ARG WI2 No.15, 2019
WI2-2019-19

地域特有の埋め込み表現を用いたイベント参加地域の推定

小久保 千裕[†,a]　　小邦 将輝[††,b]　　関 洋平[†††,c]

† 筑波大学情報学群知識情報・図書館学類　†† 筑波大学大学院図書館情報メディア研究科
††† 筑波大学図書館情報メディア系

a) *s1611503@s.tsukuba.ac.jp*　b) *s1821613@s.tsukuba.ac.jp*　c) *yohei@slis.tsukuba.ac.jp*

概要　Twitter にはユーザが参加したイベントの感想や意見が投稿される．しかし，同じ名称のイベントが複数地域で開催されている場合，どの地域で開催されたイベントに対するツイートであるかの判別は難しい．本研究では，ツイートに現れる単語の地域差に着目し，地域住民のツイートを用いて地域特有の単語埋め込み表現を作成する．そして，作成した地域特有の単語埋め込み表現を用いてツイート投稿者のイベント参加地域を推定する手法を提案する．実験では，七夕まつりを対象として，提案手法を用いた際のツイート投稿者のイベント参加地域の推定性能を既存の単語埋め込み表現を用いた手法と比較する．実験の結果，提案手法の分類性能は，比較手法と同程度の F1 値を維持しつつ再現率が向上する傾向がみられ，比較手法では判別できなかった投稿の判別を行えることを確認した．

キーワード　地域バイアス，単語埋め込み表現，位置推定，Twitter，SNS

1　はじめに

Twitter や Facebook といったソーシャル・ネットワーキング・サービス (SNS) は，多くのユーザによって，日々の出来事や意見を発信する場として利用されている．そのため，SNS の投稿を分析することにより，人々の関心やある事柄に対する意見を得られる．各地で行われるイベントについても同様であり，イベント主催者は SNS を活用することで，アンケート調査等を行うのに比べ，安価かつ手軽に参加者の意見を入手できる．

しかし，梅まつりや七夕まつり，花火大会といった同じ名称のイベントが全国各地で開催される．このような場合，ある投稿が「七夕まつり」に関する投稿であったとしても，どの地域で開催される「七夕まつり」に対する投稿であるかを判別することは困難である．そこで，SNS 上の投稿から投稿者のイベント参加地域を推定することを本研究の目的とする．

本研究では，SNS の投稿に現れる単語の地域差に着目し，地域住民の SNS 投稿を用いて地域特有の単語埋め込み表現を作成し，それを基に投稿者のイベント参加地域を推定する手法を提案する．本論文では,「七夕まつり」に関するツイートを対象として，地域特有の埋め込み表現を抽出し，ツイート投稿者のイベント参加地域の推定精度を検証した結果を示す．

2　関連研究

Zola et al.[6] は，国ごとに用いられる単語の頻度や用法が異なる点に着目し，ツイートに用いられる名詞の頻度をもとに国レベルでのユーザの位置情報の推定を行った．本研究においても，地域によって用いられる単語の頻度や用法が異なると仮定し，ツイート投稿者のイベント参加地域を推定する．一方，本研究では，名詞だけでなく，動詞，形容詞や形容動詞も用いる．

Li et al.[2] は，ツイート中に含まれる地域特有の単語を基に，ユーザの位置を推定した． Li et al. は，ユーザの位置情報を推定する際に，位置情報の推定対象とするユーザのフォロワーの位置を同様に推定することで結果の改善を図った．本研究においても，Li et al. と同じく，地域特有の表現に着目するが，地域住民のツイートから地域特有の単語埋め込み表現を取得し，位置情報の推定を行う点で異なる．

本研究では，これらの研究とは異なり，ユーザの位置を推定するのではなく，その投稿がなされた位置を推定する．また，ツイート群から地域特有の単語埋め込み表現を取得する．そして，取得した地域特有の単語埋め込み表現を基に，ツイートの埋め込み表現を抽出し，ツイート投稿者のイベント参加地域を推定する．

