1H5-5

ニューラルネットワーク言語モデルを用いた口語表現に対応した地名判定システムの構築

Development of Location-name Extraction System for Colloquial Expression using Neural Network Language Model

大谷 昭成 *1　榊 剛史 *2,*3　櫻井 彰人 *1
Akiyosi Otani　Takeshi Sakaki　Akito Sakurai

*1慶應義塾大学 Keio University
*2株式会社ホットリンク Hottolink,Inc.
*3東京大学 The University of Tokyo

地名辞書の整備は地理情報システムを構築する上で重要な課題であるが，既存の辞書や手法では，口語的な地名表現を自動的に地名と判定することは難しい．本研究では Twitter 文書内の単語の並びや構造に着目し，口語的な地名表現を判定するシステムを提案する．大規模 Twitter データから構築したニューラルネットワーク言語モデル（NNLM）を用いて，ベクトル空間内での既存地名への近接性から，ある単語が地名であるか否かを判定することを目指す．評価実験を通じて NNLM とクラスタリングを組み合わせることで，口語的な地名を抽出できる可能性を提示した．

1. はじめに

近年，スマートフォン及びソーシャルメディアの普及に伴い，地理情報処理の重要性は増大している．スマートフォンの普及により，人々は常時接続することが可能になり，またスマートフォンに搭載されている GPS によって位置情報を正確かつ容易に発信することが可能となった．そのため，現在地情報を利用した店舗推薦や観光地推薦など，O2O（Online to Offline）を目指したサービスやビジネスが増加している *1．しかし，多くの場合，ユーザの位置情報はそのスマートフォンのプラットフォーマー（Google 社及び Apple 社）やアプリ運営者しか取得することができず，広く活用することが難しくなっている．

一方，ソーシャルメディアの普及により，人々がテキスト情報をリアルタイムに発信する機会が増加した．それにより，人々の行動やコミュニケーションの記録を取得することが容易になった．普及したソーシャルメディアには位置情報付加機能（ジオタグ機能）を有していることが多く，また Foursquare や Tiwtter，Instagram などいくつかの大規模ソーシャルメディアは投稿を公開しているため，それらを容易に取得することができる．そのため，それらのソーシャルメディア上のジオタグが付与された投稿を収集し，分析する研究が，情報学，社会学の分野で増加している．これらの研究により，ジオタグが付与された大量の投稿を解析することで様々な知見が得られることが分かっている [Watanabe 11].

しかし，ジオタグが付与されている投稿の割合は非常に低い．ソーシャルメディア事業者からは公開されていないが，いくつかの研究においてジオタグ付与割合は 1% 未満であることが明らかになっている．一方，ジオタグが付与されていないものの，ある場所を一意に表す単語表現（以下，これを地名と定義する）を含む投稿が少なくない割合で存在することが分かっている．つまり，投稿内に含まれる地名を抽出し，その場所を位置情報に変換することで，ジオタグの代替となる情報を生成できる可能性がある．自然言語文から地名を抽出するアプローチは固有表現抽出（Named Entity Recognition，以下 NER）として，自然言語処理の分野で古くから行われている [Lin 04]. しかし，ソーシャルメディアの投稿は，口語的な砕けた表現を用いられることが多く，既存の NER ではうまく機能しないことが多い．また地名辞書中の地名を用いて単純マッチングするアプローチも考えられるが，人々は地名辞書の通りにソーシャルメディア上に情報を投稿するわけではない．例えば，「東京」のような一般的な地名でも「トーキョー」「とぉきょぉ」などという表現を用いる．大規模地名辞書によるマッチングを用いていると思われる Google MAPS に「トーキョー」と入力しても「ノースダコタトーキョー」がマッチしてしまう．このように，地名辞書をそのまま用いることも困難である．

そこで，我々は口語的な表現に対応した新たな地名辞書を構築することを目指す．このような地名辞書は，地名のダイレクトマッチング的なアプローチには直接適用可能であり，また，機械学習を用いた NER の正解データとして用いることができる．地名辞書を構築するためには，1. 地名表現の収集，2. 収集した地名表現の位置情報への変換という 2 つのステップが必要となる．本研究では，1 番目のステップである地名表現の収集に焦点を絞り，口語的な表現でも判定可能な地名判定手法を提案し，地名判定システムを構築する事を目指す．口語的な表現を処理するためにニューラルネットワーク言語モデル（以下，NNLM）を用いる．NNLM は近年自然言語処理の分野で注目されているアプローチであり，単語をより意味に近いベクトルで表現することを可能にする手法である．また本稿では，データ収集の容易性から Twitter を対象とする．大規模 Twitter データから構築した NNLM を用いて，「NNLM によるベクトル空間上において，地名同士は近接する」という仮説に基づき，ベクトル空間内での既存地名への近接性からある単語が地名であるか否かを判定することを目指す．

