Vol.2014-NL-217 No.14
2014/7/4

ソーシャルメディアにおける空間的近接性と時間的一貫性を考慮した地名の曖昧性解消

粟村 誉[1,a)]　荒牧 英治[1,b)]　河原 大輔[1,c)]　柴田 知秀[1,d)]　黒橋 禎夫[1,e)]

概要：近年，膨大な量の文書が Web 上に溢れるようになるにつれ，それらから有用な情報を抽出する技術が重要になってきた．特に，Twitter などのソーシャルネットワークサービス (SNS) は地域固有の情報を含むことが多いため，文書内の地名表現がどこの地名，地域を指しているかを同定することが必要となる．これまで，このような地名曖昧性解消の問題は，語義曖昧性解消の手法を利用して，語彙情報に基づいて解かれることが多く，地名特有の手がかりが使われていない．本研究では，(1) 空間的近接性と (2) 時間的一貫性の 2 つの手がかりを用いて，地名曖昧性解消の精度向上を目指す．空間的近接性は，投稿内の地名同士は距離が近いことが多いという傾向，時間的一貫性は，一連の投稿に現れる地名はそれぞれ関連性があるという傾向をとらえるために導入する．位置情報付きツイートを用いた実験によって，2 つの手がかりの有効性を確認した．

1. はじめに

近年，膨大な量の文書が Web 上に溢れるようになるにつれ，それらから有用な情報を抽出する技術が重要になってきた．特に，Twitter などのソーシャルネットワークサービス (SNS) は地域固有の情報を含むことが多いため，文書内の地名表現がどこの地名，地域を指しているかを同定することが必要となる．過去の地名曖昧性解消の研究では，語義曖昧性解消の手法に基づき，文書内の語彙情報を利用して，地名の曖昧性解消を行っている．しかし，SNS 文書などのような極端に短い文書については文書の語彙情報のみから推定を行うのは難しい場合が多い．例えば，次のツイート中の「県庁前駅」という地名がどこの「県庁前駅」であるかを単語情報のみから識別するのは難しい．[*1]

「首里駅から 県庁前駅 にきました！」

本研究では，Twitter のツイートに含まれる曖昧な地名表現についてその曖昧性を解消し，地名を同定するために，(1) 空間的近接性と (2) 時間的一貫性を用いる手法を提案する．空間的近接性は，文中に含まれる複数の地名は相互に関係があると仮定する．上記の例について，「県庁前駅」と「首里駅」の間の近接性を考える事で「沖縄県の県庁前駅」であると推定することができる．時間的一貫性は，過去のツイートは対象のツイートと関連が高いと仮定する．上記の例について直近のツイートに「沖縄県」に関する情報が含まれていれば同様に「沖縄県の県庁前駅」であると推定することができる．

本論文では，曖昧性を含む「県庁前駅」のような地名を**地名表現（LEX; Location EXpression）**と呼び，その LEX が指す位置（< 県庁前駅（沖縄県）> など）を**地名エンティティ（LE; Location Entity）**と呼ぶ．また，「県庁前駅」などのように LE を複数持つ LEX を**曖昧な LEX** と呼ぶ．LEX と LE のデータベースは日本語 Wikipedia を利用して定義する．学習データとして位置情報付きツイートを使用し，上記のデータベースを用いて自動で作成する．作成した地名曖昧性解消システムは位置情報が付与されていないあらゆるツイートに対して適用できる．

本研究の特長は次の 3 点である．

- 地名曖昧性解消のための 2 つの有効な手がかりを提案
- 位置情報付きツイートからの学習データ自動生成
- 位置情報が付与されていないツイートの LEX の LE 判別

本論文では，まず 2 節で関連研究について述べる．3 節で本研究に用いたリソースについて述べる．4 節で提案手法である空間的近接性と時間的一貫性の利用について述べる．5 節で実験と考察について述べ，6 節でまとめと今後

[1] 京都大学
[a)] awa@nlp.ist.i.kyoto-u.ac.jp
[b)] eiji.aramaki@design.kyoto-u.ac.jp
[c)] dk@i.kyoto-u.ac.jp
[d)] shibata@i.kyoto-u.ac.jp
[e)] kuro@i.kyoto-u.ac.jp