3　提案手法

提案手法は，地域を考慮した単語埋め込み表現を取得する部分とイベントに関するツイートを分類する部分から構成される．まず，地域住民のツイートを用いて，その地域特有の単語埋め込み表現を取得し，得られた単語埋め込み表現を用いて，ツイートの埋め込み表現を抽出する．そして，抽出したツイートの埋め込み表現を基に，ツイート投稿者のイベント参加地域を分類する．

図 1 に示すとおり，地域別単語埋め込み表現を取得する際には，イベント開催地域の住民であると推定されるユーザ群から発信されたツイートおよびイベントに関する内容が含まれているジオタグなしツイートを用いる．

表 1: 収集したツイートと件数

ツイートの種類	訓練データ	テストデータ	合計
仙台市民	-	-	37,776 件
平塚市民	-	-	2,644 件
ジオタグなし七夕まつり（仙台）	-	-	3,853 件
ジオタグなし七夕まつり（平塚）	-	-	1,869 件
ジオタグ付き七夕まつり	4,012 件	326 件	4,338 件

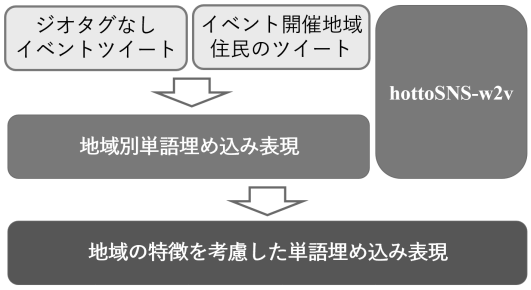

図 1: 地域の特徴を考慮した単語埋め込み表現

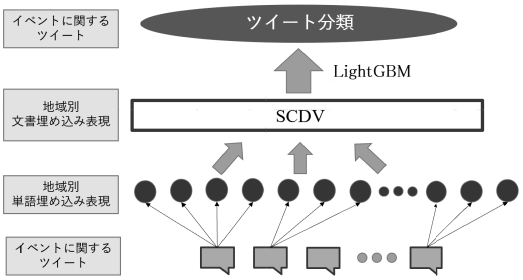

図 2: 実験の流れ

得られた地域別単語埋め込み表現と既存の単語埋め込み表現を組み合わせることで地域の特徴を考慮した単語埋め込み表現を作成する．なお，イベント開催地域の住民の収集方法については，4.1 節で述べる．また本研究では，既存の単語埋め込み表現として株式会社ホットリンクが公開している hottoSNS-w2v[1] を用いる．

続いて，イベント名をクエリとしてジオタグ付きツイートを収集し，収集したツイート群をイベント開催地域に紐づけて参加したイベントとみなす (例: 宮城県仙台市→仙台七夕まつり)．そして，地域特有の単語埋め込み表現をもとにツイートの埋め込み表現を抽出する．ツイートの埋め込み表現の抽出には，Mekala et al.[3] によって提案された Sparse Composite Document Vectors (SCDV) を用いる．本研究では，ここで抽出したツイートの埋め込み表現をもとに，ツイート投稿者のイベント参加地域を推定する．

4 実験

本章では，七夕まつりに関するツイートを対象として，ツイート投稿者のイベント参加地域の推定を行った実験 (図 2) について述べる．

[1]https://github.com/hottolink/hottoSNS-w2v.