2. 関連研究

ソーシャルメディア上の投稿を用いて位置情報を取得する研究の一つとしては，ユーザの居住地を推定する研究が行われている．ソーシャルメディアのユーザプロフィールに記載されている地名を位置情報として用いている研究や [Hecht 11]，ユーザのリンク情報や, ユーザとリンク関係にあるユーザの位置情報を用いてユーザの位置情報を推定する手法も提案されている [Backstrom 10].

一方，ソーシャルメディア上の投稿を解析する研究のうち，位置情報が必要となる代表的な研究として局所的なイベントを検出する手法が提案されてきた [Lee 11a, Lee 11b]．これらの

連絡先: 大谷 昭成，慶應義塾大学大学院理工学研究科，otani@a2.keio.jp

*1 http://www.watch.impress.co.jp/headline/docs/kyodonews/international/20150227_690503.html

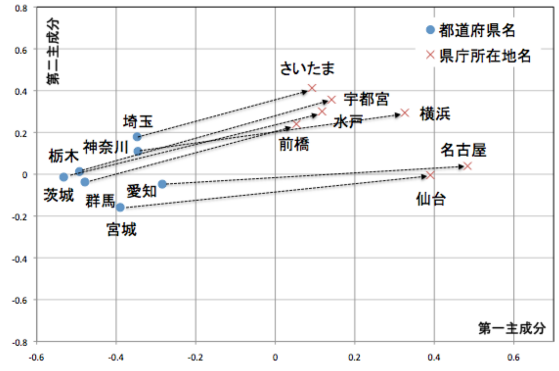

図 1: 都道府県名と県庁所在地名の関係性

研究においては，位置情報として，ツイートに付与された GPS 情報 [Lee 11b] やユーザの居住地情報 [Li 12]，特定地域の地名リストを用いている [Walther 13]．その中で，Watanabe らの研究は機械学習の手法を用いてツイートへの地名付与を実現している．[Watanabe 11]．

また固有表現抽出の手法を用いて，ソーシャルメディアの投稿から地名を抽出する手法も提案されている [Ji 09, Lin 04]．さらに Ritter らは CRF を適用することでツイートへの品詞付与エラーを減らすと共に Labeled LDA を適用することで，既存手法と比べ固有表現抽出の精度が 25% 改善することを示している [Ritter 11]．しかし，日本語を解析する際には品詞を付与する前に単語分割を行う必要があるが，既存の形態素解析ツールでは，口語的な文書において単語分割に失敗することが多い．そのため，この手法を日本語ツイートにそのまま適用することは難しい．

3. 提案手法

本研究では「しぶや」のような口語表現に対応した地名判定手法を提案することを目指す．ただし，口語表現は常に変化していくため，過去のデータを学習データとして用いる単純な機械学習的アプローチや辞書を用いたアプローチでは対応することが困難である．言い換えれば，変化がしやすい自然言語の表層的な特徴を扱うのではなく，より意味的な特徴を扱う必要があると言える．

そこで，言語の意味的な特徴を扱うために，近年注目を浴びている NNLM を適用する．NNLM では，ある単語をその周囲に出現する語から予測するようなニューラルネットワークを学習し，そのニューラルネットワークの各層の重みを単語のベクトルとする．このように構築したベクトル空間では，意味的に近い語同士の距離が近くなること，意味的に類似した 2 語関係を表すベクトル同士の差が小さくなることが知られている [Mikolov 13]．予備実験として，後述する Twitter コーパスから構築した NNLM を用いて，都道府県名と県庁所在地名をベクトル空間上にマップした図を図 1 に示す．図 1 において，都道府県名及び県庁所在地名がそれぞれ近接している点，また各都道府県とそれに対応する県庁所在地のベクトルが平行に近くなっている点が見てとれる．

本研究では，地名を「ある特定の地域を一意に表す語」と定義する．この定義の元，下記の様な仮説をおく．

> NNLM によるベクトル空間上において，地名同士は近接する

このような仮説に基づき，「NNLM によるベクトル空間上で，ある語に近接している語の多くが地名である時，その語を地名とみなす」ことにより地名判定を行う．

3.1 定式化

本提案手法を定式化すると下記の様になる．ある語 w_i の単語ベクトルを $\overrightarrow{w_i}$ と定義する．$\overrightarrow{w_i}$ から見て n 番目に近接している語を，$\mathrm{Near}\,(\overrightarrow{w_i}, n)$ と表す．また，ある単語 w_i が地名であるか否かを判定する関数 Loc を下記の様に定義する．

$$\mathrm{Loc}(w_i) = \begin{cases} 1 & w_i\text{が地名のとき} \\ 0 & w_i\text{が地名でないとき} \end{cases} \tag{1}$$

