theta_{D_j}^{(-d_k)}, & \text{if } \tilde{\theta}_{d_{kj}} = 0 \\ (1 - \alpha_{d_k}) \tilde{\theta}_{d_{kj}}, & \text{o.w.} \end{cases} \tag{4}$$

セル c_i 中にあらわれる単語 w_j の生起確率 $\theta_{c_{ij}}$ について求める．まず以下の式 (5) を計算し，訓練データにおけるセル中にあらわれる単語の生起確率 $\tilde{\theta}_{c_{ij}}$ を求める．

$$\tilde{\theta}_{c_{ij}} = \frac{\sum_{d_k \in c_i} \#(w_j, d_k)}{\sum_{d_k \in c_i} \sum_{w_l \in V} \#(w_l, d_k)} \tag{5}$$

この後の処理は式 (2) から (4) と同様に行う．それが式 (6) から式 (8) である．

$$\alpha_{c_i} = \frac{|w_j \in V s.t. \#(w_j, c_i) = 1|}{\sum_{w_j \in V} \#(w_j, c_i)} \tag{6}$$

$$\theta_{C_j}^{(-c_i)} = \frac{\theta_{C_j}}{1 - \sum_{w_l \in c_i} \theta_{C_l}} \tag{7}$$

$$\theta_{c_{ij}} = \begin{cases} \alpha_{c_i} \theta_{C_j}^{(-c_i)}, & \text{if } \tilde{\theta}_{c_{ij}} = 0 \\ (1 - \alpha_{c_i}) \tilde{\theta}_{c_{ij}}, & \text{o.w.} \end{cases} \tag{8}$$

式 (4) と式 (8) から文書中とセル中における単語の生起確率が求められたので，最後に入力文書 d_k がどの候補セル c_i と最も KL 距離が近いかを求めそのセルが出力 $\hat{c}$ となる．つまり式 (9) となる．

*1 http://nlftp.mlit.go.jp/isj/index.html

$$\hat{c} = \arg\min_{c_i \in C_{d_k}} KL(\theta_{dk} || \theta_{ci}) \quad (9)$$

$$= \arg\min_{c_i \in C_{d_k}} \sum_{w_j \in V_{dk}} \theta_{dkj} log \frac{\theta_{dkj}}{\theta_{cij}} \quad (10)$$

候補 c_i は住所 DB のエントリ数だけあり，その数は非常に多い．そのため，疎データの状態に陥りやすい．

6. 拡張手法

5. 章に示した TRIPDL を 2. 章で述べたようにセルでなく行政地区を用いて実装する場合，ある 1 文書に対応する可能性のある候補は 4. 章より，いくつかの候補住所に一意に決まる．TRIPDL ではこれらの候補住所（セル）の単語確率分布は教師がある限り必ず 1 つはあるが，今回の設定では「いくつかの」候補住所に絞られているので，どの候補住所も単語確率分布を持たない可能性がある．その場合は候補間の確率分布がない，つまり等距離の状態となり出力なしとなってしまう．これを解決するために，1 文書に対して複数の候補を学習させる方法を考える．

TRIPDL では式 (5) の学習の際，1 文書 d_k に対して 1 セル c_i を学習させる．その部分を 1 セルではなく複数のセルに拡げたのが拡張手法である．つまり式 (5) を書き換えると式 (11) になる．ここで，Widen_city() とは 1 セルを与えたときにそのセルを拡げた，複数のセルを返す関数である．

先に述べたように今回の設定ではセルではなく行政地区なので Widen_city() には大字レベルまたは市区町村レベルの住所が入力される．その場合はその住所の属する市区町村に属する全ての住所候補が返される．例えば「茨城県つくば市天久保」という大字レベルの住所が入力された場合,「茨城県つくば市」に属する住所候補（「茨城県つくば市」,「茨城県つくば市天久保」,「茨城県つくば市桜」など）が全て返される．

また，式 (12) については関数部分のみが違うだけで Widen_city() が Widen_pref() に替わっている．Widen_pref() とは市区町村でなく入力候補の都道府県に属する住所候補全てを返す関数である．拡張手法では式 (5) の替わりに式 (11) あるいは式 (12) を用いる．つまり，市区町村レベルや都道府県レベルまで正解の解釈を拡げて学習するのが拡張手法である．

$$\tilde{\theta}_{c_{ij}} = \frac{\sum_{c_m \in Widen\_city(c_i)} \sum_{d_k \in c_m} \#(w_j, d_k)}{\sum_{c_m \in Widen\_city(c_i)} \sum_{d_k \in c_m} \sum_{w_l \in V} \#(w_l, d_k)} \quad (11)$$

$$\tilde{\theta}_{c_{ij}} = \frac{\sum_{c_m \in Widen\_pref(c_i)} \sum_{d_k \in c_m} \#(w_j, d_k)}{\sum_{c_m \in Widen\_pref(c_i)} \sum_{d_k \in c_m} \sum_{w_l \in V} \#(w_l, d_k)} \quad (12)$$