4.1 実験データ

収集したツイートと件数を表 1 に示す．本実験では，仙台市民と平塚市民を対象として，地域住民のツイート収集を行った．まず，ツイプロ[2]を用いて，対象の市民であると判断されたユーザを収集する．得られたユーザのプロフィールを Twitter の Streaming API[3]を用いて取得し，Location と Description に記述されている内容から実際に対象の市民であるかを判定する．また，集められたユーザのフォロワーとフォロイーを取得し，同様に対象の市民であるかの判定を行い，ツイートの収集対象とするユーザの拡張を行った．なお，仙台市民と推定されるユーザについては，仙台で七夕まつりが開催される 8 月のツイートを収集した．この結果，リツイートとリプライを除き，合計 37,776 件のツイートを収集した．また，平塚市民と推定されるユーザについては，平塚で七夕まつりが開催されている 7 月のツイートを収集した．この結果，リツイートとリプライを除き，合計 2,644 件のツイートを収集した．

続いて，Twitter の Streaming API を用い，ジオタグ付きツイートの中からツイート本文中に「七夕まつり」もしくは「七夕祭り」という単語があるツイートを収集した．収集期間は 2015 年から 2019 年の 7 月，8 月であり，合計 4,338 件のツイートを収集した．収集したジオタグ付き七夕まつりツイートに，ジオタグに基づいて各ツイートに七夕まつりの開催地域を割り当てる．割り当てる際は国土交通省が提供している位置参照情報[4]を利用した．全 4,338 件の七夕まつり開催地域のうちツイートが 120 件以上あった「宮城県仙台市」,「神奈川県平塚市」,「愛知県一宮市」,「愛知県安城市」,「東京都杉並区」,「東京都福生市」の 6 市区と「その他の都市」でラベルを付与した．これらのツイートのうち 2015 年から 2018 年までの合計 4,012 ツイートを単語埋め込み表現を取得する際の訓練データ，2019 年の合計 326 ツイートをツイートの埋め込み表現を用いた分類の有効性をはかるテストデータとして扱う．

ジオタグ付き七夕まつりツイートと同様に Twitter の Streaming API を用いて，ツイート本文中に「七夕まつり」もしくは「七夕祭り」という単語があるツイートを

[2]https://twpro.jp
[3]https://developer.twitter.com/en.html
[4]http://nlftp.mlit.go.jp/isj/index.html

表 2: 分類結果：hottoSNS-w2v

	Precision	Recall	F1-score
宮城県仙台市	0.900	0.818	0.857
その他の都市	0.972	0.986	0.979
Macro avg.	0.936	0.902	0.918
Micro avg.	0.962	0.963	0.962

表 3: 分類結果：hottoSNS-w2v + Sendai

	Precision	Recall	F1-score
宮城県仙台市	0.833	0.909	0.870
その他の都市	0.986	0.971	0.979
Macro avg.	0.909	0.940	0.924
Micro avg.	0.965	0.963	0.954

表 4: 分類結果：hottoSNS-w2v

	Precision	Recall	F1-score
神奈川県平塚市	0.975	0.795	0.876
その他の都市	0.965	0.996	0.980
Macro avg.	0.970	0.896	0.928
Micro avg.	0.967	0.966	0.965

表 5: 分類結果：hottoSNS-w2v + Hiratsuka

	Precision	Recall	F1-score
神奈川県平塚市	0.975	0.816	0.889
その他の都市	0.968	0.996	0.982
Macro avg.	0.972	0.906	0.935
Micro avg.	0.969	0.969	0.968

収集した．収集期間は 2015 年から 2019 年の 7 月，8 月であり，そのうち同じツイート内に「仙台」という単語が含まれる合計 3,853 件のツイートを抽出した．また，同じくツイート内に「平塚」という単語が含まれる合計 1,869 件のツイートを抽出した．

4.2 実験方法

実験は以下の手順で行う．

1. 地域の特徴を考慮した単語埋め込み表現取得
2. ツイートの埋め込み表現取得
3. ツイートの分類
4. 分類結果の比較・評価

まず，対象市民のツイートとジオタグなし七夕まつりツイートを用いて，Word2Vec[5] を使用して地域別単語埋め込み表現を取得する．得られた単語埋め込み表現をその地域に特有の単語埋め込み表現であると考え，既存の単語埋め込み表現である hottoSNS-w2v と組み合わせることで地域の特徴を考慮した単語埋め込み表現を作る．単語埋め込み表現を組み合わせる際には，既存の単語埋め込み表現である hottoSNS-w2v に出現しなかった語彙を加えている．本実験では対象地域を仙台市と平塚市とし，両都市それぞれについて地域の特徴を考慮した単語埋め込み表現を作成した．

