DOI: 10.1111/tgis.12629

RESEARCH ARTICLE

Transactions in GIS WILEY

SE-KGE: A location-aware Knowledge Graph Embedding model for Geographic Question Answering and Spatial Semantic Lifting

Gengchen Mai[1] | Krzysztof Janowicz[1] | Ling Cai[1] | Rui Zhu[1] | Blake Regalia[1] | Bo Yan[2] | Meilin Shi[1] | Ni Lao[3]

[1]Department of Geography, University of California Santa Barbara, Santa Barbara, CA, USA

[2]LinkedIn Corporation, Mountain View, CA, USA

[3]SayMosaic Inc., Palo Alto, CA, USA

Correspondence
Gengchen Mai, Department of Geography, University of California Santa Barbara, Santa Barbara, CA, USA.
Email: gengchen_mai@geog.ucsb.edu

Abstract

Learning knowledge graph (KG) embeddings is an emerging technique for a variety of downstream tasks such as summarization, link prediction, information retrieval, and question answering. However, most existing KG embedding models neglect space and, therefore, do not perform well when applied to (geo)spatial data and tasks. Most models that do consider space primarily rely on some notions of distance. These models suffer from higher computational complexity during training while still losing information beyond the relative distance between entities. In this work, we propose a location-aware KG embedding model called *SE-KGE*. It directly encodes spatial information such as point coordinates or bounding boxes of geographic entities into the KG embedding space. The resulting model is capable of handling different types of spatial reasoning. We also construct a geographic knowledge graph as well as a set of geographic query–answer pairs called *DBGeo* to evaluate the performance of *SE-KGE* in comparison to multiple baselines. Evaluation results show that *SE-KGE* outperforms these baselines on the *DBGeo* data set for the geographic logic query answering task. This demonstrates the effectiveness of our spatially-explicit model and the importance of considering the scale of different geographic entities. Finally, we introduce a novel downstream task called *spatial semantic lifting* which links an arbitrary location in the study area to entities in the KG via some relations. Evaluation

on *DBGeo* shows that our model outperforms the baseline by a substantial margin.

1 | INTRODUCTION AND MOTIVATION

The term *Knowledge Graph* (KG) typically refers to a labeled and directed multi-graph of statements (called triples) about the world. These triples often originate from heterogeneous sources across domains. According to Nickel, Murphy, Tresp, and Gabrilovich (2015), most of the widely used KGs are constructed in a curated (e.g., WordNet), collaborative (e.g., Wikidata, Freebase), or auto semi-structured (e.g., YAGO (Hoffart, Suchanek, Berberich, & Weikum, 2013), DBpedia, Freebase) fashion rather than an automated unstructured approach (e.g., Knowledge Vault (Dong et al., 2014)). Despite containing billions of statements, these KGs suffer from *incompleteness and sparsity* (Dong et al., 2014; Lao, Mitchell, & Cohen, 2011; Mai, Janowicz, et al., 2019). To address these problems, many relational machine learning models (Nickel et al., 2015) have been developed for KG completion tasks, including several embedding-based techniques such as *RESCAL* (Nickel, Tresp, & Kriegel, 2012), *TransE* (Bordes, Usunier, Garcia-Duran, Weston, & Yakhnenko, 2013), *TransH* (Wang, Zhang, Feng, & Chen, 2014), HOLE (Nickel, Rosasco, & Poggio, 2016), R-GCN (Schlichtkrull et al., 2018), and TransGCN (Cai, Yan, Mai, Janowicz, & Zhu, 2019). The key idea of the embedding-based technique (Bordes et al., 2013; Cai et al., 2019; Nickel et al., 2016; Wang et al., 2014; Wang, Mao, Wang, & Guo, 2017) is to project entities and relations in a KG onto a continuous vector space such that entities and relations can be quantitatively represented as vectors/embeddings.

The aforementioned incompleteness and sparsity problems also affect the performance of downstream tasks such as question answering (Wang et al., 2018) since missing triples or links result in certain questions becoming unanswerable (Rajpurkar, Jia, & Liang, 2018). Consequently, researchers have recently focused on relaxing these unanswerable queries or predicting the most probable answers based on KG embedding models (Hamilton, Bajaj, Zitnik, Jurafsky, & Leskovec, 2018; Mai, Yan, Janowicz, & Zhu, 2019; Wang et al., 2018).

Most research on KG embeddings has neglected spatial aspects such as the location of geographic entities, despite the important role such entities play within KGs (Janowicz, Scheider, Pehle, & Hart, 2012). In fact, most of the current KG embedding models (e.g., *TransE*, *TransH*, TransGCN, R-GCN, and HOLE) ignore triples that contain *datatype properties*, and, hence, literals for dates, texts, numbers, geometries, and so forth. Put differently, properties such as `dbo:elevation`, `dbo:populationTotal`, and `dbo:areaWater`, to name but a few, are not considered during the training phase. Instead, these models strictly focus on triples with *object type properties*, leading to substantial information loss in practice. A few models do consider a limited set of datatypes. LiteralE (Kristiadi, Khan, Lukovnikov, Lehmann, & Fischer, 2019) is one example, which encodes numeric and date information into its embedding space, while MKBE (Pezeshkpour, Chen, & Singh, 2018) encodes images and unstructured texts. Therefore, in this work, we propose a novel technique which directly encodes spatial footprints, namely point coordinates and bounding boxes, thereby making them available while learning KG embeddings.