[*1] 語彙情報として「首里駅」が手がかりとなりうるが，そのためには「首里駅」と沖縄県の「県庁前駅」を含むような大量の学習データが必要である．

の課題について述べる．

2. 関連研究

本研究で扱う地名曖昧性解消は語義曖昧性解消 (WSD) と密接な関係がある．語義曖昧性解消は様々な研究がされており，SemEval[1], [2] や SENSEVAL[3], [4] などの語義曖昧性解消国際ワークショップも開催されている．地名曖昧性解消では，語義曖昧性解消の手法に加えて地名に関する情報も適切に扱う必要がある．さらに，本研究では地名に関してソーシャルメディアの情報を扱うため，様々なノイズをフィルターすることでそれらの情報を適切に扱う必要がある．まず地名曖昧性解消の 1 つである地名推定の研究について述べ，次にソーシャルメディアにおける情報抽出の研究について述べる．

2.1 地名推定

Web 文書における地名に関する曖昧性の解消は長い間研究されている．いくつかの手法が提案されているが，その中で最も単純で信頼できるとされていた手法の 1 つは IP アドレスを利用する手法である．しかし，IP アドレスに依存する場合，過去のコンテンツに使われた IP にアクセスできない，携帯端末の普及により有用さを欠くことが多いなどの問題が生じる．結果として地名推定は実際の原文テキストを用いて解決を試みる，より難しい方法に着眼されるようになった．情報検索の世界では主に Web ページ，Wikipedia の編集ログなどが地名推定の基礎として用いられてきた．これらのリソースは基本的に均質であり扱いやすいが，対照的に膨大なデータを有するソーシャルメディアのテキストはしばしばノイズを含み，扱いにくいという問題が存在している．

地名推定は，従来の方法では入力文書中の地名について，“地名の候補” と “文書中で共起する地名の候補” との実際の距離，及び曖昧性のある地名に住所階層や人口数などから算出された有名度というスコアを用いて判別を行う [5]．平野らはこの従来の手法に対して，有名度の算出方法の変更及び有名度の優先によって地名判別を行った [6]．Han らは，地名を表す単語を独自の計算法や情報利得比などによって Twitter のツイートデータから抽出した [7]．また，抽出したそれらの単語を用いて，発言された都市の推測をした場合良い結果が得られたことを述べている．彼らはこのアプローチが地名推定にとって低メモリかつ高速であること，また地域に関連した語の抽出ができることから辞書編集などに役立てられるのではないかと結論づけている．Chandra らは，Twitter のユーザーの居住地をツイートから IP アドレスや地名辞典などを使用せず推定し，リプライを基に地名推定した手法の優位性を示した [8]．

これらの研究では，地名に関する特定の表現や地名の有名度などを用いることで，それぞれ地名の推定を行っている．しかし，従来方法では，地名同士の距離などの空間的近接性，投稿時間などの時間的一貫性の一方もしくは両方を考慮した研究は行われていない．本研究では，空間的近接性及び時間的一貫性を考慮した新しい地名推定方法を提案する．

2.2 ソーシャルメディアにおける情報抽出

Twitter などのソーシャルメディアから情報の抽出を試みた研究には様々なものが存在する．近藤らは，Twitter の投稿内容からユーザー個人の興味，特徴を推定し，はてなブックマークなどからパーソナライズされた記事などを提供するポータルサイトを構築した [9]．情報の抽出は，はてなキーワードに登録された単語の有無とツイート中の URL のドメインを利用した．岡元らは，Twitter を利用し，エゴグラム分析を行ったユーザーの投稿内容からナイーブベイズ法により同様のエゴグラム分析結果が得られるかを実験した [10]．榊らは，位置情報の付与されたツイートを用いて，地震の検知をおこなった [11]．まず，「地震」や「揺れた」などの言葉を含むツイートが実際の地震の発生直後にされたものかどうかを SVM で判定し，整形したツイート群から地震発生地点の予測を行った．また，予測した結果から，ユーザーに地震の発生を知らせるシステムの開発を行った．Bollen らは，ツイート全体から社会的な気分の情報を抽出し，ぞれらの “感情” について評価したデータから N 日後の株価の変動を予測した [12]．結果として，“落ち着き” を指し示すデータから 3 日後の株価を予測できたと結論づけている．荒牧らは，インフルエンザについてのツイート全体からインフルエンザの流行予測を行った [13]．Twitter テキストは実世界を反映した情報が流れてくること，言語処理によりそれらの情報を一定の確率で抽出可能であることを示した．

このようにソーシャルメディアを用いた様々な研究が行われている．これらを用いた地名に関する研究も多く議論されているが，その多くが位置情報をツイートなどの投稿に付与された GPS 情報に頼っている．しかし，GPS 情報が付与された発言は，全発言の 1%にも満たない．[*2] 上記のような研究の精度を高めるためには，膨大な位置情報無し文書から地名の情報を抽出する手法の発展が必要不可欠である．