3.2 地名判定

本稿では，具体的には 2 つの地名判定を提案する．

- 近接語の地名割合に基づく地名判定

  ある単語について，ベクトル空間内で近接する N 語に占める地名の割合が閾値 th を超えるとき，その単語を地名と判定する．定式化に基づいて説明すると，単語 w_i に近接する N 語が地名である割合 $R_{w_i,N}$ は下記の様に表される．

$$R_{w_i,N} = \frac{\sum_{n=1}^{N} \mathrm{Loc}\,(\mathrm{Near}(\overrightarrow{w_i}, n))}{N} \tag{2}$$

  このとき，$R_{w_i,N} \geq th$ 以上の時に，単語 w_i を地名と判定する．

- 単語クラスタに基づく地名判定

  まず，ある単語集合についてベクトル空間内でクラスタリングを行い，単語クラスタを生成する．そして各単語について，その単語が含まれるクラスタに地名が含まれる時にその単語を地名と判定する．本項では極力単純な手法を用いるために，最も一般的なクラスタリング手法の一つである K-means 法を用いる．

4. データセット

本研究で用いるデータセットについて説明する．それぞれ，NNLM の生成に用いたコーパス，地名判定用の既存単語辞書である．

4.1 word2vec による言語モデル

本稿では，2013 年の 1 月〜3 月までに投稿された日本語ツイートのうちユーザ単位で 10% サンプリングされたデータを約 4 億ツイートを NNLM の入力コーパスとした．これらは株式会社ホットリンク内に蓄積されているデータである．また，NNLM としては最も代表的な手法である word2vec を用いた．NNLM を生成する時のパラメータとしては，次元数を 200，ウィンドウサイズを 5 語に設定し，さらに，ネガティブサンリング，Hierarchical Softmax 関数を適用することとした．

4.2 地名辞書データ

本稿で用いた既存地名辞書について説明する．評価実験では，下記の辞書から抽出した地名リスト全てを等価に扱うものとする．

郵便番号辞書 郵便局により提供される全国都道府県の区市町村及び大字，小字までを記述した辞書 [郵便 15]．本項では，都道府県名，区市町村名，字をそれぞれ地名として用いた．

表 1: 高頻度語を用いた地名判定結果

dataset	Accuracy	Precision	Recall	F-score
山手線	-	1.00	0.93	-
頻出語 (4311/5000)	0.86 (89/706)	0.13 (89/144)	0.62	0.27

GSK 地名施設辞書第 2 版 言語資源協会 (GSK) で販売されている GSK2012-C GSK 地名施設名辞書第 2 版を使用した [言語 15]. 施設名辞書は日本国内の美術館，博物館，テーマパーク (遊園地) の合計 1,000 件について，名称，住所，異称，緯度・経度を記述した辞書である．このうち，名称，異称を地名として用いた．

Foursquare 地名辞書 Twitter Streaming API で取得したデータから，Foursquare 経由で投稿されたもののうち 1500 万ツイートを抽出し，そこに記載された地名，施設名を地名として使用した．

はてなキーワード はてなキーワード *2 の地名カテゴリに含まれるキーワードのうち，日本国内の緯度・経度情報をもつものを地名として利用した．

.

5. 評価実験

本論文で提案する 2 つのアプローチについてそれぞれ評価実験を行った．

5.1 近接語の地名割合に基づく地名判定

各語について，近接語の地名割合に基づいて地名判定を行った．すなわち，語 w_i がある時，$R_{w_i,N} \geq th$ を満たす時に w_i を地名と見なす．ここれは $N = 10, th = 0.1$ と設定した．$th = 0.1$ と低めに設定したのは，ある地名に近接する語が地名だとしても，その地名が既存地名辞書に含まれない場合も想定し，極力再現率を高めるためである．