7. 評価実験

7.1 実験の設定

7.1.1 実験条件

教師なし学習の先行手法である POPULATION[4] と文脈参照法 [6] と MENTION_COUNT[6]，教師あり学習の先行手法である TRIPDL，拡張手法の市区町村レベル，拡張手法の都道府県レベルについて正解率・拡大正解率を測定する実験を行った．また TRIPDL と拡張手法では，以下のように素性に使う単語を変えてもそれぞれ実験を行った．数詞を除く名詞全てを使った場合，固有表現 ORGANIZATION（O）のみを使った場合，固有表現 PERSON（P）のみを使った場合，固有表現 LOCATION（L）のみを使った場合，固有表現 ARTIFACT（A）のみを使った場合，固有表現 O,P,L,A(OPLA) のみを使った場合について実験を行った．文脈参照法において参照する文字数 k を指定する必要がある．今回はコーパスの 1 記事あたりの平均文書長 $L = 1,657$ のとき，前方文脈を k_1，後方文脈を k_2 とすると $k_1 = k_2 = \frac{L}{2} = 829$ とした．今回の設定ではコーパス内にあるターゲットは入力として与えられる．また TRIPDL は本来 document geolocation の手法であるが，本研究では toponym resolution を扱うため，その間の違いを吸収しなくてはならない．そのため，入力の toponym に対してその前後 829 文字（文脈参照法の k と同様）を結合した文字列を擬似 document とし，TRIPDL で解いた．

7.1.2 データセット

評価を行なうコーパスは 3. 章に示したものを使用する．候補データ生成に用いる住所データベースは 4. 章に示したものを利用する．POPULATION に用いる人口 DB は政府が発行している人口統計データ「e-stat 統計で見る日本」*2 の調査年 2015 年のデータを用いる．MENTION_COUNT で用いる言及回数 DB を作成する際に使用する生コーパスには毎日新聞社のコーパス [7]1 年分を用いる．

正解データは人手で作成した．拡張固有表現タグ付きコーパスに付与された City タグが付く地名のうち，住所 DB を参照して複数の候補が得られるものについて，その近隣文脈を読むことで正解候補を選択した．今回は毎日新聞ジャンルのみの実験であるため，毎日新聞ジャンル中の候補が複数ある City タグ全 1732 個について正解を付与した．また，作業は基本的に一人で行ったが，アグリーメントのため全 City タグのうちランダムに抽出した 200 個をもう一人も作業を行い，どの程度二人のアノテーション結果が一致するかも検証した．その結果一致しなかったタグは全 200 個中 3 個であり，このアノテーションは一般性を確保できていると考えられる．

7.1.3 評価尺度

評価尺度には正解率と拡大正解率を用いる．正解率とは手法により選択した候補と正解ファイル中の候補とを比較し，一致していた場合に正解，そうでない場合に不正解としたときの候補が複数ある全 City のうちの正解の割合である．拡大正解率とは手法により選択した候補が属する都道府県と正解ファイル中の候補が属する都道府県とを比較し，一致していた場合に正解，そうでない場合に不正解としたときの候補が複数ある全 City のうちの正解の割合である．

7.2 実験結果

実験の結果を表 2 に示す．ここで，不正解を誤選択，正解なし，出力なしの 3 つに分けた．誤選択とは正解と出力がどちらも存在するが一致しない場合であり，使用した手法によって正しい候補を選ぶことができなかったことをあらわす．正解なしとは候補中に正解が存在しない場合である．これは合併などによる原因が考えられる．出力なしとは候補があるにも関わらず，使用した手法で出力を出せなかった場合である．また正解なしの場合は出力の有無はドント・ケアであり，出力がなくても出力なしにはカウントされない．