Geographic information forms the basis for many KG downstream tasks such as geographic knowledge graph completion (Qiu, Gao, Yu, & Lu, 2019), geographic ontology alignment (Zhu, Hu, Janowicz, & McKenzie, 2016), geographic entity alignment (Trisedya, Qi, & Zhang, 2019), geographic question answering (Mai, Yan, et al., 2019), and geographic knowledge graph summarization (Yan, Janowicz, Mai, & Zhu, 2019). In the following, we will focus on geographic logic query answering as an example and more concretely on conjunctive graph queries (CGQs) or logic queries (Hamilton et al., 2018). Due to the sparsity of information in KGs, many (geographic) queries are unanswerable *without spatial or non-spatial reasoning*. Knowledge graph embedding techniques have, therefore, been developed to handle unanswerable questions (Hamilton et al., 2018; Mai, Janowicz, et al., 2019; Mai, Yan, et al., 2019; Wang et al., 2018) by inferring new triples in the KG embedding space based on existing ones. However, since most KG embedding models cannot handle *datatype properties*, thus cannot encode

geographic information into the KG embedding space, they perform spatial reasoning tasks poorly in the KG embedding space, which in turn leads to a poor performance in handling unanswerable geographic questions

```
SELECT ?State WHERE {
 ?RiverMouth dbo:state ?State.                    (a)
 ?River dbo:mouthPosition ?RiverMouth.              (b)
 ?River dbp:etymology dbr:Alexander_von_Humboldt.   (c)
}
```

Listing 1: Query q_A: An unanswerable SPARQL query over DBpedia which includes a partonomy relation

One example of unanswerable geographic questions that can be represented as a logic query is *Which states contain the mouth of a river which is named after Alexander von Humboldt?* (Query q_A). The corresponding SPARQL query is shown in Listing 1. Running this query against the DBpedia SPARQL endpoint yields no results. In fact, two rivers are named after von Humboldt—`dbr:Humboldt _ River` and `dbr:North _ Fork _ Humboldt _ River`—and both have mouth positions as entities in DBpedia (`dbr:Humboldt _ River _ _ mouthPosition _ _ 1` and `dbr:North _ Fork _ Humboldt _ River _ _ sourcePosition _ _ 1`). However, the `dbo:state` (or `dbo:isPartOf`) relation between these river mouths and other geographic features such as states is missing. This makes Query q_A unanswerable (graph query pattern (a) in Listing 1). If we use the locations of the river mouths to perform a simple point-in-polygon test against the borders of all states in the U.S., we can deduce that `dbr:Nevada` contains both river mouths

```
SELECT ?place WHERE {
 dbr:Yosemite_National_Park dbo:nearestCity ?place.
}
```

Listing 2: Query q_B: A SPARQL query over DBpedia which indicates a simple point-wise distance relation

Another example is the query in Listing 2, which asks for *the nearest city to Yosemite National Park* (Query q_B). If the triple(`dbr:Yosemite _ National _ Park dbo:nearestCity dbo:Mariposa, _ California`) is missing from the current knowledge graph, Query q_B becomes unanswerable, while it could simply be inferred by a distance-based query commonly used in GIS. Similar cases can include cardinal directions such as `dbp:north`. All these observations lead to the following research question: how could we enable spatial reasoning via *partonomic relations*, *pointwise metric relations*, and *directional relations* in KG embedding-based systems?

One may argue that classical spatial reasoning can be used instead of direct location encoding to obtain answers to the aforementioned questions. This is partially true for data and query endpoints that support GeoSPARQL and for data sets that are clean and complete. However, in some cases even GeoSPARQL-enabled query endpoints cannot accommodate spatial reasoning due to inherent challenges of representing spatial data in KGs . These challenges stem from principles of conceptual vagueness and uncertainty (Regalia, Janowicz, & McKenzie, 2019), and are further complicated by technical limitations. In this study we aim to enable the model to perform *implicit spatial reasoning* in the hidden embedding space. Instead of performing classical spatial reasoning by explicitly carrying out spatial operations during query time, the spatial information (points or bounding boxes) of geographic

entities (e.g., *Indianapolis*) is directly encoded into the entity embeddings which are jointly optimized with relation embeddings (e.g., *isPartOf*). The trained embeddings of geographic entities encode their spatial information, while by embedding the spatial relations we also hope to capture some of their implicit semantics for simple spatial reasoning tasks. At query time, a normal link prediction process can be used to answer geographic questions and no explicit spatial reasoning is needed. More details of this example can be found in Section 7.

Existing approaches are only able to incorporate spatial information into the KG embedding space in a very limited fashion, (e.g., through their training procedures). Furthermore, they estimate entity similarities based on some form of distance measures among entities, and ignore their absolute positions or relative directions. For example, Trisedya et al. (2019) treated geographic coordinates as strings (sequences of characters) and used a compositional function to encode these coordinate strings for geographic entity alignment. In order to incorporate distance relations between geographic entities, both Mai, Yan, et al. (2019) and Qiu et al. (2019) borrowed the translation assumption from TransE (Bordes et al., 2013). For each geographic triple $s = (h, r, t)$ in the KG, where h and t are geographic entities, the geospatial distance between h and t determines the frequency of resampling this triple such that triples containing two closer geographic entities are sampled more frequently, and thus these two geographic entities are closer in the embedding space. Similarly, Yan et al. (2019) used distance information to construct virtual spatial relations between geographic entities during the KG summarization process. This data conversion process (coordinates to pairwise distances) is unnecessarily expensive and causes information loss (e.g., absolute positions and relative directional information). In this work, we explore directly encoding entity locations into a high-dimensional vector space, which preserves richer spatial information than distance measures. These location embeddings can be trained jointly with KG embeddings.

Location encoders (Chu et al., 2019; Mac Aodha, Cole, & Perona, 2019; Mai et al., 2020) refer to the neural network models which encode a pair of coordinates into a high-dimensional embedding which can be used in downstream tasks such as geo-aware fine-grained image classification (Chu et al., 2019; Mac Aodha et al., 2019; Mai et al., 2020) and Point Of Interest (POI) type classification (Mai et al., 2020). Mai et al. (2020) showed that multi-scale grid cell representation outperforms commonly used kernel based methods (e.g., the radial basis function (RBF)) as well as the single-scale location encoding approaches. Given the success of location encoding in other machine learning tasks, the question is whether we can incorporate the location encoder architecture into a KG embedding model to make it spatially explicit (Mai, Yan, et al., 2019). One initial idea is to directly use a location encoder as the entity encoder which encodes the spatial footprint (e.g., coordinates) of a geographic entity into a high-dimensional vector. Such entity embeddings can be used in different decoder architectures for different tasks. However, several challenges remain to be solved for this initial approach.

