

Word2Vec を用いた地域やランドマークの意味演算

土田 崇仁[†] 遠藤 雅樹^{††,†††} 加藤 大受^{††,††††} 江原 遙^{††} 廣田 雅春^{†††††}
横山 昌平^{††††††} 石川 博^{††}

[†] 首都大学東京 システムデザイン学部 〒191-0065 東京都日野市旭が丘 6-6

^{††} 首都大学東京大学院 システムデザイン研究科 〒191-0065 東京都日野市旭が丘 6-6

^{†††} 職業能力開発総合大学校 基盤ものづくり系 〒187-0035 東京都小平市小川西町 2-32-1

^{††††} ウイングアーク 1 s t 株式会社 〒150-0031 東京都渋谷区桜丘町 20-1 渋谷インフォスター

^{†††††} 大分工業高等専門学校 情報工学科 〒870-0152 大分県大分市大字牧 1666

^{††††††} 静岡大学 情報学部 〒432-8011 静岡県浜松市中区城北 3-5-1

E-mail: [†]tsuchida-takahito@ed.tmu.ac.jp, ^{†††}endou@uitec.ac.jp, ^{††††}kato.d@wingarc.com,

^{††}{ehara,ishikawa-hiroshi}@tmu.ac.jp, ^{†††††}m-hirota@oita-ct.ac.jp, ^{††††††}yokoyama@inf.shizuoka.ac.jp

あらまし Word2Vec とは、ニューラルネットワークを用いた言語モデルであり、「類似した文脈で利用される単語は、類似した意味を持つ」という分布仮説に基づき、単語をベクトルによって表現する定量化手法である。これにより、「東京－日本＋フランス＝パリ」のように、単語の持つ意味を加算・減算することが可能である。本研究では、この意味演算を観光分野に応用する。具体的には、東京のスカイツリーに対応する大阪のランドマークや、横浜のような港湾都市に最も近い意味を持つ関西の都市はどこか、といったような実行例を用いて、都市・地域とランドマークの意味的な関係性を捉え、ある地域のあるランドマークに対応する他地域のランドマークの抽出や、都市・地域に特徴的なランドマークを加算・減算した際、新たな知見を得ることを目指す。また、Word2Vec を用いた単語間の意味演算の有用性を検証する。

キーワード Word2Vec, テキストマイニング, 観光情報, マイクロブログ

1. はじめに

近年、SNS（ソーシャル・ネットワーキング・サービス）の急速な普及と共に、インターネット上には日常的に膨大な量の情報が発信されている。特に、代表的なマイクロブログサービスである Twitter ^(注1) は、その手軽さから国内における利用者も多く、発信された多種多様な内容の情報を活用する研究が盛んに行われている。

Twitter には、一度に投稿が可能な文字数に 140 文字という制限があり、ユーザが体験したことや、感じたことをスマートフォンやタブレット端末を用いて何処からでも気軽に投稿が可能であるという特徴がある。そのため、地域に密着した情報や、その地域を実際に訪れたユーザの感想が記載された多数の投稿があり、Twitter から観光スポットの持つ特徴や、その観光スポットを訪れる人の傾向などの観光情報を抽出する研究は盛んに行われている [1], [2], [3]。

本論文では、Twitter に投稿されたテキストに対して Word2Vec ^(注2) を適用し、観光分野に応用が可能な情報の抽出を目指す。Word2Vec とは、Tomas Mikolov ら [4] によって 2013 年に公開されたツールであり、コーパス（大規模なテキストデータ集合）を入力として与えることで、語彙空間を学習し

単語の分散表現を取得する。これにより、「東京－日本＋フランス＝パリ」のように、単語の持つ意味に対して、加算や減算などの演算を行う事が可能である。

本研究では、この意味演算を用いて、ツイートから都市・地域とランドマークの意味的な関係性を抽出する。ランドマークとは、ある特定の地域を象徴する建造物や街並み、事象のことを指し、東京都のスカイツリーや大阪市の通天閣、札幌市の時計台などが例として挙げられる。

