Vol.2021-NL-248 No.3
Vol.2021-CVIM-226 No.3
2021/5/20

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

概要：本稿では、文書ジオロケーション課題において、地理的な関係性を構造化した地理的知識グラフを取り込んだ深層学習モデルを提案する。提案手法では、まず、国内の住所および住所属性をもつ施設名のリストから地理的知識グラフを構築し、このグラフに対して TransE 法 [1] および TransE-GDR 法 [2] を適用することで地理的知識グラフの埋め込み表現を獲得する。そして、アテンション機構を用いることによって、地理的知識グラフの情報を文書中の各トークンの埋め込み表現と統合する。評価実験の結果、地理的知識グラフを取り込んだ提案手法は、従来手法に比べて高い推定精度を達成することを確認した。

1. はじめに

SNS サービスの普及により、情報の発信や共有が手軽にできるようになり、日々、多種多様なデータが大量に SNS 上に流れるようになった。これらデータのうち一部の投稿文書は、災害時における各地域の状況把握や余暇活動としての観光における計画立案など、データと地理的位置を結びつけた幅広い応用に利用されている。しかしながら、Twitter における位置情報付き投稿は全投稿の 1%にも満たないという Sloan らの報告 [3] にもあるように、地理的位置情報をメタ情報として持つ SNS 投稿は全投稿に対して相対的に極めて少量であり、上記のような応用分野に利用できる投稿データはごく一部に限られている。

このような背景から、投稿文書の投稿場所（地理的位置）を自動推定する技術が求められるようになっており、近年、研究開発が進められている [4][5][6][7][8] 。この課題は文書ジオロケーション（document geolocation）と呼ばれ、特に、Twitter の投稿である Tweet を対象とした場合はツイートジオロケーション（tweet geolocation）と呼ばれている。これらジオロケーション課題に対する先行研究が抱える共通の問題点として、推定モデルが十分な地理的知識を持ち合わせていない点が挙げられる。投稿内に地名（例えば「東京」）や所在が一意の施設名「東京タワー」が含まれており人間が見れば比較的自明な場合であっても推定モデルは推定を誤ることがある。

知識ベースは、世の中の世界知識を構造化して集約したものであり、特に、グラフ構造を採用したものは知識グラフ（knowledge graph, KG）と呼ばれる [9] 。知識グラフでは、実世界に存在する実体（エンティティ）間の関係を軸にして、関係元のエンティティ（head entity）、関係（relation）および関係先のエンティティ（tail entity）からなる三つ組 $< h, r, t >$ の形式で知識を記述する（例: < 日本, 首都, 東京 >）。ここで、三つ組の各エンティティをノード、関係をエッジと見なすことで、三つ組の集合は図 1 のようなグラフを形成する。知識グラフの中でも図のように地理的関係からなる三つ組を基に構築されたものは地理的知識グラフ (Geographical KG, GeoKG) と呼ばれる [2]。

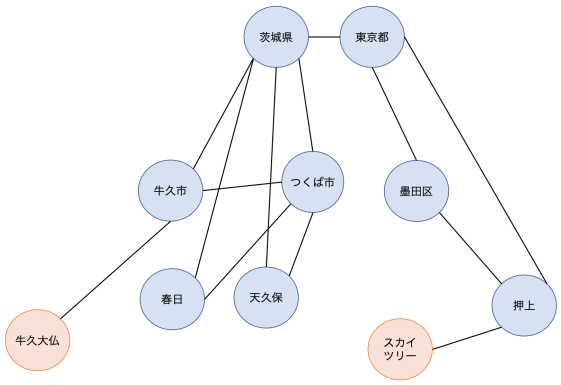

図 1 地理的知識グラフの概念図

本研究では、先述の、既存推定モデルが十分な地理的知識を持ち合わせていない点を踏まえ、深層学習ベースの文書ジオロケーションモデルに対して、地理的知識グラフを取り込んだモデルを提案し、その有効性を検証する。これまでに、（地理的ではない）一般的な知識グラフの情報を考慮した文書ジオロケーション手法の提案 [10] や、知識グラフから取得した埋め込み表現をニュースの記事分類や自然言語推論等の NLP 応用課題に利用した研究 [11] は存在しているが、地理的知識グラフを取り込んだ文書ジオロケーションモデルは我々が知る限りこれまでに提案されていない。

ⓒ 2021 Information Processing Society of Japan

本研究では、ジオロケーション推定対象を日本国内とする。そこでまず、国内の住所リストと施設名リストから地理的知識グラフを構築し、その埋め込み表現を獲得する。この際、知識グラフのエッジの方向および埋め込み表現学習の方法について比較評価をおこなう。獲得した地理的知識グラフの情報はアテンション機構を用いて深層学習ベースの文書ジオロケーションモデルに取り込む。この時、事例単位とトークン単位の２種類のアテンションを検討し、比較評価する。Twitter データを用いた評価実験の結果、地理的知識グラフを取り込んだ提案ジオロケーションモデルは、既存手法と比べて高い推定精度を達成することを確認した。

2. 関連研究

2.1 ツイートジオロケーション

単一ツイートからその投稿場所を推定する手法として、Han らは固有表現抽出器とナイーブベイズ分類器を用いる手法を提案した [6] 。また、Han らは W'NUT2016 においてツイートジオロケーションを共有タスクとして提案し、この際、英語データセットを作成している [12]。これにより、ツイートジオロケーションに関する研究が盛んになった。近年では、ニューラルネットワークを用いた推定モデルが高い精度を示しており、Lau らの文字レベルの Recurrent Convolutional Network を用いた研究 [8] や Fornaciari らのマルチタスク学習を用いた研究 [5] などがある。