表 2 より，まず POPULATION と文脈参照法と TRIPDL(名詞) を比較すると，教師なし学習である

*2 https://www.e-stat.go.jp/regional-statistics/ssdsview

表 2: 実験結果

手法	拡大正解率	正解率	正解	誤選択	出力なし	正解なし
POPULATION	57.0	56.9	985	207	388	152
文脈参照法	63.5	56.2	974	279	327	152
TRIPDL(名詞)	67.4	67.0	1,161	124	295	152
TRIPDL(O)	62.1	61.8	1,071	190	319	152
TRIPDL(P)	60.1	59.8	1,035	233	312	152
TRIPDL(L)	68.9	68.6	1,188	97	295	152
TRIPDL(A)	55.4	55.2	956	205	419	152
TRIPDL(OPLA)	68.0	67.6	1,171	114	295	152
拡張手法 市区町村 (名詞)	59.4	57.2	991	505	84	152
拡張手法 市区町村 (O)	44.8	43.0	745	735	100	152
拡張手法 市区町村 (P)	32.5	30.5	528	956	96	152
拡張手法 市区町村 (L)	59.9	58.0	1,004	492	84	152
拡張手法 市区町村 (A)	32.6	31.3	542	847	191	152
拡張手法 市区町村 (OPLA)	59.8	57.5	996	500	84	152
拡張手法 都道府県 (名詞)	61.8	53.7	930	650	0	152
拡張手法 都道府県 (O)	43.7	38.3	664	916	0	152
拡張手法 都道府県 (P)	30.8	25.6	443	1,137	0	152
拡張手法 都道府県 (L)	64.0	56.3	975	605	0	152
拡張手法 都道府県 (A)	22.2	17.8	309	1,261	10	152
拡張手法 都道府県 (OPLA)	64.2	56.8	983	597	0	152
MENTION_COUNT	72.1	66.7	1,156	364	60	152

POPULATION，文脈参照法に比べ教師あり学習である TRIPDL（名詞）は正解率が高いことがわかる．

次に，同じ TRIPDL でも使用する素性が違う場合をそれぞれ比べる．L が一番正解率が高いことがわかる．これは固有表現の中で一番 LOCATION が地理的位置推定に有効だということを示しており，直感通りの結果となった．ここで，それぞれの素性の数は，名詞:271011，O:5736，P:7933，L:16302，A:1593 である．名詞の数は他の素性に比べてかなり多くあるにもかかわらず L に正解率で及んでいないのは全ての名詞を利用するよりも固有表現 LOCATION のみを使ったほうがより小さい計算量で正解率が向上することを示していると考えられる．

次に TRIPDL と拡張手法を比較すると，TRIPDL が一番正解率がよく，市区町村，都道府県と範囲を拡げる毎に正解率が悪くなっていることがわかる．不正解の内訳から，出力なしの数は範囲を拡げる毎に確実に減っており，候補間が等距離になる問題は解決されていることがわかる．しかしそれと同時に正解は減り誤選択は増えている．

最後に一番正解率が高かった TRIPDL(L) と MENTION_COUNT を比較する．MENTION_COUNT は教師なし学習にもかかわらず，教師あり学習の TRIPDL と同じ程度の正解率をだしており，非常に興味深い結果となった．これは MENTION_COUNT では大量の教師なしデータを考慮できる点が性能へ反映されたからと考えられる．

8. おわりに

本論文では新聞記事の地理的位置推定のタスクにおける地名の曖昧性問題について，曖昧性解消に有効な素性の調査を行った．実験より，固有表現の中では LOCATION が一番有効な素性であり，全名詞も同じくらい有効ではあるが，計算量を考えると全名詞よりも LOCATION のほうが良いという結果が得られた．今後の課題と検討として，次のものが挙げられる．

- 拡張手法によって正解率が下がってしまったがその原因の追求
- MENTION_COUNT と TRIPDL(L) とのアルゴリズムの突き合わせを行い，結合方法を検討する

謝辞

本研究の一部は科研費（18K11982）の助成を受けて実施されました．

参考文献

[1] Stuart Middleton,Lee Middleton,and Stefano Modafferi.Real-time crisis mapping of natural disasters using social media.2014.

[2] 北本 朝展.G 空間 オープンな地名情報システム GeoNLP.THE JOURNAL OF SURVEY 測量,pp6-10,2014.

[3] Lianhua Chi,Kwan Hui Lim,Nebula Alam and Christopher J. Butler.Geolocation Prediction in Twitter Using Location Indicative Words and Textual Features.The 2nd Workshop on Noisy User-generated Text.2016.

[4] Michael Adrian Speriosu.Methods and Applications of Text-Driven Toponym Resolution with Indirect Supervision.THE UNIVERSITY OF TEXAS AT AUSTIN.2013.

[5] 橋本 泰一, 乾 孝司, 村上 浩司. 拡張固有表現タグ付きコーパスの構築. 情報処理学会研究報告自然言語処理,pp113-120,2008.

[6] 関 龍, 乾 孝司. 局所文脈と関連文書を用いた地名に対する地理的位置の同定. 人工知能学会,2018.

[7] 毎日新聞社.CD-毎日新聞 2004 データ集.