次にラベリングしたジオタグ付き七夕まつりツイートを用いてツイートの埋め込み表現を取得する．今回は「仙台市の単語埋め込み表現」が「仙台七夕まつり」の分類に有効か,「平塚市の単語埋め込み表現」が「湘南ひらつか七夕まつり」の分類に有効かを調査するため，ジオタグ付き七夕まつりツイートについて該当都市以外の 5 都市区は「その他の都市」として再度ラベルを付与した．ツイート本文を MeCab[1] により形態素解析し，得られた単語に 1. で取得した単語埋め込み表現を割り当てる．Sparse Composite Document Vectors（SCDV）を使用し，訓練データからツイートの埋め込み表現を取得する．SCDV は単語ベクトルから文書ベクトルを抽出する際，ベクトル空間をクラスタリングし，そのクラスタに基づいて文書ベクトルを構築する．そのため，単語埋め込み表現取得の際に得られた特徴と訓練データの特徴を生かしてツイートの埋め込み表現を取得できる．

3. では取得したツイートの埋め込み表現をもとに，テストデータの分類をおこなう．Mekala et al[3]. は，分類に Support Vector Machine を使用していたが，分類速度を考慮して，本実験では LightGBM[4] を使用して分類を行う．

本実験では，ホットリンク社の単語埋め込み表現モデルのみを用いた分類結果を比較手法として用い，本研究の提案手法である地域の特徴を考慮した単語埋め込み表現（hottoSNS-w2v+Sendai, hottoSNS-w2v+Hiratsuka）を用いた分類結果との比較・評価を行う．

4.3 実験結果

hottoSNS-w2v を用いた仙台七夕まつりの分類結果を表 2 に，仙台市の単語埋め込み表現 Sendai を用いた分類結果を表 3 に示す．また，hottoSNS-w2v を用いた湘南ひらつか七夕まつりの分類結果を表 4 に，平塚市の単語埋め込み表現 Hiratsuka を用いた分類結果を表 5 に示す．

hottoSNS-w2v+Sendai を用いた分類では，Precision は比較手法に劣るものの，Recall，F1-score ではマクロ平均値において，比較手法を上回る結果となった．hottoSNS-w2v+Hiratsuka を用いた分類でも同様に，Recall，F1-score で分類性能の向上がみられた．

5 考察

提案手法では Recall が優れた結果を示している．既存の単語埋め込み表現では得られなかった何らかの特徴を地域別単語の埋め込み表現を加えることで得られたと考えられる．

提案手法で分類できたが，比較手法では分類できなかったツイートの例を表 6 に示す．ツイート番号 274 番 118 番のツイートでは,「仙台」,「平塚」や「七夕祭り」

表 6: 提案手法でのみ検出できた七夕まつりツイートの例

ツイート番号	ラベル	ツイートの分かち書き
274	宮城県仙台市	夢のひとつ 叶う めちゃくちゃ 綺麗 七夕祭り 夏祭り 仙台 花火 浴衣デート かん ちゃん 浴衣 似合う 着崩れ 気 なる すぎる 俺氏 やる 崩れる の 不明 場
322	宮城県仙台市	仙台 七夕祭り 美容室 アクセル 本町 ホテル法華クラブ ショート アレンジ 場所
18	神奈川県平塚市	3 年ぶり 位 平塚 七夕祭り 行く 来る 七夕 飾り 他 地区 電飾 無し ある 夕方 夜 見る 方 綺麗 場所
118	神奈川県平塚市	先日 平塚 七夕まつり 行う みる 平日 かかわる 大勢 人 来場 する いる 関東 三 大祭 言う れる ある 感じ 場所 ひらつか 七夕祭り

