The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

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

Location identification using local context and relevant documents

関 龍[*1] 乾 孝司[*2]
Ryo Seki Takashi Inui

[*1]筑波大学情報学群
School of Informatics,University of Tsukuba

[*2]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 discuss the task of matching the location name appearing in the document to the geographical location on the real world. In this task, it often happens that one location name points to more than one geographical location. In order to resolve this situation, many previous methods have been proposed for selecting the best candidate from the candidates. For example, a method of outputting a candidate having the largest population was proposed. In this research, we propose a method to select the best candidate by using local context and relevant documents which are easily obtained and easily preprocessed than the external data necessary for the previous research. Through experiments, the proposed method achieves about 70% accuracy, and further improvement of accuracy is expected by handling more relevant documents.

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 10 年ほど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ない地理的位置推定における手法を扱う．その代表的な例が Speriosu[3] の人口データを利用する「POPULATION」である．この手法では人口データが必要となる．また一般に，人口が集計・管理される粒度は市区町村のレベルに留まっており，それより狭い区画を指す大字（おおあざ）レベルの人口データを全国規模で入手することは困難である．

本研究では先行研究で必要な外部データを使わず，それに準じかつ入手や扱いが容易な局所文脈や関連文書を用いた手法を提案する．文書中の局所文脈の情報を利用して曖昧性を解消する文脈参照法と関連文書を用いて曖昧性を解消する MENTION_COUNT の二つの手法である．

本論文では文書中に表れる実世界上のどこかを指す単語を「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．

図 1: 中央区に対する曖昧性

2. 関連研究

POPULATION[3] とは人口データ（以降，人口 DB）を使用する方法であり，候補のうち一番人口が多い候補を出力とする．主に文書で言及される都市は有名，つまり人口が多い都市であることが多いという前提のもとに考えられた方法である．例えば「ロンドン」と文書中にあった場合は主にイギリスのロンドンを指すことが多いがカナダのオンタリオ州にもロンドンが存在する．しかし，人口はイギリスのロンドンの方が多いためイギリスのロンドンを出力とし，カナダのロンドンを避けることができる．

文書中の地名を t，t が指す候補地（地理的位置をあらわす住所）の集合を $C^t = \{c_i^t\}$ とすると，地名 t に対する地理的位置同定の問題は次式で書くことができる．

$$\hat{c^t} = \arg\max_{c_i^t \in C^t} score(c_i^t) \quad (1)$$

POPULATION は上式のスコア関数 $score(c_i^t)$ として候補地 c_i^t の人口を返す関数を想定した手法である．

しかしこの方法はターゲットに対応した人口 DB が必要であり，対象とする地名のレベルを市区町村でなく大字レベルにまで細分化させた場合，人口 DB の量も非常に大きくなる．また，大字レベルの人口 DB を日本全国の範囲で入手しようとすると困難である．

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. 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る．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．

4. 提案手法

4.1 文脈参照法

文脈参照法は非常に直感的な方法であり，ターゲットとなる地名の前後 k 文字を参照し，そこにターゲット以外の地名があった場合にその情報を用いて地理的位置を決定する方法である．例えばターゲットとして「中央区」があり，候補として「東京都中央区」と「大阪府大阪市中央区」があるとする．このときに「中央区」の前後 k 文字中に「東京都」があった場合出力として「東京都中央区」を選択する．前後 k 文字中に複数の地名が出現することもあるが，この場合は次のような規則に従う．

- （優先度 1）後方文脈よりも前方文脈を優先させる
- （優先度 2）ターゲットに近いものを優先させる

これは文書において地名は曖昧性が無いように書かれることが多く，例えば「東京都の中央区」というように地名の前にその地名の属する地名を書くことが多いため前方文脈を優先している．また，ターゲットに近いものを優先するのは単純にその方が遠いものよりもターゲットの説明をしている場合が多いと考えたからである．この方法では前後 k 文字内にターゲット以外の地名がない場合は曖昧性は解消されない．

文脈参照法の具体的な手続きを以下に示す．まず，準備として，以下の処理をおこなう．

- 文書中の地名 t に対して，その前後文脈の文字列を取得する．
- 前節でみたように，各候補は都道府県から始まる住所である．そこで，候補を都道府県，市区町村，大字の 3 つの要素に分解する．

以上の準備のあと，各候補の要素（都道府県／市区町村／大字）が t の前後文脈に出現しているか否かを文字列マッチングによって確認し，出現していた候補を出力する．複数の候補がマッチした場合は上述の優先順位に従って出力となる候補を決定する．また，どの候補ともマッチングが取れない場合は出力なしとなる．

