B3IM2018

修士論文

自然言語文における場所参照表現のグラウンディングに関する研究

佐々木 彬

2015 年 2 月 10 日

東北大学 大学院
情報科学研究科 システム情報科学専攻

本論文は東北大学 大学院情報科学研究科 システム情報科学専攻に
修士 (情報科学) 授与の要件として提出した修士論文である。

佐々木 彬

審査委員：

乾 健太郎 教授　　（主指導教員）
篠原 歩 教授
徳山 豪 教授
岡崎 直観 准教授　（副指導教員）

自然言語文における場所参照表現のグラウンディングに関する研究*

佐々木 彬

内容梗概

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，実世界の特定の場所を指し示す表現（場所参照表現）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．しかしながら，従来の研究では場所参照表現として地名のみが対象として扱われ，施設名については考慮されていなかったという問題点があり，地名・施設名とその実際の場所を関連づけたコーパスが存在しなかったため，どのような現象がどの程度で出現するのか，といった定量的な分析がされずにいた．本研究ではその問題を解決すべく，地名・施設名を含む場所参照表現とそれが指し示す実際の場所とを関連づけたコーパスを作成し，作成したコーパス内でどのような現象が起きているのかを分析する．

キーワード

自然言語処理，地理情報処理，固有表現抽出，曖昧性解消，グラウンディング

*東北大学 大学院情報科学研究科 システム情報科学専攻 修士論文, B3IM2018, 2015 年 2 月 10 日.

目次

図目次

表目次

1　はじめに

近年，Twitter[1] 等のマイクロブログの流行により，世界中のユーザが情報発信を行える環境が整ってきた．マイクロブログは従来のブログと比較して少ない文字数で投稿されることが一般的であり，従来ブログのようなサービスを利用していなかったユーザも多く利用している．Twitter 社の報告によると，2012 年には 1 日あたりの投稿数が 4 億件を突破したこともあった [2]．

このような爆発的な普及に併せて，マイクロブログは緊急時の情報交換の場としても重要な役割を果たしつつある．例えば 2011 年 3 月 11 日の東日本大震災時には，避難所や物資，行方不明者などについての情報が Twitter 上で盛んに発信された．これらの情報は被災地のユーザが近隣の情報を得るためにも，また被災地外のユーザやマスメディアなどが被災地の状況を知るためにも，重要な情報源となっていた．

しかしながら，マイクロブログ上の情報は膨大であり，その中から人手で欲しい情報のみを抽出することは困難である．災害時などに各地域に関係する投稿を収集・分類することができれば有用であると考えられるが，マイクロブログ上からそのように地域を限定して情報を収集することは容易ではない．Twitter ではスマートフォンなどの GPS 機能により投稿に緯度・経度といった座標情報を付与することができるため，座標情報に基づき特定の地域のツイートを集めることは可能ではあるが，座標情報を付与するか否かは各ユーザの設定に依存する．Middleton ら [1] の報告によると，座標情報を付与されているツイートは全体の 1%にも満たない．このため，座標情報を利用して特定の地域に関する投稿を収集しようとしても，網羅性に欠ける．

座標情報を利用できない場合に特定の地域に関する投稿を収集するための手段として，テキスト中の表現を手がかりにすることが考えられる．例えば，仙台市に関する情報のみを収集したい場合は,「仙台市」というクエリで投稿全体を単純に検索するだけでいいのではないか，と一見すると思われる．だが，この手法では「仙台市」というキーワードがテキスト中に含まれる投稿のみしか取得できず，仙台市内の地名や施設名などに言及している投稿までも収集することは難しい．

そこで，この問題を解決するためのひとつの案として，テキスト中に含まれる，特定の場所を指し示す表現を解析する，というタスクを考える．自然言語文の中には，以下のように実世界中の座標を持つエンティティを指し示す表現がしばし

[1] http://twitter.com/

[2] https://twitter.com/TwitterAds/status/210867782361948161

ば現れる．

(1) 仙台駅 近くの ヨドバシカメラ に来ています

(1) のテキスト中では,「仙台駅」と「ヨドバシカメラ」という表現はそれぞれ,実世界中の座標を持つエンティティ「仙台駅」と「ヨドバシカメラ マルチメディア仙台」を指し示す表現である．

本研究では，自然言語文中に含まれる，実世界中の座標を持つエンティティを指し示す表現を場所参照表現と定義し，表現とそのエンティティを対応付けるという，場所参照表現のグラウンディングを行うことを最終目標に見据える．自然言語文中の場所参照表現を実世界の座標を持つエンティティと対応付けることは容易ではない．例えば (1) のテキスト中の「ヨドバシカメラ」に着目し，地名・施設名辞書中を検索したとすると,「ヨドバシカメラ マルチメディア仙台」の他に「ヨドバシカメラ 新宿西口本店」,「ヨドバシカメラ マルチメディア Akiba」,また「ヨドバシカメラ マルチメディア吉祥寺」といった複数の候補が生じる．この際に，それらの複数の候補から適切な候補を選び出す必要があるが，そのためには周辺文脈などを考慮しなければ判断不可能な場合もあるなど，非常に難しい問題となっている．

ここで，評価をするため，あるいは機械学習の訓練データとして使うためのコーパスが，現時点では存在しないという問題がある．既存研究では，扱う対象として地名のみに限定した上でテキスト中の表現と実世界のエンティティを対応付けたコーパスを作成していたが，施設名などのその他の場所参照表現までを考慮して具体的なエンティティを付与したコーパスは存在しない．

本研究の主な貢献は以下の 3 点である．

- 場所参照表現として施設名まで考慮し，テキスト中の表現と実世界のエンティティを対応付けたコーパスを作成した．
- 作成したコーパスを分析することで，施設名を含む場所参照表現をグラウンディングするにあたって，どのような問題点が存在するのかを明らかにした．
- 作成したコーパスに地名を対象としていた既存研究の曖昧性解消手法を適用することで，施設名の曖昧性解消に既存手法が有効であるかを評価した．

本論文の構成を述べる．はじめに，2節で場所参照表現の関連研究を述べる．3節，4節では，コーパスを作成するにあたってのガイドライン設計，必要なアノテーションツールなどについて議論する．5節では，ツイートデータに対して実際にアノテーションを行う．6節では，作成したコーパスを分析することで，場所参照表現のグラウンディングに必要となる知識を整理する．7節では，既存研究で用いられていた場所参照表現の曖昧性解消手法を本コーパスに適用する．8節では，4節で論じたアノテーション手順をクラウドソーシングサービスに適用するにあたり，具体的にどのような手順を踏む必要があるかを議論する．最後に9節にて，本論文のまとめを述べる．

2　関連研究

場所参照表現に関する研究は，**Document GeoLocation** と **Toponym Resolution** という 2 種類のタスクに大別される．本節では，各々のタスクの説明とともに，既存研究について述べる．

2.1　Document Level GeoLocation

Document Level GeoLocation は，Web ページ，新聞記事などをドキュメントとみなし，そのドキュメントを実世界の特定の場所と対応付ける（緯度経度情報といったジオコードを付与する），というタスクである．Pyalling ら [2] 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た．Serdyukov ら [3] は，写真投稿サイト Flickr[3] に着目し，ユーザにより記述された写真の説明文とジオコードを訓練データとして用いて機械学習を行った．Lieberman ら [4] 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 ら [5] 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．Wing ら [6] [7]，Roller ら [8] は，地球上にグリッドを作成し，各グリッドについて教師あり学習を行うことで，グリッド単位でドキュメントの対応付けを行った．

Document Level GeoLocation では，テキスト中の各々の場所を指し示す表現を解析するのではなく，ドキュメント自体に着目する．テキスト中の場所参照表現に対してではなく，ひとつのドキュメントに対してジオコードを付与するというのが，後述する Toponym Resolution と異なる点である．

2.2　Toponym Resolution

Toponym Resolution は，テキスト中の場所を指し示す表現（**toponym**，本研究では場所参照表現と呼称）について，その表現が指し示している実際の場所を判定する，というタスクである．ここで，場所参照表現の中には，同一の文字列であるにも関わらず異なる場所を指し示すものがあり，これが大きな問題とな

[3] https://www.flickr.com/

る．例えば *“London”* という場所参照表現は，イギリスのロンドンを指す場合もあれば，カナダのオンタリオ州に存在するロンドンという都市を指す場合もある．

この曖昧性を解消するべく，様々な手法が提案されている．Smith ら [9] は場所参照表現の周辺単語を考慮した曖昧性解消手法を用いた．Ladra ら [10] は人口の情報を利用し，最も人口の多い候補を選択するという手法を取り入れた．Speriosu ら [11] は，Wikipedia[4] のジオコード付きの記事を用いた Indirect Supervision を用いた学習を行った．

また，メタデータを利用する例として，Paradesi [12] 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．しかしながら，テキストデータには必ずしもジオコードのようなメタデータが付随するとは限らない．例えば，Twitter ではユーザが自身のツイートに GPS 情報を埋め込むように設定することができるが，Middleton [1] によると，ツイート全体のうち GPS 情報が付与されているツイートは 1%にも満たない．このため，GPS 情報に依存した手法は限定的なものになってしまう．

場所参照表現に関するコーパスを作成した既存研究として，Leidner らの研究 [13] が挙げられる．Leidner らはテキスト中の場所参照表現と実際の場所との対応をアノテートできるインタフェースを用意し，それを用いて **TR-CoNLL** コーパスを作成した．ただしアノテーション付与の対象は地名に限定され，施設名へのアノテートは行われていない．また，付与対象文章のドメインはニュース記事となっていた．その他に，Crane ら [14] は **CWAR** というコーパスを作成した．このコーパスもまたアノテーション付与対象は地名のみとなっており，付与対象文章のドメインは書籍であった．

これらの既存研究では場所参照表現として扱う対象を都市名，国名，大陸名といった地名に限定して取り組んでいた．しかしながら，実際には「東京タワー」「ファミリーマート」「本屋」のような施設名も特定の場所を指し示している．こういった従来考慮されていなかった施設名までを対象に見据えてコーパスを作成するというのが，本研究と既存研究との大きな差異である．

[4] http://en.wikipedia.org/

3　取り扱うべき曖昧性の種類

場所参照表現をグラウンディングするにあたって，たとえ全ての場所参照表現の文字列が地名・施設名辞書に含まれていたとしても，その文字列に曖昧性がある場合は単純にグラウンディングすることはできない．

(2) 結局川崎でご飯食べることにした

(2) の「川崎」は地名・施設名辞書に含まれるが,「北海道虻田郡真狩村字川崎」「岩手県一関市川崎町」「神奈川県川崎市」など，複数のエンティティが存在する．このような，ある文字列が，エンティティ辞書（本稿では，地名・施設名辞書）のどのエンティティにあたるものか，に関する曖昧性をエンティティの曖昧性と呼称する．

