D-037

場所に関する特徴語を利用したリアルタイム地名曖昧性解消手法

Realtime Toponym Disambiguation using Location-related Words

落合 桂一† 鳥居 大祐†
Keiichi Ochiai Daisuke Torii

1．まえがき

リアルタイムに情報が共有される Twitter や Facebook などの SNS を解析することで, 実世界で起こる様々なイベントを検出したり，注目されているスポットを抽出するなど, 今まさに起きている世の中の動向を知ることができる．これらの抽出された情報はユーザーにとっても有益なため情報提供サービスでの利用が期待される．このような位置情報サービスを提供するにあたり重要なのは，地名やスポットに関する情報を“リアルタイム”かつ“正確”に収集できることである．

Twitter で位置に関連するツイートを集める方法は大きく 2 つある．1 つはジオタグ（緯度経度）を付加して投稿されたツイートを利用する方法，もう 1 つはツイート本文をテキスト解析し地名を抽出して位置と関連付ける方法である．Cheng ら[1]によると全ツイート中 0.42%のツイートのみジオタグが付加されている．また，Kitamoto ら[2]によると文章中に地名が含まれるツイートは全体の 5〜10%である．そのため情報抽出という観点では，ツイートを多く集めるためツイートのテキスト解析を行い言及されている場所を特定する方法が有効である．テキスト解析で位置と関連付ける場合は地名曖昧性解消が課題となる．

地名の曖昧性には 2 つの種類がある[3]．1 つは Geo/Non-geo 曖昧性と呼ばれ，地名と同じ表記で地名以外の意味を持つものである．例えば「松島」という地名は人名としても使われる．もう 1 つは Geo/Geo 曖昧性と呼ばれ，表記が同じ地名が複数存在するものである．例えば「日本橋」という地名は東京と大阪に存在する．Geo/Non-geo 曖昧性の解消には CRF（Conditional Random Field）を用いた固有表現抽出を行う手法が多い[3-5]．Geo/Geo 曖昧性には「1 つのコンテクストで現れる地名は地理的に近い場所を示すことが多い」という仮定のもと，地理的に近い地名（近隣地名）を利用して曖昧性を解消する方法が多い[2-4]．既存研究ではそれぞれの曖昧性に対して個別に解消を行っているが，本研究で目的とするようなリアルタイム処理には 2 つの曖昧性を 1 つの枠組みで処理できることが望ましい．また，従来研究では曖昧性解消のために地名のみを利用しているが，Twitter は文章が短いため地名以外の単語も利用できる方がよい．

そこで本研究では，Twitter のようなリアルタイムに生成されるドキュメントを対象とし，テキストマッチングにより 2 つの地名曖昧性を 1 つの枠組みでリアルタイムに解消する手法を提案する．提案手法では二段階でツイートの抽出を行う．まず第一段階では，地名と，曖昧性解消するための単語の 2 単語が共起しているか，テキストマッチングを行い曖昧性を解消する．Geo/Non-geo 曖昧性に対しては，曖昧性解消するための単語として地名に関連する静的な文書から抽出した静的特徴語を利用して曖昧性解消を行い，Geo/Geo 曖昧性については従来手法と同様に近隣地名を利用して曖昧性を解消する．曖昧性解消の対象の地名に対してどちらの曖昧性が存在するかを予めタグ付けしておく．第一段階の抽出では適合率を重視して抽出を行っており，再現率が低い．そこで第二段階の抽出として特徴語を増やしてツイート抽出を行う．ツイートされる内容にはその場所特有のトピックが存在することが多いと考え，第一段階で抽出したツイートから地名ごとにその場所特有の単語（特徴語）を抽出し，特徴語を曖昧性解消するための単語として利用する．

† 株式会社 NTT ドコモ, NTT DOCOMO, INC.

