場所参照表現のグラウンディングに向けて

佐々木 彬 †　　五十嵐 祐貴 ‡　　渡邉 陽太郎 †　　乾 健太郎 †

東北大学

† {aki-s, yotaro-w, inui}@ecei.tohoku.ac.jp　　‡ yuki.i@s.tohoku.ac.jp

1　はじめに

テキスト中に含まれる表現を実世界と対応づけることは，自然言語処理の分野において大きな課題となっている．その中で，テキスト中に含まれる特定の場所を指し示す表現（以下，場所参照表現という）の実際の場所を特定するというタスクは，様々な応用例が考えられ，需要の大きいものとなっている．その一例として，震災時に拡散された情報の集約が挙げられる．

東日本大震災時にそうであったように，SNS 等が緊急時の情報共有の手段として用いられるようになっている．しかしながら，多くの人が状況を場所参照表現とともに発信するため，人手で状況を整理・把握することは極めて困難である．そこで，それらの場所参照表現を機械的に処理し，地図にマッピングできれば，どの場所に救助が必要な人が多いか，また，どの場所に物資提供を求めている人が多いかなどを視覚的に知ることが可能となる．

また，マイクロブログ等に出現する場所参照表現からユーザの行動を分析するプロファイリングなどにも利用できる．

以上のように様々な応用が考えられるものの，場所参照表現を地図と対応づけるというタスクは，解決の難しい問題となっている．地図と対応づけるためには，大きく以下の問題を解決する必要がある．例を図 1 に示す．

テキスト中の場所参照表現の同定

場所参照表現の同定は，固有表現抽出の対象とされてきた地名以外にも，「気仙沼駅近くの交差点」や「国道 45 号線沿いのコンビニ」などの一般名詞を伴う表現も対象となる．そのため，事前に場所参照表現を同定する必要がある．

場所参照表現が指し示している実際の場所の特定

場所参照表現には，「宮城球場」と「楽天 Kobo スタジアム宮城」のように異表記で同一の場所を指す場合や，「ヨドバシカメラ」のように同一表記で異なる場所を指す場合がある．そのため，住所データベースに登録されている場所情報が，場所参照表現と同一の表記とは限らず，単純に住所データベースから取得することは難しい．よって，表記ゆれや表記の曖昧性，文脈等を考慮したうえで，場所参照表現が指し示している場所の住所を特定する必要がある．

図 1: 場所参照表現のグラウンディング

本研究では，テキスト中に含まれる全ての場所参照表現に対して実際の場所の特定を行うにあたっての，問題分類を行う．これにより，場所参照表現の曖昧性解消へ向けた課題を分析する．

はじめに，テキスト中の場所参照表現の同定に関する既存研究について言及する．その次に，固有表現抽出器による場所参照表現の同定について評価実験を行い，それとともにテキスト中の場所参照表現の分布を調査する．また，場所参照表現が指し示している実際の場所の特定における問題分類を行い，文脈を考慮することによってはじめて特定できるような場所参照表現や，それでも特定することのできない場所参照表現などについて，例とともに述べる．

2　関連研究

場所参照表現のグラウンディングに関する既存研究として，以下のものが挙げられる．

[Pyalling 06] は，IP アドレスやドメイン名といった情報に基づき，Web サイトに対してジオコード付与を行っている．[Lieberman 10] は，一般的に知られる地名から構成される *global lexicon* と，ある特定の地域だけで使われる地名から構成される *local lexicon* という概念を用いて，ニュース記事へのジオコード付与を行った．[Cheng 10] は，アメリカのテキサス州で使われる “howdy” という単語のように，ある特定の地域で頻繁に使われる単語を手がかりとして，都市単位で Twitter 上のユーザの位置を推定した．これらは，Web サイトやニュース記事，ツイート自体にジオコードを付与するものであり，テキスト中に含れる場所参照表現まではジオコードの付与を行っていない．

