DEIM2020 J3-2(day1 p65)

イベント参加地域推定のための単語埋め込み表現の拡張

小久保千裕† 小邦 将輝†† 関 洋平†††

† 筑波大学情報学群知識情報・図書館学類 〒 305–8550 茨城県つくば市春日 1 丁目 2 番地
†† 筑波大学大学院図書館情報メディア研究科 〒 305–8550 茨城県つくば市春日 1 丁目 2 番地
††† 筑波大学図書館情報メディア系 〒 305–8550 茨城県つくば市春日 1 丁目 2 番地
E-mail: †s1611503@s.tsukuba.ac.jp, ††m-oguni@klis.tsukuba.ac.jp, †††yohei@slis.tsukuba.ac.jp

あらまし 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．イベント開催地域住民のツイートを用いて地域別の単語埋め込み表現を作成し，既存の単語埋め込み表現を拡張することで地域の特徴を反映した単語埋め込み表現を作成し，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精度を評価し，既存の単語埋め込み表現のみを用いた手法と比較する．実験の結果，提案手法の分類精度は，比較手法に比べてマクロ平均値において F1 値が上回る結果が得られることを確認した．また，地域別の単語埋め込み表現を作成する際はツイートと共に地域 Web ページを用いることで，分類精度が向上することを確認した．

キーワード 地域推定，単語埋め込み表現，Twitter，SNS

1 はじめに

1.1 本研究の目的

本研究では，Twitter[1]の投稿から，投稿したユーザが参加したイベントの開催地域を推定するために，単語埋め込み表現を拡張することを目的とする．同じ単語であっても，投稿される地域や期間によってその語の使われ方や文脈が異なると，単語埋め込み表現も変化する．そのため，コーパスの違いが単語埋め込み表現の違いになり，特徴となる．そこで，地域ごとにツイートや Web ページ等の地域リソースを収集し，地域別単語埋め込み表現を作成することで，地域リソースに出現する語の地域差を取得する．また，地域リソースにイベントに関するツイートを含めることで，イベント特有の単語の扱われ方も入手できると考える．

本研究では，イベント開催地域の住民のツイート等の地域リソースを用いて地域別単語埋め込み表現を作成し，既存の単語埋め込み表現を地域別単語埋め込み表現を用いて拡張することで地域の特徴を反映した単語埋め込み表現（以下，地域特有の単語埋め込み表現）を作成する．地域特有の単語埋め込み表現を用いることで投稿者のイベント参加地域を推定する手法を提案する．

1.2 本研究の背景

Twitter や Facebook[2]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利用されており，日々の出来事や意見が発信されている．SNS ユーザは自身が参加したイベント等についての感想や意見を投稿することが多々ある．SNS では各個人の評価や嗜好が表れるため，SNS の投稿を分析することは，人々の関心やある事柄に対する意見の入手に繋がるといえる．

一方，様々なイベントの主催者は，イベント参加者の意見を SNS の投稿から得ることにより，要望や改善点を発見することができる．しかし,「梅まつり」,「七夕まつり」や「花火大会」といった同じ名称のイベントは全国各地で開催されている．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．

本研究では SNS のうち Twitter に着目する．Twitter は世界中の多くの人に利用されており [3]，イベントに関する意見や感想などが投稿される．

ツイートに位置情報タグを付け，自らの位置を公開することもできるが，位置情報を公開するユーザは非常に少なく，位置情報付きのツイートは全体に対して非常に少ない [7], [11]．そのため，ジオタグに依らない位置情報の推定とツイートの分析が必要である．

本稿ではイベントとして「七夕まつり」に着目する.「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作成し，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

本研究では，Twitter の投稿であるツイートを使用して，地域に特徴的な単語埋め込み表現を取得し，それ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．単語埋め込み表現を組み合わせることに着目した研究および，SNS を使用したユー

1 : https://twitter.com
2 : https://www.facebook.com/
3 : https://investor.twitterinc.com/financial-information/annual-reports/