2．関連研究

単語の地理的な局所性を利用した場所の特定に関する従来研究として，Cheng ら[1]のユーザー位置推定に関する研究，長岡ら[5]の実世界の位置情報類推に関する研究がある．Cheng ら[1]の研究では，ツイートに含まれる単語とそのツイートを行ったユーザーの位置情報を用いて単語の地理的な分布を作成し，ある地域に特有の単語を抽出する．そして，抽出した地域に特有の単語を利用してユーザーの位置を推定する．この研究では，地域に特有の単語抽出を行っているが曖昧性解消は行っていない．

長岡ら[5]の研究では，ブログを対象として地名と関連の強い単語を地名との共起度および単語の一般性をもとに生成し，主題となる場所を判断するために関連語を利用している．この研究では地名の曖昧性解消には従来研究と同様に地名のみを利用しており，地名の関連語を利用した曖昧性解消は行っていない．

3. 提案手法

3.1 提案手法の概要

提案手法の流れを図 1 に示す．処理は大きく分けて以下の 4 つの要素から構成される．(1) オフラインでの静的特徴語辞書の生成, (2) オンラインでの第一段階の地名曖昧性

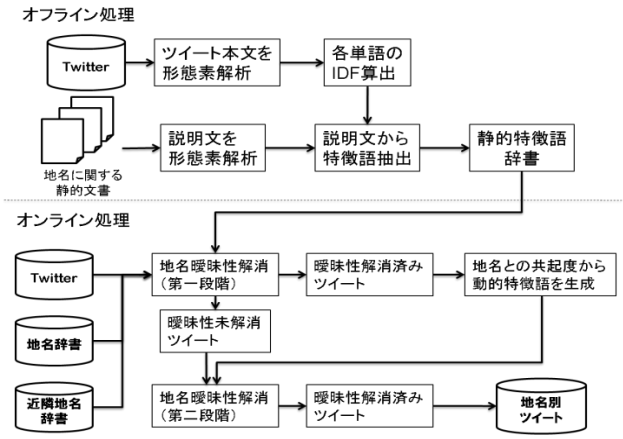

図 1 提案手法の概要

解消，(3) 動的特徴語辞書の生成，(4) 動的特徴語辞書を利用した第二段階の地名曖昧性解消．

3.2 静的特徴語の生成と曖昧性解消

静的特徴語は，対象の地名についての Wikipedia や観光案内の Web ページ，観光情報データベースのような，SNS に比べ静的な文書から抽出を行う．ここでは，1 つの地名に対して複数の特徴語を抽出する．

次に，生成した地名ごとの特徴語，および従来手法と同様に近隣地名を利用して地名の曖昧性解消を行う．本研究では地名辞書に予め対象となる地名と同名の地名が存在するか（Geo/Geo 曖昧性），地名以外の意味で利用されるか（Geo/Non-geo 曖昧性）を示すタグ付けを行っている．曖昧性解消の際には，このタグを参照し Geo/Geo 曖昧性の場合は近隣地名辞書を利用して曖昧性解消を行う．Geo/Non-geo 曖昧性の場合は特徴語を利用して曖昧性解消を行う．

3.3 動的特徴語の生成と曖昧性解消

前節で説明した曖昧性解消を行って抽出されたツイートから地名と関連する動的特徴語を抽出する．動的特徴語の指標としては地名と単語のダイス係数を利用する．

$$\frac{2|X \cap Y|}{|X| + |Y|} \quad (1)$$

ここで，X は地名を含むツイート，Y はある単語を含むツイートである．ダイス係数が閾値以上の場合に動的特徴語として利用する．

次に動的特徴語を利用して曖昧性を解消する．ここでは地名を含み第一段階で曖昧性解消されなかったツイートを対象に，地名と動的特徴語を含むツイートを抽出する．

4. 評価実験

4.1 実験環境

提案手法による地名曖昧性解消の性能を評価するため曖昧性解消の第一段階および第二段階に対してそれぞれ実験を行った．どちらの実験にも共通で地名辞書および地名に関する静的文書として，NTT ドコモで提供している「ご当地ガイド」という観光向けアプリに利用されている観光スポットデータと観光スポットごとの説明文を利用した．ここから静的特徴語の生成を行った．近隣地名辞書として，全国の駅名データ，市町村名データを利用した．

第一段階の曖昧性解消についての実験では，2011/12/30～2012/1/2 のツイートを対象とした．提案手法により抽出されたツイートを人手で内容を確認し正誤を判断した．

第二段階の曖昧性解消の性能評価のため 2013/5/2 の 1 日分のツイートを対象として実験を行った．実験では 5/2 のツイートから 1 時間ごとに関連語を生成し，地名と動的特徴語を共に含むツイートを 2013/5/2 のツイートから抽出した．性能評価としてツイート数と精度を算出した．精度は曖昧性解消されたツイートから 100 ツイートをランダムで抽出し，人手で内容を確認し正誤を判断した．