また地域を象徴する事象の例としては、青森市のねぶた祭りや徳島市の阿波踊などが挙げられ、地域を代表する祭事もランドマークの一種であると言える。そこで、本論文では、観光の目的となる建造物や祭事をランドマークと定義する。しかし、地域とランドマークの関係性を考慮した際、ランドマークという括りの中にも、意味的な違いは存在していると考えられる。例えば、東京都においてスカイツリーが持つ意味に対して、大阪市において通天閣が持つ意味と、札幌市において時計台が持つ意味とではどちらが近い意味を持つのかということを考えたとき、その地を訪れる人の目的や建造物自体の外観から、スカイツリーと時計台よりもスカイツリーと通天閣の方が互いに近い意味を持っていると考えられる。このような判定を機械的な手法によって実現するため、Word2Vec を用いて都市・地域やランドマークの意味演算を行う。また、この意味演算の有用性を検証する。

以下、本論文の構成を説明する。2 章では、本研究の関連研

(注1) : <http://twitter.com/>

(注2) : <http://code.google.com/p/word2vec/>

究について説明し、本研究の位置付けを明白にする。3章では、提案手法について説明し、4章では実験結果を示し考察を行う。最後に、5章でまとめと今後の課題について説明する。

2. 関連研究

2.1 観光情報の抽出に関する研究

我々の知る限り、Word2Vecを用いて観光情報の抽出を行った研究はない。一方、TwitterなどのSNSから観光情報の抽出を行っている研究は多数あるので、これらを解説する。

石野ら[5]は、機械学習を用いてブログ集合から旅行に関するブログエントリのみを自動検出し、観光情報の抽出を行っている。具体的には、ブログテキストから、地域名と土産物の対と、地域名と建造物の対を抽出している。その結果、高い適合率、再現率が得られており、旅行ブログエントリが観光情報の有益な情報源であることを示している。

また、長谷川ら[6]は、観光地域に関する記述の時間的遷移を明らかにするため、Twitterから地域特徴語辞書を構築する手法を提案している。具体的には、地域名との共起を利用して建造物の特徴語をツイートから抽出し、さらに時間、空間の連続性を考慮することで、時期ごとの変化や、周辺地域との関連性を捉えた観光地域の特徴語辞書を生成する手法を提案している。その結果、キーワード検索や一定期間の全ツイートを取り出す検索より、観光体験の高精度な組織化が可能になることを検証している。

以上に記述した研究では、テキスト中において地域名と共起する単語を用いて観光情報を抽出している。本論文では、この共起関係を用いて都市・地域とランドマークの意味的な関係性を取得することで観光分野への応用が可能な情報の抽出を目指す。

2.2 Word2Vec

Word2Vecとは、Tomas Mikolovら[4]によって提唱されたニューラルネットワークを用いた、単語の分散表現の計算手法である。これはskip-gramモデルと呼ばれる言語モデルによって単語の分散表現を取得する。具体的には、コーパス中の文脈において特定の単語と共起する単語、すなわち付近に存在する単語の出現確率を最大にするようなパラメータ調整を行うことで、単語の密なベクトルを生成する。以後、単語 w のベクトルを $Vec(w)$ と表記する。このモデルによって構築された語彙のベクトル空間では、類似した意味を持つ単語のベクトル同士は近く、相違した意味を持つ単語のベクトル同士は遠くに割り当てられ、単語間の距離計算や単語の持つ意味の加算・減算が可能である。例えば、 $Vec(\text{“東京”}) - Vec(\text{“日本”}) + Vec(\text{“フランス”}) \approx Vec(\text{“パリ”})$ のような演算が成り立つ。これは「東京」、すなわち「日本の首都」という概念から「日本」という国の概念を減算し、「フランス」という国の概念を加算することで「パリ」、すなわち「フランスの首都」という概念を算出している。本論文では、この意味演算を用いてマイクロブログに投稿されたテキストから観光情報の抽出を試みる。