ザの位置情報推定についての研究について述べる．

2.1 単語埋め込み表現を組み合わせることに着目した研究

単語埋め込み表現は通常広大なコーパスから語と語の関連性を学習し，高い次元のベクトルで単語を表現する．したがって，単語埋め込みの学習機構が異なると，同じ単語であってもそれぞれ付与されているベクトルの表す意味は異なってしまう．

Yin et al. [8] は，様々な埋め込み表現モデルのパフォーマンスの違いから，それぞれの単語埋め込み表現を組み合わせることで，多様なタスクに対応できる単語埋め込みを学習できると考え，その手法を提案した．ここでは，「メタ埋め込み」が存在すると仮定し，いくつかの語彙の単語埋め込みとの変換を学習することで新たな単語埋め込みを構築した．

Sarma et al. [1] は，ドメインに固有の単語埋め込みが対象ドメインからのデータでのみ訓練されることに着目し，ドメイン固有の埋め込み表現の特異性を組み合わせる方法を提案した．Domain Adapted（DA）と呼ばれる単語埋め込みは，対応する単語埋め込みを整列させることによる相関を分析することで形成される．この結果，DA は分類タスクにおいて汎用の単語埋め込みとドメイン固有の埋め込みの両方を上回ることを示した．

本研究では，地域住民のツイート群から取得した地域別の単語埋め込み表現と既存の単語埋め込み表現のモデルを組み合わせる事によって，地域の特徴を反映した単語埋め込み表現を作る．これらの手法を参考にしながら，異なる学習機構から作られた単語埋め込み表現を組み合わせることで，課題に適応できると考える．

2.2 SNS を使用したユーザの位置情報推定についての研究

SNS ユーザは自分の身近な事柄について投稿することが多い．そのため，ユーザの投稿を分析し，投稿に現れる単語の傾向からそのユーザの位置を推定する手法が提案されている．Zola et al. [9] 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，Twitter の投稿である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．Google Trend に現れる単語を用いることで，ユーザが興味のある国を推定することができ，またそのユーザの位置情報の推定につなげている．

一方，Li et al. [3] は，マイクロブログ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．地域特有の単語とは，その単語が地域に由来するものであり，また，その地域にアイデンティティのあるユーザ群が他の単語よりもその単語を使用する可能性が高い語である．さらに，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．

これらの研究は，投稿中の単語とユーザのプロフィール情報から，ユーザの位置を推定したものである．ここで扱われている位置情報は，ユーザの主な活動地域を表している．本研究では，ユーザの主な活動地域ではなく，特定のイベントに限定するが，ツイートを投稿した位置を直接推定する．

SNS における投稿は，他の文書に比べて一文が短く，文章数も少ないことが多い．そのため，ただ単語の頻度を分析するだけでは，出現する単語が非常に少ないために位置の推定が難しい．そこで，単語そのものではなく，単語や文書の埋め込み表現を用いてユーザの位置情報を推定する研究がなされている．奥村ら [10] の研究では，ツイートから位置推定をする際の埋め込み表現は，FastText，Doc2Vec に比べて，Latent semantic Indexing（LSI）が最も良好な結果が出た．単語埋め込み表現を用いることで，ツイート分類タスクにおいて良好な結果が得られる．本研究では，地域に特有の単語埋め込み表現を作成し，それを用いることでイベントに関するツイートの埋め込み表現を取得し，ツイートを分類することでユーザ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．提案手法の概要を図 1 に示す．地域イベントについて，イベント名を手がかりとしてツイートを収集し，ツイート群の特徴から該当ツイートを投稿したユーザが参加したイベントの参加地域を推定する．ツイートの特徴を取得する際には，単語埋め込み表現を用いる．

本研究では，単語の扱われ方の地域差に注目する．まず，イベント開催地域ごとにツイートを収集し，収集したツイート群から地域別の単語埋め込み表現を構築する．このとき，ツイートは 1 件の投稿につき最大 140 字と文字数が制限されているため，単語埋め込み表現モデルを訓練する際のデータが十分とは言えない．そこで，地域の Web ページを訓練対象に加えることで，単語埋め込み表現の質を改善し，参加地域推定の精度向上を目指す．

