様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １１ 日現在

機関番号：１２１０２
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１２１３７

研究課題名（和文）メタ情報に頼らない高被覆旅行記ジオロケーション技術の開発

研究課題名（英文）Developing Metadata-independent High Coverage Geolocation for Travelogues

研究代表者
乾　孝司（INUI, Takashi）

筑波大学・システム情報系・准教授

研究者番号：６０３９７０３１

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究課題は、観光ビックデータを活用した包括的観光ニーズ調査の対象データとして、観光旅行記SNSデータの利活用を促進するため、SNS投稿内容から高精度に地理的位置を特定する技術を開発することを目的とする。主な研究成果として以下が挙げられる。(1) TransE法をベースに地理的近接性を保持した埋め込み表現学習手法を開発し、地理的知識グラフを構築した。(2) Wikipedia アンカリンク情報に基づいて単語の地理的位置を特定する程度をあらわす指標（地理的特定性指標）を開発した。(3) 47都道府県レベルの文書ジオロケーション課題の性能評価を実施し、開発手法の有効性を確認した。

研究成果の学術的意義や社会的意義
本研究課題は、地方都市部の観光振興事業において利活用が困難であった観光旅行記SNSビックデータに対して、当該データの利活用促進を実現するための高被覆地理的位置特定技術の開発を目的としたものである。本研究課題で得られた成果を活用することにより、包括的観光ニーズ調査の調査可能地点を地方都市部を含む全国各所の観光地へと拡大させ、従来よりも調査の質を向上させることができると期待される。

研究成果の概要（英文）：The goal of this research project is to develop technology that accurately identifies the geographical locations from the content of SNS posts in order to promote the utilization of travelogue SNS data as target data for comprehensive tourism needs surveys using tourism big data. The main research achievements include the following: (1) Developed an embedding representation learning method based on the TransE approach that maintains geographical proximity and constructed a geographical knowledge graph. (2) Developed a measure (geographical specificity index) that indicates the degree to which the geographical location of words can be identified based on Wikipedia anchor link information. (3) Conducted performance evaluations of the 47 prefecture-level document geolocation tasks and confirmed the effectiveness of the developed methods.

研究分野：自然言語処理

キーワード：文書ジオロケーション技術　地理的知識グラフ　埋め込み表現学習　ランドマーク抽出　エンティティリンキング　Data Augmentation

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

観光振興には、地方創生の礎として強い期待が寄せられており、国および自治体を挙げて多くの取り組みがなされている。近年では、観光ニーズの包括的な調査へ観光ビックデータを活用するための方法が模索され始めており、その中でも、SNS 投稿データは観光者の具体的な観光体験（例：寺で宿坊体験に参加し、精進料理を食べた等）を把握するための不可欠な情報源として強い期待が寄せられている。しかしながら現状では、(1)投稿内容の地理的な位置情報を位置メタ情報から取得しているが、(2)地理的位置を示すメタ情報は個人の所在を示す情報であり非公開であることが多く、位置メタ情報が公開付与されたデータは相対的に極めて少数である。以上のことから、特に、地方創生のターゲット・エリアとなるべき地方都市部の観光振興事業において観光ビックデータを十分に活用することが困難である。

２．研究の目的

本研究では、地理的位置情報が不明な観光旅行記 SNS ビッグデータを対象にして、メタ情報に頼らず､投稿内容から高精度にその地理的位置を特定する技術の確立を目指す。これによって、観光ニーズ調査の被覆対象地域を国内全域へ拡大し、全国的に観光振興事業を加速させることを目標とする（図１）。