2.3 Word2Vecを用いた研究

近年、Word2Vecを用いた研究は盛んに行われている。野沢

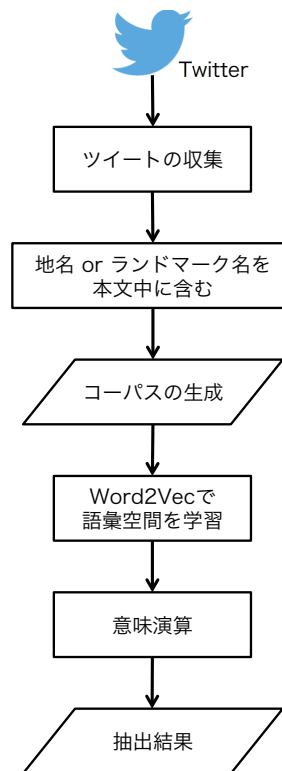


図1 抽出手順

ら[7]は、レシピ投稿検索サイトに投稿されたレシピデータに含まれる、調理手順が記載されたテキストを用いてコーパスを作成し、Word2Vecによって語彙空間を学習することで、特定の食材に対する他の食材の類似度を算出している。そして、その算出結果を用いて「ベーコン→ウィンナー」や「牛乳→豆乳」のように代替が可能な食材のペアを発見している。また、調理手順からコーパスを作成する際、調理ステップの分割手法や単語の抽出方法を変えた複数の言語モデルを用いることで適切なコーパス生成手法を評価している。

また、久保田ら[8]は、Word2Vecを用いた手法と係り受け構造を用いた手法によって単語の概念的類似度を抽出し、両手法と人手で作成された既存の意味辞書との一致率を比較することで、日本語におけるWord2Vecの有用性を定量的に評価している。その結果、両手法共にコーパスの影響を受けたとしても、約一割程度の一致率が確認され、さらにWord2Vecを用いた類似度算出手法では、より日本人の分類感覚に近い単語が上位に上がることが確認されている。

本論文では、Word2Vecの持つ大きな機能である単語間の意味演算を用いて都市・地域とランドマークの意味的な関係性の抽出を行うと共に、この意味演算の有用性を検証する。

3. 提案手法

本章では、Twitterに投稿されたテキストに対してWord2Vecを適用し、都市・地域とランドマークの意味的な関係性を抽出する手法について説明する。提案手法の概要を図1に示す。

3.1 提案手法の概要

はじめに、TwitterAPI^(注3)を用いてツイートを収集する。次に、Word2Vecを用いて語集空間を学習し、都市・地域やランドマークの名称のベクトルを生成する。しかし、前述した通りTwitter上には膨大なツイートが存在する。そのため、無作為に収集したツイートをWord2Vecに学習させるのではなく、その中から都市・地域やランドマークの名称を本文中に含むツイートを抽出し、これらの名称を多く含むコーパスを学習させることで、有意義なベクトルの生成を図る。最後に、生成したベクトルを用いて意味演算を行い、都市・地域とランドマークの意味的な関係性を抽出する。以下、3.2節で都市・地域やランドマークの名称を含むキーワードリストを作成する手法について説明した後、3.3節でツイートを収集する手法とコーパスを作成する手法について説明する。

3.2 都市・地域やランドマークの名称を含む

キーワードリストの作成

本研究では、平成26年から国土交通省が提供している観光資源データ^(注4)を用いて、都市・地域やランドマークの名称を含むキーワードリストを作成する。後に、このキーワードリストに含まれるキーワードを用いて、ツイートの抽出を行う(3.3節)。観光資源データは「観光資源名」や「所在地住所」などのカラムによって構成されており、全国各地の建造物や祭事、名物などが収録されている。収録されているデータの例を表1に示す。その中の「種別名称」カラムには、「建造物」や「食」、「年中行事」など、観光資源の大まかな種別が記載されている。本研究では、観光の目的となる建造物や街並み、祭事を表すと考えられる、「神社・寺院・教会」、「建造物」、「郷土景観」、「集落・街」、「城跡・城郭・宮殿」、「庭園・公園」、「史跡」、「動植物園・水族館」、「年中行事」となっている観光資源を対象とする。