また,「川崎」が場所参照表現としてではない使われ方をする場合もある．

(3) 大阪、川崎、新宿とかなり濃くてハードな３日間をすごしました。

(4) 川崎戦、前半は 0-0 で終了。しかし東京はなかなか高い位置でボールを奪えず、シュートも少ない前半でした。

(5) 川崎ちゃんとやっと来年のツアー相談。

(6) 川崎から南武線に乗って立川まで行きました。

上記の例のそれぞれの「川崎」について，(3) は地名として，(4) については文脈よりサッカークラブの「川崎フロンターレ」として，(5) は人名として，そして (6) は「川崎駅」として用いられていると判断できる．これらのように，ある文字列が，地名・施設名等の場所を指す表現であるか，また，そうである場合はどのサブクラス (県名・駅名・店舗名など) に当たるものか，に関する曖昧性をクラスの曖昧性と呼称する．

4　コーパス設計

3 節にて議論したように，場所参照表現と実際の場所との対応をアノテートしたコーパスを作成するにあたって，場所参照表現に付随する問題である，エンティティの曖昧性とクラスの曖昧性に注意する必要がある．

クラスの曖昧性に関しては，既存の固有表現タグ付きコーパスが参考になると考えられる．日本語の固有表現タグ付きコーパスとしては，IREX ワークショップ実行委員会が公開しているコーパス [15]，拡張固有表現タグ付きコーパス [16] が存在し，テキスト中のどの範囲の文字列が固有表現であるか，またその固有表現のクラスが何であるか，といったアノテーションが人手で付与されている．しかしながらいずれのコーパスにも，各固有表現が指す具体的なエンティティまでは付与されていない．

本節では，アノテート対象を場所参照表現に限定したうえで，従来の固有表現タグ付きコーパスで行われていたクラスの付与に加えて具体的なエンティティの付与を行うことを目的とし，コーパス設計の枠組みを議論する．また，従来の固有表現タグ付きコーパスでは固有名詞に限定したアノテートが行われていたが，場所参照表現には「コンビニ」や「病院」といった普通名詞も存在し，具体的なエンティティを付与できる場合があると考えられるため，固有名詞に加えて普通名詞もアノテート対象とする．

以上を踏まえたうえで，以下の要件を満たす検討を行った．

- 各工程を単純化するために工程を分解し，将来コーパス作成にクラウドソーシングを容易に利用できるようにする
- 各工程でのエラー要因を確認しやすくする

検討により，アノテート作業者（アノテーター）の行うタスクは **Mention Detection**（言及抽出），**Entity Resolution**（エンティティ解決）の 2 種類となった．以下，各タスクについての説明を記述する．

4.1　Mention Detection(言及抽出)

与えられたテキストのうち，どの部分文字列がタグ付与の対象であるかを指定したうえで，4.3 節で述べたタグセットから適切なタグを付与する．ここで，指定する部分文字列としては固有名詞ないし普通名詞，またその連続を対象とする．

4.2 Entity Resolution(エンティティ解決)

Mention Detection(言及抽出) によりタグを付与した文字列に対して，可能であれば具体的なエンティティを付与する．この際，付与するエンティティは地名・施設名辞書から選択する．

場所参照表現によっては，複数のエンティティを対応付けることが適切である場合もある．

- 都内ヨドバシカメラで完売ってどう言うことなの…？

この例の「ヨドバシカメラ」は1つの店舗ではなく，東京都内の複数の店舗を指し示していると考えられる．そのため,「ヨドバシカメラ 新宿西口本店」,「ヨドバシカメラ マルチメディア新宿東口」,「ヨドバシカメラ マルチメディア Akiba」，…,「ヨドバシカメラ マルチメディア錦糸町」というエンティティを全て付与する必要がある．ただし，以下のように付与すべきエンティティが膨大になってしまう場合，備考欄にその旨を記述することとする．

- 来年中にセブンイレブン全店で販売

この例の場合は，備考欄に「セブンイレブン全店舗」などと記述する．これは，アノテートコストを考慮しての対処である．

また，適切なエンティティが地名・施設名辞書中に見つからない場合もある．これは地名・施設名辞書のカバレッジの問題であるため，具体的なエンティティを付与せずに，備考欄に「辞書になし」などといった注釈を付与する．

加えて，エンティティを付与できた場合には，エンティティを選択する際に利用した手がかりを備考欄に記述する．ここで記述した手がかりに基づき，6節でエンティティの曖昧性解消に必要な手がかりを整理する．

4.3 アノテーションに用いたタグセット

本研究のコーパス作成時に用いるタグセットを表1に示す．以下，各々のタグの説明を記述する．

4.3.1 LOC(地名)

都道府県，市区町村，大字などの行政区域に対して，本タグを付与する．

- 横浜行きたすぎてやばい

表 1: アノテーションに用いたタグセット

タグ	具体例	説明	エンティティに対応付けるか
LOC(地名)	埼玉県 仙台市 神保町	都道府県，市区町村，大字などの行政区域	○
FAC(施設名)	仙台駅 九州大学 ファミリーマート	具体的な場所を持った施設	○
RAIL(鉄道路線名)	京浜東北線 田園都市線	具体的な路線名称	今後対応付ける予定
ROAD(道路名)	4 号線 東北道	具体的な道路名称	今後対応付ける予定
ORG(組織名)	政府 情報処理学会 火山学会	場所として言及されていない複数の人間からなる組織の名前	対応付けない
GEN(総称表現)	病院 コンビニ	施設名のうち総称的に述べられている表現	対応付けない
FIC(架空の地名)	洞窟 おとぎの国	現実世界に存在しないが，仮想的な場所の概念を表している表現	対応付けない
AMB(クラスが曖昧)		クラスが上記のものに当てはまらないが，地名・施設名である可能性を否定できない場合	対応付けない

- 新宿を久しぶりに闊歩した
- 九州上陸する頃には 950hpa ぐらいになってるんじゃないかな

4.3.2 FAC(施設名)

現実世界中で具体的な場所を持っている施設に対して，本タグを付与する．

- 思いつきで行った USJ から帰宅
- ゲストハウスまでもう少しやけど眠たい

- シメに マック 行って帰り途中

4.3.3 RAIL(鉄道路線名)

具体的な鉄道路線に対して，本タグを付与する．

- 京浜東北線 川崎で人身事故
- 仙山線 が熊を轢き遅延
- 山手線、止まったあああああああ!!!!!

4.3.4 ROAD(道路名)

具体的な道路名に対して，本タグを付与する．

- 国道 47 号線、事故？
- 東名高速 通ります！
- 今日の 常磐道 空いてる

4.3.5 ORG(組織名)

場所として言及されていない，複数の人間からなる組織の名前に本タグを付与する．

- 白泉社 新入荷
- 相対性理論 のレコほしいな
- ベガルタ仙台 の移籍加入・退団情報をまとめました

4.3.6 GEN(総称表現)

施設名のうち，総称的に述べられている表現に本タグを付与する．

- たまに 高層マンション のベランダに布団干してる人いるよね
- お盆って 病院 あいてる？
- 最近の コンビニ のコーヒーはクオリティ高いな〜

4.3.7　FIC(架空の地名)

漫画，ゲーム，小説などに現れる，架空の地名・施設名に本タグを付与する．

- ガスグスタフ火山洞窟 をクリアした！
- 杜王町 を舞台にした漫画『ジョジョの奇妙な冒険』第 4 部
- 国立魔法大学附属第一高校 に行きたかった

4.3.8　AMB(クラスが曖昧)

アノテーターがアノテート時に付与対象文字列がどのクラスであるかを文脈から判別できなかった場合，本タグを付与する．

- 郡上八幡 思い出したー
- 予想外に 秋山 ガッツリだね
- 大宮 とかかな？

4.4　アノテーション付与対象

4.3 節にて定義したタグセットに従い，テキスト中の付与対象部分文字列に対してタグを付与する．この際，具体的な定義として関根の固有表現階層 7.1.0[5][17] を参考とした．

4.5　地名・施設名辞書

地名・施設名辞書を構築するにあたって，各種オープンデータ，Web 上データベースを用いた．各辞書種別とそのエントリ数を表 2 に示す．

4.6　コーパスアノテーションのためのツール

本節で述べるコーパス作成手順においては **Mention Detection**(言及抽出)，**Entity Resolution**(エンティティ解決) の 2 つのタスクを行うこととなるが，この際にアノテーションのためのツールが効率面で重要となる．そこで本研究では，

[5]https://sites.google.com/site/extendednamedentityhierarchy/

表 2: 各辞書種別，エントリ数

辞書種別	情報源	エントリ数
県・市区町村名・大字	街区レベル位置参照情報	147774
ランドマーク	Yahoo!ロコ	4989652

コーパスアノテーションのためのツール開発を行った．開発したアノテーションツールの全体図を図 1 に示す．アノテーションツールはウェブブラウザ上で動作し，左右に分割された 2 つのペインで構成されている．以下，実際のアノテーション手順に従って，アノテーションツールの詳細を説明する．

図 1: コーパスアノテーションのためのツールの全体図

アノテーションツール読み込み時の初期状態は，図 2 のようになる．これはアノテーション付与前のツイートの一覧表示であり，図 1 でいうところの左側のペインに表示される．ここで各ツイートの左側に位置する「edit」というボタンをクリックすると，図 3 のウィンドウがポップアップ表示される．アノテーターはこのウィンドウ内のテキスト中の，タグ・エンティティ付与対象文字列の範囲をドラッグで選択する．

図 2: ツイート一覧表示画面

左のペインで以上の操作を行うと，右のペインに図 4 の画面が表示される．画面上部のボタンはタグの一覧を表している．また，その下には備考欄を設けてあり，アノテート時に備考として別途記述すべき内容があれば，ここに書き記す．さらにその下には,「東京」という文字列で地名・施設名辞書を検索した結果を表示している．なおここで，検索の際に内部で ElasticSearch[6] を用いることで，検索結果出力の高速化を図っている．アノテーターは，この検索結果中に付与すべきエンティティが見つかった場合，そのエンティティのチェックボックスをクリックすることで選択する．また，もし検索結果中に付与すべきエンティティが見つからない場合，アノテーター自身で検索クエリを入力する必要がある．画面最下部の「自治体 search」という箇所に検索クエリを入力すると地名辞書からの検索結果が，また「施設 search」という箇所に検索クエリを入力すると施設名辞書からの検索結果が表示されるようになっている．エンティティを付与する際には，

[6] http://www.elasticsearch.org/

図 3: ポップアップ表示内のタグ・エンティティ付与対象文字列の選択

備考欄にそのエンティティを選択した手がかりを記入する．最後に，画面上部から適切なタグのボタンを押すことにより，左のペインのポップアップウィンドウで選択した範囲の文字列に対して，タグと具体的なエンティティが付与される．