2.2 NLP 応用課題における知識グラフの活用

近年では、自然言語処理の応用課題における性能向上を目指して知識グラフが活用されている。Miyazaki ら [10] は、あるユーザが投稿したツイート集合からそのユーザの所在を推定するユーザジオロケーション（user geolocation）課題において、知識グラフ内のエンティティと単語分散表現を対応させることで、ツイートの情報と知識グラフの情報を統合する手法を提案した。また、Annervaz ら [11] は、文書分類や自然言語推論などの自然言語処理応用課題において、深層学習ベースの応用モデルに知識グラフから得た分散表現を取り込む汎用的な枠組みを提案した。

2.3 知識グラフと埋め込み表現学習

知識グラフの欠損値を自動補完する知識グラフ補完 (KB completion) 課題などにおいて、知識グラフ（の要素）の埋め込み表現を学習する手法が提案されている。代表的な埋め込み表現学習手法として、Bordes らが提案した TransE がある [1]。TransE では、知識グラフ内の三つ組 $< h, r, t >$ に対して、$h + r = t$ が成り立つように埋め込み表現を学習する。このような手法は、構造ベースの埋め込み手法と呼ばれており、TransE の改良手法も幾つか提案されている [13][14]。

本研究では、知識グラフの埋め込み表現学習手法として TransE[1] と、TransE を地理的地理的グラフ向けに改良した TransE-GDR[2] を採用する。そこで、以下で両手法の概要を説明する。

TransE では以下の目的関数を最小化することで埋め込み表現を学習する。

$$L = \sum_{h,t} \sum_{h',t'} max(0, f_r(h,t) + \gamma - f_r(h',t')) \quad (1)$$

ここで、h と t は知識グラフ内のある三つ組を構成しているエンティティ対である。また、h' と t' は学習時に用いられる負例であり、三つ組を構成しないエンティティ対である。この負例は、知識グラフ内に存在する三つ組のどちらか一方のエンティティをランダムに変更することで作成される。また、式 (1) 中の $f_r(\cdot)$ は評価関数であり、以下のように定義される。

$$f_r(h,t) = -||l_h + l_r - l_t||_p \quad (2)$$

ここで、l はエンティティおよびそれらの間の関係をあらわすベクトルである。この式自体は L_p ノルムを計算しており、具体的には L_1、もしくは L_2 ノルムが用いられる。

TransE-GDR は、TransE から式 (1) の目的関数を改良した手法である。地理的知識グラフ内の三つ組において、エンティティが地名である時、エンティティ間の地理的な距離を計算できる。そこで、TransE-GDR では、この距離に関する情報を目的関数に加えた。

$$L_{geo} = \sum_{h,t} \sum_{h',t'} max(0, f_r(h,t) + \gamma - w_{geo} f_r(h',t')) \quad (3)$$

この式中の w_{geo} が距離情報である。w_{geo} は以下で定義される。

$$w_{geo} = \frac{1}{|log_{10} \frac{dis(h,t)+\theta}{dis(h',t')+\theta}| + 1} \quad (4)$$

ここで、$dis(h,t)$ がエンティティ間の距離を表している。θ は対数の分母がゼロにならないための定数である。$dis(h,t)$ は以下のように定義される。

$$dis(h,t) = \sqrt{(h_x - t_x)^2 - (h_y - t_y)^2} \quad (5)$$

h_x, t_x はエンティティの経度、h_y, t_y はエンティティの緯度である。

3. GeoKG

3.1 GeoKG の構築

3.1.1 動機

本節では、文書ジオロケーションモデルに取り込む地理

的知識グラフ（GeoKG）の構築について述べる。GeoKG の構築は、Qiu ら [2] を参考にして進める。Qiu ら [2] は GeoKG を対象とした知識グラフ補完課題に取り組んだ際、異なる情報源から２種類の GeoKG を構築している。ひとつは DBpedia[*1]を情報源としたもので、DBpedia から地理的な知識を選別して用いている。この DBpedia 版 GeoKG には「河口」や「ラジオ局の場所」など、仔細な関係タイプが定義されている点が特徴である。もうひとつは GADM[*2]を情報源としたものである。GADM 版 GeoKG は、地理に特化したデータ集が情報源となっており、隣接関係や包含関係といった地域間の地理的位置関係をあらあす関係タイプのみで構成されている。本研究では、投稿文書で言及された地域をあらわす地名に関して、地理的位置関係を取り込みたいため、GADM 版 GeoKG の構築方法を参考にする。ただし、GADM では日本国内の市区町村レベルまでのデータしか扱えないため、GADM の代わりに国土交通省が発行している住所集を情報源とする。さらに、文書ジオロケーションでは、「東京タワー」のようなランドマークに関する言及が重要となるため、SNS から施設名のリストを収集して情報源を補強する。詳細は以下の通りである。

3.1.2 住所エンティティと施設エンティティ

本研究で構築する GeoKG では、住所エンティティか施設エンティティが知識グラフのノードとなる。

住所エンティティは、国土交通省が発行している「街区レベル参照情報」[*3] から住所情報を抽出して作成する。抽出した住所情報を、都道府県レベル、市区町村レベル、字（あざ）レベルの３層構造で格納し、各レベルの要素（「茨城県」や「つくば市」）を住所エンティティとする。

施設エンティティは、Twitter 投稿データから施設名を自動収集することで作成する。まず、Twitter Streaming API により 2014 年から 2015 年のジオタグが付与されたツイートを収集する。Twitter では、Instagram[*4]や Foursquare[*5]などとのサービス連携によりツイートに位置情報を付与した場合は「ツイート内容 @ 施設名 URL」という形式の投稿内容になる。そこで、このような投稿に対して、正規表現によるマッチング処理によって簡易的に施設名を取得する。ノイズを排除するため、このようにして取得した施設名のうち、５回以上、正規表現にマッチングしていた施設名を施設エンティティとする。また、取得元ツイートに付与されたジオタグ（緯度経度情報）を逆ジオコーディングすることで施設名の住所情報を取得しておく。この住所情報は GeoKG のエッジを作成する際に参照する。