次に、対象とした観光資源名を形態素解析し、抽出された名詞をキーワードリストに追加する。観光資源名を形態素解析する理由は、観光資源の正式名称ではなく、通称でツイートを抽出するためである。例えば、観光資源データには、「東京スカイツリー」のように正式名称で記載されているが、通称は「スカイツリー」であり、形態素解析すると「東京」と「スカイツリー」に分割される。そのため、「スカイツリー」でツイートを抽出することが可能になる。

また、「所在地住所」カラム内のデータから市区町村名を抽出しキーワードリストに追加する。都道府県名についても、キーワードリストに追加する。また、Wikipediaのランドマークに関する記事^(注5)や、TripAdvisor^(注6)を参照することで、京都タワーや神戸ポートタワーなど、主要なランドマークであると言えるが観光資源データに収録されていなかったものに関しては、観光資源名を手で抽出し、キーワードリストに追加する。

表1 観光資源データの例

観光資源名	都道府県コード	種別名称	所在地住所
摩周湖	01	湖沼	弟子屈町
八甲田山	02	山岳	青森市
秋田竿燈まつり	05	年中行事	秋田市旭北
伊香保温泉	10	温泉	渋川市
東京スカイツリー	13	建造物	墨田区押上 1-1-2
箱根駅伝	13	芸能・興行・イベント	千代田区
兼六園	17	庭園・公園	金沢市兼六町 1
青木ヶ原の樹海	19	植物	富士河口湖町精進
善光寺	20	神社・寺院・教会	長野市元善町 491
川床の京料理	26	食	京都市上京区
道頓堀	27	集落・街	大阪市中央区
軍艦島	42	郷土景観	長崎市高島町端島

3.3 ツイートの抽出とコーパスの作成

3.2節で作成したキーワードリストを用いてツイートの抽出を行う。具体的には、キーワードリストに含まれる都市・地域やランドマークの名称を本文中に少なくとも1つ含むツイートを全て抽出する。抽出したツイートをコーパスに用いて実験を行う。

4. 実験

4.1 データセット

本研究では、2015年3月11日から同年10月28日までの期間に日本国内で投稿されたジオタグ付きツイートを収集した。その結果、収集されたツイートは約1,005,610,000件であった。さらにこの中から対象とするツイートを抽出しコーパスを作成する。また、ツイートを抽出する際、国土交通省が提供している観光資源データを用いる。

4.2 ツイートの抽出結果と正規化

収集したツイートの中から、都市・地域やランドマークの名称を含むツイートを抽出した。その結果、約115,610,000件のツイートが抽出された。また、都市・地域やランドマークの名称を含まないツイートを同件数、抽出した。

実験を行う前に、抽出したツイートに対して、正規化の処理を施した。具体的には、アルファベット、名詞に含まれない数字、「@」、「#」などの記号を除去した。また、「!」、「?」、「♪」、「☆」など、文章の末尾に添えられることの多い記号は、全て「。」に置換した。さらに、「!!!」のように同じ記号が連続して用いられている場合、「。。」に置換するのではなく、「。」に置換した。

4.3 Word2Vecの設定

Word2Vecを用いてベクトル空間を構築する際、モデル構築のためのハイパーパラメータを設定する必要がある。ハイパーパラメータの設定により、意味演算の精度は変化する[9]。まず、生成するベクトルの次元数は400次元に設定し、モデル構築と正規化手法には、Word2Vecがデフォルトとして設定しているskip-gramモデルと階層的ソフトマックスをそれぞれ用いる。

また、実験において、windowをベクトル空間構築の際のパラメータとする。windowの説明図を図2に示す。このパラメータは、コーパス中の文脈において着目する単語の前後何単語ま

(注3) : <http://apps.twitter.com>

(注4) : http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-P12-v2_2.html

(注5) : <http://ja.wikipedia.org/wiki/ランドマーク>

(注6) : <http://www.tripadvisor.jp/>

海外 / に / 向け / た / 適切 / な / 情報
 w_{t-3} w_{t-2} w_{t-1} w_t w_{t+1} w_{t+2} w_{t+3}
← window →

図2 windowの説明図. この例の場合, その値は2.