最後に，既存の単語埋め込み表現モデルと地域別単語埋め込み表現を組み合わせることで，手がかりとなる語彙の不足を解消し，地域の特徴を考慮した単語埋め込み表現モデルを構築する．そして，得られた単語埋め込み表現をもとにツイートの埋め込み表現を構築し，ツイートの埋め込み表現に基づいて発信したユーザの参加地域を推定する．

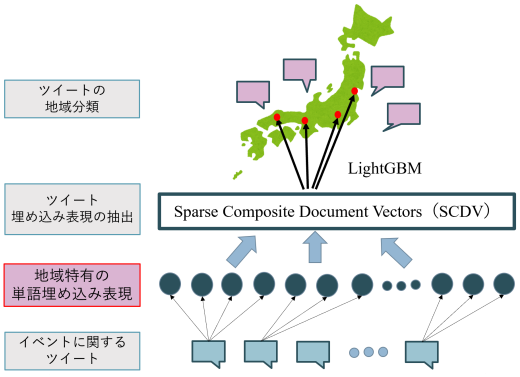

図 1 提案手法の概要

3.1 地域を考慮した単語埋め込み表現の構築

まず，イベント開催地域の住民のツイート，開催地域のイベントに関するツイートと開催地域の Web ページを用いて，地域別の単語埋め込み表現を取得する．得られた地域別単語埋め込み表現を用いて既存の単語埋め込み表現を拡張することで地域特有の単語埋め込み表現を構築する（図 2）．

なお，イベント開催地域の住民のツイート等の収集方法については，4.2 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[4] を用いる．hottoSNS-w2v は日本語 SNS，日本語 Wikipedia などの大規模なコーパスから学習された単語埋め込み表現モデルである．本研究では hottoSNS-w2v を地域別単語埋め込み表現を用いて拡張することで分類性能の向上を目指す．

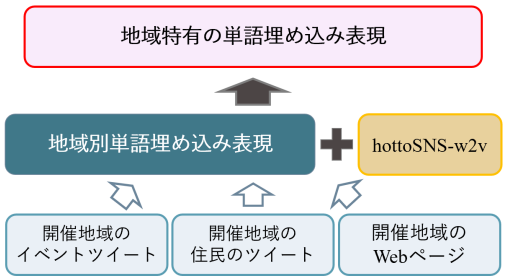

図 2　地域の特徴を考慮した単語埋め込み表現

hottoSNS-w2v を拡張する際は，地域別単語埋め込み表現に存在する語彙のうち hottoSNS-w2v には存在しない語彙に着目する．地域別単語埋め込み表現を用いて，hottoSNS-w2v には存在しない新出単語の類似語をコサイン類似度により抽出し，その中から hottoSNS-w2v の語彙集合に含まれている語を選択する．類似度の閾値は，予備実験から 0.8 とした．

選択した語のベクトルを hottoSNS-w2v から取り出し，新出単語のベクトルとしてそのまま付与する（図 3）．図中の W_x は地域別単語埋め込みにしか存在しない語彙を表し，追加対象語彙である．また，W_y は地域別単語埋め込み表現中においての W_x の類似語であり，hottoSNS-w2v の語彙集合に存在する語句を表している．以上の手続きにより，地域別単語埋め込み表現を使って既存の単語埋め込み表現である hottoSNS-w2v を拡張し，地域の特徴を考慮した単語埋め込み表現を構築する．

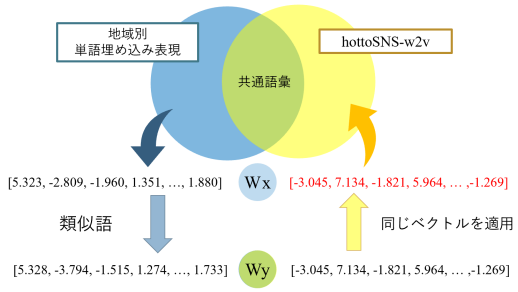

図 3　2 つの単語埋め込み表現の融合

4：https://github.com/hottolink/hottoSNS-w2v.