[*1] https://www.dbpedia.org/
[*2] https://gadm.org/
[*3] https://nlftp.mlit.go.jp/cgi-bin/isj/dls/_choose_method.cgi
[*4] https://www.instagram.com/
[*5] https://foursquare.com/

Qiu ら [2] 以外の既存 GeoKG として、住所情報のみから構築されたもの [15] や、ユーザ参加型で地理情報を作成するオープンストリートマップから抽出した施設名と住所情報から構築されたもの [16] がある。これらと比較すると、本研究の GeoKG は、Twitter データから施設名を収集して利用している点に特徴がある。このように施設名を収集することで、GeoKG には Twitter ユーザが実際に訪問した施設名が入り、必要性の高いエンティティに絞って GeoKG に登録することができる。

3.1.3 地理的関係タイプ

つぎに、上述のエンティティをノード要素とする三つ組知識の構築について説明する。ある２つのノードから三つ組知識を形成することは、当該ノード間に知識グラフのエッジを作成することと同義であるので、ここでは、GeoKG のエッジの観点から説明するが、例を示す際は三つ組で例を示す。

本研究では、Qiu ら [2] を参考にして、包含 (ispartof), 所在 (islocatedin), 近接 (adjacency), 施設間近接 (landmark_adjacency) の４種類の関係タイプを定義し、ノード間にエッジを作成する。以下でそれぞれの関係タイプについて説明する。

- **包含関係（ispartof）**：
  ある住所エンティティがほかの住所エンティティに含まれた地域である場合、関係タイプが包含関係のエッジをノード間に張る。この時、事前に住所エンティティを 国レベル (0), 都道府県レベル (1), 市区町村レベル (2), 字レベル (3) にレベル分けしておき、地域レベルに応じて包含関係を３種類に細分する。例えば「茨城県に含まれているつくば市」に対応する三つ組は、＜つくば市, ispartof1, 茨城県＞となり、関係先エンティティのレベル番号が付いた関係タイプとなる。
- **所在関係（islocatedin）**：
  ある施設エンティティがある住所エンティティがあらわす地域に所在している場合、関係タイプが所在関係のエッジをノード間に張る。施設エンティティの所在は先述した逆ジオコーディングで得た住所とする。所在関係でも包含関係と同様に関係先のレベルを考慮する。例えば、「芝公園に所在している東京タワー」を表す際は、＜東京タワー, islocatedin3, 芝公園＞となる。ここで、芝公園は東京都港区の町名であり、字レベル (3) である。また、この時、包含関係を辿ることで、(東京タワー, islocatedin2, 港区), (東京タワー, islocatedin1, 東京都) および (東京タワー, islocatedin0, 日本) も同時に作成する。
- **近接関係（adjacency）**：
  ２つの住所エンティティがあらわす地域が地理的に近接している場合、関係タイプが近接関係のエッジをノード間に張る。ここで、本州の隣接都道府県の県庁

表 1 構築した地理的知識グラフの統計情報

	単方向グラフ	双方向グラフ
総エンティティ	240,328 件	240,328 件
住所エンティティ	181,297 件	181,297 件
施設エンティティ	59,031 件	59,031 件
関係タイプ	9 個	18 個
総三つ組	1,854,940 件	3,709,880 件
包含関係の三つ組	536,670 件	1,073,340 件
所在関係の三つ組	224,539 件	449,078 件
近接関係の三つ組	112,107 件	224,214 件
施設間近接の三つ組	981,724 件	1,963,448 件

所在地間の平均距離 104.32km を地理的に近接しているかどうかの目安の基準値とし、エンティティ間の距離が基準値以内である場合を近接しているとした。近接関係のエッジは、地域レベルが同レベルのエンティティのノード間にのみ作成した。エンティティ間の距離はユークリッド距離とする。この際、都道府県レベルと市区町村レベルのエンティティでは都道府県庁の所在地および市区町村役所の所在地の緯度経度情報を利用した。字レベルのエンティティでは、街区レベル参照情報に記載の緯度経度情報を利用した。

- **施設間近接関係（landmark_adjacency）**：
  2 つの施設エンティティが同じ住所エンティティに所在している場合、関係タイプが施設間近接関係のエッジをノード間に張る。施設間近接関係のエッジは、地域レベルが字レベル (3) の住所までが同じ場合のみ作成する。例えば、東京タワーと増上寺は同一字内である芝公園に所在しているため、< 東京タワー,landmark_adjacency, 増上寺 > となる。

以上によって、ノードとエッジを構成し、1 つの地理的知識グラフを構築する。説明の都合上、これを単方向の地理的知識グラフと呼ぶ。

3.1.4 双方向グラフ

本研究で採用する TransE 系の知識グラフの埋め込み表現学習手法では、エッジの向きを考慮する。このことを踏まえ、先程述べた単方向グラフ中のすべてのエッジに対して、それらの逆方向エッジを考え、それらを加えた知識グラフも構築する。これを双方向の地理的知識グラフと呼ぶ。逆方向エッジを考えるにあたり、すべての関係タイプをさらに 2 分割した forward 関係タイプと backward 関係タイプを用意した。例えば、単方向グラフの要素となる三つ組 < 東京タワー, islocatedin3, 芝公園 > は、双方向では、< 東京タワー, islocatedin3_forward, 芝公園 > と < 芝公園, islocatedin3_backward, 東京タワー > となる。

表 1 に、構築した単方向 GeoKG と双方向 GeoKG のそれぞれの統計情報を示す。2 つの知識グラフでは、ノードは共通であり、ノード数は同一である。一方で、エッジは逆方向エッジが双方向 GeoKG に加わるため、三つ組数には 2 倍の差がある。

表 2 エンティティ予測実験用データの内訳（三つ組数）

	単方向 GeoKG	双方向 GeoKG
学習データ	1,487,645	2,971,596
開発データ	183,648	369,142
評価データ	183,647	59,894