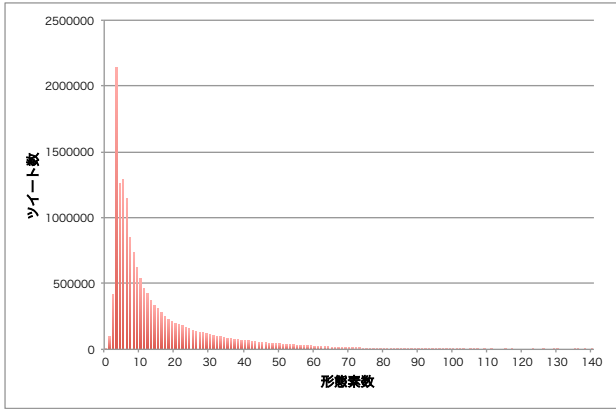


図3 形態素数に対するツイート数

でを考慮するのかを決定する. すなわち, window 値が大きいくほど, 着目する単語から離れた単語を考慮した学習を行うことが可能だが, 学習に要する時間は長くなる. ベクトル間の類似度はコサイン類似度を用いて算出する.

4.3.1 window の設定

window は, 考慮する文脈の長さを決定するハイパーパラメータであり, その値を window 値とする. window 値は適切に設定する必要がある. また, 上限が存在せず, 大きな値を設定することも可能ではあるが, 過剰に大きな window 値は意味演算の精度を悪化させることが考えられる.

過剰に大きな window 値が, 意味演算の精度を悪化させる理由を簡単に説明する. 1章で述べたように, Twitter には, 一度に投稿が可能な文字数に 140 文字という制限があり, ユーザは簡潔な文章を投稿する傾向にある. 例えば, 「スカイツリーからの夜景めっちゃ綺麗だったなあー. はあー明日からまた学校ですよ。」のようなツイートがあったとする. このツイートを形態素解析すると, 「スカイツリー/から/の/夜景/めっちゃ/綺麗/だっ/た/なあ-/././はあ/ー/明日/から/また/学校/です/よ./。」となる. このツイートは, 一文目と二文目で, 一切関係のない事柄について言及している. このとき, window 値を 20 に設定すると, 「スカイツリー」のベクトルを生成する際, 二文目の単語を全て考慮することになってしまう. そのため, Twitter の投稿からベクトル空間を構築する際, window 値の設定は重要であると考えられる.

そこで, window 値に上限値 L を設定し, L 個のベクトル空間を構築する. そして, それぞれについて意味演算を行い, その精度を評価することで, window の最適値を検証する.

まず, L を設定するため, 都市・地域やランドマークの名称を含むツイートを全て形態素解析し, 形態素数を計数した. その結果を図3に示す. 形態素数の平均値は 14 であり, L をこの値より十分に大きい 20 に設定する.

次に, 意味演算の精度を評価する手法について述べる. 本

表2 ベンチマークの例

$Vec(\text{“東京タワー”}) - Vec(\text{“東京”}) + Vec(\text{“京都”}) \simeq Vec(\text{“京都タワー”})$
$Vec(\text{“東照宮”}) - Vec(\text{“日光”}) + Vec(\text{“伊勢”}) \simeq Vec(\text{“伊勢神宮”})$
$Vec(\text{“兼六園”}) - Vec(\text{“金沢”}) + Vec(\text{“東京”}) \simeq Vec(\text{“六義園”})$
$Vec(\text{“春日大社”}) - Vec(\text{“奈良”}) + Vec(\text{“出雲”}) \simeq Vec(\text{“出雲大社”})$
$Vec(\text{“姫路城”}) - Vec(\text{“姫路”}) + Vec(\text{“京都”}) \simeq Vec(\text{“二条城”})$