3.2 イベントに関するツイートの地域分類

まず，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する．収集したツイート群に付与されているジオタグ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，神奈川県平塚市→湘南ひらつか七夕まつり)．

そして，前節で取得した地域特有の単語埋め込み表現を利用して，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. [4] によって提案された Sparse Composite Document Vectors (SCDV) を用いる．まず，Mekala et al. の手法に基づき，単語埋め込み表現をクラスタリングし，そのクラスタに基づいて単語埋め込み表現を再構築する．さらに，再構築された単語埋め込み表現をツイートの各単語に割り当てることでツイートに対する埋め込み表現を抽出する．本研究では，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．分類器は，決定木アルゴリズムに基づいた勾配ブースティングのフレームワークである Light GBM [5] を用いる．

4 実験: 地域の特徴を考慮した単語埋め込み表現を用いたツイート分類

本実験では，提案する地域の特徴を考慮した単語埋め込み表現を用いて，ツイートの投稿者が参加したイベントの開催地域の推定タスクに対して，その有効性を検証する．具体的なイベントとしては，七夕まつりを対象とする．

4.1 実 験 方 法

本稿では,「七夕まつり」をイベント名とし，その中でも「仙台七夕まつり」と「湘南ひらつか七夕まつり」に注目することで，仙台市と平塚市の地域別単語埋め込み表現を取得し，仙台市と平塚市の地域の特徴を考慮した単語埋め込み表現の構築を行う．評価用データのツイートは，本文に「七夕まつり」または「七夕祭り」を含むジオタグ付きツイートを収集し，地域の特徴を考慮した単語埋め込み表現を利用して，ツイートを分類し，分類の F1 値，精度，再現率を評価する．以下に詳細を示す．

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

まず，(1) では，イベント開催地域住民のツイートと開催地域の七夕まつりツイート，開催地域の Web ページを用いて，Word2Vec [6] を使用して地域別単語埋め込み表現を構築する．仙台市と平塚市について構築した地域別単語埋め込み表現は，これ以降の論文でそれぞれ Sendai，Hiratsuka として参照する．また，既存の単語埋め込み表現である hottoSNS-w2v と地域別単語埋め込み表現とを組み合わせることで，地域特有の単語埋め込み表現を構築する．組み合わせる際には，地域別単語埋め込み表現にしか存在しない語彙に着目し，その類似語のベクトルを割り当てる．

次に，(2) では，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．形態素解析器として MeCab [2]，辞書として NEologd [12] を使用してツイート本文を形態素解析し，得られた単語に (1) で構築した単語埋め込み表現を割り当てる．さらに，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．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は「その他の都市」としてラベルを付与した．本実験では，3.4 節で述べた通り，LightGBM [5] を使用して分類を行う．

(4) では，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）を用いた分類結果を，hottoSNS-w2v のみを用いた分類結果と比較し，F1 値，精度，再現率で評価する．

4.2 実験データ

収集したデータと件数を表 1，表 2 に示す．収集したデータは，以下のとおりである．

（1） ジオタグ付きイベントツイート（分類タスク評価用）
（2） イベント開催地域住民のツイート
（3） 開催地域のイベントツイート
（4） 開催地域の Web ページ

これら 4 種類のデータそれぞれの収集方法と前処理について，述べる．なお，品詞や地名の判定には，形態素解析器として MeCab，辞書として NEologd を使用した．

まず，ツイートの地域分類タスクの評価のために，ジオタグ付きイベントツイートを収集した．収集には Twitter の Streaming API を用いた．また，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．なお，収集期間は 2015 年から 2019 年の 7 月，8 月である．次に，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た．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 [5] を利用した．この投稿位置をユーザのイベントの参加地域とみなす．収集したツイートに対象地域である「宮城県仙台市」と,「神奈川県平塚市」の 2 都市と「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 年のツイートをツイートの埋め込み表現を用いた地域分類の有効性を評価するテストデータとした．