3.1.5 埋め込み表現学習

上記によって作成された単方向 GeoKG、および、双方向 GeoKG それぞれの地理的知識グラフに対して、2.3 節で示した 2 つの埋め込み表現学習手法 TransE と TransE-GDR をそれぞれ適用することでそれらの埋め込み表現を獲得する。結果として、4 種類のグラフ埋め込み表現が獲得される。

3.2 エンティティ予測による評価

3.1.5 項の方法で獲得した 4 種類のグラフ埋め込み表現を知識グラフ補完の標準評価手法であるエンティティ予測課題によって評価する。エンティティ予測課題とは、ある三つ組に対して、関係元エンティティもしくは関係先エンティティのどちらかを隠した状態の評価用三つ組を新たに用意し、評価用三つ組の隠されたエンティティを知識グラフ（の埋め込み表現）に予測させる課題である。この時、知識グラフは、式 (2) の評価関数により予測候補となる各エンティティのスコアを求め、スコアの小さい順にランク付けしたエンティティ・リストを出力するものとする。

正解判定の際、予測されたエンティティが以下のような関係にある場合は、予測結果が正解エンティティと一致していない場合でも正解とみなすことにした。例えば、三つ組 < つくば市, ispartof1, 茨城県 > から関係元エンティティである「つくば市」を隠した評価用三つ組の出力として、「牛久市」を出力した場合、実世界の関係としては「つくば市」と「牛久市」の双方とも茨城県内の市区町村であるので、このような事例は正解とみなすことにした。

評価実験に用いるデータは、3.1 節で構築した知識グラフの要素となる三つ組データである。これを 8：1：1 の割合で分割し、学習データ、開発データ、評価データとした。学習データと開発データは、知識グラフ構築とその埋め込み表現学習の際に利用される。また、評価データから評価用三つ組を作成した。データセットの内訳を表 2 に示す。双方向 GeoKG の評価データが上記の割合よりも少なくなっているが、これは、分割された評価データから評価用三つ組を作成したあと、その三つ組とは逆方向エッジをもつ三つ組が学習データあるいは開発データに存在する場合、その評価用三つ組は採用しなかったためである。

評価指標には、MeanRank（MR）および hit@K を用いる。MR は、正解のエンティティが出力リストの何番目に出現するかの順位の平均である。hit@K は出力リストの K

表 3 単方向 GeoKG に対する実験結果

モデル	関係タイプ	MR(h)	MR(t)	hit@10(h)	hit@10(t)	hit@1(h)	hit@1(t)
TransE	all	2611.6	1044.3	0.454	0.665	0.217	0.370
	ispartof	3997.0	26.3	0.146	0.790	0.021	0.555
	islocatedin	4388.2	2058.2	0.376	0.664	0.348	0.429
	adjacency	3.9	3.9	0.946	0.935	0.349	0.334
	landmark_adjacency	1766.6	1466.2	0.578	0.568	0.276	0.263
TransE-GDR	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
	adjacency	15.2	14.6	0.745	0.742	0.000	0.000
	landmark_adjacency	43.3	71.6	0.968	0.970	0.431	0.429

表 4 双方向 GeoKG に関する実験結果

モデル	関係タイプ	MR(h)	MR(t)	hit@10(h)	hit@10(t)	hit@1(h)	hit@1(t)
TransE	all	36.1	46.8	**0.979**	**0.954**	**0.476**	**0.558**
	ispartof	31.8	63.2	0.953	0.875	0.181	0.570
	islocatedin	293.7	299.1	0.936	0.836	0.230	0.256
	adjacency	3.9	4.0	0.950	0.949	0.188	0.220
	landmark_adjacency	3.6	9.2	0.996	0.996	0.624	0.620
TransE-GDR	all	**24.4**	**28.6**	0.960	0.950	0.363	0.457
	ispartof	54.2	71.1	0.961	0.929	0.253	0.648
	islocatedin	83.2	100.6	0.936	0.894	0.198	0.362
	adjacency	10.8	11.0	0.703	0.726	0.000	0.000
	landmark_adjacency	7.6	6.1	0.980	0.979	0.445	0.440

番目以内に正解エンティティが出現した割合である。

単方向 GeoKG の実験結果を表 3、双方向 GeoKG の実験結果を表 4 に示す。各表の関係タイプをまとめた結果（表中の all）において、TransE と TransE-GDR を比較し、結果の良い方を太字で示す。表中の all の行に注目すると、表 3 の単方向の場合は、全ての指標において TransE-GDR の方がよい結果となっていることがわかる。関係タイプ別に見ると、隣接関係（adjacency）では TransE-GDR よりも TransE が良い結果となっているが、それ以外の関係タイプでは安定して TransE-GDR の方が良い結果であった。

つぎに、表 4 の双方向の場合は、MR 指標では TransE-GDR が良いが、hit@K では TransE が良い結果となっている。このことから、TransE では正解となるエンティティを適切に予測できる場合は上位（10 位以内）に順位付けできるが、誤ってしまう場合は順位を非常に落としてしまうと考えられる。一方で、TransE-GDR は、hit@10 の値が TransE よりも低いことから 10 位以内ではないが、安定的に正解となるエンティティを高い順位で予測できていると考えられる。また、単方向 GeoKG と双方向 GeoKG では使用データが異なるため、厳密な比較はできないが、全体的には、双方向 GeoKG の方が良い評価値を得ていることが確認できる。

以上の結果を踏まえ、これ以降の議論では地理的知識グラフとして、双方向 GeoKG に対して TransE-GDR で埋め込み表現学習をしたものを採用し、これを文書ジオロケーションモデルに取り込むことを考える。

4. GeoKG を取り込んだ文書ジオロケーション手法

前節で述べた GeoKG を既存の文書ジオロケーションモデルに取り込む方法について述べる。本研究では、文書ジオロケーションモデルとして deepgeo[8] を採用し、このモデルに対してアテンション機構を用いることで、GeoKG の知識を取り込む。この時、入力文書の事例単位でアテンション処理を施す方法と、入力文書内のトークン単位でアテンション処理を施す方法の 2 種類の方法を検討する。