3. リソース

3.1 LEX データベース

まず本研究で扱う LEX と LE を定義する必要がある．ここで，Wikipedia において GIS 情報を持つ LEX と LE に注目した．本研究では LEX と LE のデータベースを **LEX データベース**と呼び，Wikipedia の GIS データのタイプに

[*2] Semiocast によると，Twitter の位置情報付きツイートは全体の 0.77%しか存在しない．


県庁前駅 (兵庫県)

県庁前駅（けんちょうまええき）は、兵庫県神戸市中央区下山手通にある神戸市営地下鉄西神・山手線の駅である。駅番号はS04。開業当時は「山手（県庁前）駅」であったが、後に現在の名称に変更されている。

目次 [非表示]
1 利用可能な路線
1.1 鉄道
1.2 路線バス
2 駅構造
3 利用状況
4 駅周辺
5 歴史
6 その他
7 隣の駅
8 関連項目
9 外部リンク

利用可能な路線 [編集]

鉄道 [編集]
- 西日本旅客鉄道（JR西日本）
  - 東海道本線（JR神戸線） - 元町駅

県庁前駅*

左手、駅入口（県庁南）

けんちょうまえ - Kenchōmae

◂S03 三宮 (0.9km)　(1.1km) 大倉山 S05▸

所在地	神戸市中央区下山手通5丁目 北緯34度41分27.29秒 東経135度11分1.89秒
駅番号	S04
所属事業者	神戸市交通局
所属路線	西神・山手線 （正式路線名は山手線）
キロ程	2.2km（新神戸起点） 谷上から9.7km
駅構造	地下駅
ホーム	2層式 2面2線
開業年月日	1985年（昭和60年）6月18日
備考	1993年『山手(県庁前)駅』から改称。

この表について [表示]

図 1　Infobox
Fig. 1　Infobox information


県庁前駅 (沖縄県) 座標: 北緯26度12分51.79秒 東経127度40分45.68秒

この記事は検証可能な参考文献や出典が全く示されていないか、不十分です。
出典を追加して記事の信頼性向上にご協力ください。（2014年2月）

県庁前駅（けんちょうまええき）は、沖縄県那覇市久茂地1丁目にある沖縄都市モノレール線（ゆいレール）の駅である。

当駅が沖縄県の現在の代表駅・中心駅であるが（現在計画中の那覇 - 名護間の鉄道計画が実施されれば変わる）、他県の代表駅に比べるとモノレールの途中駅であるため小さい。計画当時の仮称は「御成橋駅（おなりばしえき）」であった。

目次 [非表示]
1 歴史
2 駅構造
2.1 のりば
2.2 駅設備
3 駅周辺
4 バス路線
5 その他
6 隣の駅
7 関連項目

県庁前駅

県庁前駅（背後はパレットくもじ）

けんちょうまえ - Kenchō-Mae

◂旭橋 (0.58km)　(0.72km) 美栄橋▸

所在地	沖縄県那覇市久茂地1丁目 北緯26度12分51.79秒 東経127度40分45.68秒
駅番号	7
所属事業者	沖縄都市モノレール
所属路線	■沖縄都市モノレール線
キロ程	5.91km（那覇空港起点）
駅構造	高架駅
ホーム	1面2線
乗車人員 -統計年度-	4,466人/日（降車客含まず） -2011年-
開業年月日	2003年（平成15年）8月10日

図 2　座標情報
Fig. 2　Latitude/Longitude information

よる以下の 2 種類の方法で LEX データベースを獲得した．

- Infobox
- Wikipedia 記事内の座標情報

3.1.1　Infobox

Infobox は図 1 に示すような Wikipedia の記事に用いられているメタテンプレートの 1 つである．地名についての記事の場合，所在地や緯度・経度が記載される場合がある．このようなエントリを全て LE として抽出した．

日本語 Wikipedia において抽出を行った結果，759 個の LEX について 884 個の LE が得られた．

3.1.2　Wikipedia 記事内の座標情報

Wikipedia 記事内には，図 2 に示すように地名の記事上部に緯度・経度情報が記述される事が多い．このような GIS 情報を含む地名記事から LE として抽出した．

日本語 Wikipedia においてこの方法を適用した結果，17,140 個の LEX について 17,426 個の LE が得られた．

上記二種類のデータベースを併合し，重複を取り除く事で本研究の LEX データベースを作成した．合計 17,724 個の LEX と 18,256 個の LE を獲得した．表 1 に「県庁前駅」の LE を示す．表 2 に LEX と LE の頻度を LE 数ごとに示す．表 2 から，994 個の LE に相当する 462 個の曖昧な LEX が得られたことがわかる．