[Amitay 04] は，“Cairo” や “Paris” などの，地球上に複数存在するような地名の曖昧性解消を行った．[Paradesi 11] は，位置情報サービスなどにより付与されたジオコードを手がかりとして，ツイートに含まれている場所参照表現へのジオコード付与を行った．これらの既存研究では場所参照表現として，対象を地名や固有名詞のみに限定していた．しかしながら，実際には固有名詞の場所参照表現以外にも，普通名詞を伴う場所参照表現も多く存在すると考えられるため，それらについても考慮する必要がある．

本研究では，固有表現抽出による場所参照表現の同定の評価とともに，固有名詞と普通名詞の場所参照表現の分布を調査し，そのうえで場所参照表現が指し示している実際の場所をいかにして特定できるかを分析する．

3 固有表現抽出による場所参照表現同定の評価

まず，場所参照表現の同定を固有表現抽出の問題として解く．ここで，固有表現抽出器により同定できる場所参照表現は固有名詞に限られるため，評価の対象も固有名詞のみとする．

3.1 訓練データ

訓練データには，拡張固有表現タグ付きコーパス [橋本 08] を用いる．ここで，タグについて，拡張固有表現階層 [Sekine 02] のうち，場所参照表現となりうるもののみを残し，その他のタグを O とする．タグは地名[1]，組織名[2]，施設名[3]を用い，サブクラスは集約する．

固有表現抽出の手法として，条件付確率場 [Lafferty 01] を用い，タグとそれに付随する IOB タグを採用して系列ラベリングを行う．条件付確率場の実装として，CRFsuite [Okazaki 07] を用いる．

3.2 素性

条件付確率場の素性として，前後 2 個までの周辺形態素（表層形，原形，品詞，品詞細分類），読み，発音，文字種，ランドマーク辞書（完全一致），ランドマーク辞書（類似度），住所辞書（完全一致）を用いる．

ランドマーク辞書には，Web ページをクロールして抽出した 21,333,845 件のランドマーク名と，対応する住所が含まれる．住所辞書には，64,498 件の都道府県名，市区町村名が含まれる．これらの辞書により，形態素自体がランドマーク名や都道府県名，市区町村名となっているときに，より正確に場所参照表現であると認識できるようにする．

ランドマーク辞書（完全一致），住所辞書（完全一致）については，対象の形態素が辞書に完全一致で含まれる場合は True，そうでない場合は False となる．ランドマーク辞書（類似度）は，対象の形態素とのコサイン類似度が 0.8 以上の項目がランドマーク辞書に含まれる場合は True，そうでない場合は False となる．これは，テキスト中のランドマーク名は，ランドマーク辞書のものと比較して一部が省略されるなどして異なる表記となる場合が考えられるためである．ここで，コサイン類似度によるランドマーク辞書からの高速な類似文字列検索のために SimString [岡崎 10] を用いる．以下，ランドマーク辞書（完全一致）を *landmark*，ランドマーク辞書（類似度）を *landmark_sim*，住所辞書（完全一致）を *address* と表記する．

3.3 テストデータ

テストデータとして，株式会社ホットリンクより提供されたツイートデータを用いる[4]．ツイートデータをテストデータとした理由は，Twitter 等の SNS 上では，自らの現在地について発信する例や，待ち合わせ場所の連絡など，多くの場所参照表現が含まれると考えられるためである．このツイートデータは，2011 年 3 月 11 日から 2011 年 3 月 29 日までの約 2 億 1 千万のツイートを含む．ここで，場所参照表現を実際の分布に従って取るために，ツイート本文のみをランダムに約 5000 文抽出し，人手により場所参照表現を含む文のみを抜き出す．そのうえで，文中の場所参照表現にアノテートを行う．約 5000 文中，場所参照表現を含むツイートは 400 ツイート存在した．それらの中には複数の場所参照表現を含むツイートもあり，固有名

[1]地名その他，温泉名，GPE その他，地区町村名，郡名，都道府県州名，国名，地域名その他，大陸地域名，国内地域名，地形名その他，山地名，島名，河川名，湖沼名，海洋名，湾名

[2]組織名その他，法人名その他，企業名，企業グループ名，政府組織名

[3]施設名その他，施設部分名，遺跡名その他，古墳名，GOE その他，公共機関名，学校名，研究機関名，取引所名，公園名，競技施設名，美術博物館名，動植物園名，遊園施設名，劇場名，神社寺名，停車場名，電車駅名，空港名，港名