以下ではまず、deepgeo の概要を説明し、その後、2 種類のアテンション処理について説明する。

4.1 deepgeo

Lau らが提案した deepgeo[8] は、近年、高い性能が報告されている深層学習ベースの文書（ツイート）ジオロケーションモデルのひとつである。deepgeo では、入力ツイートが与えられた時、ツイートテキストに加え、ツイートの投稿時刻や投稿者のプロフィール情報など、複数の特徴量に対して、特徴量の特性に合わせたサブネットワークを構築し、最終的にそれらを統合することで、地理的位置推定をおこなう。本研究では、ツイートテキストに対応するサブネットワークに対して、アテンション機構を導入することにより、ツイートテキスト特徴量に GeoKG の知識を取

り込んでいく。そこで、以降では、ツイートテキストに対応するサブネットワークである Lau らの Text_Network について詳細を説明する。

Text_Network は、文字ベースの recurrent convolutional ネットワーク [17] にセルフアテンションが付いた構成をしている。まず、ツイートは文字系列に分解された後、bi-directional LSTM によって文字ごとの中間表現を得る。その後、畳み込み層により（文字列レベルの）トークン中間表現 g_t を得る。さらに、これに対して、ウィンドウサイズを P とした max-over-time pooling を適用し、$\hat{g}_j$ を得る。ここで、$\hat{g}_j$ はツイートテキストの文字数を T としたときに、$T-P+1$ 個生成される。

第 t 番目の文字分散表現を x_t とした時のここまでの処理を式であらわすと以下の通りである。

$$h_t^f, h_t^b = bi\text{-}LSTM(x_t) \tag{6}$$

$$\hat{x_t} = h_{t-1}^f \oplus x_t \oplus h_{t+1}^b \tag{7}$$

$$g_t = ReLU(Conv(\hat{x_t})) \tag{8}$$

$$\hat{g_j} = max(g_t, g_{t+1}, \ldots, g_{t-P+1}) \tag{9}$$

ここで、式 (7) の $\oplus$ はベクトルの連結を表す演算であり、式 (8) の $Conv$ は畳み込み層を適用する関数である。つづいて、さきほどの $\hat{g}_j$ に対してセルフアテンションを適用し、最終的にツイート全体のベクトル表現 f_{text} を得る。

$$\alpha_j = v^T tanh(W_v \hat{g_j}) \tag{10}$$

$$\mathbf{a} = softmax(\alpha_0, \alpha_1, \ldots, \alpha_{T-P}) \tag{11}$$

$$f_{text} = \sum_{j=0}^{T-P} \mathbf{a}_j \hat{g_j} \tag{12}$$

ここで、式 (10) の v と W_v は重みパラメータであり、それぞれのサイズは、畳み込み層のフィルタ数 O と $O \times O$ である。Text_Network によって得られる f_{text} は、ほかのサブネットワークの出力と統合され、さいごに deepgeo の出力（地理的位置）y を得る。以下の式は、サブネットワークとして、投稿時間サブネットワークの出力 f_{time} と投稿者のプロフィールにある居住地サブネットワークの出力 f_{loc} を連結した場合の式である。本研究では、この 3 種類のサブネットワークからなる deepgeo を採用する。

$$\hat{f} = f_{text} \oplus f_{time} \oplus f_{loc} \tag{13}$$

$$o = tanh(W_o \hat{f}) \tag{14}$$

$$y = softmax(o) \tag{15}$$

4.2 事例単位アテンションによる知識の取り込み

事例単位（ツイート単位）のアテンションによる知識の取り込み方法について説明する。この方法では、さきほど述べた Text Network からの出力 f_{text} に対して GeoKB の知識を取り込む。この方法は Annervaz ら [11] が提案した方法である。Annervaz らは、任意のモデルへ知識を取り込む一般的な枠組みとして提案しているが、我々は彼らの方法を文書ジオロケーションモデルに GeoKB を取り込む目的で使用する。

Annervaz 法では、まず、Text Network からの出力ベクトル f_{text} の次元と GeoKG から獲得した埋め込み表現の次元を合わせるために、以下の線形変換を適用する。ここで、W は $m \times n$ の重み行列であり、m は GeoKG から獲得した埋め込み表現の次元、n は f_{text} の次元数である。

$$C = ReLU(W f_{text}) \tag{16}$$

GeoKG から獲得した埋め込み表現について、エンティティと関係タイプを個別に処理するために、式 (16) により得られる出力をコピーし、エンティティ用の C_E と関係タイプ用の C_R を用意する。

GeoKG に含まれている i 番目の三つ組知識に注目したとき、関係元エンティティの埋め込み表現ベクトル e_i に対して、以下のようにアテンションを施し、エンティティの代表表現 $\hat{e}$ を得る。ここで、$|E|$ は GeoKG に含まれているエンティティ数である。

$$\alpha_{e_i} = \frac{exp(C_E^T e_i)}{\sum_{j=0}^{|E|-1} exp(C_E^T e_j)} \tag{17}$$

$$\hat{e} = \sum_{i=0}^{|E|-1} \alpha_{e_i} e_i \tag{18}$$

また同様に、関係タイプの埋め込み表現ベクトル r_i に対して、以下のようにアテンションを施し、関係タイプの代表表現 $\hat{r}$ を得る。ここで、$|R|$ は GeoKG に含まれている関係数である。

$$\alpha_{r_i} = \frac{exp(C_R^T r_i)}{\sum_{j=0}^{|R|-1} exp(C_R^T r_j)} \tag{19}$$

$$\hat{r} = \sum_{i=0}^{|R|-1} \alpha_{r_i} r_i \tag{20}$$

TransE では $h + r = t$ となるように学習しているので、GeoKB 全体の代表ベクトル F を先程求めた $\hat{e}$ と $\hat{r}$ を用いて、$F = \hat{e} \oplus \hat{r} \oplus (\hat{e} + \hat{r})$ とする。