本研究では，図 1 や図 2 の < 県庁前駅（兵庫県）> や < 県庁前駅（沖縄県）> など，丸括弧のついた地名を 1 つの LE，丸括弧を外した「県庁前駅」などをそれらの共通の LEX とする．

表 1　LEX「県庁前駅」の LE
Table 1　LEs for the LEX “Kencho-mae Station”

LE	緯度	経度
県庁前駅（兵庫県）	34.69	135.18
県庁前駅（千葉県）	35.60	140.12
県庁前駅（富山県）	36.69	137.20
県庁前駅（広島県）	34.39	132.45
県庁前駅（愛媛県）	33.84	132.76
県庁前駅（高知県）	33.55	133.53
県庁前駅（沖縄県）	26.21	127.67

表 2　LEX データベースの統計情報
Table 2　Statistics of LEX database

LE 数	LEX	LE
1	17,262	17,262
2	412	824
3	38	114
4	8	32
5	2	10
7	2	14
計	17,724	18,256

3.2　地名曖昧性解消コーパス

LEX 曖昧性解消の学習・評価のためには，LEX と LE が対応付けられているデータが必要となる．これを位置情報付きの Twitter データから抽出する．例えば，「県庁前駅に集合しよう」というツイートがあった際に，そのツイート

の位置情報が沖縄県を指していれば，そのツイートにおける「県庁前駅」は＜県庁前駅（沖縄県）＞を指していると考えることができる．従って，LEX を含む位置情報付きツイートに対して以下の手順で LE を付与する．

- STEP 0（前処理）：ツイートデータの準備
  2011 年 7 月 15 日〜2012 年 7 月 31 日にかけて位置情報付きのツイートデータの収集を行った．この時，テキストが重複するツイートは取り除いた．それぞれのツイートデータは以下のような項目を含む．
  - ツイート ID
  - 日付
  - （ユーザ ID）
  - ユーザ名
  - 言語
  - （ツイートソフト）
  - テキスト
  - 緯度
  - 経度

  ただし，ユーザ ID とツイートソフトは 2012 年 5 月 13 日から 2012 年 7 月 31 日までの期間のみに含まれる．
- STEP 1：LEX を含むツイートの抽出
  3 節で作成した LEX データベースに基づき曖昧な LEX を含むツイートを抽出する．曖昧でない LEX を含むツイートは対象のツイートとしては使用しないが，4.3 節で述べる時間的一貫性の手がかりとしては使用する．この手順では LEX の文字列がツイートに含まれているかを判断し，そのツイートを各 LEX ごとにまとめる．複数の曖昧な LEX が含まれている場合は，それぞれの LEX について対象とした．例えば，「県庁前駅から元町駅に行きます」というツイートは「元町駅」と「県庁前駅」の両方についてそれぞれ使用した．
- STEP 2：LE の付与
  STEP 1 で抽出したツイート中の曖昧な LEX に LE を付与する．ツイートの位置情報と対象とする LEX の LE の位置情報に基づいて付与を行う．ある LE についてこの 2 つの距離が短い時，そのツイート内の LEX はその LE を指しているとした．例えば，「県庁前駅」という曖昧な LEX を含むツイートの位置情報が＜県庁前駅（沖縄県）＞と近い場合，この「県庁前駅」は＜県庁前駅（沖縄県）＞を指しているとした．本研究では，この LE を判断する距離の閾値を 10 km とした．つまり，ツイートの位置情報と LE の位置情報との距離が 10 km 以内である場合，このツイート内の曖昧な LEX はこの LE を指しているとし，それ以外の場合はこのツイートを破棄した．10 km 以内にある LE が複数ある場合，最も近い LE を付与した．

上記の手順で約 18 万ツイートを学習・評価データとして獲得した．この時，462 個の曖昧な LEX のうち，1 ツイート以上得られた LEX は 353 個であった．距離の計算は緯度 1 度：111.319km，経度 1 度：91.187km として行った．

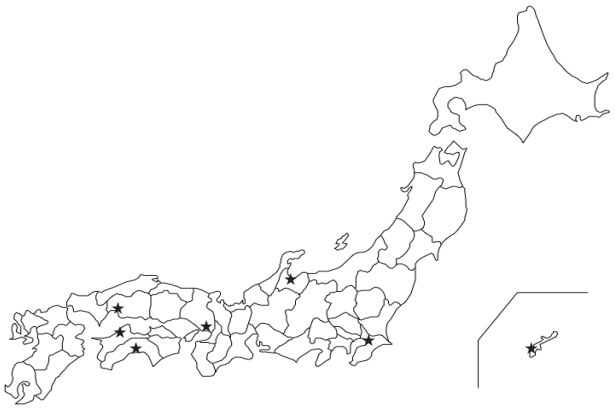

図 3 県庁前駅の位置
Fig. 3 Locations of “Kencho-mae Station”