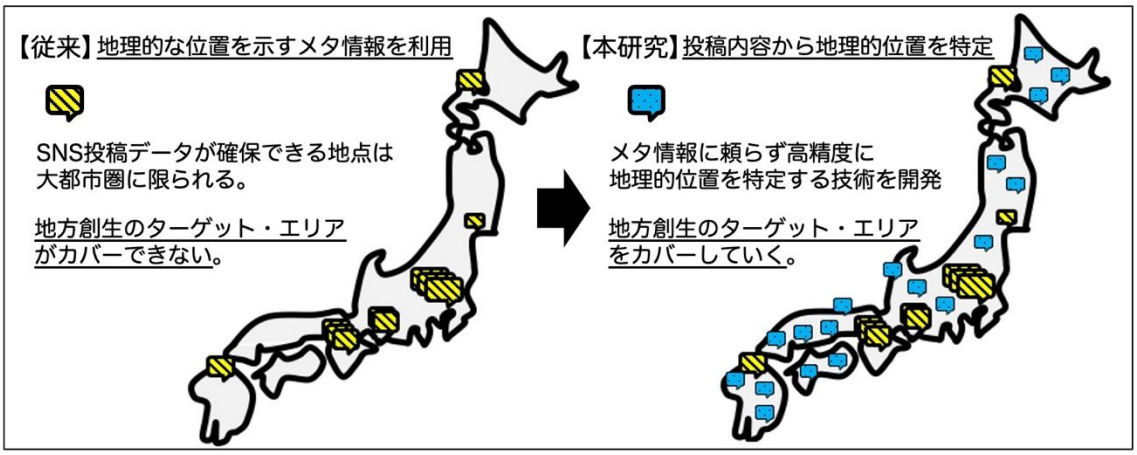

図１　観光ビックデータを用いた観光ニーズ調査における被覆対象のちがい

３．研究の方法

位置メタ情報を持たない旅行記 SNS ビックデータに対して、投稿内容からその地理的位置を特定する課題は文書の著者や執筆時期を推定する文書属性推定課題の中のひとつである文書の地理的位置推定（文書ジオロケーション、Document Geolocation、DG）課題の特殊形である。そこで本研究では、申請者の専門領域である自然言語解析の技術に立脚した DG 課題に関連する２つの要素技術（地理的近接性を保持した埋め込み表現学習および地理的特定性を有するランドマーク抽出）を新たに開発することを軸にして、これらに基づいて最終的にメタ情報に頼らず、投稿内容から高精度に旅行記 SNS ビックデータの地理的位置情報を特定する技術を確立することを目的とする。軸となる主要技術の詳細を以下に示す。

地理的近接性を保持した埋め込み表現学習：
近年の DG 課題は深層学習手法が主流であり、単語を実数値ベクトルで表現する埋め込み表現学習が用いられる。標準的な埋め込み表現学習は意味的類似性に基づくため、例えば「東京」と「大阪」の関連度が強くなる等、都市の規模に基づいて学習が進むことになり、本研究課題で重要となる都市間の地理的近接情報が効果的に学習されない。そこで本研究では新たに、都市間の地理的近接情報を保持できる埋め込み表現学習手法を開発する。

地理的特定性を有するランドマーク抽出：
DG 課題では地理的特定性を有するランドマークが重要な手がかりとなる。例えば「三社祭」は東京・浅草神社の神事であり「三社祭」は「東京・浅草」という地理的特定性を持つランドマークである一方、「収穫祭」や「文化祭」のように言語表記上は同じ「○○祭」でも地理的特定性を持つランドマークとはならないケースもあり、両者を適切に区別して処理しなければならない。そこで本研究では新たに、文書内言及から地理的特定性を有するランドマーク名を正確に抽出する手法を開発する。

４．研究成果

単語に対する一般的な埋め込み表現学習では、単語の周辺文脈情報に基づいて学習が進むため分布仮説に従った意味情報が単語ベクトルに反映されるが、この方法では本研究課題が注目する都市間の地理的近接性に関する情報を単語ベクトルに埋め込むことが出来ない。そこで本研究では、３つ組関係知識の埋め込み表現学習手法である TransE 法をベースにして、地理的近接性を保持した埋め込み表現学習手法を開発した。提案手法では、日本国内の住所地名および、ある住所地名に所在している施設名をエンティティとし、また、これらのエンティティ間に包含関係、所在関係、近接関係、施設間近接関係の４タイプの地理的関係タイプを定義した上で、エンティティ間の地理的位置が近接しているほど類似した埋め込み表現が学習される。提案手法によって獲得された地理的知識グラフ（GeoKG）をエンティティ予測課題により評価した結果を表１に示す。双方向グラフの場合、islocatedin 関係を除けば、安定して良好な結果であることがわかる。