最後に、この F を Text Network からの出力 f_{text} に取り込む。

$$F' = ReLU(W_k F) \tag{21}$$

$$f_{text+KG} = F' \oplus f_{text} \tag{22}$$

ここで、W_k は $u \times 3m$ の重み行列であり、u は Text Network からの出力ベクトル f_{text} の次元数である。

文書ジオロケーションの学習、推定を実行をする際は、この $f_{text+KG}$ を f_{text} の代わりに用いる。その他は deepgeo と同じである。

表 5　データサイズ

データ区分	データサイズ
学習データ	200,000 件
開発データ	4,000 件
評価データ	7,000 件

4.3　トークン単位アテンションによる知識の取り込み

Annervaz 法では、ツイート単位の処理となるため、ツイートの内容ごとに応じた知識の取り込み制御が難しい。そこで、次に、トークン単位のアテンションによる知識の取り込み方法について検討する。

deepgeo の Text Network は処理の途中で、4.1 節の式 (8) における g_t で表される文字列をトークンごとにまとめ上げた表現を獲得する。この g_t と GeoKG から獲得した埋め込み表現に対してアテンションを適用することで、より細かく知識を取り込む。4.2 節の式 (16) から式 (22) までのアテンションを計算し重み付きベクトルと連結するまでの処理を $knowledge\_infuse(\cdot)$ と定義すると、トークン毎のアテンションは以下の式で表される。

$$G_t = knowledge\_infuse(g_t) \oplus g_t \tag{23}$$

$$\hat{G_t} = ReLU(W_g G_t) \tag{24}$$

ここで、W_g は重み行列であり、式 (22) からの出力の次元を元の表現 g_t の次元に合わせるために適用する。

4.1 節の式 (9) における g_t を式 (24) により得られた $\hat{G_t}$ で置き換えることで、GeoKG から得られた知識をトークン毎に取り込んだ表現の獲得を行う。

5.　評価実験

5.1　実験設定

日本国内を対象として、文書ジオロケーション実験をおこなう。データセットとして、平川ら [18] が作成した日本語 Twitter データセットを用いた。本データセットは、日本国内から投稿されているもののみを対象とし、2014 年 1 月から 2015 年 12 月の間で収集されたもので、キーワードマッチングにより観光関係のトピックに絞り込まれた Tweet で構成されている。収集された各データセットのサイズを表 5 に示す。地理的位置として、日本国内の 47 都道府県を正解クラスとして採用した。ツイートに付与されたジオタグが示す経度緯度情報を逆ジオコーディングすることで正解となる都道府県の情報を取得した。ジオロケーションモデルの出力が正解クラスと等しい場合を正しく推定できたと考え、分類精度（評価データのうち正しく正解クラスを推定できた割合）で性能を評価した。

提案手法の有効性を検証するため、２つの既存手法と提案手法の性能を比較する。既存手法のひとつは、4.1 節で述べた deepgeo である。もうひとつは、deepgeo を改良したインジケータ付 deepgeo[18] である。インジケータ付 deepgeo は、あらかじめ地名を表す文字列の両端に特別なトークンを挿入する（インジケータを付与）ことでモデルに地名に関する情報を与える手法である。平川ら [18] の論文では、インジケータの付与に、地名辞書を用いる手法と MeCab[*6] を用いる手法の２手法を提案しているが、本研究ではこれらのうち、MeCab インジケータ付 deepgeo を比較手法として用いる。

4.2 節で説明した Annervaz 法 [11] では、アテンション処理の計算負荷の低減や勾配消失問題への対策として、4.2 節の手法を直接適用するのではなく、GeoKG の各要素（各エンティティ）の埋め込み表現ベクトルの替わりにそれらをクラスタリングして得られたクラスタベクトルを用いている。すなわち、クラスタリング後の i 番目のクラスタベクトルを ε_i とした時、4.2 節の式 (17) と式 (18) は以下の式に変更される。

$$\alpha_{\varepsilon_i} = \frac{exp(C_E^T \varepsilon_i)}{\sum_{j=0}^{|E|} exp(C_E^T \varepsilon_j)} \tag{25}$$

$$\hat{\varepsilon} = \sum_{i=0}^{|E|} \alpha_{\varepsilon_i} \varepsilon_i \tag{26}$$

本研究でも Annervaz ら [11] に従い、事例単位アテンションおよびトークン単位アテンションによって知識を取り込む際はクラスタリング処理を挟むことにした。クラスタリング手法には、k-means および Deep Compositional Code Learning(DCC)[19] を用いた。DCC は、単語埋め込み行列の圧縮のために提案された手法である。クラスタ数は、都道府県数の 47 を基準とし、その倍数となるように設定した。ただし、DCC の実装の都合上、クラスタ数を偶数に揃える必要があったため奇数値に対しては１加えることで偶数にした。

ジオロケーションモデルの各ハイパーパラメータは表 6 のように設定した。性能を比較する deepgeo とインジケータ付 deepgeo のハイパーパラメターは、平川ら [18] の設定に従う。

5.2　結果と考察

実験結果を表 7 に示す。各列の最良値を太字で示している。提案手法と比較手法の間で符号検定を実施し、deepgeo との間で有意水準 5% で有意差が認められた結果に「*」、有意水準 1%で有意差が認められた結果に「**」を付けている。また、インジケータ付 deepgeo との間で有意水準 5% で有意差が認められた結果に「+」、有意水準 1%で有意差が認められた結果に「++」を付けている。インジケータ付 deepgeo との間で有意差がある場合は「*」および「**」の表示は省略した。また、分類精度の上限の参考値として、人手で分類した場合の分類精度もあわせて示す。本実験データの中には非常に短いツイートも多く、人間が見ても地理的位置を推定することが困難な事例が少なからず含