First, point location encoding can handle pointwise metric relations such as distance (e.g., `dbo:nearestCity`) as well as directional relations (e.g., `dbp:north`, `dbp:south`) in KGs, but it is not easy to encode regions which are critical for relations such as containment (e.g., `dbo:isPartOf`, `dbo:location`, `dbo:city`, `dbo:state`, and `dbo:country`). For example, in Query q_A, the location encoder can encode `dbr:Yosemite_National_Park` and `dbo:Mariposa,_California` as two high-dimensional embeddings based on which distance relations can be computed since the location embeddings preserve the relative distance information between locations (Mai et al., 2020). However, point locations and location embeddings are insufficient to capture more complex relations between geographic entities such as containment as these require more complex spatial footprints (e.g., polygons). This indicates that we need to find a way to represent geographic entities as *regions* instead of points in the embedding space based on location encoders, especially for large-scale geographic entities such as `dbr:California`, which is represented as a single pair of coordinates (a point) in many widely used KGs. We call this the *scale effect* to emphasize the necessity of encoding the spatial extents of geographic entities instead of points, especially for large-scale geographic entities.

The second challenge is how to seamlessly handle geographic and non-geographic entities together in the same entity encoder framework. Since the location encoder is an essential *component* of the entity encoder, how should we deal with non-geographic entities that do not have spatial footprints? This is a non-trivial problem. For example, in order to weight triples using distance during KG embedding training, Qiu et al. (2019) constructed a geographic KG

which only contains geographic entities. Mai, Yan, et al. (2019) partially solved the problem by using a lower bound l as the lowest triple weight to handle non-geographic triples. However, this mechanism cannot distinguish triples involving both geographic and non-geographic entities from triples that only contain non-geographic entities.

The third challenge is how to capture the spatial and other semantic aspects at the same time when designing a spatially-explicit KG embedding model based on location encoders. The embedding of a geographic entity is expected to capture both its spatial (e.g., spatial extent) and other semantic information (e.g., type information) since both of them are necessary to answer geographic questions. Take Query q_A in Listing 1 as an example. Intuitively, to answer this query, the spatial information is necessary to perform partonomical reasoning to select geographic entities which *contain* a given river mouth, while type information is required to filter the answers and get entities with type *state*. Therefore, we need both spatial and type information encoded in the entity embeddings to answer this question. The traditional KG embedding models fail to capture the spatial information, which leads to lower performance in geographic question answering.

Finally, thanks to the inductive learning nature of the location encoder, another interesting question is how to design a spatially-explicit KG embedding model so that it can be used to infer new relations between entities in a KG and any arbitrary location in the study area. We call this task *spatial semantic lifting* by analogy with traditional *semantic lifting*, which refers to the process of associating unstructured content to semantic knowledge resources (De Nicola, DiMascio, Lezoche, & Tagliano, 2008). For example, given any location $\mathbf{x}_i$, we may want to ask *which radio station broadcasts at* $\mathbf{x}_i$, that is, to infer `dbo:broadcastArea`. None of the existing KG embedding models can solve this task.

In this work we develop a spatially-explicit KG embedding model, *SE-KGE*, which directly solves those challenges. The contributions of our work are as follows:

1. We develop a spatially-explicit KG embedding model (*SE-KGE*), which applies a location encoder to incorporate spatial information (coordinates and spatial extents) of geographic entities. To the best of our knowledge, this is the first KG embedding model that can incorporate spatial information, especially spatial extents, of geographic entities into the model architecture.
2. *SE-KGE* is extended to an end-to-end geographic logic query answering model which predicts the most probable answers to unanswerable geographic logic queries over a KG.
3. We apply *SE-KGE* to a novel task, spatial semantic lifting. Evaluations show that our model can substantially outperform the baseline by 9.86% on AUC and 9.59% on APR for the *DBGeo* data set. Furthermore, our analysis shows that this model can achieve implicit spatial reasoning for different types of spatial relations.

The remainder of this article is structured as follows. We briefly summarize related work in Section 2. We then discuss basic concepts in Section 3. In Section 4 we formalize the query answering and spatial semantic lifting task. Then, in Section 5, we give an overview of the logic query answering task before introducing our method. In Section 6 we describe the *SE-KGE* architecture. We then summarize our experiments and evaluations in Section 7. Finally, we provide a conclusion in Section 8.

2 | RELATED WORK

In this section, we briefly review related work on KG embeddings, query answering, and location encoding.

2.1 | Knowledge graph embedding

Learning KG embeddings is an emerging topic in both the Semantic Web and Machine Learning fields. The idea is to represent entities and relations as vectors or matrices within an embedding space such that these distributed

representations can be easily used in downstream tasks such as KG completion and question answering. Many KG embedding models have been proposed such as *RESCAL* (Nickel et al., 2012), *TransE* (Bordes et al., 2013), and TransH (Wang et al., 2014). Most of these approaches cannot handle triples with datatype properties nor triples involving spatial footprints.

The only KG embedding methods considering distance decay between geographic entities are Qiu et al. (2019) and Mai, Yan, et al. (2019). Mai, Yan, et al. (2019) computed the weight of each geographic triple $s = (h, r, t)$ as $\max\left(\ln \frac{D}{\mathrm{dis}(h,t)+\varepsilon}, l\right)$, where h and t are geographic entities, and D is the longest (simplified) Earth surface distance. ε is a hyperparameter to avoid zero denominator and l is the lowest edge weight we allow for each triple. As for non-geographic triples, l is used as the triple weight. Then this KG is treated as an undirected, unlabeled, edge-weighted multi-graph. An edge-weighted PageRank is applied on this multi-graph. The PageRank score for each node/entity captures the structure information of the original KG as well as the distance decay effect among geographic entities. These scores are used in turn as weights to sample the entity context from the 1-degree neighborhood of each entity which is used in the KG embedding training process. As for Qiu et al. (2019), the distance decay effect was deployed in a triple negative sampling process. Given a triple $s = (h, r, t)$ in the KG, each negative triple $s' = (h', r, t')$ of it was assigned a weight based on:

$$w_{\mathrm{geo}} = \left(1, +, |, \log_{10}, \frac{\mathrm{dis}(h,t)+\theta}{\mathrm{dis}(h',t')+\theta'}, |\right)^{-1} \tag{1}$$

where θ is a hyperparameter to avoid a zero denominator. w_{geo} is used in the max-margin loss function for the embedding model training. Note that non-geographic triples are not considered in Qiu et al. (2019). We can see that, instead of directly encoding an entity's location, they rely on some form of distance measures as weights for triple resampling. This process is computationally expensive and does not preserve other spatial properties such as direction. In contrast, our work introduces a direct encoding approach to handle spatial information.