表１　エンティティ予測の結果

	関係タイプ	MR(h)	MR(t)	hit@10(h)	hit@10(t)	hit@1(h)	hit@1(t)
単方向 GeoKG	all	264.1	**139.9**	0.828	0.838	0.301	0.417
	ispartof	487.6	56.7	0.681	0.718	0.165	0.491
	islocatedin	830.8	686.7	0.596	0.589	0.197	0.400
	near	15.2	14.6	0.745	0.742	0.000	0.000
	landmark_near	43.3	71.6	0.968	0.970	0.431	0.429
双方向 GeoKG	all	**98.0**	303.8	**0.855**	**0.894**	**0.560**	**0.634**
	ispartof	243.5	650.5	0.742	0.793	0.480	0.612
	islocatedin	215.8	948.1	0.542	0.751	0.301	0.615
	near	8.0	7.8	0.822	0.824	0.000	0.000
	landmark_near	4.2	5.1	0.990	0.989	0.726	0.722

上記の地理的知識グラフに対してアテンション処理を施すことで、既存の文書ジオロケーションモデル（deepgeo）に地理的知識グラフの情報を取り込み、その有効性を検証した。アテンション処理では、２種類の方法を検討した。１つは、入力文書の単位でアテンション処理を施す文書単位アテンションであり、もう１つは、文書内の部分ごとにアテンション処理を施すトークン単位アテンションである。評価結果を表２に示す。deepgeo 単体モデルと比べて、地理的知識グラフを取り込んだモデルは高い分類精度を達成している。アテンション処理の比較では、文書単位よりもトークン単位のアテンションの方が分類精度が高いことが確認できる。

表２　地理的知識グラフを考慮した文書ジオロケーション実験の結果

モデル	クラスタ数	分類精度	平均分類精度
deepgeo	-	0.663	-
MeCab インジケータ付 deepgeo	-	0.677	-
文書単位アテンション	kmeans48	0.678**	0.672
	kmeans94	0.671	
	kmeans240	0.673*	
	kmeans470	0.667	
文書単位アテンション	DCC48	0.678**	0.677
	DCC94	0.676**	
	DCC240	0.675**	
	DCC470	0.678*	
トークン単位アテンション	kmeans48	0.685^{++}	0.676
	kmeans94	0.685^{++}	
	kmeans240	0.681**	
	kmeans470	0.681**	
トークン単位アテンション	DCC48	$\mathbf{0.695}^{++}$	**0.688**
	DCC94	0.689^{++}	
	DCC240	0.684**	
	DCC470	0.685^{++}	
人手（参考上限値）	-	0.767	-

次に、地理的特定性を有するランドマークを抽出するため、地理的エンティティに対して地理的特定性を推定する手法を開発した。本手法では、Wikipedia のエンティティ（ページ）と、Wikipedia ページ内にあらわれるエンティティをあらわす単語がもつアンカリンク情報に注目し、(1)ある特定のエンティティへリンクされるアンカ文字列の異なり数に基づく地理的曖昧性、および(2)ある特定のエンティティへリンクされるアンカ文字列の集合に対して、そこから当該のエンティティへリンクされるリンク割合に基づいて求められる名称専有性という２種類の指標値を求める。

日本語 Wikipedia データを知識源とすることで、43,189 件の地理的エンティティに対して地理的特定性を求めたところ、ある程度人間の直感に沿う妥当な指標値を推定できていることを確認した。推定例を表３に示す。例えば、「八坂神社」という名称をもつ神社は日本国内に複数存在するため、「八坂神社」の地理的曖昧性は比較的高い値となっていた。また、国内の「八坂神社」のうち、京都市に所在する「八坂神社」は総本山であり他よりも認知度が高いため、名称専有性も比較的高い値となっていた。ほかの例として、「いろは坂」という名称の道路は日本国内に２箇所しか存在せず、「いろは坂」の地理的曖昧性は低い値となっていた。また、２箇所のうち日光市の「いろは坂」はもう１件と比べると認知度が非常に高いため、名称専有性は高い値となっていた。