4. 提案手法

Twitter のツイートにおける地名曖昧性解消の手法を提案する．本研究では，機械学習アルゴリズム SVM を用いてツイート内の LEX の LE を自動的に識別する．SVM 分類器は LEX ごとに作成した．各 SVM 分類器はそれぞれ以下の素性を含む．

4.1 ベースライン素性

ベースライン素性として以下の 2 つの素性を使用する．
(1) 語彙的素性：ツイートの形態素
(2) マジョリティ素性：LE の頻度

4.2 空間的近接性

空間的近接性として，対象の LEX と同一ツイート内の曖昧でない LEX について，それぞれの LE 間の距離を素性として用いる．本手法が適用できる典型的な例を次に示す．

(1)　首里駅から 県庁前駅 まで約 20 分

テキスト中の「県庁前駅」という曖昧な LEX には，図 3 に示すように以下の 7 つの LE が存在している．
(1)＜県庁前駅（兵庫県）＞
(2)＜県庁前駅（千葉県）＞
(3)＜県庁前駅（富山県）＞
(4)＜県庁前駅（広島県）＞
(5)＜県庁前駅（愛媛県）＞
(6)＜県庁前駅（高知県）＞
(7)＜県庁前駅（沖縄県）＞
今回の例での「県庁前駅」がどの LE を指すかの判別は語彙情報からは困難である．しかし，LEX と同一のツイートに含まれる曖昧でない地名との関連性を考慮することで，LEX 曖昧性解消の手がかりとなる．一般に，関連性の強

い LEX は対象の LEX と共に用いられることが多い．SVM によって暗黙に考慮される可能性はあるが，期待はできない．従って，本研究では 2 地名間の距離を関連性として明示的に用いた．この距離が短い場合，その 2 地名は関連が強いとした．上の「県庁前駅」の例の場合，< 首里駅 > は < 県庁前駅（沖縄県）> とは比較的距離が近いが，< 県庁前駅（千葉県）> とは近くない．従って，この「県庁前駅」の LE は < 県庁前駅（沖縄県）> であると推定できる．

この空間的近接性をツイートに適用する際，まずツイートに別の LEX が含まれているかを調べる．もしその LEX が曖昧でない場合，この曖昧でない LE と対象の曖昧な LEX の各 LE との距離を比較し [*3]，その距離の大きさに応じた素性を付与する．もし含まれる別の LEX が曖昧な場合，その LEX が指す LE が決定できないため空間的近接性の素性として使用しない．

例えば，「県庁前駅」という曖昧な地名表現を対象としている場合に，テキスト中に「首里駅」という曖昧でない地名表現が含まれているとする．「県庁前駅」の全ての LE（< 県庁前駅（沖縄県）>，< 県庁前駅（千葉県）> など）について < 首里駅 > との距離を計算し，< 県庁前駅（沖縄県）> とは 0〜10 km の距離，< 県庁前駅（愛媛県）> とは 500〜1000 km の距離であった場合，この 2 つを別々の要素として素性化する．対象とする曖昧な地名表現の LE 数を l，指定した距離の種類を d 種類とすると，近接性の素性数は ld となる．この d は 5.1 節で述べる距離の設定数である．

4.3 時間的一貫性

ここまで LE 推定に対象のツイートのみについて考慮してきた．しかし，対象のツイートが短すぎるために，曖昧性解消の手がかりがほとんど含まれないことがある．従って本研究では，対象のツイートの直前 t 時間までのツイートを考慮する．これらの過去のツイートからベースライン素性や空間的近接性素性も抽出する．

本手法が適用できる典型的な例を次に示す．

(2)　県庁前駅 にきました！

このユーザーの直近 3 発言を以下に示す．

(3)　今から飛行機だ〜！沖縄楽しみ！

(4)　沖縄到着〜！

(5)　ゆいレールにのって首里駅を目指します！

このような場合，過去の発言を問題対象として同時に扱うことで有用な情報を抽出できる．例えば，“沖縄” という単語は < 県庁前駅（沖縄県）> と関連を持つ語であり，さらに「首里駅」は < 県庁前駅（沖縄県）> と距離が近い．これらの理由から，この「県庁前駅」は < 県庁前駅（沖縄県）> である可能性が高いと判断できる．

この時，極端に古いツイートについては対象のツイートとほとんど関連性を持たないため，参照する時間の閾値 t を決める必要がある．これについては 5.1 節で述べる．