4.2 MENTION_COUNT

MENTION_COUNT は式 (1) のスコア関数 $score(c_i^t)$ として，関連文書集合 D 中での候補地 c_i^t の言及回数を返す関数を用いる手法である．事前に各候補地の言及回数を保存した言及回数データベース（言及回数 DB）を作成しておき，地理的位置同定の際はこの言及回数 DB を参照することでスコア値を得る．

4.3 言及回数 DB

言及回数 DB とは候補とその言及回数の対を保存したデータベースである．生コーパスを使用し，文脈参照法を一記事の範囲に対して行うことによって作成した．ここで使用するコーパスは固有表現抽出などの処理が加えられておらず，真にテキストのみのデータであるから非常に入手が簡単である．本論文で用いる外部データはこのコーパスのみであるが，先行研究で用いるような人口 DB や座標データに比べると入手は簡単であり，前処理もシンプルであるため使いやすいと言える．

候補地の言及回数を数えるには，関連文書中において「大阪府大阪市中央区」のように，候補間で曖昧性を持たないように記述されていることが理想であるが，そのような記述は地名の言及の一部にすぎない．そこで，都道府県レベルに注目し，また，文脈を一つの記事全体とした文脈参照法を適用することで近似的に各候補地の言及回数を求める．手順としてはまず，関連記事に出現する都道府県名を記事毎にリスト化しておく．次に住所 DB から地名をひいていき，全ての記事に対して検索をかける．ある記事にその地名があった場合，先の都道府県リストを見に行き，リスト中の都道府県に属するターゲットの

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

候補があったらその候補をインクリメントする．全ての記事についてこれを行ったらひとつの地名に対しての処理は終わりである．これを全ての地名について行えば言及回数 DB が完成する．

5. 評価実験

5.1 実験の設定

5.1.1 実験条件

先行手法である POPULATION，提案手法である文脈参照法と MENTION_COUNT，文脈参照法を使用して出力なしとなった場合に POPULATION を使う手法 (文脈参照法+POPULATION)，文脈参照法を使用して出力なしとなった場合に MENTION_COUNT を使う手法 (文脈参照法+MENTION_COUNT) の 5 つを比較する実験を行なう．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．ま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．

5.1.2 データセット

評価を行なうコーパスは拡張固有表現タグ付きコーパス [4] の新聞ジャンルを用いる．候補データ生成に用いる住所データベースは 3 節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．言及回数 DB を作成する際に使用する生コーパスとしては毎日新聞社のコーパス [5] 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．今回は新聞ジャンルのみの実験であるから，新聞ジャンル中の候補が複数ある City タグ全 234 個について正解を付与した．

5.1.3 評価尺度

評価尺度には正解率を用いる．これ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．

5.2 実験結果

5 つの手法それぞれの正解率を表 2 に示す．この結果に対して，ターゲットとなる地名が大字レベルのものとそれ以外*3に分割して再評価した結果を表 3 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．誤選択とは正解候補と出力候補がどちらも存在するが一致しない場合であり，提案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．正解なしとは正解候補が存在しない場合であり，コーパス中の地名と住所 DB の地名とが一致せずに正解候補を正しく選択できなかったものである．これは合併などによる原因が考えられる．出力なしとは出力候補が存在しない場合であり，提案手法によって候補選択が行えなかった状態である．また正解なしの場合は出力候補の有無はドント・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，先行手法の POPULATION よりも提案手法の文脈参照法や MENTION_COUNT の方が正解率が高いことがわかる．また，文脈参照法+POPULATION や文脈参照法+MENTION_COUNT は POPULATION や MENTION_COUNT 単体よりも大きく正解率が上がっている．つまり，文脈参照法に他の手法を組み合わせることで性能が向上すると言える．文脈参照法+POPULATION と文脈参照法+MENTION_COUNT では正解率がほぼ同じであることもわかる．表 3 より，大字レベルでは POPULATION の正解数 0 に対して文脈参照法と MENTION_COUNT は表 2 とくらべてあまり正解率が変わっていない．これは人口 DB が大字レベルのものはなく，POPULATION では市区町村レベルの曖昧性解消しかできないことを示してる．

以上の結果より，POPULATION，文脈参照法，MENTION_COUNT の 3 つでは文脈参照法が単体では一番性能が良く，文脈参照法では出力がない場合に他の手法を適用する組み合わせ手法を用いればより性能が上がるとわかった．組み合わせる手法として POPULATION と MENTION_COUNT を用いたが，どちらも同等の性能であり，人口 DB を用意できない場合に言及回数 DB を代わりに用いれば同程度の精度が出せるということがわかった．また，曖昧性を解消する地名のレベル（市区町村，大字）によって有効な手法が違うということもわかった．