4.7　アノテート時の留意点に関する検討

コーパスを作成する前に，アノテーションのガイドラインを明確にする必要がある．そこでガイドライン策定のために，2 名のアノテーターで独立に 200 件のツイートをアノテートし，アノテーター間でアノテーション結果が揺れる事例を分析した．その結果より，本研究におけるアノテート時の留意点を検討した．検討の結果を，付録 B に記述する．

図 4: タグ・エンティティ選択画面

4.8 マイクロブログ上のテキストを扱うにあたって，判明した問題

本研究では，既存研究で行われていた Leidner ら [13] によるニュース記事ドメインのテキストへのアノテート，Speriosu ら [11] による書籍ドメインのテキストへのアノテートと異なり，Twitter というマイクロブログ上のテキストへのアノテートを行う．ここで，4.7 節に記述したように 2 名のアノテーターが事前に 200 件のツイートをアノテートした際に，マイクロブログの性質によるいくつかの困難が見えてきた．本小節では，マイクロブログ上のテキストに含まれる場所参照表現をグラウンディングするにあたって，どのような固有の問題があるのかを述べる．

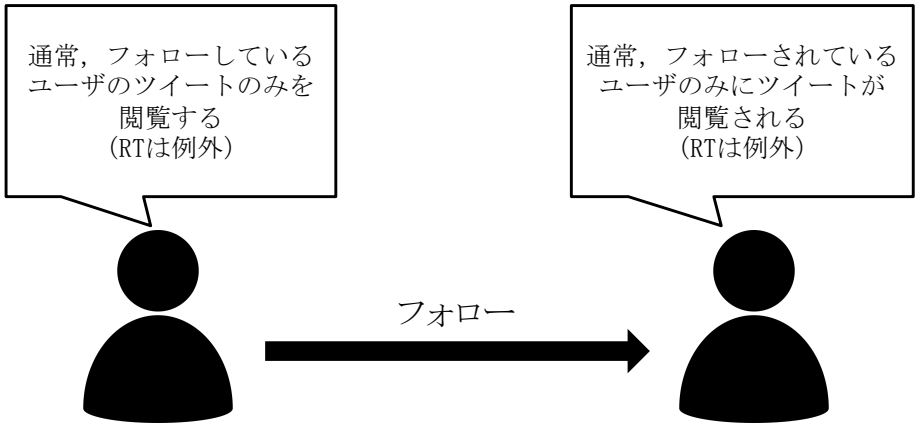

図 5: Twitter のフォローという概念

4.8.1 限定されたユーザへの情報発信

ニュース記事や書籍と異なり，マイクロブログ上では，著者が限られた読者を想定してテキストを記述することが多い．例えば本研究でアノテート対象としている Twitter ではフォローという概念（図 5）があり，フォローしているユーザがリツイート（他者のツイートの引用）をする場合は例外であるが，基本的に各ユーザは自らがフォローしているユーザのツイートのみを閲覧することとなる．このような背景があるため，ツイートを発信するユーザも，自らのツイートがフォロワー（自分をフォローしているユーザ）にのみ閲覧される，という想定で記述することがある．この現象は，ニュース記事や書籍のような，不特定多数に向けて記述されているテキストとの大きな違いを生み出している．

また，Twitter 上にはフォロー・フォロワーという概念に加えて，さらに限定的に特定のユーザに対して発信する，リプライ（返信）という概念もある．これは，ツイートの先頭に「@返信先のユーザ名」という記述をすることにより，フォロー・フォロワーという概念と関係なく，その特定のユーザに対してツイートを発信する，というものである．

以上のように限定されたユーザへの発信が行われる場合，発信者と受信者の間である背景知識が共有されているという前提で，場所参照表現が用いられる場合がある．この場合，アノテーターを含む第三者からはその場所参照表現が実際に指し示しているエンティティを判別できない，ということに繋がってしまう．

(7) @*** 学校 で待ってるからはやくよくなってね!!!

(7) の例は，ツイートの著者がある特定のユーザに向けてリプライ（返信）をしている．この例では,「学校」という場所参照表現はある具体的なエンティティを指し示していると考えられるが，ツイートの著者とリプライ（返信）先のユーザはそれを想起できると思われるものの，第三者から見て判断することはできない．本研究では 4.9 節に示すランダムサンプリングサブコーパス，フィルタードサブコーパスの 2 種類のコーパスを作成するが，この際にはリプライ（返信）をあらかじめ除去するという処理を加えている．

4.8.2　1 ツイートあたりの文字数制約

Twitter には，1 ツイートあたり最大 140 文字まで記述できる，という文字数の制約がある．これもまた，ニュース記事や書籍のテキストにはない特徴である．この制約は場所参照表現の記述にも影響を与える．

(8) 遅ればせながら明けました 2015。年越しは 東京 D でカウコンという名のマッチコンで年を越し、キンキさんの神々しさとかわいい後輩達に眼福し、光一さんのギリギリ派閥発言にうおおーとなってきました。ほんとに全員集まりたいという希望叶う日が来ますよーに。

(8) の例では，ツイートの著者は複数文から 1 ツイートを構成している．ここでは，その 140 文字という制限を超えないようにするための工夫か,「東京ドーム」を「東京 D」と省略して記述している．

また，このような文字数の制約があることから，複数ツイートに分けて記述を行う例もある．

(9) 映画「パシフィック・リム」も「ベイマックス」も日本じゃ永久に作れないんですよ。発生してくる文脈も背景も違いすぎるから。あれらは生まれる時から「世界」を相手にするために、世界中から才能を総動員し、世界規模の富をかき集め作られる…であるが故に、”元ネタ”それ自体は作れないという矛盾。

(10) （承前）何故なら、”元ネタ”＝完全なるオリジナル作品…というのは、つまるところは個人、たった一人の狂気にも等しい「執着」からしか生まれないから。最初から「世界」を相手に圧倒的に売り上げて投資を回収するために、そういう文脈の元で失敗を許されない作品とは、根本的に相反する存在なので。

この例では，(9) というツイートの直後に，(10) というツイートを発信することで，Twitter の 1 ツイートあたり 140 文字という制約を超えて，1 つの話題を発信している．(9)，(10) では「(承前)」という記述により (10) が直前に発信された (9) の続きであることを明示しているが，この記述の仕方はユーザによって異なり，何も記述せずに複数のツイートにより 1 つの話題を発信するユーザも多い．

このように複数のツイートにより 1 つの話題が発信される現象は，場所参照表現のグラウンディングにも影響を与える場合がある．

(11) 仙台駅 なう

(12) 今から 駅 の中の 本屋 向かう

(13) 本屋 でマンガ買ってきた

(11) から (13) が，連続したツイートとして発信されていたとする．ここで，(13) の「本屋」という場所参照表現は，著者は特定のエンティティを指しているものの，第三者からは (13) のテキストを見ただけでは特定することが不可能である．また，周辺のツイートとして (12) までを考慮に入れると,「本屋」が「駅」の中の「本屋」であると判断できるが，これでも「駅」が具体的にどのエンティティを指しているか特定できないため，不十分である．さらにツイートを遡って，(11) までを考慮することによって初めて，(13) の「本屋」が「仙台駅」の中の「本屋」である，という判断をすることが可能となる．

人間はこのように一連のツイートの流れを考慮して判断を行うが，これはコンピュータによって場所参照表現をグラウンディングする際にも不可欠である．本研究のアノテーションではフィルタードサブコーパス，ランダムサンプリングサブコーパスのどちらも，収集した各ツイートの周辺ツイートについては取得していない．今後コーパスの拡充を行う際には，ユーザ単位で直近最大数百ツイートを取得する，といった手法を考えている．そのうえで，ユーザごとに取得したツイート全体をひとつのドキュメントとみなし，アノテーションの際にはドキュメント全体を考慮することで，140 文字という制限に因むツイートあたりの情報量の少なさを克服し，より多くの場所参照表現にエンティティを付与できるのではないかと期待できる．

4.8.3 テキストの崩れた表記

マイクロブログ上のテキストは，ニュース記事や書籍のテキストに比べて崩れた表記が多く含まれることが知られている．

(14) フォロワー1900人いったよ wwwwwwwwwww うは wwwwwwwwwww テンションあがる wwwwwwwwwwww

(15) まぢで笑 またききますわぁー (￣▽￣)

(14) や (15) のように，顔文字が含まれるテキストや,「まぢで」といった崩れた表記が含まれるテキストは，自然言語処理を行うにあたって非常に大きな障壁となる．場所参照表現のグラウンディングを行う際にも，崩れた表記が問題となる場面がある．

(16) でぃずにー たのしーーーー

(17) わたしも とーきょー まいごになったわー

(16) の「でぃずにー」は「東京ディズニーランド」という施設名を，(17) の「とーきょー」は「東京」という地名をそれぞれ指していると思われるが，自然言語処理において形態素解析器として多くの研究で用いられる MeCab[7] [18] を用いても，形態素解析に失敗してしまう．今後，実際に場所参照表現のグラウンディングを行う際には，既存の自然言語処理ツールをどのように利用すれば本研究の目的に適しているか，という点に留意し，検討を行いたい．

4.8.4 BOT の存在

Twitter 上には，あらかじめ組み込まれたツイートを自動的に発信する，BOT という機能を持つクライアント（PC やスマートフォンなどから Twitter を利用するにあたって，Twitter の公式ウェブサイト以外から利用するためのクライアントソフトウェア）が存在する．

(18) 時刻は、 16 時 36 分 を過ぎました。

(19) 東京の現在 (12/02 05:15) の天気は Partly Cloudy(12.2℃) です.