また，以下では地域別の単語埋め込み表現を構築する上で収集した 3 種類のコーパスについて説明する．まず，イベント開催地域住民のツイートについては，仙台市民と平塚市民を対象として，ツイプロ [6] を用いて，対象地域の住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[7] 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．Location はユーザの所在地を記載するフィールドであり，Discription は自らの紹介を自由に記述できるフィールドである．仙台市民と判定されたユーザについては，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，平塚市民と判定されたユーザについては，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．ツイート群のうち，安藤ら [13] の手法を参考に bot や宣伝と思われるツイートは除去し，また，リツイートも除去した．

次に，開催地域のイベントツイートについては，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」「平塚」という単語が含まれるツイートをそれぞれ抽出し，地域のイベントツイートとした．地域のイベントツイートも同様に bot や宣伝と思われるツイートとリツイートは除去した．

最後に，開催地域の Web ページについては，検索エンジンの Google[8] を利用して，都市名をクエリとしてヒットした上位 150 件の Web ページを使用した．収集した Web ページを URL をもとにスクレイピングし，本文のみを取り出す．また，収集したページの内容のうち名詞が 8 割を超える場合は除去した．

4.3 実験結果

提案手法である仙台市の特徴を考慮した単語埋め込み表現 hottoSNS-w2v+Sendai を用いた分類結果を表 3 に，比較手法である hottoSNS-w2v を用いた仙台七夕まつりの分類結果を表 4 に示す．また，提案手法である平塚市の特徴を考慮した単語埋め込み表現 hottoSNS-w2v+Hiratsuka を用いた分類結果を表 5 に，比較手法の hottoSNS-w2v を用いた湘南ひらつか七夕まつりの分類結果を表 6 に示す．

hottoSNS-w2v+Sendai を用いた分類（表 3）では，F1 値, 精度，再現率 の全てについて，マクロ平均値が比較手法を上回る結果となった．hottoSNS-w2v+Hiratsuka を用いた分類（表 5）では僅かであるが，hottoSNS-w2v+Sendai の場合と同様に，F1 値, 精度，再現率 で分類性能の向上がみられた．

4.4 考　　察

提案手法では，比較手法で扱った hottoSNS-w2v に地域の特徴によって語彙を追加した．語彙を追加する際は，追加する語彙に類似している語を算出し，そのベクトルを割り当てた．そ

5 : http://nlftp.mlit.go.jp/isj/index.html

6 : https://twpro.jp
7 : https://developer.twitter.com/en.html
8 : https://www.google.co.jp/

表 1　データセット：地域分類タスクの評価

データの種類	訓練データ	テストデータ	合計
ジオタグ付き七夕まつりツイート	4,014 件	501 件	4,515 件

表 2　データセット：地域別単語埋め込み表現の構築

データの種類	宮城県仙台市	神奈川県平塚市	合計
開催地域住民のツイート	19,563 件	2,583 件	22,146 件
開催地域のイベントツイート	3,305 件	1,676 件	4,981 件
開催地域の Web ページ	133 件	124 件	257 件

表 3　分類結果：hottoSNS-w2v+Sendai

	F1 値	精度	再現率
宮城県仙台市	0.913	0.985	0.852
その他の都市	0.984	0.972	0.997
Macro avg.	0.949	0.978	0.924

表 4　分類結果：hottoSNS-w2v

	F1 値	精度	再現率
宮城県仙台市	0.851	0.940	0.778
その他の都市	0.974	0.959	0.990
Macro avg.	0.912	0.949	0.884

表 5　分類結果：hottoSNS-w2v+Hiratsuka

	F1 値	精度	再現率
神奈川県平塚市	0.869	0.984	0.778
その他の都市	0.977	0.959	0.996
Macro avg.	0.923	0.972	0.887

表 6　分類結果：hottoSNS-w2v

	F1 値	精度	再現率
神奈川県平塚市	0.861	0.994	0.765
その他の都市	0.977	0.957	0.987
Macro avg.	0.919	0.970	0.882

の結果，SCDV で文書埋め込み表現を取得する際に変化があったと考えられる．SCDV はクラスタを作成し，そのクラスタリング結果を考慮して新たな文字埋め込み表現を作成する文埋め込み表現取得方法である．単語埋め込み表現を拡張させることで，クラスタリングに影響を与えることができる．