5. 実験と考察

5.1 実験設定とデータ

4 節で提案した手法を用いてそれぞれの LEX ごとに SVM 分類器を作成した．分類器によってツイートに含まれるある曖昧な LEX の LE を特定する．地名曖昧性解消問題は多クラスの識別問題であるため，one-versus-the-rest を手法として用いた．正解データとして，3.2 節で得られたコーパスから，それぞれの曖昧な LEX について，10 ツイート以上に含まれる 354 種類の LEX のみを対象とした 70,184 ツイートを使用した．SVM には TinySVM の 2 次の多項式カーネルを使用した．[*4] 日本語の単語分割には日本語形態素解析器 JUMAN を使用した．[*5]

精度は，“正解ツイート数 / 対象のツイート数” によって計算する．さらに，それぞれの LEX について含むツイート数が 10〜100，100〜1,000，1,000〜であるものについて分類し，それぞれの場合についての精度も比較した．

5.2 比較手法

本研究では次の 4 つの方法を比較する．

- ベースライン (B): この手法では (1) 語彙的素性と (2) マジョリティ素性の 2 つについてのみ使用する．語彙的素性は形態素の原型を素性として使用した．ここで用いた語彙的素性は頻度上位 100,000 の形態素のみ使用し，語の頻度を語彙的素性として使用した．
- +空間的近接性 (+SP): この手法ではベースラインと空間的近接性を使用する．空間的近接性素性は 4 節で述べたように，対象の LE と別の曖昧でない LEX（LE）との距離によって付与する．表 3 に示すように 4 種類の距離セットについて検討した．
- +Temporal Consistency (+TC): この手法ではベースラインと時間的一貫性を使用する．4.3 節で述べたように，過去のツイートを最大 3 ツイートまで参照し，素性を付与する．ある一定の時間以上古いツイートについては無視する．この時間について表 3 に示すように 4 種類の時間セットについて検討した．
- +空間的近接性+時間的一貫性 (+SP+TC): この手法では上記の全ての素性を使用する．空間的近接性素性は時間的一貫性によって取得した過去のツイートに対しても素性付与を行う．

[*3] 曖昧でない LE が複数ある場合，それぞれについて考慮する．

[*4] http://chasen.org/~taku/software/TinySVM/

[*5] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

表 3 空間的近接性と時間的一貫性の詳細設定
Table 3 Settings for SP and TC

手法	設定
SP (デフォルト: 0〜10, 100〜500 km)	+10〜100 km
	+10〜50, 50〜100 km
	+10〜100, 500〜1000 km
	+10〜50, 50〜100, 500〜1000 km
TS	無期限
	〜24 h
	〜12 h
	〜6 h
	〜3 h
	〜1 h

5.3 実験結果と考察

表 4 に全ての手法についての結果を示す．ここで，近接性，一貫性はそれぞれ次の値を用いた．

- 0〜10, 10〜100, 100〜500 km
- 6 h 以内

また，MB はマジョリティベースラインで，対象とする曖昧な LEX について，一番頻度の多い LE を全てシステム出力として返す方法である．

表 4 に示す通り，近接性及び一貫性の素性により LE 推定の正解率が向上した．特に，近接性素性に関しては，曖昧な LEX ごとのツイート数に関わらず比較的大きな正解率の向上が見られた．一貫性素性は，ツイート数が少ない LEX に対しては正解率は向上したが，ツイート数が多い LEX に関しては効果は少なく，ベースラインを下回った．近接性と一貫性の両方を考慮した場合，ツイート数の少ない LEX に対しては正解率が大きく向上し，MB に比べて 7.13%の正解率の向上が見られた．

また，結果の妥当性を調べるために検定を行った．検定には，符号検定を用いた．表 4 にそれぞれのベースラインに対して，有意水準 5%で検定を行った場合に優位性が示せた結果には †，有意水準 1%で検定を行った場合に優位性が示せた結果には ‡ を付与した．これにより，特にツイートの少ないときに，それぞれの手法についてベースラインとの優位性を示せた．また，ツイート数によらず全体についての正解率もベースラインに対して優位性を示すことができた．

一貫性素性について，ツイートの多い LEX については正解率は向上しなかった．原因の 1 つとして，実験に用いたツイートデータに様々な偏りが見られたことが挙げられる．今回の例では LEX を含み，かつ位置情報の付与されたツイートのみについて扱ったため，LEX ごとの正解 LE が偏ってしまっている例が多く見られた．これは MB の精度が 85%を超えていることからも分かる．ツイート数が多い LEX ほど，大きな偏りが見られた．また，位置情報付与付きのツイートは，Foursquare など位置情報連携サービスを用いたものが多く，LE ごとのテキストに偏りが生じた．この結果，多くの場合，マジョリティを考慮するだけで高い正解率を出すことができてしまうため，他の素性を用いた場合での大きな正解率の向上は期待できない．本研究では，使用したツイートは位置情報付きの LEX を含むものであったが，位置情報の付いていない一般的なツイートに対して実験を行うことで，より明確な精度の向上が示せると考えられる．