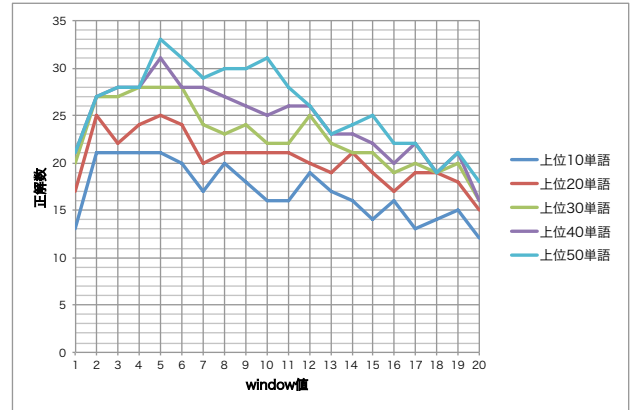


図4 window 値に対する正解数

表3 $Vec(\text{“スカイツリー”}) - Vec(\text{“東京”}) + Vec(\text{“大阪”})$ の演算結果

順位	演算結果	類似度
1	彦根城	0.323
2	通天閣	0.317
3	浅草	0.311
4	隅田川	0.310
5	マリンタワー	0.309
6	ハルカス	0.292
7	丸亀城	0.292
8	ライトアップ	0.291
9	三社祭	0.284
10	名古屋城	0.283

論文では, $Vec(\text{“東京タワー”}) - Vec(\text{“東京”}) + Vec(\text{“京都”}) \simeq Vec(\text{“京都タワー”})$ のような演算例を 5 人の協力者と共に人手で定め, その中から 60 例の演算を選出した. これをベンチマーク^(注7)として評価に用いる. ベンチマークの例を表2に示す.

まず, 左辺の演算をそれぞれのベクトル空間で行い, 演算結果のベクトルに近いベクトルを持つ単語を出力する. そして, 右辺の単語が上位 10 単語以内に出力された演算を正解とし, その数を計数した. 同様に, 上位 20, 30, 40, 50 単語以内に出力された演算を正解とし, その数を計数した. その結果を図4に示す.

検証の結果, window 値が 20 のときに正解数が最も少ないことから, 大きな window 値が意味演算の精度を悪化させることが確認できた. また, window 値が 1 のときから 5 のときにかけて, 意味演算の精度が向上しており, 6 以降は低下していくことが確認できた. 以上のことから, window 値を 5 に設定し, 実験を行う.

5. 実験結果と考察

はじめに, $Vec(\text{“スカイツリー”}) - Vec(\text{“東京”}) + Vec(\text{“大$

(注7): 同類物を比較, 評価する際に用いる, 数量的, 又は質的な性質.

表 4 $Vec(\text{スポット名}) - Vec(\text{“横浜”}) + Vec(\text{“神戸”})$ の演算結果

スポット名 順位	ベイブリッジ		横浜中華街		みなとみらい		マリインタワー	
	演算結果	類似度	演算結果	類似度	演算結果	類似度	演算結果	類似度
1	明石海峡大橋	0.465	南京町	0.542	三ノ宮	0.566	明石海峡大橋	0.498
2	明石大橋	0.434	三ノ宮	0.476	三宮	0.505	三ノ宮	0.477
3	三ノ宮	0.429	三宮	0.439	南京町	0.494	ジャーマンビアフエスティバル	0.448
4	神戸タワー	0.427	異人館	0.394	須磨	0.469	インフィオラータ神戸	0.447
5	ジャーマンビアフエスティバル	0.400	板宿	0.390	異人館	0.468	南京町	0.434
6	三宮	0.373	王子動物園	0.383	旧居留地	0.461	異人館	0.431
7	神戸駅	0.362	中華街	0.381	北野	0.456	神戸タワー	0.415
8	神戸大橋	0.357	帝廟	0.372	ハーバーランド	0.454	三宮	0.415
9	神戸港	0.357	ジャーマンビアフエスティバル	0.369	ジャーマンビアフエスティバル	0.452	神戸駅	0.411
10	旧ハンター住宅	0.355	神戸駅	0.361	桃山台	0.451	ルアンエール	0.405