2.2 | Query answering

Compared to link prediction (Bordes et al., 2013), query answering (Hamilton et al., 2018; Mai, Janowicz, et al., 2019; Wang et al., 2018) focuses on a more complex problem since answering a query requires a system to consider multiple triple patterns together. Wang et al. (2018) designed an algorithm to answer a subset of SPARQL queries based on a pretrained KG embedding model. However, this is not an end-to-end model since the KG embedding training and query answering process are separated. Hamilton et al. (2018) proposed an end-to-end logic query answering model, *GQE*, which can answer conjunctive graph queries. CGA (Mai, Janowicz, et al., 2019) further improved *GQE* by using a self-attention-based intersection operator. In our work, we will utilize *GQE* and CGA (Mai, Janowicz, et al., 2019) as the underlying logic query answering baseline. We provide an overview of logic query answering in Section 5.

2.3 | Location encoding

Generating representations of points/locations that can benefit representation learning is a long-standing problem in machine learning. There are many well-established methods such as the *kernel trick* (Schölkopf, 2001) widely used in support vector machine classification and regression. However, these location representation methods use the positions of training examples as the centers of Gaussian kernels and thus need to memorize the training examples.

Kejriwal and Szekely (2017) proposed a graph embedding approach to representing GeoNames locations as high-dimensional embeddings. They converted the locations in GeoNames into a weighted graph where locations are nodes and the weight of each edge is computed based on the distance between two locations. Then a GloVe (Pennington, Socher, & Manning, 2014) word embedding model is applied on this generated graph to obtain the embedding for each location. Despite its novelty, this model is in a transductive learning setting, which means that if new locations are added, the weighted graph has to be regenerated and the whole model needs to be retrained. In other words, this embedding approach cannot be easily generalized to unseen locations. This calls for inductive learning(Hamilton, Ying, & Leskovec, 2017a) models.

Recently, a location encoding technique (Chu et al., 2019; Mac Aodha et al., 2019; Mai et al., 2020) has been proposed to directly encode a location (pair of coordinates) $\mathbf{x}$ as a high-dimensional vector which can be incorporated into multiple downstream tasks. As shown by Mai et al. (2020), the advantages of location encoding are that: (1) it can preserve absolute position information as well as relative distance and direction information between locations; and (2) it does not need to memorize the positions of training examples as all kernel-based methods do (Schölkopf et al., 1997); (3) in contrast to many transductive learning models, it is an inductive learning model (Battaglia et al., 2018) which can encode any location/point no matter whether it appears in the training data set or not.

In theory, we can adopt any location encoder (Chu et al., 2019; Mac Aodha et al., 2019; Mai et al., 2020) to capture the spatial information of each geographic entity e_i in a knowledge graph $\mathcal{G}$. In this work, we utilize the *Space2Vec* (Mai et al., 2020) location encoder, which is inspired by Nobel Prize-winning neuroscience research about grid cells (Abbott & Callaway, 2014) as well as the position encoding module of the Transformer model (Vaswani et al., 2017). *Space2Vec* first encodes a location $\mathbf{x}$ as a multi-scale periodic representation $PE(\mathbf{x})$ by using sinusoidal functions with different frequencies and then feeds the resulting embedding into an N-layer feed-forward neural network $\mathbf{NN}()$:

$$\text{LocEnc}^{(x)}(\mathbf{x}) = \mathbf{NN}(PE(\mathbf{x})). \tag{2}$$

The advantages of such a location encoder compared to previous work (Chu et al., 2019; Mac Aodha et al., 2019) are that: (1) it can be shown that location embeddings from *Space2Vec* are able to preserve global position information as well as relative distance and direction; and (2) the multi-scale representation learning approach outperforms traditional kernel-based methods (e.g., RBF) as well as single-scale location encoding approaches (Chu et al., 2019; Mac Aodha et al., 2019) for several machine learning tasks. In the following, we will use $\text{LocEnc}^{(x)}()$ to denote the *Space2Vec* model.

3 | BASIC CONCEPTS

Definition 1 (Geographic Knowledge Graph)A geographic knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a directed edge and node labeled multigraph where $\mathcal{V}$ is the set of entities/nodes and $\mathcal{E}$ is the set of directed edges. Any directed and labeled edge will be called a triple $s = (h, r, t)$ where the nodes become heads $h \in \mathcal{V}$ and tails $t \in \mathcal{V}$, and the role label $r \in \mathcal{R}$ will be called the relationship between them. The set of triples/statements contained by $\mathcal{G}$ is denoted by $\mathcal{T}$, and $\mathcal{R}$ denotes as the set of relations (predicates, edge labels) in $\mathcal{G}$. Each triple can also be represented as $r(h, t)$, or $r^{-1}(t, h)$, where r^{-1} indicates the inverse relation of r. Domain(r) and Range(r) indicate the domain and range of relation r.

$\Gamma()$: $\mathcal{V} \rightarrow \mathcal{C}$ is a function which maps an entity $e \in \mathcal{V}$ to a unique type $c \in \mathcal{C}$, where $\mathcal{C}$ is the set of all entity types in $\mathcal{G}$.[1]

The geographic entity set $\mathcal{V}_{pt}$ is a subset of $\mathcal{V}$ ($\mathcal{V}_{pt} \subseteq \mathcal{V}$). $\mathcal{PT}(\cdot)$ is a mapping function that maps any geographic entity $e \in \mathcal{V}_{pt}$ to its geographic location (coordinates) $\mathcal{PT}(e) = \mathbf{x}$, where $\mathbf{x} \in \mathcal{A} \subseteq \mathbb{R}^2$. Here $\mathcal{A}$ denotes the bounding box containing all geographic entities in the studied knowledge graph $\mathcal{G}$. We call it the study area.

$\mathcal{V}_{pn}$ is a subset of $\mathcal{V}_{pt}$ ($\mathcal{V}_{pn} \subseteq \mathcal{V}_{pt}$) which represents the set of large-scale geographic entities whose spatial extent cannot be ignored. In this work, we use a bounding box to represent a geographic entity's spatial footprint. $\mathcal{PN}(\cdot)$ is a mapping function defined on $\mathcal{V}_{pn}$ that maps a geographic entity $e \in \mathcal{V}_{pn}$ to its spatial extent $\mathcal{PN}(e)$ and $\mathcal{PN}(e) = [\mathbf{x}^{\min};\mathbf{x}^{\max}] \in \mathbb{R}^4$. In the vector concatenation above, $\mathbf{x}^{\min}, \mathbf{x}^{\max} \in \mathcal{A} \subseteq \mathbb{R}^2$ indicate the southwest and northeast point of the entity's bounding box.