という語が出現しているのにも関わらず，他の語が多いため比較手法ではその地域であると判定されなかった例である．322 番の「本町」とは仙台市内の地名であり，「本町商店街」が存在する．七夕飾りが飾られる商店街ではないが，七夕まつり開催期間中に豪華な七夕飾りが飾られる通りの近くである．また，湘南ひらつか七夕まつりでは夕刻になると七夕飾りに取り付けられた電飾でライトアップされ，これが湘南ひらつか七夕まつりの特徴である．18 番にあらわれる「電飾」は，他の地域では特徴として扱われないが，平塚市の特徴が出ている語であると考えられる．

そして,「仙台」や「平塚」などの対象とした地名が含まれていながらその地域のツイートではない投稿に関しても判別ができている例がある．したがって，ただ地名をもとに分類しているわけではないことが窺える．

このように，比較手法では判別できなかった地域の特徴が提案手法では考慮できているといえる．

6 おわりに

今回の実験では，ホットリンク社のモデル hottoSNS-w2v に地域別の単語埋め込み表現を加えることにより，hottoSNS-w2v のみを単語埋め込み表現として用いた比較手法よりも F1-score において同程度，再現率において上回る分類結果が得られることが確認できた．今後は，地域の特徴を落とさずにイベント中の特徴を捉えるために，イベントに関する内容が出現しやすいと考えられるイベント開催日前後のツイートのみを扱うことで出現する単語の煩雑さを抑えることを検討する．ユーザの拡張や収集期間を検討し，語彙数を増やす必要もあると考えられる．また，今回は地域の特徴を考慮した単語埋め込み表現を取得する際に，hottoSNS-w2v に出現しなかった語を加えた．得られた地域の特徴を考慮した単語埋め込み表現をどのように既存の単語埋め込み表現に加えるのかは検討する必要がある．

今回は仙台と平塚の七夕まつりに限定して実験を行ったが，七夕まつりは全国各地で行われている．しかし，地域住民の少ない地域の場合，得られる開催地域住民のツイートも更に少なくなってしまうことが考えられる．そのため，得られたツイートが少ない場合でも地域の特徴を反映できるように，手法を改善していく必要がある．

謝辞

本研究の一部は，科学研究費補助金基盤研究 B（課題番号 19H04420）の助成を受けて遂行された．

本研究では，株式会社ホットリンクから提供された日本語大規模 SNS+Web コーパスによる単語分散表現モデルを使用した．ここに深く感謝いたします．

参考文献

[1] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto. Apply-ing Conditional Random Fields to Japanese Morphological Analysis. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP2004). Association for Computa-tional Linguistics, 2004, p. 230-237.

[2] Chuanyang Li，Xiuqin Lin，Bin Wu，and Chuan Shi. Location Inference Using Microblog Text and Friendships. Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). 2014，p. 778 -784.

[3] Dheeraj Mekala，Vivek Gupta，Bhargavi Paranjape，Harish Karnick. SCDV: Sparse Composite Document Vectors using soft clustering over distributional representations. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Association for Computational Linguistics，2017，p. 659-669.

[4] Qi Meng，Goulin Ke，Taifeng Wang，Wei Chen，Qiwei Ye，Zhi-Ming Ma，and Tie-Yan Liu. 2016. A Communication-Efficient Parallel Algorithm for Decision Tree. In Advances in Neural Information Processing Systems. 1271–1279.

[5] Tomas Mikolov,Tlya Sutskever, Kai Chen, Greg Corrado, and Jeffery Dean. Distributed Representations of Words and Phrases and Their Compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS 2013). 2013, p. 3111-3119.

[6] Paola Zola，Paulo Cortez，and Maurizio Carpita. Twitter user geolocation using web country noun searches. Decision Support Systems. 2019，vol. 120，p.50-59.