表３　地理的特定性の推定例（amb は地理的曖昧性、exc は名称専有性）

	名称専有性 高	名称専有性 低
地理的曖昧性 高	八坂神社 (京都市) amb: 21, exc: 0.78	中央区 (大阪市) amb: 135, exc: 0.11
地理的曖昧性 低	いろは坂 (日光市) amb: 2, exc: 0.92	竹下通り (福岡市) amb: 2, exc: 0.082

関連した調査として、地理的エンティティ以外のエンティティに対する名称専有性の指標値分析を実施した。その結果、エンティティと何等かの関係がある地域において名称専有性が高くなる傾向があることを確認した。例えば、特産品の例では、以下のように品名に地域名を冠する場合、冠さない場合を問わず、その特産品で有名な地域の指標値が高くなっていた。以上の結果から、開発した地理的特定性指標は、文書ジオロケーション課題における入力文書中の単語の特徴付けに利用できることが示唆される。

- 水沢うどん（群馬県:1)、関あじ（大分県:1）
- 牛タン（宮城県:0.522)、玉子焼（兵庫県:0.741）

上記の結果を受け、次に、地理的特定性を利用した文書ジオロケーション手法を開発した。地方都市部においてデータ確保が難しい問題に対応するために、データ拡張手法を適用することを考え、データ拡張時に地理的特定性の情報を利用する。具体的には、図２のように、文書中の地理的エンティティをそのエンティティの所在地情報に変換することでデータ拡張を実現する。その際、所在地候補が複数存在する場合があるため、この候補の順位付けに地理的特定性指標のうちの名称専有性の値を採用し、この値を参照することで信頼性の高いデータのみを拡張する。

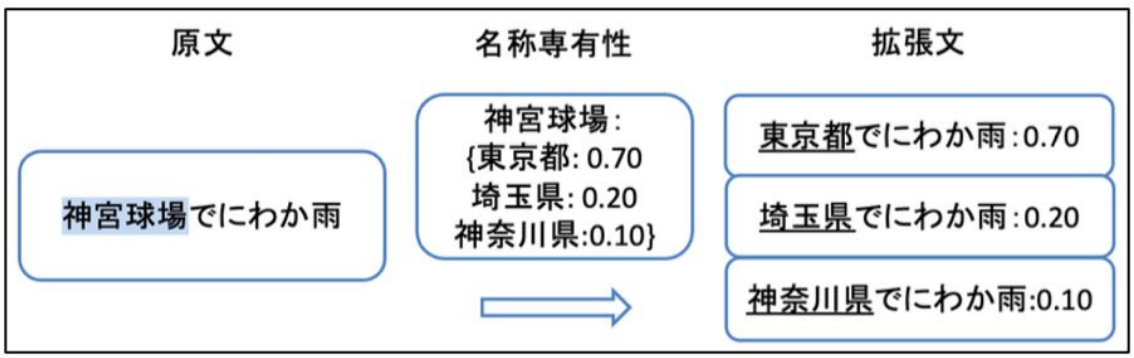

図２　データ拡張の例

データ拡張後のデータ数を表４に示す。事例あたり１事例の拡張（k=1）を許す場合では、平均８事例の拡張に成功した。また、拡張数を３事例まで許す（k=3）と平均 21 事例が拡張できていた。次に、データ拡張を適用した実験結果を表５に示す。ここでは地方部の性能を確認するため、表２のようにデータ全体の要約ではなく、都道府県ごとの精度を示している。上側の５件は訓練データ数が多い県であり、大都市を含む地域が並んでいることがわかる。一方、下側の５件は訓練データ数が少ない県である。まず表全体の傾向として、既存モデル deepgeo と