Note that in many existing KGs, a triple can include a datatype property (e.g., `dbo:abstract`) implying that the tail is a literal. In line with related work (Bordes et al., 2013; Nickel et al., 2015, 2016; Wang et al., 2017), we do not consider triples of this kind here in general. However, we do consider datatype properties about the spatial footprints of geographic entities implicitly by using $\mathcal{PT}(\cdot)$ or $\mathcal{PN}(\cdot)$.2 While we do not model them directly as triples, we use the spatial footprints of geographic entities as input features for the entity encoder.

Definition 2 (Conjunctive Graph Query (CGQ)) A conjunctive graph query (or logic query) is a query $q \in Q(\mathcal{G})$ that can be written as:

$$q = V_?.\exists V_1, V_2, \ldots, V_m: b_1 \wedge b_2 \wedge \ldots \wedge b_n, \tag{3}$$

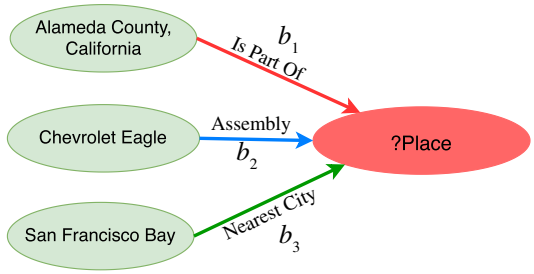
where $b_i = r_i(e_k, V_l)$, with $V_l \in \{V_?, V_1, V_2, \ldots, V_m\}, e_k \in \mathcal{V}, r \in \mathcal{R}$, or $b_i = r_i(V_k, V_l)$, with $V_k, V_l \in \{V_?, V_1, V_2, \ldots, V_m\}, k \neq l, r \in \mathcal{R}$.

Here $Q(\mathcal{G})$ is the set of all conjunctive graph queries that can be asked over $\mathcal{G}$. $V_?$ denotes the target variable of Query q (*target node*) which will be replaced with the answer entity a, while $V_1, V_2, \ldots, V_m$ are existentially quantified bound variables (*bound nodes*). $\{e_k | e_k \text{ in } q\}$ is a set of anchor nodes and b_i is a basic graph pattern in this CGQ. We define the dependency graph of q as the graph with basic graph pattern $\{b_1, \ldots, b_n\}$ formed between the anchor nodes $\{e_k | e_k \text{ in } q\}$ and the variable nodes $V_?, V_1, V_2, \ldots, V_m$ (Figure 1). Each conjunctive graph query can be written as a SPARQL query.3

Note that the dependency graph of q represents computations on the KG and is commonly assumed to be a *directed acyclic graph* (DAG) (Hamilton et al., 2018) where the entities (*anchor nodes*) e_k in q are the source nodes and the target variable $V_?$ is the unique sink node. This restriction makes the logic query answering task in line with the usual question answering setup (e.g., semantic parsing (Berant, Chou, Frostig, & Liang, 2013; Liang, Berant, Le, Forbus, & Lao, 2017)).

Definition 3 (Geographic Conjunctive Graph Query (GCQG))A conjunctive graph query $q \in Q(\mathcal{G})$ is said to be a geographic conjunctive graph query (GCGQ) if the answer entity a corresponding to the target variable $V_?$ is a geographic entity, that is, $a = \varphi(\mathcal{G}, q) \wedge a \in \mathcal{V}_{pt}$, where $\varphi(\mathcal{G}, q)$ indicates the answer when executing Query q on $\mathcal{G}$. We denote all possible GCGQs on $\mathcal{G}$ by $Q_{\text{geo}}(\mathcal{G}) \subseteq Q(\mathcal{G})$.

An example geographic conjunctive query q_C is shown in Figure 1 whose corresponding SPARQL query is shown in Listing 3. The corresponding natural language question is *[Which city in Alameda County, California is the assembly place of Chevrolet Eagle and the nearest city to San Francisco Bay?]*. This query is especially interesting since it includes a non-spatial relation (`dbo:assembly`), a pointwise metric spatial relation (`dbo:nearestCity`), and a partonomy relation (`dbo:isPartOf`). Note that executing each basic graph pattern in Query q_C over DBpedia will yield multiple answers. For example, b_1 will return all subdivisions of Alameda County, California. b_2 matches multiple assembly places of Chevrolet Eagle, such as `dbr:Oakland_California`, `dbr:Oakland_Assembly`, and `dbr:Flint_Michigan`. Interestingly, `dbr:Oakland_Assembly` should be located in `dbr:Oakland_California`, while there is no relationship between them in DBpedia except for their spatial footprints, from which it can be inferred that they are close to each other. b_3 will return three entities4: `dbr:San_Francisco`, `dbr:San_Jose`, `_California` and `dbr:Oakland_California`. Combining



$$?Place : Is\ Part\ Of\ (Alameda\ County, ?Place) \wedge$$
$$Assembly\ (Chevrolet\ Eagle, ?Place) \wedge$$
$$Nearest\ City(San\ Francisco\ Bay, ?Place)$$

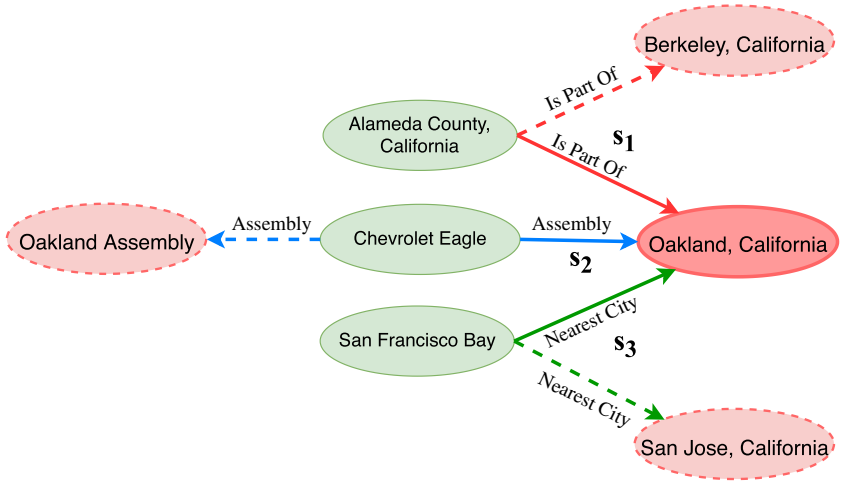


FIGURE 1 Query q_C. (Top) Conjunctive graph query and directed acyclic graph of the query structure corresponding to the SPARQL query in Listing 3. b_1, b_2, and b_3 indicate three basic graph patterns in Query q_C. `?Place` is the target variable shown as the red node, while the three green nodes are anchor nodes. There is no bound variable in this query. (Bottom) The matched underlining knowledge graph patterns represented by solid arrows. s_1, s_2, and s_3 indicate the matched triples for b_1, b_2, and b_3 respectively for Query q_C