(20) お昼ですお兄様！

[7] https://code.google.com/p/mecab/

BOT の種類は様々で，現在の時刻を発信する (18) のようなもの，天気予報を発信する (19) のようなもの，アニメのキャラクターのセリフを発信する (20) のようなものなどがある．5 節で記述したように本研究でコーパスを作成する際には，実際に人が発信しているツイートに限定するため，BOT のツイートを除去する処理を行っている．BOT ツイートの除去手法として，BOT 機能を持つクライアントを排除するためのブラックリストを作成する手法が考えられるが，事前調査の結果，BOT のクライアント名が自動生成されている事例が散見された．よって 5 節では，実際に人が発信していると判断されたツイートを元にして，BOT 機能を持たないクライアントからなるホワイトリストを作成し，BOT ツイートを除去している．

ただし，場所参照表現のグラウンディングを行うにあたって，BOT を除去する必要が必ずしもあるとは言えない．5 節ではアノテーションのコストを極力下げるために BOT を除去する処理をかけていたが，実際にグラウンディングを行う際，(19) のような天気予報ツイートが必要であるか不要であるかは，どのような応用目的で場所参照表現のグラウンディングを行っているのかに依存すると考えられる．

4.8.5　架空の場所参照表現

ニュース記事ドメインのテキストなどに見られないマイクロブログ上のテキスト固有の問題として，現実世界には存在しない，架空の場所参照表現が挙げられる．

(21) 国立魔法大学附属第一高校 に行きたかった

(22) サザエさん一家が 福岡 から 東京 に引っ越してきた

(21)，(22) は，架空の場所参照表現について言及している例である．ここで注意したいのは，(21) の場所参照表現「国立魔法大学付属第一高校」は現実世界に実在しない場所参照表現であることから地名・施設名辞書にマッチしないため特にこれといった対処をする必要がないと考えられるが，(22) の「福岡」と「東京」は，それぞれ現実世界にも存在する場所参照表現となっている，という点である．これらの場所参照表現をグラウンディングする必要があるか否かについてもタスク依存になると考えられるが，例えば情報抽出を行うにあたって，現実世界に即さない情報についてはノイズとなってしまう恐れがある．

4.9　アノテーション対象データ

アノテーション対象データとして，本研究では Twitter 上のツイートデータを扱う．本研究では以下の 2 種類の手法でアノテーション対象データを収集し，各々にアノテートすることとした．なお，各々のコーパスについて，実際に人が発信しているツイートに限定するために，BOT と思われるツイートの除去を行っている．

4.9.1　ランダムサンプリングサブコーパス

アノテーションを行う際に，バイアスをかけずにツイートを収集するためには，完全にランダムにツイートを抽出することが望ましい．これにより，ツイート全体における場所参照表現の各種分布を測ることが可能となる．そのため，本研究では純粋にツイートデータからランダムサンプリングすることによる，ランダムサンプリングサブコーパスをはじめに作成した．

4.9.2　フィルタードサブコーパス

単純にツイートをランダムサンプリングしてしまうだけでは，場所参照表現を含まないツイートが大量に抽出されてしまうという問題がある．アノテーションを行うにあたって，ツイートの著者とアノテーターに共通の知識が共有されていることが望ましいと考えられる．例えば，宮城県で主に生活する Twitter ユーザのツイート中には，宮城県内の話題が多く現れると考えられ，それに伴い同様に，ツイート内の場所参照表現もまた宮城県内のものが多くなるのではないかと想定される．その場合，アノテーターが宮城県在住者であるほうが，より正確にアノテートできるのではないかと期待される．そのような理由から，以下の条件によってフィルタリングを行った，フィルタードサブコーパスを作成した．はじめに，4.5 節で作成した「県・市区町村名・大字」辞書中のエントリが複数含まれるツイートをフィルタリングする．次に，それらのツイートのうち，アノテーターの在住都道府県に含まれる市区町村名を少なくとも一個以上含むツイートのみをフィルタリングする．これらのフィルタリング操作により，場所参照表現を含まないツイートの割合が大幅に減り，さらにアノテーターにとって比較的アノテートしやすい，土地勘を利用できるツイートを多く取得することができた．

5　コーパスに対するアノテーション

4 節では，コーパスに対してアノテーションを行うにあたってのガイドラインを策定した．本節では，策定したガイドラインに基づき実際にアノテーションを行った結果を報告する．

アノテーション対象データとして，4.9 節で述べた，ランダムサンプリングサブコーパスとフィルタードサブコーパスを作成した．ランダムサンプリングサブコーパスは，4.9.1 節で述べた手法により収集した 10,000 ツイートから構成される．また，フィルタードサブコーパスは，4.9.2 節で述べた手法により収集した 1,000 ツイートから構成される．ツイートデータ収集の対象としては，2014 年に投稿されたツイートを用いた．なお，アノテート作業者（アノテーター）は，2 名からなる．

本節でははじめに，2 名のアノテーター間のアノテーションの一致度合いを 5.1 節にて述べる．そのうえで，5.2 節にてフィルタードサブコーパスに対するアノテーション，5.3 節にてランダムサンプリングサブコーパスに対するアノテーションの結果をそれぞれ述べる．

5.1　アノテーションの一致度合い

コーパスの品質を測るために，本小節では 2 名のアノテーターによるアノテーションの一致度合いを測る．そのために，フィルタードサブコーパスの内から 200 ツイートをランダムに選択し，それらに 2 名のアノテーターが独立にアノテート作業を行った．なお本小節では，4.3 節で述べた全てのタグを用いてアノテート作業を行っている．

5.1.1　Mention Detection（言及抽出）

Mention Detection（言及抽出）タスクについての，アノテーター間のアノテーションの一致度合いを測る．はじめに，2 名のアノテーターによって 200 ツイートに付与されたアノテーションを，文字単位で IOB2 コーディングへ変換する．例として,「仙台駅に行く」というテキストに付与されるアノテーションを IOB2 コーディングへ変換すると以下のようになる．

仙 B-FAC
台 I-FAC
駅 I-FAC
に O
行 O
く O

O タグが付与されている文字は，エンティティを指しているというアノテーションが行われなかった文字である．そのうえで，2 名のうち一方のアノテーション結果を正解とみなして，もう一方のアノテーションの精度，再現率，F 値を測定する，という手法を用いた．この結果を表 3 に示す．IOB2 コーディングへ変換された 2 名のアノテーターのアノテーションを文字単位で比較した場合の Cohen's Kappa は 0.892 であった．また，2 名のアノテーター両者が O タグを付与した文字を除いて計算した Cohen's Kappa は 0.785 であった．

表 3 から，タグによって一致率に差異があることがわかる．LOC(地名) タグは非常に高い一致率となっている一方，FAC(施設名) タグ，ORG(組織名) タグの一致率はやや低い．これは，テキストの著者が具体的な場所を意識しているか否かの判断が難しい事例があり，アノテーターの判断に揺れが生じるためである．

(23) これでもう 大学図書館 から取り寄せてもらわなくていいのね

(23) の「大学図書館」は，テキストの著者が施設名として用いているか，組織名として用いているかを判断することが難しい．今後さらに大規模にアノテーションを行うにあたって，タグの判断についてのさらなる検討を行いたい．

5.1.2 Entity Resolution（エンティティ解決）

Entity Resolution（エンティティ解決）タスクについての，アノテーター間のアノテーションの一致度合いを測る．ここでは，アノテーター間で場所参照表現に付与したエンティティについて，双方の付与したエンティティの誤差距離に基づき議論を行う．

アノテーションの一致度合いを測る対象として用いている 200 ツイートに含まれる文字列のうち，2 名のアノテーターがどちらも地名・施設名辞書中のエンティティ情報を付与した文字列は 243 件であった．これらの文字列について，付与されたエンティティ情報に含まれる座標情報に基づき誤差距離を求めた．その結果，

表 3: 2 名のアノテーター間のタグの一致率

タグ	精度	再現率	$F_{\beta=1}$
LOC(地名)	90.16% (174/193)	96.67% (174/180)	93.30
FAC(施設名)	84.09% (74/ 88)	72.55% (74/102)	77.89
RAIL(鉄道路線名)	100.00% (9/ 9)	56.25% (9/ 16)	72.00
ROAD(道路名)	66.67% (2/ 3)	40.00% (2/ 5)	50.00
ORG(組織名)	84.75% (50/ 59)	81.97% (50/ 61)	83.33
GEN(総称表現)	50.00% (4/ 8)	57.14% (4/ 7)	53.33
AMB(クラスが曖昧)	16.67% (1/ 6)	100.00% (1/ 1)	28.57
FIC(架空の地名)	0.00% (0/ 1)	0.00% (0/ 0)	0.00
Overall	86.01% (504/586)	88.11% (504/572)	87.05

2 名のアノテーター間の誤差距離は平均 1,648 メートル，最大値 72,101 メートル，中央値 0 メートルとなっていた．これより，2 名のアノテーターが付与したエンティティの一致度が非常に高いことがわかる．誤差距離計測対象の 243 件中，199 件（81.9%）のエンティティはアノテーター双方で同一のエンティティを付与していた．

一方，一部の文字列に対してはアノテーター間で付与したエンティティの誤差距離が大きくなっていた．以下に，誤差距離が大きくなっていた文字列の例を示す．

(24) (誤差 70.8km) 江坂周辺、[淡路 A:LOC/兵庫県淡路市 B:FAC/淡路駅 (大阪市東淀川区)] 周辺、西中島南方周辺、新大阪周辺でバイト見つけたいよね、

(25) (誤差 68.9km) 原木シイタケのホダ木処分対象地域ってことは、まずは [福島 A:FAC/福島第一原子力発電所 B:LOC/福島県福島市] の事故で風評被害じゃないんだよ。

(26) (誤差 8.6km) もう [東京 A:LOC/東京都 B:FAC/東京駅] ついた、やっぱ新幹線は速いなぁ。

(24) の例では，アノテーター A は「江坂」,「西中島」,「新大阪」というテキスト内の各表現をそれぞれ地名であると解釈し,「淡路」についてもその周辺の地名である兵庫県淡路市であると判断している．一方アノテーター B は，それらの表現が駅名であると判断し,「淡路」についても駅名である，という推論を行ってい

た．(25) の例の「福島」については，アノテーター A は周辺文脈より「事故」が福島第一原子力発電所の事故を指している，という推論を経て付与を行っていたが，アノテーター B はそのような推論を行わず，字義通りの地名であると判断していた．(26) の例では,「東京」を字義通りの地名であると判断するか，新幹線の停車駅であると判断するか，という点でアノテーションに差異が生じた．

以上のように，2 名のアノテーターが付与したエンティティ同士の誤差距離が大きくなっていた例を見ると，各アノテーターの持つ背景知識や，各アノテーターが文脈からどの程度推論を行うか，といった要素によってアノテーションに揺らぎが生じる，ということがわかった．アノテーションの一致度合いを高めるためには，用いる背景知識の程度，推論の度合いについて，あらかじめ取り決める必要があるのではないかと考えられる．この点については，今後より大規模にアノテーションを行うにあたって，検討したい．

5.2 フィルタードサブコーパスに対するアノテーション結果

フィルタードサブコーパスに対するアノテーションについては，2 名のアノテーターで合計 1,000 ツイートに対して行った．アノテーションに用いたタグと，該当するエンティティの分布を表 4 に示す．

5.3 ランダムサンプリングサブコーパスに対するアノテーション結果

ランダムサンプリングサブコーパスに対するアノテーションについては，2 名のアノテーターで合計 10,000 ツイートに対して行った．アノテーションに用いたタグと，該当するエンティティの分布を表 5 に示す．なお，ランダムサンプリングサブコーパスに対するアノテーションの際には，4.3 節に示したタグのうち，ORG(組織名) タグ以外を用いている．これは，フィルタードサブコーパスに比べてランダムサンプリングサブコーパスのアノテート対象ツイート数が多く，ORG(組織) タグを用いないことにより，アノテーションコストの削減（アノテート作業に要する時間の短縮）が見込めたためである．

表 4: フィルタードサブコーパスに付与されたタグの分布．LOC(地名)，FAC(施設名) タグの集計中の括弧内は，(辞書中にアノテートすべきエンティティが存在せず，付与できなかった表現数/文脈から付与すべきエンティティが判断できなかった表現数/ひとつ以上のエンティティを付与することができた表現数) を表す．