[*6] http://taku910.github.io/mecab/

表 6 ハイパーパラメータ

ネットワーク	パラメータ	値
全体	バッチサイズ	32
	エポック数	10
	ドロップアウト率	0.2
	学習率	0.001
	最終表現の次元数	400
Text Network	最大長	300
	埋め込み次元	200
	ウィンドウ幅	10
	フィルター数	400
知識グラフの埋め込み表現	エンティティベクトルの次元数	100
	リレーションベクトルの次元数	100
	クラスタ数	[48, 94, 240, 470]

表 7 実験結果

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	-
事例単位アテンション	kmeans48	0.684^{++}	0.680
	kmeans94	$\mathbf{0.686}^{++}$	
	kmeans240	0.677**	
	kmeans470	0.672*	
事例単位アテンション	DCC48	0.682**	0.674
	DCC94	0.665	
	DCC240	0.674**	
	DCC470	0.673*	
トークン単位アテンション	kmeans48	0.680**	0.673
	kmeans94	0.675**	
	kmeans240	0.656	
	kmeans470	0.680**	
トークン単位アテンション	DCC48	0.685^{++}	**0.683**
	DCC94	0.685^{++}	
	DCC240	0.681**	
	DCC470	0.681**	
人手 (参考上限値)	-	0.767	-

まれている。そのため、参考上限値は 0.8 に満たない結果となっている。

表 7 から、提案手法は、すべての設定で deepgeo よりも高い分類精度を達成しており、また、一部を除くほぼすべての設定でインジケータ付 deepgeo よりも高い分類精度を達成していることがわかる。

提案手法におけるクラスタリング手法およびクラスタ数を変更した際の性能の違いについては、明らかな傾向は見受けられない。クラスタ数を減らすことで GeoKG 内の知識がより抽象化された形でジオロケーションモデルに取り込まれていると考えられるが、今回の実験ではこれに関する明瞭な影響は観察されなかった。事例単位アテンション（kmeans、クラスタ数 94）において、最良の分類精度を達成している。しかし、事例単位アテンションで kmeans クラスタリングを使うと、クラスタ数によって分類精度が比較的激しく増減している。最良値は事例単位アテンション（kmeans、クラスタ数 94）であるが、トークン単位アテンション（DCC）は、どのクラスタ数でも安定して高い分類精度を保っている。表の最右列に、各設定における、クラスタ数を変えた時の分類精度の平均値を示す。この平均値でみると、クラスタ数に対してはトークン単位アテンション（DCC）がもっとも安定して良い結果を達成していることがわかる。

次に、事例が地名（住所の一部）や施設名を含むか否かという条件と分類精度の間の関係を調査した。この調査では評価データを

- (a) 地名と施設名の両方を含む事例
- (b) 地名を含まず施設名を含む事例
- (c) 地名を含むが施設名を含まない事例
- (d) 地名と施設名の両方を含まない事例

ⓒ 2021 Information Processing Society of Japan

表 8 人手サンプリングによるデータ分割評価（カテゴリあたり 100 件収集）

	分類精度		
	deepgeo	Mecab インジケータ付 deepgeo	トークン単位アテンション (DCC48)
地名と施設名の両方を含む	0.86	0.90	0.89
地名を含まず施設名を含む	0.56	0.55	0.59
地名を含むが施設名を含まない	0.65	0.75	0.73
地名と施設名の両方を含まない	0.39	0.39	0.38

表 9 提案手法の成功事例

ツイートテキスト	正解	予測 deepgeo	予測 インジケータ付 deepgeo	予測 トークン単位アテンション
お土産に駅弁こうた。ふつうのがなくて 金のかけ紙のプレミアムなんだけど、どう プレミアムなのか知らないままこうてしもた @ 中央線 小淵沢駅 https://t.co/dB6BnuimF0	山梨県	奈良県	神奈川県	山梨県
おやすみなさい！ @ ホテル成島イン https://t.co/lKPqVoQAtT	群馬県	千葉県	東京都	群馬県
遅い母の日に、黒川温泉(の近く) にきたー #夏休み @ 旅館 湯之迫 https://t.co/rSSa61anpY	熊本県	大分県	静岡県	熊本県
仕事中に観光 #中津 #青の洞門 @ 青ノ洞門 https://t.co/b2WhoRWEsH	大分県	東京都	東京都	大分県

のように４つのカテゴリに分割し、各カテゴリでの分類精度を再評価した。ただし、自動では正確なカテゴリ分割処理が難しいため、人手でカテゴリ判定をおこなうことにし、各カテゴリの事例が 100 件になるまで作業した。

この結果を表 8 に示す。提案手法は、さきほどの議論を踏まえての安定性の良いトークン単位アテンション（DCC）を取り上げて評価する。また、クラスタ数は DCC の中で最も性能が良かった 48 とする。比較手法はさきほどと同じ２手法である。この表から、まず、どの手法においても、「(a) 地名と施設名の両方を含む事例」では分類精度が高くなり、反対に、「(d) 地名と施設名の両方を含まない事例」では分類精度が低くなることが確認できる。この結果は、人間の直感に沿うものと考えられる。また、手法間の性能差はこれらの２カテゴリではなく、地名か施設名のどちらかを含む場合（(b) および (c)）において生じているようである。deepgeo と比較すると、インジケータ付 deepgeo と提案手法はカテゴリ (c) での改善幅が大きく、インジケータ付 deepgeo と提案手法の比較では、カテゴリ (b) による改善が確認できる。以上の結果を整理すると、提案手法ではカテゴリ (d) では効果が期待できないものの、地名や施設名が１件以上含まれている事例においては、GeoKG 内の知識がうまくジオロケーション推定に反映されていることが伺える。特に、提案手法で取り込んでいる GeoKG では施設名が住所と関係性をもつグラフという形で構造化されており、このことがカテゴリ (b) の性能改善としてあらわれていると考えられる。