these three basic graph patterns will yield one answer, `dbr:Oakland, _ California`. In our KG, both triples s_1 (see Figure 1) and s_2 are missing, which makes Query q_C an unanswerable geographic query

.

```
SELECT ?place  WHERE{
 ?place dbo:isPartOf dbr:Alameda_County,_California.      (1)
 dbr:Chevrolet_Eagle dbo:assembly ?place.                 (2)
 dbr:San_Francisco_Bay dbo:nearestCity ?place.            (3)
}
```

Listing 3: Query q_C: A geographic conjunctive query which is rewritten as a SPARQL query over DBpedia including both non-spatial relations and different types of spatial relation.

4 | PROBLEM STATEMENT

In this work, we focus on two geospatial tasks: *geographic logic query answering* and *spatial semantic lifting*.

Task 1 (Logic Query Answering)Given a geographic knowledge graph $\mathcal{G}$ and an unanswerable conjunctive graph query $q \in Q(\mathcal{G})$ (i.e., $\varphi(\mathcal{G},q) = \emptyset$), a query embedding function $\mathbf{\Phi}_{\mathcal{G},\theta}(q): Q(\mathcal{G}) \rightarrow \mathbb{R}^d$, which is parameterized by θ, is defined to map q to a vector representation of d dimension. The most probable answer to q is the entity nearest to $\mathbf{q} = \mathbf{\Phi}_{\mathcal{G},\theta}(q)$ in the embedding space:

$$a' = \arg\max_{e_i \in \mathcal{V}} \Omega(\mathbf{\Phi}_{\mathcal{G},\theta}(q), \mathrm{Enc}(e_i)) = \arg\max_{e_i \in \mathcal{V}} \Omega(\mathbf{q}, \mathbf{e}_i) \tag{4}$$

Here $\mathbf{e}_i = \mathrm{Enc}(e_i) \in \mathbb{R}^d$ is the entity embedding of e_i produced by an embedding encoder Enc(). $\Omega(\cdot)$ denotes the cosine similarity function:

$$\Omega(\mathbf{q}, \mathbf{e}_i) = \frac{\mathbf{q} \cdot \mathbf{e}_i}{\| \mathbf{q} \| \| \mathbf{e}_i \|} \tag{5}$$

Note that q can be a geographic query or non-geographic query (i.e., $q \in (Q(\mathcal{G}) \setminus Q_{\mathrm{geo}}(\mathcal{G})) \vee Q_{\mathrm{geo}}(\mathcal{G})$). *Geographic logic query answering* refers to a logic query answering process over $Q_{\mathrm{geo}}(\mathcal{G})$. The query embedding function $\mathbf{\Phi}_{\mathcal{G},\theta}(q)$ is constructed based on all three components of *SE-KGE* without any extra parameters: Enc(), $\mathcal{P}()$, and $\mathcal{I}()$ (i.e., $\theta = \{\theta_{\mathrm{Enc}}, \theta_{\mathcal{P}}, \theta_{\mathcal{I}}\}$).

Task 2 (Spatial Semantic Lifting)Given a geographic knowledge graph $\mathcal{G}$ and an arbitrary location $\mathbf{x} \in \mathcal{A} \subseteq \mathbb{R}^2$ from the current study area $\mathcal{A}$, and a relation $r \in \mathcal{R}$ such that $\mathrm{Domain}(r) \subseteq \mathcal{V}_{pt}$, we define a spatial semantic lifting function $\mathbf{\Psi}_{\mathcal{G},\theta_{\mathrm{ssl}}}(\mathbf{x},r): \mathcal{A} \times \mathcal{R} \rightarrow \mathbb{R}^d$, which is parameterized by θ_{ssl}, to map $\mathbf{x}$ and r to a vector representation of d dimension (i.e., $\mathbf{s} = \mathbf{\Psi}_{\mathcal{G},\theta_{\mathrm{ssl}}}(\mathbf{x},r) \in \mathbb{R}^d$). A nearest-neighbor search is utilized to search for the most probable entity $e' \in \mathcal{V}_{pt}$ so that a virtual triple can be constructed between location $\mathbf{x}$ and e' (i.e., $r(\mathbf{x}, e')$), where:

$$e' = \arg\max_{e_i \in \mathcal{V}} \Omega(\mathbf{\Psi}_{\mathcal{G},\theta_{\mathrm{ssl}}}(\mathbf{x},r), \mathrm{Enc}(e_i)) = \arg\max_{e_i \in \mathcal{V}} \Omega(\mathbf{s}, \mathbf{e}_i) \tag{6}$$

The spatial semantic lifting function $\mathbf{\Psi}_{\mathcal{G},\theta_{\mathrm{ssl}}}(\mathbf{x},r)$ consists of two components of *SE-KGE* without any extra parameter: Enc() and $\mathcal{P}()$ (i.e., $\theta_{\mathrm{ssl}} = \{\theta_{\mathrm{Enc}}, \theta_{\mathcal{P}}\}$). This spatial semantic lifting task is related to the link prediction task (Lao et al., 2011) which is commonly used in the KG embedding literature (Bordes et al., 2013; Cai et al., 2019; Nickel et al., 2016). The main difference is that instead of predicting links between entities in the original knowledge graph $\mathcal{G}$ as link prediction does, spatial semantic lifting links an arbitrary location $\mathbf{x}$ to $\mathcal{G}$. *Since none of the existing KG embedding models can directly encode locations, they cannot be used for spatial semantic lifting.*

5 | LOGIC QUERY ANSWERING

Before introducing our *SE-KGE* model, we will first give an overview of how previous work (Hamilton et al., 2018; Mai, Janowicz, et al., 2019) tackled the logic query answering task with KG embedding models. Generally speaking, a logic query answering model is composed of three major components: entity encoder Enc(), projection operator $\mathcal{P}()$, and intersection operator $\mathcal{I}()$.