地名判定候補としては 2 つの単語セットを用意した．一つは明らかに地名とわかる JR 山手線全 29 駅（「新宿」「代々木」など,「駅」は除く），もう一つはコーパスとした Twitter 本文中に出現する単語のうち，頻度上位 1001 位〜6000 位の単語，計 5000 語である．なお，頻度上位 1〜1000 位を除いたのは，これらは「http」「これ」「あれ」などの汎用的な語であり，地名は殆ど含まれないと考えたためである．評価実験結果を表 1 に示す. なお表 1 において，Accuracy は地名が地名，非地名が非地名と正しく判定された割合，Precision は地名と判定されたものに本当の地名が占める割合，Recall は，全地名のうち実際に地名と判定された割合，F-score は Precision と Recall の調和平均である．表 1 より，山手線，頻出語の結果において Recall は共に高く，地名を実際に地名と判定する割合は高い．しかし，地名で無い語を地名と判定してしまう割合も高い．これは閾値設定が低すぎるためだと思われる．しかし，多くの地名を収集するためには，閾値設定を低くする必要がある．結果として，本手法は実用的ではないと考えられる．

なお山手線で地名判定に失敗したのは「田端」「大塚」であった.「田端」は「田端でバタバタ」というハッシュタグが当時の Twitter ユーザの一部で流行してたために，その影響により「バタバタ」「バタ」「田端で」などのような単語が近接語に多く出現し，地名と判定されなかった．また「大塚」は近接語の多くに人名が含まれてしまったために，地名と判定されなかった．

*2 http://d.hatena.ne.jp/keyword/

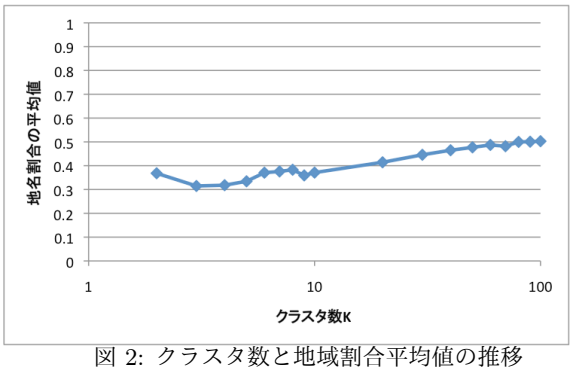

図 2: クラスタ数と地域割合平均値の推移

5.2 単語クラスタに基づく地名判定

次に単語クラスタに基づく地名判定を行った．本実験においては，クラスタリングを行うためにある程度の語数が必要である．そこで，山手線 29 駅と各駅の近接語 100 語を抽出し，そこから重複した語を除いて，全部で 1577 語を判定対象とした．クラスタリングの結果の評価方法としては，下記の様に行う．

1. 各クラスタに地名が含まれる割合（地名割合）を算出する
2. 各クラスタの地名割合が閾値 th より大きいクラスタを地名クラスタ，小さいクラスタを非地名クラスタとする
3. 全クラスタのうち，地名クラスタにあたるクラスタの地域割合の平均値 Avg_cl を算出する

このような地域割合の平均値 Avg_cl が高いほど，地名がよくまとまっていると考えられる．K-means 法におけるクラスタ数 K を 1〜10 まで 1 刻み，20〜100 まで 10 刻みで推移させ，各 K ごとに 10 回試行を行い，Avg_cl を平均した値を図 2 にプロットした．図 2 より $K = 50$ 程度で大体 Avg_cl が収束することがわかる．次に $K = 50$ での結果の一部を表 2 に示す．地名割合が 0.00 のクラスタでは，記号や氏名のみがクラスタに含まれている．地名割合が 0.80 のクラスタは当然殆ど地名，それも地理的に近い地名がまとまっている．また地名割合が 0.30〜0.50 程度のクラスタでも，地名辞書に含まれない地名がまとまっている．例えば，ID9 のクラスタでは，原宿，渋谷付近でのランドマークがまとまっており，ID10 のクラスタではコンサート会場がまとまっている（ZeppNamba，ダイホ＝ダイヤホールは名古屋のコンサートホールである）．このように地名割合が 0.5 以上はもちろんのこと，それよりも低くても，殆どは地名辞書に含まれない地名がまとまっていた．また，地名割合 0.1 未満のクラスタの殆どは，地名以外の語がまとまっていた．

つまり，地名割合が低いクラスタを除くことで，地名のみを抽出できる可能性があると言える．また，ダイホ（ダイヤホール），ララポ（ららぽーと），ドムジャ（イオンモール（旧ジャスコ）名古屋ドーム前 ）など，口語特有の表現も地名と判定できる可能性も示せた．さらに，同じクラスタ内には地理的に近い語がまとまっている場合も多く，緯度経度推定への応用も考えられる．