表 4 全手法の比較
Table 4 Main results

ツイート数 (合計)	手法	正解数	精度
10〜100 (4,891)	MB	4,171	0.8528
	B	4,485	0.9170
	+SP	4,515	0.9231 ‡
	+TC	4,491	0.9182 ‡
	+SP+TC	**4,520**	**0.9241** ‡
100〜1,000 (25,758)	MB	22,477	0.8726
	B	24,725	0.9599
	+SP	**24,752**	**0.9609**
	+TC	24,708	0.9592
	+SP+TC	24,737	0.9604
1,000〜 (39,535)	MB	36,896	0.9332
	B	39,041	0.9875
	+SP	**39,054**	**0.9878**
	+TC	39,036	0.9874
	+SP+TC	**39,054**	**0.9878**
10〜 (70,184)	MB	63,544	0.9054
	B	68,251	0.9725
	+SP	**68,321**	**0.9735** ‡
	+TC	68,235	0.9722
	+SP+TC	68,311	0.9733 †

†：有意水準 5%で B に対して有意差ありと判定されたもの
‡：有意水準 1%で B に対して有意差ありと判定されたもの

表 5 に LE 間の距離の比較による結果を示す．表 5 に示す通り，LEX ごとのツイート数が少ないときは+500〜1000 km の素性，多いときは+10〜50, 50〜100 km の素性のときに正解率が向上した．ツイート数が少ないときは，より多くの情報を必要とするため 500〜1000 km の素性に効果が見られたと考えられる．また，ツイート数が多いときは，情報の追加よりも情報の細分化が必要であるため，10〜100 km の素性の代わりに 10〜50, 50〜100 km の素性を用いた結果，正解率が向上したと考えられる．

表 6 に参照するツイートの期間の比較による結果を示す．表 6 に示す通り，一貫性によるツイートの参照期間は時間による大きな違いは見られなかった．しかし，24 時間

表 5 空間的近接性の距離による比較
Table 5 Comparison of SP features

ツイート数 (合計)	近接性設定	正解数	精度
10∼100 (4,891)	+10∼100 km	4,515	0.9231
	+10∼50, 50∼100 km	4,513	0.9227
	+10∼100, 500∼1000 km	4,519	0.9239
	+10∼50, 50∼100, 500∼1000 km	**4,520**	**0.9241**
100∼1,000 (25,758)	+10∼100 km	24,752	0.9599
	+10∼50, 50∼100 km	**24,758**	**0.9612**
	+10∼100, 500∼1000 km	24,744	0.9606
	+10∼50, 50∼100, 500∼1000 km	24,746	0.9607
1,000∼ (39,535)	+10∼100 km	39,053	0.9878
	+10∼50, 50∼100 km	**39,069**	**0.9882**
	+10∼100, 500∼1000 km	39,064	0.9881
	+10∼50, 50∼100, 500∼1000 km	39,065	0.9881

表 6 時間的一貫性の時間による比較
Table 6 Comparison of TC features

ツイート数 (合計)	一貫性設定	正解数	精度
10∼100 (4,891)	indefinite	4,429	0.9055
	∼24 h	4,491	0.9182
	∼12 h	4,493	0.9186
	∼6 h	4,491	0.9182
	∼3 h	**4,494**	**0.9188**
	∼1 h	4,493	0.9186
100∼1,000 (25,758)	indefinite	24,694	0.9587
	∼24 h	24,700	0.9589
	∼12 h	24,709	0.9593
	∼6 h	24,708	0.9592
	∼3 h	24,718	0.9596
	∼1 h	**24,725**	**0.9599**
1,000∼ (39,535)	indefinite	38,988	0.9862
	∼24 h	38,988	0.9873
	∼12 h	**39,036**	**0.9874**
	∼6 h	**39,036**	**0.9874**
	∼3 h	39,033	0.9873
	∼1 h	39,034	0.9873

以上の場合の正解率は比較的低く，さらに，時間を指定せず過去のツイートを取得する場合は LEX ごとのツイート数に関わらず正解率が最も低くなった．このことから，時間的一貫性の時間考慮の妥当性が示された．

6. おわりに

本研究では，ソーシャルメディアにおけるユーザーの投稿に含まれる LEX について，その LE を推定した．その際に，空間的近接性と時間的一貫性を考慮することによって，LE の推定精度を向上させる方法を提案した．結果として，