1. *Entity encoder Enc()*. This represents each entity as a high-dimensional vector (embedding).

2. *Projection operator* $\mathcal{P}()$. Given a basic graph pattern $b = r(e_i, V_j)$ (or $b = r(V_i, V_j)$) in a CGQ q, while the subject embedding $\mathbf{e}_i$ (or $\mathbf{v}_i$) of entity e_i (or variable V_i) is known beforehand, $\mathcal{P}()$ *projects* the subject embedding through a relation-specific matrix to predict the embedding of V_j.
3. *Intersection operator* $\mathcal{I}()$. This integrates different predicted embeddings of the same variable (e.g., V_j) from different basic graph patterns into one single embedding to represent this variable.

Given these three neural network modules, any CGQ q can be encoded by following their DAG query structures such that the embedding of the unique target variable $V_?$ for each query can be obtained: $\mathbf{v}_?$. We call it *query embedding* $\mathbf{q} = \mathbf{\Phi}_{\mathcal{G},\theta}(q) = \mathbf{v}_?$ for CGQ q. Then the most probable answer is obtained by a nearest-neighbor search for $\mathbf{q}$ in the entity embedding space (see Equation 4). Our work will follow the same model component setup and query embedding computing process. However, neither Hamilton et al. (2018) nor Mai, Janowicz, et al. (2019) considered encoding spatial information of geographic entities into the entity embedding space, which is the core contribution of our work. Moreover, we extend the current model architecture such that it can also be applied to the spatial semantic lifting task. This new task cannot be handled by previous work. In the following, we will use $.^{(GQE)}$ and $.^{(CQA)}$ to indicate that these are model components used by Hamilton et al. (2018) and Mai, Janowicz, et al. (2019):

5.1 | Entity encoder

In general, an entity encoder aims to represent any entity in a KG as a high-dimensional embedding so that it can be fed into subsequent neural network modules. The normal practice shared by most KG embedding models (Bordes et al., 2013; Cai et al., 2019; Mai, Janowicz ,& Yan, 2018; Mai, Yan, et al., 2019; Nickel et al., 2016; Qiu et al., 2019; Schlichtkrull et al., 2018; Wang et al., 2014) is to initialize an embedding matrix randomly where each column indicates an embedding for a specific entity. The entity encoding becomes an *embedding lookup* process, and these embeddings will be updated during the neural network backpropagation during training time.

Previous work has demonstrated that most of the information captured by entity embeddings is type information (Hamilton, Ying, & Leskovec, 2017b; Hamilton et al., 2018). So Hamilton et al. (2018) and Mai, Janowicz, et al. (2019) took a step further and used a type-specific embedding lookup approach. We call the resulting module entity feature encoder $\mathsf{Enc}^{(c)}()$.

Definition 4 (Entity Feature Encoder $\mathsf{Enc}^{(c)}()$)Given any entity $e_i \in \mathcal{V}$ with type $c_i = \Gamma(e_i) \in \mathcal{C}$ from $\mathcal{G}$, the entity feature encoder $\mathsf{Enc}^{(c)}()$ computes the feature embedding $\mathbf{e}_i^{(c)} \in \mathbb{R}^{d^{(c)}}$ which captures the type information of entity e_i by using an embedding lookup approach:

$$\mathbf{e}_i^{(c)} = \mathsf{Enc}^{(c)}(e_i) = \frac{\mathbf{Z}_{c_i}\mathbf{h}_i^{(c)}}{\| \mathbf{Z}_{c_i}\mathbf{h}_i^{(c)} \|_{L^2}} \tag{7}$$

Here $\mathbf{Z}_{c_i} \in \mathbb{R}^{d^{(c)} \times |\mathcal{C}|}$ is the type-specific embedding matrix for all entities with type $c_i = \Gamma(e_i) \in \mathcal{C}$. $\mathbf{h}_i^{(c)}$ is a one-hot vector such that $\mathbf{Z}_{c_i}\mathbf{h}_i^{(c)}$ will perform an embedding lookup operation which selects an entity feature embedding from the corresponding column. $\| \cdot \|_{L^2}$ denotes the L^2-norm.

Both Hamilton et al. (2018) and Mai et al. (2019) use $\mathsf{Enc}^{(c)}()$ as their entity encoder:

$$\mathsf{Enc}^{(\mathrm{GQE})}(e_i) = \mathsf{Enc}^{(\mathrm{CGA})}(e_i) = \mathsf{Enc}^{(c)}(e_i) \tag{8}$$

Figure 2 is an illustration of their approach. Note that this encoder does not consider the spatial information (e.g., coordinates and spatial extents) of geographic entities, which leads to poorer performance for answering geographic logic queries. As for our *SE-KGE* model, we add an additional entity space encoder, $\mathsf{Enc}^{(x)}()$, to handle this (see Definition 7).

5.2 | Projection operator

The projection operator is utilized to do link prediction: given a basic graph pattern $b = r(h_i, V_j)$ in a CGQ q with relation r in which h_i is either an entity e_i (an anchor node in q) or an existentially quantified bound variable V_i, the projection operator $\mathcal{P}()$ predicts the embedding $\mathbf{e}'_i \in \mathbb{R}^{d^{(c)}}$ for variable V_j. Here, the embedding of h_i can be either the entity embedding $\mathbf{e}_i = \text{Enc}^{(c)}(e_i)$ or the computed embedding $\mathbf{v}_i$ for V_i which is known beforehand. Both Hamilton et al. (2018) and Mai, Janowicz, et al. (2019) share the same projection operator $\mathcal{P}^{(\text{GQE})} = \mathcal{P}^{(\text{CGA})}$ by using a bilinear matrix $\mathbf{R}_r \in \mathbb{R}^{d^{(c)} \times d^{(c)}}$:

$$\mathbf{e}'_i = \begin{cases} \mathcal{P}^{(\text{GQE})}(e_i, r) = \mathcal{P}^{(\text{CGA})}(e_i, r) = \mathbf{R}_r \text{Enc}^{(c)}(e_i) = \mathbf{R}_r \mathbf{e}_i & \text{if input } = (e_i, r) \\ \mathcal{P}^{(\text{GQE})}(V_i, r) = \mathcal{P}^{(\text{CGA})}(V_i, r) = \mathbf{R}_r \mathbf{v}_i & \text{if input } = (V_i, r) \end{cases} \tag{9}$$

$\mathbf{R}_r$ can also be a bilinear diagonal matrix as DisMult (Yang, Yih, He, Gao, & Deng, 2015) whose corresponding projection operator is denoted by $\mathcal{P}^{(\text{GQE}_{\text{diag}})}$.