阪)」という意味演算を行った。演算結果を表 3 に示す。演算の結果、大阪に存在するランドマークが二つ出力された。一つ目は、2 位に出力された「通天閣」である。「スカイツリー」、「通天閣」はそれぞれ、東京と大阪を代表する展望施設であり、人気の観光スポットである。

二つ目は、6 位に出力された「ハルカス」である。これは大阪市阿倍野区に存在する、「あべのハルカス」を指す単語であると考えられる。「スカイツリー」は、2012 年 5 月 22 日に開業された東京都を代表する建造物であり、日本一高い電波塔である^(注8)。また、「あべのハルカス」は、2014 年 3 月 7 日に開業された大阪府を代表する建造物であり、日本一の高層ビルである^(注9)。このように、二つのスポットは近年開業されたばかりの建造物であり、建造物の種類ごとに、それぞれ日本一の高さであることで有名である。また、「スカイツリー」には「ソラマチ」、「あべのハルカス」には「近鉄本店」という大型ショッピングモールが併設されており、観光者の訪問の目的も近いことが考えられる。

しかし、1 位の類似度が 0.323 と低く、「通天閣」は 0.317、「ハルカス」は 0.292 という値を示している。そのため、東京における「スカイツリー」に最も近い大阪のランドマークは、「通天閣」や「あべのハルカス」であるが、その類似度は高くなく、と考えられる。

次に、横浜、神戸という二つの港湾都市に着目し、意味演算を行った。具体的には、 $Vec(\text{スポット名}) - Vec(\text{“横浜”}) + Vec(\text{“神戸”})$ という式中のスポット名に、横浜に存在するランドマークを代入し、神戸に存在するランドマークの抽出を試みた。演算結果を表 4 に示す。はじめに、「ベイブリッジ」を代入すると、1 位に「明石海峡大橋」が出力された。「ベイブリッジ」と「明石海峡大橋」は、それぞれの都市を代表する吊り橋である。次に、「横浜中華街」を代入すると、1 位に「南京町」が出力された。「横浜中華街」と「南京町」は、それぞれの都市を代表する中華街である。次に、「みなとみらい」を代入すると、8 位に「ハーバーランド」が出力された。「みなとみらい」は、横浜市西区と同市中区にまたがる再開発地区であり、「ハーバーランド」は神戸市中央区にある再開発地区である。二つの再開発地区は、共に沿岸地域に存在し、東の「みなとみらい」、西の「ハーバー

表 5 $Vec(\text{“ジャーマンビアフエスティバル”}) - Vec(\text{“神戸”}) + Vec(\text{“横浜”})$ の演算結果

順位	演算結果	類似度
1	赤レンガ倉庫	0.604
2	山崎千裕	0.591
3	川崎	0.587
4	磯子区	0.586
5	レジデントアーティスト	0.578
6	ランドマークタワー	0.573
7	ラチッタ	0.572
8	プティステーグリハーサルバレエレッスン	0.568
9	ポニカバー	0.562
10	栄区	0.554
11	ヨコハマサイクルスタイル	0.541
12	グルメンタ	0.539
13	ヨコハマフューリングスフェスト	0.539
14	コスモワールド	0.537
15	みなとみらい	0.518
16	井上陽水	0.508
17	氷川丸	0.500
18	横浜赤レンガ倉庫	0.496
19	ナミヒラユコ	0.486
20	ワールドポーターズ	0.485

ランド」と呼ばれ、注目を集めていた。次に、「マリインタワー」を代入すると、7 位に「神戸タワー」が出力された。これは神戸市中央区に存在する、「神戸ポートタワー」を指す単語であると考えられる。「マリインタワー」、「神戸ポートタワー」はそれぞれ、横浜と神戸を代表する展望施設であり、人気の観光スポットである。しかし、「神戸タワー」は、「ベイブリッジ」を代入したときに 4 位に出力されており、原因として、「神戸タワー」という単語の持つ意味をベクトルで正しく表現できていないことが考えられる。以上の演算結果から、横浜、神戸には、複数の類似するランドマークが存在していることが分かった。