- 近接性の考慮による正解率の向上
- ツイート数の少ない LEX に対して，一貫性の考慮による正解率の向上
- 近接性，一貫性両方の考慮による 7.13%の正解率の向上（マジョリティベースライン比）

の 3 つを示すことができた．

今後の課題としては，まず，ツイート数が多い LEX についての，一貫性の考慮における LE 推定の正解率の向上が挙げられる．

本研究では曖昧でない LEX についてのみ，対象の LEX との距離を求めることで素性付与を行った．しかし，ツイート内に複数の曖昧な LEX が含まれている場合に，事前に曖昧な LEX の LE の推定ができていれば，対象の LEX の近接性素性として用いることができる．これを考慮することで正解率を向上させることができると考えられる．

また，本研究では曖昧な LEX の LE の推定を行ったが，地名推定には他にもいくつかの問題がある．その 1 つとして，「ユーザーが発言に含まれる LEX の場所に実際にいるか」という問題が挙げられる．この問題の解決も，位置特化型のアプリケーションなどの有益な情報の収集において非常に重要である．今後の課題として，本研究で提案した空間的近接性，及び空間的一貫性等を考慮することで，この問題にも取り組んでいきたい．

参考文献

[1] Carpuat, M.: NRC: A Machine Translation Approach to Cross-Lingual Word Sense Disambiguation (SemEval-2013 Task 10), *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, Association for Com-

putational Linguistics, pp. 188–192 (online), available from ⟨http://www.aclweb.org/anthology/S13-2034⟩ (2013).

[2] Aramaki, E., Imai, T., Miyo, K. and Ohe, K.: UTH: Semantic Relation Classification using Physical Sizes, *Proceedings of the Association for Computational Linguistics (ACL2007) Workshop on Semantic Evaluation (SemEval2007)*, pp. 464–467 (2007).

[3] Kilgarriff, A.: SENSEVAL: An exercise in evaluating word sense disambiguation programs, *Proceedings of the International Conference on Language Resources and Evaluation (LREC1998)*, pp. 581–588 (1998).

[4] Kurohashi, S.: SENSEVAL2 Japanese Translation Task, *Proceedings of SENSEVAL2*, pp. 37–40 (2001).

[5] Li, H., Srihari, R. K., Niu, C. and Li, W.: InfoXtract Location Normalization: A Hybrid Approach to Geographic References in Information Extraction, *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1*, HLT-NAACL-GEOREF '03, Vol. 6, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 39–44 (online), DOI: 10.3115/1119394.1119400 (2003).

[6] 平野徹，松尾義博，菊井玄一郎：地理的距離と有名度を用いた地名の曖昧性解消 (自然言語処理, 一般セッション, 人工知能と認知科学)，全国大会講演論文集，Vol. 70, No. 2, pp. '2–85'–'2–86'（オンライン），入手先 ⟨http://ci.nii.ac.jp/naid/110006865351/⟩ (2008).

[7] Han, B., Cook, P. and Baldwin, T.: Geolocation Prediction in Social Media Data by Finding Location Indicative Words, *Proceedings of COLING 2012*, pp. 1045–1062 (online), available from ⟨http://www.aclweb.org/anthology/C12-1064⟩ (2012).

[8] Chandra, S., Khan, L. and Muhaya, F.: Estimating Twitter User Location Using Social Interactions–A Content Based Approach, *Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom)*, pp. 838–843 (online), DOI: 10.1109/PASSAT/SocialCom.2011.120 (2011).

[9] 近藤直人，内田 理：Twitter を用いた個人の特徴抽出とその情報提供ポータルサイト構築への応用，言語処理学会第 20 回年次大会 (NLP2014)，北海道 (2014).

[10] 岡本拓馬，松本和幸，吉田 稔，北 研二：ナイーブベイズ法を用いた Twitter による性格推定，言語処理学会第 20 回年次大会 (NLP2014)，北海道 (2014).

[11] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, Vol. 10, New York, NY, USA, ACM, pp. 851–860 (online), DOI: 10.1145/1772690.1772777 (2010).

[12] Bollen, J., Mao, H. and Zeng, X.: Twitter mood predicts the stock market, *Journal of Computational Science*, Vol. 8, pp. 1–8 (2011).

[13] 荒牧英治，増川佐知子，森田瑞樹：Twitter Catches the Flu:事実性判定を用いたインフルエンザ流行予測，情報処理学会研究報告. SLP, 音声言語情報処理， Vol. 2011, No. 1, pp. 1–8（オンライン），入手先 ⟨http://ci.nii.ac.jp/naid/110008584112/⟩ (2011).