タグが付与された表現数	フィルタードサブコーパス
LOC(地名)	977 (68/8/901)
FAC(施設名)	356 (51/19/286)
RAIL(鉄道路線名)	61
ROAD(道路名)	7
ORG(組織名)	208
GEN(総称表現)	32
FIC(架空の地名)	3
AMB(クラスが曖昧)	18
ツイート数	1000
総文字数	69806

5.4 エンティティを付与できなかった事例の考察

5.2 節と 5.3 節の結果より，施設名中で付与すべきエンティティが判断できなかった事例の割合が，フィルタードサブコーパスでは 356 件中 19 件（5.3%），ランダムサンプリングサブコーパスでは 517 件中 273 件（52.8%）と極端に異なることがわかる．本小節では，これらの事例について考察する．

(27) とりあえず サークル K すね

(28) まだ 会社 に着いていない

(27) の「サークル K」は施設名であると判断できるものの，店舗名でありエンティティの候補が膨大である．それにもかかわらず，文脈中に地名などの手がかりが出現していないため,「サークル K」という場所参照表現に具体的なエンティティを付与することはできなかった．(28) の「会社」は，普通名詞の場所参照表現である．店舗表現と同様，普通名詞の場所参照表現もまた候補となるエンティ

表 5: ランダムサンプリングサブコーパスに付与されたタグの分布．LOC(地名)，FAC(施設名) タグの集計中の括弧内は，(辞書中にアノテートすべきエンティティが存在せず，付与できなかった表現数/文脈から付与すべきエンティティが判断できなかった表現数/ひとつ以上のエンティティを付与することができた表現数) を表す．

タグが付与された表現数　ランダムサンプリングサブコーパス	
LOC(地名)	406 (14/94/298)
FAC(施設名)	517 (41/273/203)
RAIL(鉄道路線名)	25
ROAD(道路名)	3
GEN(総称表現)	65
FIC(架空の地名)	24
AMB(クラスが曖昧)	3
ツイート数	10000
総文字数	332739

ティの数が非常に多い．今回作成したコーパス中に出現した普通名詞の場所参照表現について，付録 A に示した．

これらのように，地名と比較して，施設名の場合には大量のエンティティが候補となる場合がある．特にチェーン店のような店舗名や，普通名詞の場合，手がかりが一切なければ判断は不可能となる．

6　エンティティの曖昧性解消に必要な手がかりの整理

アノテートしたコーパス中に出現する場所参照表現には，3 節にて述べたクラスの曖昧性，エンティティの曖昧性を持つものがある．本小節では，5 節で作成したコーパスを用いて，エンティティの曖昧性解消に必要な手がかりの整理を行う．なお，クラスの曖昧性（場所参照表現が地名であるか，施設名であるか）と，境界（テキスト中での場所参照表現の位置）は既知であると仮定する．

必要な手がかりの整理を行うにあたって，4.9.1 節で作成したランダムサンプリングサブコーパスを用いる．ランダムサンプリングサブコーパスには，ただ 1 つのエンティティを付与された場所参照表現が 436 件（地名が 267 件，施設名が 169 件）存在した．これらの 436 件の場所参照表現のエンティティの曖昧性解消を行うにあたってどのような手がかりが必要となるかを人手で調査した結果を，表 6 に示す．なお，表 6 では，曖昧性解消のために複数の手がかりが必要となる場所参照表現は複数回集計している．以下，場所参照表現の曖昧性解消に必要となる手がかりと，コーパス中でその手がかりを必要とした場所参照表現について記述する．

6.1　場所参照表現の表層にマッチするエンティティが一つのみであり，エンティティの曖昧性がない

(29) 北浦和に置き去りにされる仕事

(30) 北海道当たれば行ってみようかな〜

(31) NHK スタジオパークにたくさん台あったので思わずコンプリート

(29) から (31) の場所参照表現「北浦和」,「北海道」,「NHK スタジオパーク」は，それぞれマッチする地名・施設名辞書中のエンティティが一つのみとなっているため，エンティティ曖昧性を解消する必要がない．

6.2　テキスト中の他の地名

(32) 松島堪能の 1 日目でした。明日は呉に南下します

(33) 釧路の市場で買ってきた鮭の燻製をはみはみしながら、旅行中に撮ったビデオ見てる。

(32) の「松島」，(33) の「市場」はどちらも，マッチするエンティティが地名・施設名辞書中に複数存在するためエンティティの曖昧性があるが，それぞれテキスト中の他の地名を手がかりとすることで，一つのエンティティを選択することができる．例えば (33) の例ではテキスト中の「釧路」という地名に曖昧性がないことから,「市場」が釧路和商市場というエンティティを指している，という判断ができる．

6.3 テキスト中の他の施設名

(34) お台場 デート w 10 年ぶりの ジョイポリス 楽しかったー

(35) 渋谷 の ヒカリエ にて発見

(34) の「ジョイポリス」は「お台場」という施設名から，(35) の「渋谷」は「ヒカリエ」という施設名から，それぞれエンティティの曖昧性を解消することができる．

6.4 人口情報

(36) 今日はこんと 恵比寿 出勤やぁ

(37) パソコン、カチャカチャするお仕事、三宮 で探してる！

地名については，Ladra ら [10] の手法のように辞書中に人口の情報が含まれれば，それを手がかりにできる場合がある．(36) の「恵比寿」，(37) の「三宮」はそれぞれマッチするエンティティが辞書中に複数存在するが，最も人口の多いエンティティを選択することで曖昧性を解消することができる．ただし，施設名の場合には人口という概念が存在しないため，施設名の曖昧性を解消する際にはこの手がかりは利用できない．

6.5 場所参照表現の表層と辞書中のエンティティの表層の表記揺れ

(38) 観覧車と ランドマークタワー みるたび横浜ーってなる

(39) リアルに ハマスタ の声援が凄い

(40) 都 内誰かいないんですか

(38) の「ランドマークタワー」，(39) の「ハマスタ」，(40) の「都」は，それぞれ「横浜ランドマークタワー」,「横浜スタジアム」,「東京都」という辞書中のエンティティと対応付けられた．しかしながら，これらは場所参照表現の表層と辞書中のエンティティの表層に表記揺れがあり，単純にマッチさせることはできない．これに当てはまる事例は，地名の場合は 267 件中 3 件（1.1%）であるのに対して，施設名の場合は 169 件中 76 件（45.0%）と非常に多い．

6.6 プロフィール情報

(41) 学校 行かなくていい気がしてきた

(42) こんにちは！今日は大九州物産展から大分「ぶんごや」の人気 No.1、そして第 5 回からあげグランプリ金賞受賞のからあげをご紹介！外はカリっと中はジューシー、あふれる旨さをご堪能くださいませ。【本館6 階・催し会場にて。9/29(月) まで】

(41) の場所参照表現「学校」は，テキストからは具体的なエンティティを判断することができないが，Twitter のプロフィール欄にユーザの所属する学校名が記述されていたことから，具体的なエンティティを付与することが可能であった．また (42) の場所参照表現「本館」については，(42) を投稿したユーザがある百貨店のアカウントであることから，判断が可能となっていた．

6.7 背景知識

(43) 両国 のチケットがどんどん無くなる中、何度誘っても最後まで来ると言わなかった家族

(44) 秋山さんが 福岡 で舞ってくれてよかった！

(43) の場所参照表現「両国」は，アノテーターが「両国国技館はライブの会場としてよく用いられる」という背景知識を持っていたため,「両国国技館」という具体的なエンティティを付与することができた．また，(44) の「福岡」は,「福岡ソフトバンクホークスが優勝した」,「福岡ソフトバンクホークスの監督は秋山幸二である」,「福岡ソフトバンクホークスの本拠地は福岡ヤフオク!ドームである」という複数の背景知識から,「福岡ヤフオク!ドーム」という具体的なエンティティであると判断された．このように，場所参照表現のエンティティ曖昧性解消のた

めに複雑な背景知識を要する場合，アノテーター間でもその背景知識の有無によりアノテートが大きく揺れることを考慮すると，グラウンディングを行うことは非常に難しいのではないかと考えられる．

6.8 添付された画像

(45) 久々の 田中そば店 で久々の#めーん部 【画像添付】

(46) 京セラ なうなう〜 【画像添付】

(45) と (46) には，各々画像が添付されていた．(45) の場所参照表現「田中そば店」は，添付された画像中の容器から，エンティティの曖昧性解消を行うことができた．また (46) の場所参照表現「京セラ」は，添付された画像にドーム型の野球場が見えたため,「京セラドーム大阪」という具体的なエンティティが付与された．

6.9 添付された URL

(47) 久々に来た。 本当ならこの 店 は… http://***

(47) の場所参照表現「店」は，テキストからは具体的なエンティティを判断することができないが，併記された URL のリンク先に具体的な店舗名が記述されていたため，曖昧性を解消できた．

表 6: ランダムサンプリングサブコーパスに含まれる場所参照表現のエンティティの曖昧性解消を行うにあたって必要となる手がかりの分布

手がかり	該当数（地名）	該当数（施設名）	合計
(1) 場所参照表現の表層にマッチするエンティティが一つのみであり，エンティティの曖昧性がない	118(44.2%)	123(72.8%)	241(55.3%)
(2) テキスト中の他の地名	6(2.2%)	7(4.1%)	13(3.0%)
(3) テキスト中の他の施設名	4(1.5%)	6(3.6%)	10(2.3%)
(4) 人口情報	140(52.4%)	2(1.2%)	142(32.6%)
(5) 場所参照表現の表層と辞書中のエンティティの表層の表記揺れ	3(1.1%)	76(45.0%)	79(18.1%)
(6) プロフィール情報	0(0.0%)	3(1.8%)	3(0.7%)
(7) 背景知識	1(0.4%)	7(4.1%)	8(1.8%)
(8) 照応関係にある場所参照表現	1(0.4%)	1(0.6%)	2(0.5%)
(9) 添付された画像	1(0.4%)	3(1.8%)	4(0.9%)
(10) 添付された URL	0(0.0%)	1(0.6%)	1(0.2%)

7　既存のエンティティ曖昧性解消手法に基づく評価

本節では，5 節で作成したコーパスに対して既存のエンティティ曖昧性解消手法を適用し，地名のみに対する性能と施設名まで含めた場合の性能を比較する．ここで，施設名に関する問題を簡潔に明らかにするために既存手法の中でも比較的単純な手法を用いることが望ましい．本節では，Speriosu [19] の POPULATION，MINDIST という手法に加え，それらを組み合わせた POPULATION+MINDIST という手法を適用する．

7.1　POPULATION

本小節では，本節で適用する既存手法の POPULATION について説明する．この手法は，テキスト（ドキュメント）中に含まれる場所参照表現のそれぞれの候補エンティティの人口情報を参照することで，最も人口の多い候補エンティティを最適なエンティティとする．ここで，施設名には人口情報が定義されていない