[4]http://www.hottolink.co.jp/press/936

表 1: 固有表現抽出による場所参照表現同定の評価

素性	precision	recall	F_1-score
基本素性，*landmark*	51.96% (53/102)	23.77% (53/223)	32.62%
基本素性，*landmark*，*landmark_sim*	50.00% (75/150)	33.63% (75/223)	40.21%
基本素性，*landmark*，*address*	51.49% (52/101)	23.32% (52/223)	32.10%
基本素性，*landmark*，*landmark_sim*，*address*	51.59% (81/157)	36.32% (81/223)	42.63%

表 2: 固有表現抽出器により付与したタグ数

タグ	場所参照表現	非場所参照表現	合計
地名	26	34	60
組織名	8	23	31
施設名	47	19	66

詞の場所参照表現は 223 個，普通名詞の場所参照表現は 236 個存在した．

3.4 実験結果

固有表現抽出器による場所参照表現同定の評価結果を表 1 に示す．結果より，*landmark_sim*，*address* の素性は実際に有効であると考えられる．しかしながら，固有表現抽出器を用いても同定できていない場所参照表現は多く，場所参照表現同定の難しさがわかる．

次に，最も性能の優れていた全素性を用いたモデルについて，付与したタグ別にどの程度場所参照表現となっているかを評価した．結果を表 2 に示す．

また，固有表現抽出器によりタグを付与できなかった場所参照表現の例を以下に示す．人手により場所参照表現とされた箇所に下線を付与する．

(1) よく 所沢 近辺の スタバ に出現するよ！

(2) 思い出横丁 の「つるかめ食堂」にいきたくなってきた

(3) 神田 の もといし で晩ごはんにつけ麺並あつ盛り。

(1) の「スタバ」は,「スターバックス」の略称である．しかしながら，ランドマーク辞書には略称は含まれないため，タグを付与できなかったと考えられる．(2) の「思い出横丁」は複数の店舗が建ち並ぶ通りの通称である．このような，ランドマーク名でも住所名の一部でもないような場所参照表現については本研究で用いた辞書に含まれないため，辞書の拡張が必要と考えられる．(3) の「もといし」についても，ランドマーク辞書に含まれないためタグを付与できていなかった．

4 場所参照表現のグラウンディングに向けて

テストデータに含まれる場所参照表現について，人手で確認し，実際の場所を特定することが可能であるか，不可能であるかを分類する．以下，実際の場所を特定することが可能である場所参照表現を「特定可能な場所参照表現」と呼び，実際の場所を特定することが不可能である場所参照表現を「特定不可能な場所参照表現」と呼ぶ．人手により場所参照表現とされた箇所に下線を付与する．

表 3: 場所参照表現のグラウンディングにおける問題分類

	固有名詞	普通名詞
場所参照表現のみで特定可能	155	0
局所的な文脈で特定可能	26	4
大局的な文脈で特定可能	2	3
特定不可能	40	229
合計	223	236

また，テストデータ中の固有名詞の場所参照表現，普通名詞の場所参照表現について，各々の分類に該当したものの個数を表 3 に示す．

4.1 特定可能な場所参照表現

4.1.1 場所参照表現のみで特定可能なもの

(4) 彼のピッチングが 甲子園球場 で見られることを祈っています

(4) の場所参照表現「甲子園球場」については，一意に定まるユニークな場所参照表現となっている．よって，場所参照表現のみで実際の場所を特定することが可能である．

4.1.2 局所的な文脈で特定可能なもの

(5) 心斎橋 ついた。アップルストア に向かう。

(6) 今、新田東 の コスモ で灯油販売してます

(7) ソニーは 宮城県白石市 の 子会社 の操業を見合わせるそうです。

(5) の場所参照表現「アップルストア」だけでは複数の候補が考えられるものの，文内の「心斎橋ついた。」という表現より，これは「アップルストア心斎橋店」であると特定することが可能である．(6) も同様に，局所的な文脈により特定可能なものとなっている．(7)

の場所参照表現「子会社」は普通名詞であるが，文内の「ソニー」という表記と,「宮城県白石市」という場所参照表現より,「ソニー白石セミコンダクタ」であると特定することが可能である．

4.1.3 大局的な文脈で特定可能なもの

(8) 福島 避難所 で患者１４人死亡 http://...