表 7 は hottoSNS-w2v+Sendai の単語埋め込み表現のクラスタリング結果である．“仙台のツイート”と判断されたツイートに出現する語彙が多く含まれるクラスタが形成されていることから，提案手法ではツイートの埋め込み表現を取得する際に地域性を考慮できていると考えられる．

提案手法のみで抽出できた七夕まつりツイートの例を表 8 に示す．ツイート番号 440 に出現する「瑞鳳殿」と「ナイト」という語彙は hottoSNS-w2v のみでクラスタリングした場合，異なるクラスタとなる．しかし，hottoSNS-w2v+Sendai では同じクラスタに分類されている．このように，語彙を追加する事によってクラスタリング結果に変化を与えていると言える．

また，hottoSNS-w2v+Sendai と hottoSNS-w2v+Hiratsuka の違いは地域別単語埋め込み表現を作成する際のコーパスサイズである．より多くのツイートを使うことで単語埋め込み表現モデルをより頑強なものにできる．提案手法では，新出単語にベクトルを付与する際，ベクトルの類似度をもとに対応付けできると考えられる語彙を算出した．そのため，作成される地域別単語埋め込み表現が結果を左右すると考えられる．地域別単語埋め込み表現を取得する際の，地域の Web ページの有効性とコーパスサイズについては，5 章の実験を通して検証を行う．

5　実験：開催地域の Web ページの有効性の検証

提案手法では，地域別単語埋め込み表現を作成する際，イベント開催地域住民のツイートと開催地域の Web ページをコーパスとして利用している．本実験では，開催地域の Web ページとイベント開催地域住民のツイートの両方を用いた場合とイベント開催地域住民のツイートのみを用いた場合の分類性能の比較を行い，開催地域の Web ページをコーパスの一部とすることの有効性を調査する．

5.1　実験方法とデータ

提案手法では，地域の特徴を掴みつつ，単語埋め込み表現を強化するという目的で開催地域の Web ページを単語埋め込み表現の作成に用いた．本実験では，イベント開催地域住民のツイートのみを使って単語埋め込み表現を作成した場合と開催地域の Web ページのみを使って単語埋め込み表現を作成した場合のイベントツイート分類を行なう．結果をイベント開催地域住民のツイートと開催地域の Web ページの両方を使って単語埋め込み表現を作成した場合とを比較する．

単語埋め込み表現を作成する際に使うデータは，4 章と同じものを扱う（表 2）．また，本実験において，対象とするイベントは「仙台七夕まつり」と「湘南ひらつか七夕まつり」とする．本実験では，仙台のイベント開催地域住民のツイートのみを使った単語埋め込み表現（SenTweet，HiraTweet）と開催地域の Web ページのみを使った単語埋め込み表現（SenWeb，HiraWeb）を構築した．それぞれについて，既存の単語埋め込み表現と組み合わせた埋め込み表現を作成する．表 1 のデータを用いて地域分類をし，F1 値, 精度, 再現率 により評価する．

5.2　実 験 結 果

hottoSNS-w2v+SenTweet を用いた分類結果を表 9 に，hottoSNS-w2v+HiraTweet を用いた分類結果を表 10 に示す．また，hottoSNS-w2v+SenWeb を用いた分類結果を表 11 に，hottoSNS-w2v+HiraWeb を用いた分類結果を表 12 に示す．

hottoSNS-w2v+SenTweet（表 9）を用いた分類では，

表 7　hottoSNS-w2v+Sendai のクラスタリング結果

クラスタ番号	語彙数	追加語彙数	語彙の例
0	449	98	遊園地，映画鑑賞，時代背景，七
1	1,086	189	ショッピングセンター，ハム，朝ごはん
:	:	:	:
22	4,470	800	災害時，仙台駅東口，七夕，折鶴
:	:	:	:
49	870	165	ポスト，クルトン，悲観，クルトン