ため，本手法は地名のみを対象とする．また，候補エンティティに人口情報の定義されているエンティティがひとつも含まれないような地名を評価の対象外とする評価方法と，そのような地名については候補エンティティからランダムに選択するという評価方法の 2 通りで評価した．

7.2 MINDIST

本小節では，本節で適用する既存手法の MINDIST について説明する．この手法は，テキスト（ドキュメント）中に含まれる場所参照表現のそれぞれの候補エンティティ同士の距離を算出することで，距離の総和が最小になるような候補を最適なエンティティとする．MINDIST の擬似コードを Algorithm 1 に示す．

7.3 POPULATION+MINDIST

Speriosu [19] で提案されていた POPULATION，MINDIST に加え，本節では POPULATION+MINDIST という手法もまた適用する．この手法は，地名のうち POPULATION の対象となるものを先に POPULATION で一意に定め，POPULATION で定められなかった地名，施設名に対して MINDIST を適用する，という手法である．この手法を用いる理由は，単純に MINDIST を適用するのではなく人口情報で一部の地名を先に当てることにより，施設名に対する予測精度も高められるのではないかと考えられたためである．

7.4 場所参照表現の候補エンティティの選択

7.2 節で述べた MINDIST を適用するにあたり，5 節で人手で付与したエンティティの他に，各場所参照表現に対して付与されるべき具体的なエンティティの候補を列挙する必要がある．本節では，場所参照表現としてアノテートされた文字列により 4.5 節で作成した地名・施設名辞書に対して部分一致検索を行い，その結果マッチしたエンティティをその場所参照表現の候補エンティティとする．なお，テキスト中のどの範囲が場所参照表現であるかは，アノテーションに基づいて与える．

Algorithm 1 MINDIST

```
for テキスト ∈ コーパス do
  for 場所参照表現 i ∈ テキスト do
    overallmin ← ∞
    for 候補エンティティ a ∈ 場所参照表現 i の候補エンティティ集合 do
      totaldist ← 0
      for 場所参照表現 j ∈ テキスト do
        if 場所参照表現 i ≠ 場所参照表現 j then
          min ← ∞
          for 候補エンティティ b ∈ 場所参照表現 j の候補エンティティ集合
          do
            dist ← distance( 候補エンティティ a, 候補エンティティ b)
            if dist < min then
              min ← dist
            end if
          end for
          totaldist ← totaldist + min
        end if
      end for
      if totaldist < overallmin then
        overallmin ← totaldist
        推定候補 ← 候補エンティティ a
      end if
    end for
    場所参照表現 i のシステム出力エンティティ ← 推定候補
  end for
end for
```

7.5　評価対象

本節での評価対象は，5 節で作成したランダムサンプリングサブコーパスとする．ここで，7.2 節で述べたように本節で用いる手法の MINDIST はテキスト中に複数の場所参照表現が含まれることを前提としているため，ランダムサンプリングサブコーパスのうち，エンティティを付与された場所参照表現が複数含まれるテキストのみを対象とする．この制約を満たすテキストはランダムサンプリングサブコーパスの 10,000 テキスト中の 64 テキストであり，場所参照表現は 151 個含まれている．なお，MINDIST を適用する際にはテキスト中の場所参照表現の二つ組みそれぞれの，全ての候補エンティティ同士の距離を求める必要があるが，この際に二つ組みのそれぞれの候補エンティティ数が数万件となっている場合には数億回の距離演算が必要になるなど，実行時間が膨大になってしまうケースがある．このため，既存手法を適用した際に実行時間が 1 分以上であったテキストは，評価の対象外とした．

7.6　評価指標

評価指標として，Speriosu ら [19] と同様のものを用いる．まず，手法により選択した候補エンティティと人手により付与されたエンティティが完全に一致している場合に正解として，精度の評価を行う．ただし，この評価指標では人手により付与されたエンティティと異なるが近距離の候補エンティティを選択した場合と，遠距離の候補エンティティを選択した場合のどちらであっても，ただ単に不正解としてしまうという問題点がある．そのため，手法により選択した候補エンティティと人手により付与されたエンティティの誤差距離を求め，その平均値，中央値についてもまた，評価指標として取り入れる．加えて，誤差距離が 161km（約 100mile）に正解とする，A_{161} という評価指標を用いる．

7.7　評価結果

5 節で作成したランダムサンプリングサブコーパスに POPULATION，MINDIST，POPULATION+MINDIST を適用した結果を表 7 に示す．

表 7: ランダムサンプリングサブコーパスに対する POPULATION，MINDIST，POPULATION+MINDIST の評価

	地名				施設名				合計			
	平均値	中央値	精度	A_{161}	平均値	中央値	精度	A_{161}	平均値	中央値	精度	A_{161}
POPULATION (人口定義なしは対象外)	25.7	0.0	88.3	94.8	—	—	—	—	25.7	0.0	88.3	94.8
POPULATION (人口定義なしはランダム選択)	187.7	0.0	77.8	87.9	—	—	—	—	187.7	0.0	77.8	87.9
MINDIST	187.2	2.9	35.4	69.7	44.6	0.2	23.1	88.5	138.2	0.9	31.1	76.2
POPULATION + MINDIST	65.0	0.0	78.8	87.9	30.5	0.2	23.1	90.4	53.1	0.0	59.6	88.7

7.8 考察

表 7 より，POPULATION，MINDIST，POPULATION+MINDIST の各手法をランダムサンプリングサブコーパスに適用した結果の考察を行う．

まず，POPULATION を適用したとき，地名については非常に高い精度で当てられていることがわかった．これは既存手法と同様の結果となっており，本研究で作成したランダムサンプリングサブコーパスでも地名に関しては人口情報を利用することが非常に有力な手がかりになるということが確認された．

次に，MINDIST を単純に適用した場合，地名に比べて施設名の精度が低くなっていることがわかる．これは，地名辞書と施設名辞書のサイズの違いが理由であると考えられる．4.5 節にて述べたように，辞書中のエントリ数は，施設名が地名の 30 倍以上と非常に多くなっている．したがって，ひとつの場所参照表現に対する候補エンティティの数も地名に比べて施設名のほうが多くなる傾向にある．これにより，地名に比べて施設名の精度が低くなっているのであると推察される．一方，誤差距離の平均値，中央値，A_{161} は各々地名に比べて施設名に対する性能のほうが良好であった．これもまた，地名辞書と施設名辞書のサイズの違いが理由であると思われる．施設名辞書のエントリ数が多いということは，それだけ座標も密集しているということになる．このため，単純に各場所参照表現の候補エ

ンティティ同士の距離によりエンティティを選択する MINDIST を適用すると，正解エンティティと近い位置にはあるが異なるエンティティを選択してしまうという傾向にあるのだと考えられる．

また，POPULATION+MINDIST では，地名については誤差距離，精度，A_{161} ともに大幅に向上している．一方，人口が最大の地名を MINDIST の際に用いているにもかかわらず，施設名については人口情報を使わなかった場合に比べて性能の向上が小さいということがわかる．これは，たとえ地名について高精度で当てることができていたとしても，結局のところ密集している施設の中から適切な候補を選ぶという必要があるため，単純に MINDIST を適用した場合と同様に，正解エンティティと近い位置にはあるが異なるエンティティを選択してしまうという問題が生じてしまうためだと思われる．

既存手法で不正解となった例を示す．

(48) 帰りに 横浜 の ヨドバシ でも寄って？レンズでも買って行こうかなと考えていたのだけれど…。

(48) の例では,「横浜」という場所参照表現に「横浜市」という正解のエンティティを付与できていたが,「ヨドバシ」という場所参照表現には「ヨドバシカメラ マルチメディア横浜」という正解のエンティティではなく,「お忍び 桜個室めぐり 美の邸 (vino - tei) 横浜ヨドバシ相鉄駅前店」という誤ったエンティティを付与してしまっていた．これは,「お忍び 桜個室めぐり 美の邸 (vino - tei) 横浜ヨドバシ相鉄駅前店」が「ヨドバシカメラ マルチメディア横浜」と距離的に近いエンティティであり，なおかつ「ヨドバシ」という文字列で部分一致検索されてしまうためである．

8　クラウドソーシングサービスを利用したアノテーションに向けて

5 節で行ったアノテーションをより大規模にするための一手段として，クラウドソーシングサービスを利用することが考えられる．日本国内のクラウドソーシングサービスとして，クラウドワークス [8] やランサーズ [9]，Yahoo!クラウドソーシング [10] などがある．これらのサービスでは，国内の不特定多数の作業者に安価に大量の作業を依頼することが可能である．しかしながら，クラウドソーシングサービスを利用する際には，それらのサービスの制約や性質を理解したうえで依頼しなければならない．そこで本節では，クラウドソーシングサービスを利用するにあたり考慮すべき点を考察し，具体的な方法を検討する．

8.1　クラウドソーシングサービスを利用するにあたって考慮すべき点

8.1.1　作業時のユーザインタフェースの制約

クラウドソーシングサービスでは，作業者が作業を行う際のユーザインタフェースに制約がある．このため，4.6 節のアノテーションツールをそのまま作業者に提示することができない．よって，4 節で述べた Mention Detection（言及抽出）タスク，Entity Resolution（エンティティ解決）タスクのそれぞれを，クラウドソーシングサービスに適した形式で作業者に提示する必要がある．

8.1.2　不特定多数の作業者への作業の分配

5 節で行ったアノテーションでは作業者が限られていたが，クラウドソーシングサービスを利用する場合には作業者は不特定多数となる．この際，4.9.2 節で述べたように，アノテーション付与対象のテキストの著者と作業者に共通の知識が共有されていることで，作業に必要な時間が抑えられ，効率的な作業を行えるものと想定されるため，可能な限りアノテーション付与対象のテキストを適切な作業者に分配するようにできれば望ましい．

[8] http://crowdworks.jp/
[9] http://www.lancers.jp/
[10] http://crowdsourcing.yahoo.co.jp/

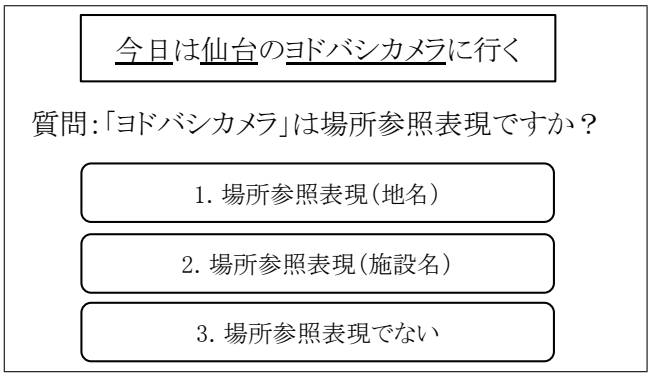

図 6: クラウドソーシングサービス上での Mention Detection タスク

8.2 具体的な方法の検討

8.2.1 ユーザインタフェースの検討

4 節で述べた Mention Detection（言及抽出）タスク，Entity Resolution（エンティティ解決）タスクをクラウドソーシングサービスに適した形式で提示するための検討を行う．