次に、上述した四つの演算結果において、共通して出力された、「ジャーマンビアフエスティバル」に着目し、 $Vec(\text{“ジャーマンビアフエスティバル”}) - Vec(\text{“神戸”}) + Vec(\text{“横浜”})$ という意味演算を行った。演算結果を表 5 に示す。演算の結果、13 位に「ヨコハマフューリングスフェスト」が出力された。「ジャーマンビアフエスティバル」は、神戸ハーバーランドで開催されるオクトーバーフェスト^(注10)であり、「ヨコハマフューリングスフェスト」は、横浜赤レンガ倉庫で開催されるオクトーバー

(注8) : <http://www.tokyo-skytree.jp>

(注9) : <http://www.abenoharukas-300.jp>

(注10) : ドイツのバイエルン州ミュンヘン市で開催されている祭事。近年、日本においても開催されている。

フェストである。このように、地域を代表するイベントを抽出することも可能であることが分かった。

以上の結果から、ある地域のあるランドマークに対応する他地域のランドマークの抽出を意味演算を用いて行うことができた。また、地域を代表するイベントの抽出に、意味演算を適用することが可能であることが分かった。

6. おわりに

本論文では、Word2Vec を用いて都市・地域とランドマークの意味的な関係性を捉え、ある地域のあるランドマークや観光スポットに対応する他地域のランドマークや観光スポットの抽出を行った。その結果、「東京」と「スカイツリー」の関係が「大阪」と「あべのハルカス」の関係に近いことなど、いくつかの関係性の抽出に成功した。

今後の課題として、 $Vec(\text{“マリンタワー”}) - Vec(\text{“横浜”}) + Vec(\text{“神戸”})$ という意味演算を行ったときに出力された「明石海峡大橋」など、演算結果として不自然な単語を、例えば、「ビル」や「橋」といったランドマークの種類や、そのランドマークの所在地などを認識することで、自動的に除去し、抽出効率を向上させることが考えられる。また、コーパスのサイズや、その他ハイパーパラメータの調整により、意味演算の精度を向上させることが挙げられる。精度の向上により、例えば、日本各地の観光スポットを、広く知られている東京のランドマークとの類推で旅行者に提示する、などといった観光分野への応用が可能になると考えられる。

謝 辞

本研究（の一部）は傾斜的研究費（全学分）学長裁量枠戦略的研究プロジェクト戦略的研究支援枠「ソーシャルビッグデータの分析・応用のための学術基盤の研究」による。

文 献

- [1] 佐伯圭介, 遠藤雅樹, 廣田雅春, 倉田陽平, 横山昌平, 石川博. 外国人 twitter ユーザの観光訪問先の属性別分析. 第 7 回データ工学と情報マネジメントに関するフォーラム, C4-3, 2015.
- [2] 新井晃平, 新妻弘崇, 太田学. Twitter を利用した観光ルート推薦の一手法. 第 7 回データ工学と情報マネジメントに関するフォーラム, G7-6, 2015.
- [3] 免田哲矢, Kryssanov, VV, 林勇吾, 小川均. Twitter を用いたリアルタイム情報収集による観光地情報推薦システム. 情報処理学会第 73 回全国大会, Vol. 3, p. 9, 2011.
- [4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.
- [5] 石野亜耶, 難波英嗣, 竹澤寿幸. 旅行ブログエントリからの観光情報の自動抽出. 第 22 巻, pp. 667–679. Japan Society for Fuzzy Theory and Intelligent Informatics, 2010.
- [6] 長谷川馨亮, 馬強, 吉川正俊. Twitter からの地域特徴語辞書の構築とその観光情報検索への応用. 第 6 回データ工学と情報マ

ネジメントに関するフォーラム, B3-4, 2014.

- [7] 野沢健人, 中岡義貴, 山本修平. word2vec を用いた代替食材の発見手法の提案. 第 114 巻, pp. 41–46. 電子情報通信学会, 2014.
- [8] 久保田豊久, 若林啓. 統計的意味論に基づく概念的類似度獲得手法の評価. 第 7 回データ工学と情報マネジメントに関するフォーラム, A2-2, 2015.
- [9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.