4.2 実験結果

第一段階の曖昧性解消の実験結果を表 1 に示す．どちらの曖昧性に対しても高い精度で曖昧性解消できている．表 2 に第二段階の曖昧性解消において，動的特徴語を抽出する指標となるダイス係数の閾値を変化させ，その際に抽出されたツイート数および精度を示す．ダイス係数の閾値を下げた場合も精度を保って曖昧性解消を行えると言える．表 3 に動的特徴語を使って抽出されたツイートの例を示す．青字が観光スポット名称で，赤字が動的特徴語である．場所のトピックが動的特徴語として抽出されており，それが曖昧性解消に効果があることがわかる．

5. まとめと今後の課題

本稿では Twitter のようなリアルタイムに生成されるドキュメントを対象とし，テキストマッチングによりリアルタイムに地名の曖昧性を解消する手法を提案した．提案手法に対して評価実験を行い有効性を確認した．

今後の課題として，地名に対して Geo/Geo 曖昧性と Geo/Non-geo 曖昧性があるかのタグ付けを自動化すること，観光スポット名以外の地名に対する提案手法の有効性を確認することが挙げられる．

表 1　第一段階の曖昧性解消の実験結果

	Geo/Geo	Geo/Non-geo
抽出数	1624	225
正解数	1434	210
精度	0.88	0.93

表 2　第二段階の曖昧性解消の実験結果

ダイス係数閾値		0.5	0.2	0.1	0.05	0.025	0.01
Geo/Geo	抽出数	188	713	1335	2657	4268	5771
	精度	0.99	0.93	0.87	0.9	0.96	0.96
Geo/Non-geo	抽出数	192	320	970	1516	2964	3951
	精度	0.99	0.97	0.96	0.9	0.96	0.97

表 3　動的特徴語を利用して抽出されたツイート例

【成功事例】	FLAKE でのチケット販売に 5/26 に円山公園音楽堂で行われる五味やタンテ、YeYe、UNCHAIN 等が出演する RAINBOW'S END 2013 の前売り取り扱いを追加！その他販売中のチケットはコチラ http://t.co/W5KNqGLg9z
	RT @username: 5 月 1 日、有志が申し入れ☆中日新聞:中央公園　伐採中止を　市民団体要請　県は変更予定なし:北陸発:北陸中日新聞から(CHUNICHI Web) http://t.co/2UPF1H50D3
	★JR おでかけネット★ 大興善寺つつじまつり　<2013 年 4 月 17 日～5 月 6 日／大興善寺>　〈ｺﾒﾝﾄ〉別名「つつじ寺」として知られる大興善寺では、4 月中旬～5 月中旬にかけて約 5 万本のつつじが咲き乱れます。イベント期間中は、つつじ..http://t.co/bzdaw2dGwR
【失敗事例】	RT @username: いよいよ宮城でも終盤を迎える桜を追って宮城県北部に行ってきました！　伊達政宗公ゆかりの岩出山城跡「城山公園」は強い風のせいかほとんど葉桜になっていましたが政宗公も見たであろう城跡高台からの景色はとてもきれいでした (´●×･) htt…

参考文献

[1] Z. Cheng, J. Caverlee and K. Lee, “You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users,” CIKM, pp.759-768, 2010

[2] A. Kitamoto and T. Sagara, "Toponym-based geotagging for observing precipitation from social and scientific data streams," GeoMM '12 Proceedings of the ACM multimedia 2012 workshop on Geotagging and its applications in multimedia, pp.23-26

[3] E. Amitay, N. Har'El, R. Sivan and A. Soer. “Web- a-where: Geotagging web content,” SIGIR, pp. 273-280, 2004.

[4] Leidner, J.L. “Toponym Resolution in Text – Annotation, Evaluation and Applications of Spatial Grounding of Place Names,” Doctoral Dissertation. University of Edinburgh.

[5] 長岡 諒, 松本 光弘, 沼尾 正行, 栗原 聡. “Web における実世界の位置情報類推に関する研究” 人工知能学会全国大会, 2009.