表 2: K=50 でのクラスタリング結果

ID	地名割合	クラスタ
0	0.80	泉岳寺 折尾 ひばりが丘 堀ノ内 京急川崎
1	0.20	秋葉原 サンシャインシティ UDX アソビットシティ オトトリズム ベルサール サンシャイン アルパ
5	0.00	森岡 深沢 上村 川原 岩瀬 小川 岩崎 良一 清水 吉田 寛和 村田
9	0.56	トーキョータイヤキ キディランド ラフォーレ キャットストリート パルコ グランフロント 代官山 原宿
10	0.35	ZeppNamba ダイホ マリンメッセ 代々木体育館 SDD フェスティバルホール 日本ガイシ 大阪城ホール
27	0.23	ヘップ ララポ 新大久保 鶴橋 ギュカル ウミエ ドムジャ
45	0.00	!★【～ !☆>【～ !◆>【～ 全席 【～ !◇【～ !★>【～ !◆» ◇>【～ ☆>

6. 終わりに

本稿では，NNLM を用いてツイートのような口語的な砕けた表現に対して，地名か否かを判定する 2 つのアプローチを試みた．

1 つ目においては，ごく単純なアプローチとして，NNLM によるベクトル空間内で判定対象表現と近接している語に既存地名辞書の語が存在しているか否かによって地名判定を行った．実用的な精度には達しなかったものの，単純な手法である程度の精度が得られることが分かった．2 つ目は，1 つ目のアプローチを発展させ，NNLM によるベクトル空間内で判定対象表現と近接している語群をクラスタリングし，地名となる語とそれ以外の語で異なるクラスタに含まれるか否かの分析を行った．結果として，クラスタ内の地名割合が低いクラスタを除くことで，地名のみを抽出できる可能性を示した．また，同じクラスタに含まれる語は地理的に近いことも多かったため，緯度経度の推定にも活用可能であるかもしれない．

今後はこれらのアプローチを詳細化し，精度を高めるとともにジオコーディングを行うために，各地名の緯度・経度情報を推定する手法を提案していきたい．

参考文献

[Backstrom 10] Backstrom, L., Sun, E., and Marlow, C.: Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity, in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 61–70, ACM Press (2010)

[Hecht 11] Hecht, B., Hong, L., Suh, B., and Chi, E. H.: Tweets from Justin Bieber's Heart: the Dynamics of the Location Field in User Profiles., in *Proceedings of the 2011 Annual Conference on Human factors in Computing Systems*, CHI '11, pp. 237–246, ACM Press (2011)

[Ji 09] Ji, R., Xie, X., Yao, H., and Ma, W.-Y.: Mining City Landmarks from Blogs by Graph Modeling, in *Proceedings of the Seventeen ACM International Conference on Multimedia*, MM '09, p. 105, ACM Press (2009)

[Lee 11a] Lee, C.-H., Yang, H.-C., Chien, T.-F., and Wen, W.-S.: A Novel Approach for Event Detection by Mining Spatio-Temporal Information on Microblogs, in *Proceedings of International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '11, pp. 254–259, IEEE (2011)

[Lee 11b] Lee, R., Wakamiya, S., and Sumiya, K.: Discovery of Unusual Regional Social Activities using Geotagged Microblogs, *World Wide Web*, Vol. 14, No. 4, pp. 321–349 (2011)

[Li 12] Li, R., Lei, K. H., Khadiwala, R., and Chang, K.-C.: TEDAS: A Twitter-based Event Detection and Analysis System, in *IEEE 28th International Conference on Data Engineering*, ICDE '12, pp. 1273–1276, IEEE (2012)

[Lin 04] Lin, J. and Halavais, A.: Mapping the Blogosphere in America, in *Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference*, Vol. 18 (2004)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed Representations of Words and Phrases and Their Compositionality, in *Advances in Neural Information Processing Systems*, pp. 3111–3119 (2013)

[Ritter 11] Ritter, A., Clark, S., Mausam, , and Etzioni, O.: Named Entity Recognition in Tweets: An Experimental Study, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 1524–1534, ACL (2011)

[Walther 13] Walther, M. and Kaisser, M.: Geo-spatial Event Eetection in the Twitter Stream, in *Proceedings of the 35th European conference on Advances in Information Retrieval*, ECIR'13, pp. 356–367, Springer-Verlag (2013)

[Watanabe 11] Watanabe, K., Ochi, M., Okabe, M., and Onai, R.: Jasmine: a Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs, in *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pp. 2541–2544, ACM (2011)

[言語 15] 言語資源協会：GSK 地名施設名辞書第 2 版, `http://www.gsk.or.jp/catalog/gsk2012-c/` (2015)

[郵便 15] 郵便局：郵便番号データ, `http://www.post.japanpost.jp/zipcode/download.html` (2015)