さいごに、比較手法では誤るが、提案手法では正しく予測できていた例を表 9 に示す。ツイートテキストにおいて、インジケータが付与されていた文字列に赤い下線、構築した GeoKG 内のエンティティに存在していた文字列に青い下線を引いている。インジケータ付 deepgeo と提案手法で比較すると、いずれの事例も、地名や施設名に対してインジケータが付与されているにも関わらず、インジケータ付 deepgeo では予測を誤っている。対して、提案手法は予測に成功しており、表 8 の結果も含め、GeoKG の知識を取り込む本研究の手法が住所やそれ以外の手掛かりとなり得る単語により対応できるようになっていることが確認できる。

6. おわりに

本論文では、住所および住所属性をもつ施設名の２種類の地理的エンティティに対して、それらの間の関係性を構造化した地理的知識グラフを準備し、この知識を深層学習ベースの文書ジオロケーションモデルに取り込む手法を提案した。エンティティ予測課題による地理的知識グラフの評価では、エンティティ間の双方向関係を考慮したグラフの方が単方向関係のグラフよりも地理的知識を効果的に表現できていることを確認した。また、日本国内の都道府県を推定対象とした文書ジオロケーション評価実験の結果、地理的知識グラフを取り込んだ提案手法は、既存手法よりも高い推定精度を達成することを確認した。

本稿では、評価実験を通して、提案手法に対する定量評

ⓒ 2021 Information Processing Society of Japan

価の結果を中心に報告した。今後は、より詳細な定性分析を実施することで、提案手法の挙動特性を把握していきたい。例えば、アテンション重みに注目することで、自明な手掛かりがない事例における重要トークンの分析などが挙げられる。また、今回は一つの地理的知識グラフを選択して文書ジオロケーションに取り込んだが、地理的知識グラフの大きさや構造が文書ジオロケーションに与える影響などについても考察を深めていきたい。

参考文献

[1] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko. “Translating Embeddings for Modeling Multi-Relational Data”. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’13, p. 2787–2795, Red Hook, NY, USA, 2013. Curran Associates Inc.

[2] P. Qiu, Jialiang Gao, L. Yu, and F. Lu. Knowledge Embedding with Geospatial Distance Restriction for Geographic Knowledge Graph Completion. *ISPRS Int. J. Geo Inf.*, Vol. 8, p. 254, 2019.

[3] L. Sloan, J. Morgan, W. Housley, M. Williams, A. Edwards, P. Burnap, and O. Rana. “Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter”. *Sociological Research Online*, Vol. 18, No. 3, p. 7, 2013.

[4] L. Chi, K. H. Lim, N. Alam, and C. J. Butler. “Geolocation Prediction in Twitter Using Location Indicative Words and Textual Features”. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp. 227–234, 2016.

[5] T. Fornaciari and D. Hovy. “Geolocation with Attention-Based Multitask Learning Models”. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 217–223, 2019.

[6] B. Han, P. Cook, and T. Baldwin. “Geolocation Prediction in Social Media Data by Finding Location Indicative Words”. In *Proceedings of COLING 2012*, pp. 1045–1062, 2012.

[7] G. Jayasinghe, B. Jin, J. Mchugh, B. Robinson, and S. Wan. ”CSIRO Data61 at the WNUT Geo Shared Task”. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp. 218–226, 2016.

[8] J. H. Lau, L. Chi, K. Tran, and T. Cohn. “End-to-end Network for Twitter Geolocation Prediction and Hashing”. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 744–753, 2017.

[9] Lisa Ehrlinger and Wolfram Wöß. Towards a Definition of Knowledge Graphs. 09 2016.

[10] Taro Miyazaki, Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. Twitter Geolocation using Knowledge-Based Methods. In *NUT@EMNLP*, 2018.

[11] K. M. Annervaz, S. Chowdhury, and A. Dukkipati. Learning beyond datasets: Knowledge Graph Augmented Neural Networks for Natural language Processing. In *NAACL-HLT*, 2018.

[12] B. Han, A. Rahimi, L. Derczynski, and T. Baldwin. “Twitter Geolocation Prediction Shared Task of the 2016 Workshop on Noisy User-generated Text”. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp. 213–217, 2016.

[13] Z. Wang, J. Zhang, J. Feng, and Z. Chen. “Knowledge Graph Embedding by Translating on Hyperplanes”. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI’14, p. 1112–1119. AAAI Press, 2014.

[14] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 687–696, Beijing, China, July 2015. Association for Computational Linguistics.

[15] Yi Zhang, Yong Gao, LuLu Xue, Si Shen, and KaiChen Chen. A common sense geographic knowledge base for GIR. *Science in China Series E: Technological Sciences*, Vol. 51, No. 1, pp. 26–37, 2008.

[16] Jiaoyan Chen, Shumin Deng, and Huajun Chen. Crowdgeokg: Crowdsourced geo-knowledge graph. In *China Conference on Knowledge Graph and Semantic Computing*, pp. 165–172. Springer, 2017.

[17] S. Lai, L. Xu, K. Liu, and J. Zhao. “Recurrent Convolutional Neural Networks for Text Classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2267–2273, 2015.

[18] 平川冬尉, 乾孝司. 日本語地理的位置推定課題におけるインジケータ付 deepgeo 法の提案と評価. 人工知能学会全国大会論文集, Vol. JSAI2020, pp. 3Rin473–3Rin473, 2020.

[19] Raphael Shu and Hideki Nakayama. Compressing word embeddings via deep compositional code learning. *arXiv preprint arXiv:1711.01068*, 2017.

ⓒ 2021 Information Processing Society of Japan

正誤表

下記の箇所に誤りがございました．お詫びして訂正いたします．

訂正箇所	誤	正
1 ページ 5 行目	(著者名欄の空白)	平川冬尉[1,a)]　乾孝司[1]
1 ページ フットノート	(空白)	1 筑波大学大学院 システム情報工学研究群 情報理工学位プログラム Graduate School of Science and Technology Degree Pro-grams in Systems and Information Engineering Master'sProgram,University of Tsukuba a) hirakawa@mibel.cs.tsukuba.ac.jp

ⓒ2021 Information Processing Society of Japan