比較すると、データ拡張を施した提案手法では佐賀県を除くすべて都道府県で性能が改善していることがわかる。訓練データ数上位県だけでなく、訓練データ数の下位県においても有効であることが確認できる。さらに、拡張データ数（k）に注目すると、上位県では k=1 の場合のみ改善する傾向があるのに対し、下位県では k=1 以外でも性能が改善している。さらに上位県よりも下位県の方が改善幅は大きく、地方都市部を含むような訓練データ数下位県において、提案手法の有効性が顕著に確認できた。

表４　実験におけるデータ拡張後のデータ数

k の値	拡張後の評価データ数	事例あたりの平均拡張数
$k=1$	59,128	8.45
$k=2$	114,405	16.34
$k=3$	147,799	21.11

表５　データ拡張を考慮した文書ジオロケーション実験の結果（都道府県別）

都道府県	訓練データ数	deepgeo	提案手法			差
			$k=1$	$k=2$	$k=3$	
東京都	27,333	0.749	**0.768**	0.767	0.767	**0.019**
大阪府	20,973	0.771	**0.775**	0.771	0.769	**0.004**
北海道	14,376	0.846	**0.850**	0.849	0.849	**0.005**
京都府	11,438	0.583	**0.586**	0.579	0.579	**0.003**
愛知県	11,037	0.634	0.668	**0.671**	**0.671**	**0.036**
佐賀県	965	**0.462**	**0.462**	**0.462**	**0.462**	0.000
秋田県	960	0.635	**0.692**	**0.692**	**0.692**	**0.058**
福井県	816	0.538	**0.577**	**0.577**	**0.577**	**0.038**
島根県	699	0.571	**0.600**	**0.600**	**0.600**	**0.029**
鳥取県	530	0.406	**0.563**	**0.563**	**0.563**	**0.156**

最後に、本研究課題では、文書ジオロケーションに地理的近接性を保持した埋め込み表現学習に基づく地理的知識グラフの情報を取り込むことで精度が向上することを確認した。また、地理的特定性を有するランドマークを適切に抽出するために地理的特定性指標を開発し、その指標値を文書ジオロケーションにおけるデータ拡張に利用することを試みた。その結果、既存モデルよりも精度が改善することを確認した。提案手法は、特に、訓練データ数の少ない地域での有効性が顕著であった。以上から、当初研究目的で述べた項目について概ね達成できたと言える。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 平川冬尉，乾孝司	４．巻 Vol.63, No.12
２．論文標題 地理的知識グラフを取り込んだニューラル文書ジオロケーションモデル	５．発行年 2022年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 1870-1883
掲載論文のDOI（デジタルオブジェクト識別子） 10.20729/00222743	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 かげ山宗一，乾孝司
２．発表標題 位置属性を有しない事物に対する地理的特定性の分析
３．学会等名 言語処理学会第29回年次大会（NLP2023）
４．発表年 2023年

１．発表者名 かげ山宗一，乾孝司
２．発表標題 言及に対する地理的特定性指標の提案と文書ジオロケーションへの適用
３．学会等名 情報処理学会自然言語処理研究会(NL-253-19)
４．発表年 2022年

１．発表者名 平川冬尉，乾孝司
２．発表標題 地理的知識グラフを取り込んだニューラル文書ジオロケーションモデル
３．学会等名 情報処理学会自然言語処理研究会(NL-248-3)
４．発表年 2021年

１．発表者名 Tomoki Okugawa and Takashi Inui
２．発表標題 Utilizing Word Embedding Representations in Word Sense Analysis of Japanese Spelling Variants
３．学会等名 The 27th International Conference on Asian Language Processing（国際学会）
４．発表年 2023年

１．発表者名 山本祐耶，乾孝司
２．発表標題 地理的エンティティ情報が与えられた文書ジオロケーションモデルの有効性検証
３．学会等名 言語処理学会第30回年次大会（NLP2024）
４．発表年 2024年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関