文脈参照法+MENTION_COUNT の結果の例を次に示していく．コーパスは簡単のためターゲットを含む一文のみを示す．

まず，正解の例を示す．上がコーパス，下が住所 DB のクエリで，太字になっている候補が出力である．この例では文脈参照法により，ターゲットである「内幸町」の直前にある「東京」を用いて「東京都_千代田区_内幸町」を出力することができ，正解となる．

> （古沢　由紀子）池田弘子さん １７日にＮＩＥ講習会 中学校教員を対象にＮＩＥ（教育に新聞を）活動の実践例を報告し、活用に役立てる「第７回ＮＩＥ講習会」が １７日午後２時３０分から、東京・内幸町の日本プレスセンターで開催される。

> 内幸町 **東京都_千代田区_内幸町** 富山県_富山市_内幸町

誤選択の例を次に示す．この例ではターゲットである「宇都宮」の前にある「新潟」を用いて「新潟県_十日町市_宇都宮」を選択してしまい，誤選択となる．なお，$k_1 < 9$ であれば MENTION_COUNT を使用するため正解となる（言及回数 DB では宇都宮に対し栃木県_宇都宮市が一番言及回数が多い）．

> ７月に秋田、新潟、盛岡、９月に宇都宮、東京でも開く。

> 宇都宮 **新潟県_十日町市_宇都宮** 栃木県_宇都宮市

MENTION_COUNT による正解の例を次に示す．この例ではターゲットである「横浜」の前後に曖昧性を解消できる地名が全くないため文脈参照法では解決できずに MENTION_COUNT を使用する．言及回数 DB では「横浜」に対し「神奈川県_横浜市」が一番言及回数が多いためそれが出力となり，正解となる．

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

*3 都道府県レベルでは曖昧性が生じないため，ここに該当する事例はすべて市区町村レベルの地名である

表 2: 5 手法の正解率

	正解	誤選択	正解なし	出力なし	正解率
POPULATION	109	51	32	81	39.9
文脈参照法	178	39	32	24	65.2
MENTION_COUNT	137	104	32	0	50.2
文脈参照法+POPULATION	196	42	32	3	71.8
文脈参照法+MENTION_COUNT	194	47	32	0	71.1

表 3: 5 手法の正解率（大字レベル・市区町村レベル）

	大字レベル					市区町村レベル				
	正解	誤選択	正解なし	出力なし	正解率	正解	誤選択	正解なし	出力なし	正解率
POPULATION	0	5	0	53	0	109	46	32	28	50.7
文脈参照法	42	13	0	3	72.4	136	26	32	21	63.3
MENTION_COUNT	30	28	0	0	51.7	107	76	32	0	49.8
文脈参照法+POPULATION	42	13	0	3	72.4	154	29	32	0	71.6
文脈参照法+MENTION_COUNT	43	15	0	0	74.1	151	32	32	0	70.2

日本では昨年２月に大阪市で初めて誕生、以後神戸、横浜、北九州、金沢、富山県高岡、長野県上田の各市と東京都で設立された。

横浜 高知県_高知市_横浜 福岡県_福岡市_西区_横浜 **神奈川県_横浜市**

以下に，言及回数 DB の横浜に対する候補を記述する．神奈川県_横浜市が言及回数が一番多いため選ばれたとわかる．

神奈川県_横浜市 175
福岡県_福岡市_西区_横浜 21
高知県_高知市_横浜 11

6. おわりに

本論文では地名に対する地理的位置の同定を行う際に発生する地名の曖昧性問題について，文脈参照法と MENTION_COUNT の提案を行った．先行研究の手法では地名に対応する人口 DB が必要であったが，提案手法ではこれらよりも入手が簡単で前処理も簡単な関連文書，具体的には新聞の生コーパスを用いることで曖昧性の解消を行った．実験により先行手法 POPULATION の正解率が 40%程度に対し提案手法である文脈参照法では正解率が 65%，文脈参照法+MENTION_COUNT では精度が 71% 程度になることがわかった．

今後の課題と検討として，次のようなものが挙げられる．

- 対象とするコーパスの拡大
- 言及回数 DB 作成において，都道府県レベルを市区町村レベルにまで範囲を小さくしたり扱うデータ（生コーパス）量をより大きくすることで言及回数 DB の精度を高める
- コーパスと住所 DB のサンプル年度の整合をとる
- 他ジャンルのコーパスでの実験

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

[4] 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.

[5] 毎日新聞社.CD-毎日新聞 2004 データ集.