(9) ラウンドワン で火事あったのか。

(8) では，テキストに付随する URL のページを辿ることによって，普通名詞の場所参照表現である「避難所」を具体的に特定することが可能である．(9) については，このツイートが発信された時間付近のニュースより，特定することが可能である．

4.2 特定不可能な場所参照表現

4.2.1 複数の候補があり，絞り込むことができないもの

(10) ヨドバシ とかもう営業してるの？

(11) コンビニ にものすごいミネラルウォーターコーナーが出来てる！

(12) 自宅 停電開始。

(10) については，固有名詞の場所参照表現「ヨドバシ」から複数の候補が考えられるものの，(5) や (6) のように局所的な文脈で特定することはできず，大局的な文脈にも手がかりとなる情報が含まれないため，実際の場所を特定することは不可能である．(11) と (12) には，普通名詞の場所参照表現が含まれるが，(10) と同様に局所的，大局的な文脈で実際の場所を特定することは不可能である．

4.3 考察

場所参照表現のグラウンディングにおける問題を分類した結果，固有名詞の場所参照表現には，その表現のみで実際の場所を特定可能なものが多く見受けられた．しかしながら，文脈を考慮しなければ特定できない場所参照表現も存在するため，それらの文脈を考慮して実際の場所を特定する枠組みが必要であると考えられる．

また，普通名詞については，文脈を考慮すれば特定できる例も一部存在したが，大部分はそれだけでは特定不可能なものとなっていた．これらについて特定するためには，場所参照表現の発信者単位でより詳細に分析し，特定のための手がかりを取得することが必要であると考えられる．

5 おわりに

本研究では，場所参照表現のグラウンディングに向けた固有表現抽出器による場所参照表現の同定について評価実験を行い，問題の分類を行った．また，固有名詞，普通名詞を含めた，場所参照表現のグラウンディングの課題分析を行った．今後の課題として，本研究で分類・整理した問題を解決する枠組みの提案が挙げられる．

謝辞

本研究は，RISTEX 社会技術研究開発センターの研究開発活動「コミュニティがつなぐ安全・安心な都市・地域の創造」の一環として行われた．

参考文献

[Amitay 04] Amitay, Einat, Nadav Har'El, Ron Sivan, and Aya Soffer. "Web-a-where: geotagging web content." Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004.

[Cheng 10] Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. "You are where you tweet: a content-based approach to geo-locating twitter users." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[橋本 08] 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，自然言語処理研究会報告 (NL-188-17),pp. 113—120, 2008.

[Lafferty 01] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

[Lieberman 10] Lieberman, Michael D., Hanan Samet, and Jagan Sankaranarayanan. "Geotagging with local lexicons to build indexes for textually-specified spatial data." Data Engineering (ICDE), 2010 IEEE 26th International Conference on. IEEE, 2010.

[Okazaki 07] Okazaki, Naoaki. "CRFsuite: a fast implementation of conditional random fields (CRFs)." URL http://www. chokkan. org/software/crfsuite (2007).

[岡崎 10] 岡崎直観，辻井潤一．"高速な類似文字列検索アルゴリズム." 情報処理学会創立 50 周年記念全国大会 URL http://www.chokkan.org/software/simstring/ (2010).

[Paradesi 11] Paradesi, Sharon Myrtle. "Geotagging Tweets Using Their Content." FLAIRS Conference. 2011.

[Pyalling 06] Pyalling, Alexei, Michael Maslov, and Pavel Braslavski. "Automatic geotagging of Russian web sites." Proceedings of the 15th international conference on World Wide Web. ACM, 2006.

[Sekine 02] Sekine, Satoshi, Kiyoshi Sudo, and Chikashi Nobata. "Extended Named Entity Hierarchy." LREC. 2002.