はじめに，Mention Detection タスクをクラウドソーシングサービスに適した形式に変換することを考える．ここで問題となるのが，4.6 節でタグ・エンティティを付与する対象の文字列範囲をドラッグで選択する部分である．クラウドソーシングサービスでタスクを提示する際に，ユーザインタフェースの制約上，テキスト中の文字列をドラッグで選択するという操作を作業者に行わせることができない．このため，図 6 のような提示方法を考えている．タスクを提示するにあたって，はじめに元のテキストに対して形態素解析をかける．そのうえで，品詞を名詞と解析された形態素，またその連接の全てについて，作業者に対して「[名詞（名詞連接）] は場所参照表現ですか？」という問いかけを行い，作業者は場所参照表現であるか否かを回答する．この際に作業者は，場所参照表現である場合には 4.3 節で説明したタグの中から，その場所参照表現に適したタグを選択する．このような提示方法により，Mention Detection タスクについては，4.6 節で述べた作業と同一の作業内容をクラウドソーシングサービス上で実現できると考えている．

次に，Entity Resolution タスクをクラウドソーシングサービスに適した形式に変換することを考える．ここでは，4.6 節で述べた地名・施設名辞書からの検索が問題となる．4.6 節で作成したツールであれば，場所参照表現として選択し

た文字列を用いて地名・施設名辞書から検索した結果をウィンドウ内の別ペインに表示することができ，これによりユーザとアノテーションツールの間で対話的にやり取りすることが可能であった．しかしながら，クラウドソーシングサービスを利用する場合，そのような対話的な検索機能を取り入れることが仕様上不可能である．このため，図 7 のような提示方法を考えている．はじめに，Mention Detection タスクで「地名」あるいは「施設名」のタグが付与された各文字列について，ElasticSearch により事前に候補エンティティを作成する．そのうえで，候補エンティティの上位数件から数十件を候補エンティティとして提示し,「この地名（施設名）に適切なエンティティは以下のどれですか？」という問いかけを行う．ここで，この方法では場所参照表現に対応する具体的なエンティティが提示した候補エンティティ内に含まれないという場合もあり得るため,「その他」という選択肢を設け，該当するエンティティの名称・住所をウェブ検索などにより作業者が別途付与する，という対処が必要になる．このような提示方法により，Entity Resolution タスクについてもクラウドソーシングサービス上で実現できると考えている．

8.2.2 作業分配方法の検討

アノテーション付与対象のテキストをクラウドソーシングサービス上の適切な作業者に分配するための手法を検討する．8.1.2 節で述べた「アノテーション付与対象のテキストの著者と作業者に共通の知識が共有されている」場合に適切な作業者に分配されているとみなすと，アノテーション付与対象のテキストの著者と作業者の居住地が近いほど好ましいと考えられる．ここで，作業前に事前に居住地を質問することにより作業者の居住地を知ることはできるが，アノテーションが付与される前のテキストには著者の居住地が明示的には記されていない．そのため，前もってアノテーション付与対象のテキストの著者の居住地を推定する必要がある．

この際，手法のひとつとしてツイートに付与されたジオコードを利用することが考えられる．アノテーション付与対象をジオコードが付与されているツイートに限定することで，そのジオコードの指す地点に居住地が近い作業者に分配することが可能となる．ただし，2 節で述べたように，ツイート全体のうちジオコードが付与されているツイートは 1%にも満たない．そこで他の手法として，ツイートの著者のプロフィールを利用することが考えられる．Twitter ユーザのプロフィール欄には,「現在地」という項目があり，各ユーザは自らの居住地を記述すること

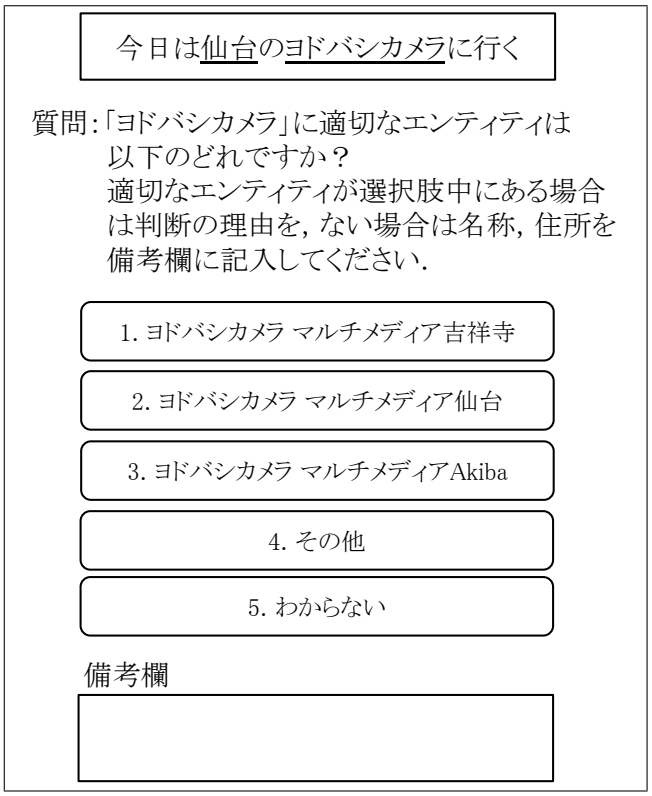

図 7: クラウドソーシングサービス上での Entity Resolution タスク

ができる．Hecht ら [20] は,「現在地」欄に適切な地名情報を記入しているユーザは全体の 66%であるという調査結果を示している．よって，ジオコードが付与されているツイートを用いなくとも,「現在地」欄に基づき著者の居住地を推定することが可能である．また，地域ごとに特徴的な単語を分析することにより Twitter ユーザの居住地を推定する Cheng ら [5] の手法を用いることも有用ではないかと考えられる．

9　まとめ

本論文では，ツイートデータに対して地名・施設名を含む場所参照表現へのタグ・具体的なエンティティの付与をしたコーパスを作成した．

具体的には，3節，4節にてアノテーションのためのガイドラインの設計について議論した．この中で，Mention Detection（言及抽出）と Entity Resolution（エンティティ解決）という2種類のタスクにアノテーションの工程を分割することで，各工程の単純化が実現し，また各工程でのエラー要因を容易に確認できるという，理想的なアノテーションの枠組みを構築できた．また，4.8 節にて議論したように，アノテートの際にマイクロブログ上のテキスト特有の問題があることがわかった．

5節では，4節で策定したガイドラインに基づき，実際にアノテーションを行った．ここで2名のアノテーターによるアノテーションの一致度合いを測った結果，Mention Detection（言及抽出）と Entity Resolution（エンティティ解決）のそれぞれのタスクで，アノテーター間で高い一致率となっており，策定したガイドラインがアノテーション作業に適切なものとなっていることが確認された．その一方で，各アノテーターの持つ背景知識や，各アノテーターが文脈からどの程度推論を行うか，といった要素によってアノテーションに揺らぎが生じるということもわかった．

6節では，エンティティの曖昧性解消を行うにあたってどのような手がかりが必要となるかを調査した．その結果，地名に関しては人口情報を用いることでエンティティの曖昧性を解消できる事例が全体の半数程度と大きな割合を占めることがわかった．また，施設名に関しては，略称のような場所参照表現の表層と辞書中のエンティティの表層の表記揺れが大きな問題となっていた．

7節では，既存のグラウンディング手法を本研究で作成したコーパスに適用し，地名のグラウンディングと施設名のグラウンディングの評価実験を行った．この結果，地名辞書と施設名辞書のサイズの違いにより，施設名のほうが比較的低い精度となることが確認された．これは，たとえ地名を人口情報により高精度で当てた場合であっても変わらず，施設名のグラウンディングの難しさが明白になった．

8節では，クラウドソーシングサービスを利用して大規模にアノテートするにあたって，どのような点を考慮しなければならず，具体的にどのような方法をとれるか，ということを検討した．そこで，クラウドソーシングサービスでは 4.6 節で述べたようなツールを利用することができないため，クラウドソーシングに適した提示方法を考案した．

最後に，今後の課題をまとめる．まず，本研究ではアノテート対象のツイートをツイート単位でランダムに取得していたが，実際にはユーザ単位でツイート集合を取得する必要があると考えられる．これにより，ツイートあたりの 140 文字という制限に因む情報量の少なさを克服し，より多くの場所参照表現にエンティティを付与できるのではないかと期待できる．

次に，どの程度の背景知識を用い，またどの程度の推論を経てアノテートを行うのかを検討する必要がある．クラウドソーシングサービスを利用して大規模にアノテートすることで，異なる背景知識を持つ不特定多数のアノテーターのアノテーションを収集することができると考えられる．それらのデータを利用して，アノテーションガイドラインのさらなる洗練に取り組みたい．

加えて，使用する辞書の拡充にも併せて取り組む予定である．本研究では地名・施設名へのエンティティ付与を行っていたが，鉄道路線名・道路名に対しては行っていなかった．実際にはエンティティの曖昧性を解消する際に鉄道路線名・道路名を利用することも考えられるため，これらの辞書を追加する．また，場所参照表現の表層と辞書中のエンティティの表層の表記揺れへの対応として，Wikipedia 等の外部リソースを利用した愛称・略称の獲得にも取り組みたい．

謝辞

本研究を進めるにあたり，ご多忙の中ご指導いただきました主指導教員の乾健太郎教授に深く感謝します．研究内容について多くのご助言をいただきました岡崎直観准教授に深く感謝します．審査委員を引き受けていただきました，篠原歩教授，徳山豪教授に深く感謝します．本研究の多くの部分を共同で研究していただき，本論文を執筆するにあたり相談に応じていただいた松田耕史研究員に深く感謝します．研究室生活の多くの場面でお世話になりました，八巻智子秘書に深く感謝します．

最後になりますが，本研究を行うにあたって多数のご意見，ご指摘をいただきました研究室の皆様に深く感謝します．

参考文献

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

[2] Alexei Pyalling, Michael Maslov, and Pavel Braslavski. Automatic geotagging of russian web sites. In *Proceedings of the 15th international conference on World Wide Web*, pp. 965–966. ACM, 2006.

[3] Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 484–491. ACM, 2009.

[4] Michael D Lieberman, Hanan Samet, and Jagan Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pp. 201–212. IEEE, 2010.

[5] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768. ACM, 2010.

[6] Benjamin P Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 955–964. Association for Computational Linguistics, 2011.

[7] Benjamin Wing and Jason Baldridge. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 336–348, Doha, Qatar, October 2014. Association for Computational Linguistics.

[8] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical*

Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1500–1510. Association for Computational Linguistics, 2012.

[9] David A Smith and Gregory Crane. Disambiguating geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries*, pp. 127–136. Springer, 2001.