表 8　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
440	宮城県仙台市	今日から 3 日間仙台は七夕祭り 夜は瑞鳳殿がライトアップされていると言うので来てみた #仙台七夕#瑞鳳殿七夕ナイト #瑞鳳殿
47	神奈川県平塚市	ちょうど 1 年ぶりに平塚へ。七夕祭り、朝早く行ったのでまだ空いてました。 平塚八幡宮も参拝。

表 9　分類結果：hottoSNS-w2v+SenTweet

	F1 値	精度	再現率
宮城県仙台市	0.872	0.956	0.802
その他の都市	0.977	0.963	0.992
Macro avg.	0.925	0.959	0.897

表 10　分類結果：hottoSNS-w2v+HiraTweet

	F1 値	精度	再現率
神奈川県平塚市	0.845	0.984	0.740
その他の都市	0.974	0.952	0.997
Macro avg.	0.910	0.968	0.869

表 11　分類結果：hottoSNS-w2v+SenWeb

	F1 値	精度	再現率
宮城県仙台市	0.896	0.945	0.851
その他の都市	0.981	0.972	0.990
Macro avg.	0.939	0.958	0.921

表 12　分類結果：hottoSNS-w2v+HiraWeb

	F1 値	精度	再現率
神奈川県平塚市	0.853	0.984	0.753
その他の都市	0.976	0.954	0.997
Macro avg.	0.914	0.969	0.875

hottoSNS-w2v のみを使った分類（表 4）と比べて F1 値, 精度，再現率の全てにおいて，マクロ平均値が比較手法を上回る結果となった．hottoSNS-w2v+Sendai を使った分類（表 3）と比べると，F1 値, 精度，再現率の全てにおいて SenTweet が下回る結果となった．hottoSNS-w2v+HiraTweet を用いた分類（表 10）では hottoSNS-w2v のみを使った分類（表 6），hottoSNS-w2v+Sendai を使った分類（表 5）の双方に対して，F1 値, 精度，再現率 が下回る結果となった．

hottoSNS-w2v+SenWeb を用いた分類（表 11）では，提案手法である hottoSNS-w2v+Sendai を用いた分類（表 3）と同程度の分類結果を示している．hottoSNS-w2v+HiraWeb を用いた分類（表 12）では，比較手法として用いた hottoSNS-w2v のみを用いた分類（表 6）より結果が悪いが，hottoSNS-w2v+HiraTweet を用いた分類（表 10）よりも僅かに良い結果を示している．

5.3 考　　察

仙台七夕まつりを判定する場合では，hottoSNS-w2v+Sendai を用いた分類（表 3）と hottoSNS-w2v+SenWeb を用いた分類（表 11）は同程度の分類結果ではあるが，hottoSNS-w2v のみを用いた分類（表 4）と hottoSNS-w2v+SenTweet 用いた分類（表 9）と比べて良い結果を示している，開催地域の Web ページをコーパスとして利用することで分類性能を向上させたといえる．

一方で，hottoSNS-w2v+HiraWeb を用いた分類（表 12）と hottoSNS-w2v+HiraTweet を用いた分類（表 10）では結果が向上しなかった．表 13 は, それぞれのコーパスについてその語彙数を示したものであり, 括弧内は語彙のうち hottoSNS-w2v には出現していない語彙数である．すなわち，括弧内の語彙数が今回の手法で hottoSNS-w2v に追加対象とする単語の数である．HiraWeb について作成された単語埋め込み表現を確認すると，HiraWeb から hottoSNS-w2v へ加えた語彙は 726 語であり，hottoSNS-w2v の語彙 2,067,629 語に比べて非常に少ないため，2 つの単語埋め込み表現を組み合わせた効果が得られなかったと考えられる．

6　お わ り に

Twitter などの SNS ではユーザが参加したイベントに対する感想や意見などが投稿されることが多いため，イベントの主催者はその投稿をみることでイベントに対する意見を収集することができる．しかし，同じような名称のイベントが全国各地で開催されている場合，どの投稿が自分の開催したイベントについてのものであるか判断するのは難しい．そこで，本研究では，Twitter のデータを使用し，各地域ごとの特徴を捉えることで投稿がどの地域で行われたイベントについてのものであるかを推定した．