[10] Susana Ladra, Miguel R Luaces, Oscar Pedreira, and Diego Seco. A toponym resolution service following the ogc wps standard. In *Web and Wireless Geographical Information Systems*, pp. 75–85. Springer, 2008.

[11] Michael Speriosu and Jason Baldridge. Text-driven toponym resolution using indirect supervision. In *ACL (1)*, pp. 1466–1476, 2013.

[12] Sharon Myrtle Paradesi. Geotagging tweets using their content. In *FLAIRS Conference*, 2011.

[13] Jochen L Leidner. An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, Vol. 30, No. 4, pp. 400–417, 2006.

[14] Gregory Crane. The perseus digital library. 2000. `http://www.perseus.tufts.edu/hopper/`.

[15] Satoshi Sekine and Yoshio Eriguchi. Japanese named entity extraction evaluation: analysis of results. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pp. 1106–1110. Association for Computational Linguistics, 2000.

[16] 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.

[17] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *LREC*, 2002.

[18] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *EMNLP*, Vol. 4, pp. 230–237, 2004.

[19] Michael Adrian Speriosu. Methods and applications of text-driven toponym resolution with indirect supervision. 2013.

[20] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 237–246. ACM, 2011.

付録

A　コーパス中に出現した普通名詞の場所参照表現

店，本屋，公演，会社，稽古場，部屋，ゲストハウス，電気屋，会場，カラオケ，コンビニ，大学，家，アパート，スーパー，ネカフェ，学校，温泉，SA，お化け屋敷，カレー屋，ショップ，マンション，映画館，駅，空港，研究室，研究所，市役所，歯医者，事務所，耳鼻科，自宅，実家，小児科，図書館，美容院，病院，満喫，幼稚園，寮，銀行

B　アノテートの際の留意点

B.0.3　地名・施設名辞書中に付与すべきエンティティが見つからない場合の対処

(49) [練馬区 LOC/東京都_練馬区] に [インコカフェ FAC/NULL (NOTE=辞書に存在しない)] あったのしらなかった

(50) [紀伊半島 LOC/NULL (NOTE=辞書に存在しない)] 以西のどこかを通ること必至

(49) の「インコカフェ」という表現は，周辺文脈より東京都練馬区に存在するインコカフェ（施設）であることを考慮すると,「FUKUROKOJI cafe」というエンティティであると判断できる．しかしながらこのエンティティは地名・施設名辞書中にしないことから，対応付けることができない．また，(50) の「紀伊半島」という地名についても，今回用いた地名・施設名辞書中には含まれない表現であった．本研究では，このような場合には備考欄 (NOTE) に「辞書に存在しない」という旨を記述することとする．

B.0.4　日本国外の地名・施設名への対処

(51) 今日は [アメリカ LOC/NULL (NOTE=日本国外)] 最後の夜

(52) [アンコールワット FAC/NULL (NOTE=日本国外)] 行ってみたい

(51) の「アメリカ」は日本国外の地名であり，(52) の「アンコールワット」は日本国外の施設名である．今回用いている地名・施設名辞書は日本国内のエントリのみから構成されていることから，これらの日本国外の地名・施設名は含まれない．そのため，備考欄に日本国外のエンティティである旨を記述することとする．

B.0.5　現存しない場所参照表現への対処

(53) [薩摩藩 LOC/NULL (NOTE=現存しない)] の盟友、大久保利通や [長州藩 LOC/NULL (NOTE=現存しない)] の木戸孝允と並び、「維新の三傑」と称される。

(53) の「薩摩藩」と「長州藩」はどちらも現存しない地名となっている．このような場所参照表現は地名・施設名辞書に含まれないため，備考欄に現存しない旨を記述することとする．

B.0.6　特定不能な表現の取り扱い

(54) [コンビニ FAC/NULL (NOTE=特定不能)] にドライトマト売ってて良かった

(55) 英語の課題 [学校 FAC/NULL (NOTE=特定不能)] に置いてきた気がする

(56) [川内 LOC/NULL (NOTE=特定不能)] に着きました

(54)，(55) の「コンビニ」「学校」は文脈から特定不能な施設名であり，(56) の「川内」は文脈から特定不能な地名である．このような場所参照表現については，備考欄に特定不能である旨を記述することとする．

B.0.7　接尾辞を含む表現の取り扱い

(57) 高校サッカー選手権大会の [宮城県 LOC/宮城県] 代表

(58) [佐賀県 LOC/佐賀県] 知事選挙（1 月 11 日投・開票）

(57) の「宮城県代表」，(58) の「佐賀県知事」はそれぞれ,「宮城県」「佐賀県」という場所参照表現に接尾辞として「代表」「知事」という接尾辞が付随しているものと考えられる．このように，場所参照表現に接尾辞が付随する場合，接尾辞より前の部分文字列にタグ，エンティティを付与することとする．

B.0.8　組織名と施設名の区分

(59) [ローソン ORG] キャンペーン詳細発表

(60) 来週は朝イチで [ローソン FAC/NULL (NOTE=特定不能)] 行くわ

(59) の「ローソン」は組織 (企業) 名として用いられていると判断できるため，施設名タグは付与しない．一方，(60) の「ローソン」は「場所としての側面」に着目していると判断できるため，施設名タグを付与する．このように，全く同じ文字列をアノテートする場合であっても，文脈に依存していずれのタグを付与するかが分かれる場合がある．

B.0.9　省略された表現の取り扱い

(61) いま [東京 D FAC/東京ドーム] なんですか？

(62) クリスマスは [TDL FAC/東京ディズニーランド] に行ったほうが運気がいい人

(61) の「東京 D」は「東京ドーム」，(62) の「TDL」は「東京ディズニーランド」のそれぞれ略称であると判断できる．このように，テキスト中の略称から具体的なエンティティを想起できる場合，そのエンティティを付与することとする．

B.0.10　話者が誤って記述したと思われる場所参照表現の取り扱い

(63) [仙台液 FAC/仙台駅] を過ぎると急に雪景色

(63) の「仙台液」は，話者が「仙台駅」を誤って表記してしまったものだと判断できる．このように，テキスト中の表現にスペルミス，語表記が含まれていた場合，もしも元の表現が推測できるのであれば，推測したエンティティを付与することとする．

B.0.11　イベント表現の取り扱い

(64) コミケ離脱！

(65) M3 は日程が合わなそうだな…残念ながら。

(64) の「コミケ」，(65) の「M3」はそれぞれ，イベント名を指していると判断できる．ここで，(64) の「コミケ」は「東京国際展示場」で開催されるイベントであるが，今回のアノテーションでは，一時的に開かれているイベントについてはアノテーション対象外とし，タグも具体的なエンティティも付与しないこととする．

B.0.12　照応関係の取り扱い

(66) 今日は 上野動物園 に行って浅草花やしきに行くハードプレイをしました。
　　動物園 出たとこで猫が枯葉をベッドにしていてびっくりしました。

(66) の「動物園」という場所参照表現は，前方の「上野動物園」という場所参照表現と照応関係にある．ここでは「上野動物園」という場所参照表現は具体的なエンティティを付与できることから，同一エンティティを「動物園」にも付与することが可能となる．

B.0.13　地方表現の取り扱い

(67) [関東 LOC/関東] 遠征頑張る

(68) [九州 LOC/九州] ではなじみのある車両ですな！

(67) の「関東」，(68) の「九州」は，自治体名というわけではないが，それぞれ日本の特定の地域を指している．よって，今回のアノテーションではタグ，具体的なエンティティをそれぞれ付与することとする．

B.0.14　施設内の部屋・設備などの取り扱い

(69) 会場：[宝塚大学新宿キャンパス FAC/宝塚大学_新宿キャンパス]2 階 202 号室

(70) [アキバ LOC/秋葉原] の [ヨドバシ FAC/ヨドバシカメラマルチメディア Akiba] はトイレがキレイ

(69) の「202 号室」は「宝塚大学新宿キャンパス」内の一部屋を指しており，また (70) の「トイレ」は「ヨドバシ」内の設備を指している．このように，特定の施設内の部屋・設備などについては，本研究で用いている地名・施設名辞書にも含まれない程度に粒度の細かいエンティティとなるため，タグも具体的なエンティティも付与しないこととする．

B.0.15　場所参照表現に付随する，位置関係などを示す表現の取り扱い

(71) [東銀座駅 FAC/東銀座駅] から徒歩 3 分

(72) 100m 先の [スーパー FAC/NULL 特定不能] までだな

(71) の「から徒歩 3 分」という表現，(72) の「100m 先の」という表現はそれぞれ，「東銀座駅」「スーパー」に付随する，位置関係を示す表現となっている．ただし，本研究では固有名詞ないし普通名詞，またその連続をアノテートの対象とするため，これらの表現はアノテートの対象外となる．よって，タグも具体的なエンティティも付与しない．

B.0.16　住所表現の取り扱い

(73) [宮城県仙台市青葉区立町２－８ LOC/NULL (NOTE=住所を直に記述)] 行きたい

(73) の例では，住所表現がテキスト中に直接記述されている．このとき，場合によっては住所表現の指す地点に存在する施設をエンティティとして付与することも考えられるが，実際にはその地点に施設が無いことや，一般住宅の住所であることなどが起こり得る．そのような場合を考慮し，本研究では (73) のような住所表現には具体的なエンティティを付与せず，備考欄に住所が直に記述されている旨を書き留めることとする．

発表文献一覧

受賞一覧

- 第 26 回日本リスク研究学会年次大会 優秀発表論文賞
  ツイッター分析に基づく福島県産桃に対する風評の実態解明とその対策 (岡崎直観，佐々木彬，乾健太郎，阿部博史，石田望)

国内会議・研究会論文

- 船木洋晃, 佐々木彬, 岡崎直観, 乾健太郎, 深田陽介, 竹下隆一郎, 田森秀明, 野澤博. インターネット上の当選運動・落選運動の分析. 人工知能学会第 28 回全国大会, 1K3-2, 2014 年 5 月.

- 佐々木彬, 五十嵐祐貴, 渡邉陽太郎, 乾健太郎. 場所参照表現のグラウンディングに向けて. 言語処理学会第 20 回年次大会, pp.177-180, 2014 年 3 月.

- 渡邉陽太郎, 佐々木彬, 五十嵐祐貴, 岡崎直観, 乾健太郎. 実世界指向情報構造化支援のための情報抽出技術. 言語処理学会第 20 回年次大会, pp.1003-1006, 2014 年 3 月.

- 岡崎直観, 佐々木彬, 乾健太郎, 阿部 博史, 石田 望. ツイッター分析に基づく福島県産桃に対する風評の実態解明とその対策. 第 26 回日本リスク研究学会年次大会, 中央大学（東京都）, 2013 年 11 月.

- 佐々木彬, 水野淳太, 岡崎直観, 乾健太郎. 機械学習に基づくマイクロブログ上のテキストの正規化. 人工知能学会第 27 回全国大会, 4B1-4, 2013 年 6 月.