提案手法は，イベント開催地域ごとに収集したユーザのツイート，地域イベントに関するツイートと地域 Web ページを用いて，地域別単語埋め込み表現を作成し，既存の単語埋め込

表 13　コーパスの種類と語彙数

	仙台市	平塚市
ツイート＋ Web ページ	48,356 語（9,113 語）	16,413 語（2,653 語）
ツイートのみ	41,386 語（7,483 語）	13,054 語（1,953 語）
Web ページのみ	12,408 語（1,988 語）	4,674 語（762 語）

全語彙数　（うち固有の語彙数）

み表現モデルである hottoSNS-w2v と組み合わせて利用することで，イベント名が出現しているツイートを各開催地域に分類する手法を提案した．単語埋め込み表現を組み合わせる際は hottoSNS-w2v に存在しない未知語を地域別単語埋め込み表現から加えた．ツイートに埋め込み表現を付与する際は SCDV を使用した．

宮城県仙台市と神奈川県平塚市を対象とし，「七夕まつり」について言及しているツイートを用いて，提案手法の有効性を検証するためにイベント参加地域の分類実験をおこなった．また，地域別単語埋め込み表現間の差による分類結果への影響を調べるために，地域リソースの比較実験をおこなった．

今回の実験では，hottoSNS-w2v に地域別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 値 において上回る分類結果が得られることが確認できた．また，地域別の単語埋め込み表現を作成する際はツイートのみをコーパスとして使用するよりも，地域 Web ページを共に用いることで，分類結果が向上することが確認できた．2 つの単語埋め込み表現を組み合わせる際に地域単語埋め込み表現から類似語を算出するため，地域別単語埋め込み表現の精度が良いことが一つの条件であるとわかった．

今回は「仙台七夕まつり」と「湘南ひらつか七夕まつり」の２つの地域に着目したが，七夕まつりは全国各地で行われている．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．

また，単語埋め込み表現の組み合わせ方は検討が必要である．ただ未知語に類似語のベクトルを適用させるだけではなく，未知語と類似語のベクトルの違いを考慮してベクトルを付与できるとよい．

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．

文　　献

[1] Prathusha Kameswara Sarma, Yingyu Liang, and Bill Sethares. Domain adapted word embeddings for improved sentiment classification. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pp. 51–59, Melbourne, July 2018. Association for Computational Linguistics.

[2] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[3] Chuanyang Li, Xiuqin LIn, Bin Wu, and Chuan Shi. Location inference using microblog text and friendships. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM 2014, pp. 778–784. IEEE, 2014.

[4] Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape, and Harish Karnick. SCDV : Sparse composite document vectors using soft clustering over distributional representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 659–669, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[5] Qi Meng, Guolin Ke, Taifeng Wang, Wei Chen, Qiwei Ye, Zhi-Ming Ma, and Tie-Yan Liu. A communication-efficient parallel algorithm for decision tree. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pp. 1279–1287. Curran Associates, Inc., 2016.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

[7] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[8] Wenpeng Yin and Hinrich Schütze. Learning word meta-embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1351–1360, Berlin, Germany, August 2016. Association for Computational Linguistics.

[9] Paola Zola, Paulo Cortez, and Maurizio Carpita. Twitter user geolocation using web country noun searches. *Decision Support Systems*, Vol. 120, pp. 50 – 59, 2019.

[10] 奥村貴俊, 彌冨仁. Tweet 解析によるユーザ埋め込み表現を用いた都道府県レベルでの位置推定. 言語処理学会第 24 回年次大会, pp. 73–76. 言語処理学会, 2018.

[11] 橋本康弘, 岡瑞起. 都市におけるジオタグ付きツイートの統計（特集人と環境に見る高次元データフローの生成と解析）. 人工知能学会誌, Vol. 27, No. 4, pp. 424–431, 2012.

[12] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き辞書 mecab-inpadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会, pp. 875–878. 言語処理学会, 2017.

[13] 安藤有生, 関洋平. 市民のツイートを用いた分散表現に基づく地名に対する市民の関心の傾向の可視化. 知能と情報, Vol. 30, No. 6, pp. 804–814, 2018.