In *SE-KGE*, we extend the projection operator $\mathcal{P}()$ so that it can be used in the spatial semantic lifting task (see Definition 8).

Figure 3 uses the basic graph pattern $b_2 = \text{Assembly}(\text{Chevrolet Eagle}, ?\text{Place})$ in Figure 1 as an example to demonstrate how to do link prediction with $\mathcal{P}^{(\text{GQE})}() = \mathcal{P}^{(\text{CGA})}()$. The resulting embedding $\mathbf{e}_{?2}$ can be treated as the

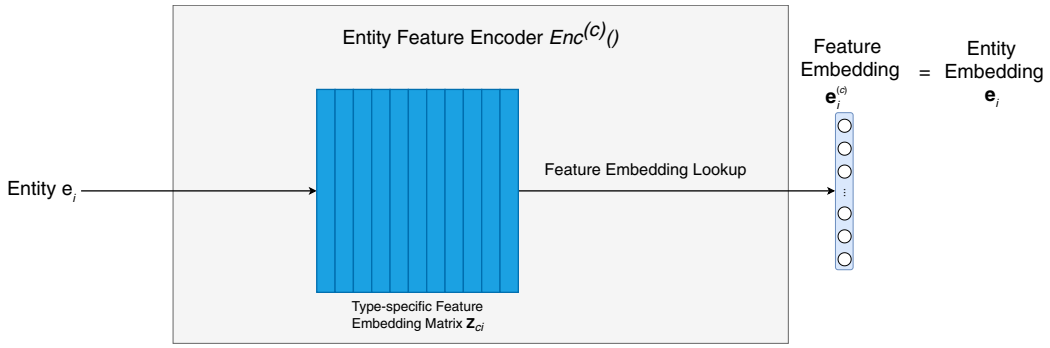


FIGURE 2 Entity encoder used by Hamilton et al. (2018) and Mai et al. (2019a)

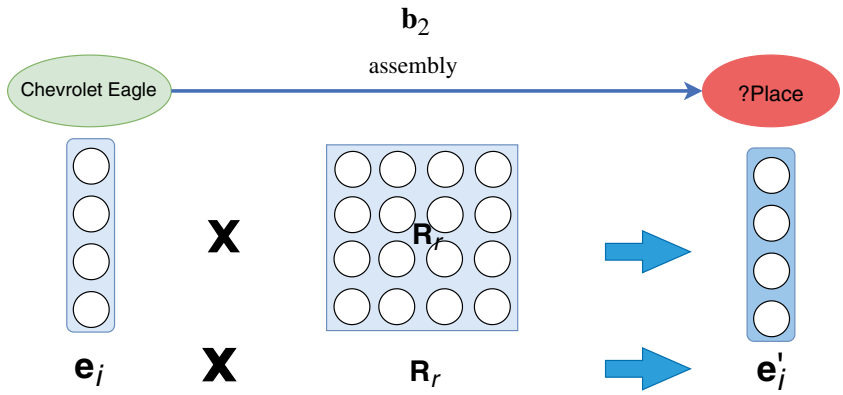


FIGURE 3 Illustration of projection operator $\mathcal{P}^{(\text{GQE})}() = \mathcal{P}^{(\text{CGA})}()$ used by Hamilton et al. (2018) and Mai et al. (2019)

prediction of the embedding of variable `?Place`. By following the same process, we can predict the embedding of the variable `?Place` from the other two basic graph patterns b_1 and b_3: $\mathbf{e}_{?1}$ and $\mathbf{e}_{?3}$.

5.3 | Intersection operator

The intersection operator $\mathcal{I}()$ is used to *integrate* multiple embeddings $\mathbf{e}_{1?}, \mathbf{e}_{2?}, \ldots, \mathbf{e}_{i?}, \ldots, \mathbf{e}_{n?}$ which represent the same (bound or target) variable $V_?$ in a CGQ q to produce one single embedding $e_?$ to represent this variable. Figure 4 illustrates this idea by using CGQ q_C in Figure 1 as an example where $\mathbf{e}_{?1}$, $\mathbf{e}_{?2}$ and $\mathbf{e}_{?3}$ indicate the predicted embedding of `?Place` from three different basic graph patterns b_1, b_2 and b_3. The intersection operator *integrates* them into one single embedding $\mathbf{e}_?$ to represent `?Place`. Since `?Place` is the target variable of q, $\mathbf{e}_?$ is the final *query embedding* we use to do a nearest-neighbor search to obtain the most probable answer (see Task 1). More formally, we have the following definition.

Definition 5 (Intersection Operator $\mathcal{I}()$)Given a set of n different input embeddings $\mathbf{e}_{1?}, \mathbf{e}_{2?}, \ldots, \mathbf{e}_{j?}, \ldots, \mathbf{e}_{n?}$, the intersection operator $\mathcal{I}()$ produces one single embedding $\mathbf{e}_?$:

$$\mathbf{e}_? = \mathcal{I}(\{\mathbf{e}_{1?}, \mathbf{e}_{2?}, \ldots, \mathbf{e}_{j?}, \ldots, \mathbf{e}_{n?}\}) \tag{10}$$

The intersection operator $\mathcal{I}()$ represents the logical conjunction in the embedding space. Any permutation-invariant function can be used here as a conjunction such as elementwise mean, maximum and minimum. We can also use any permutation-invariant neural network architecture (Zaheer et al., 2017) such as Deep Sets (Zaheer et al., 2017). GQE (Hamilton et al., 2018) used an elementwise minimum plus a feed-forward network as the

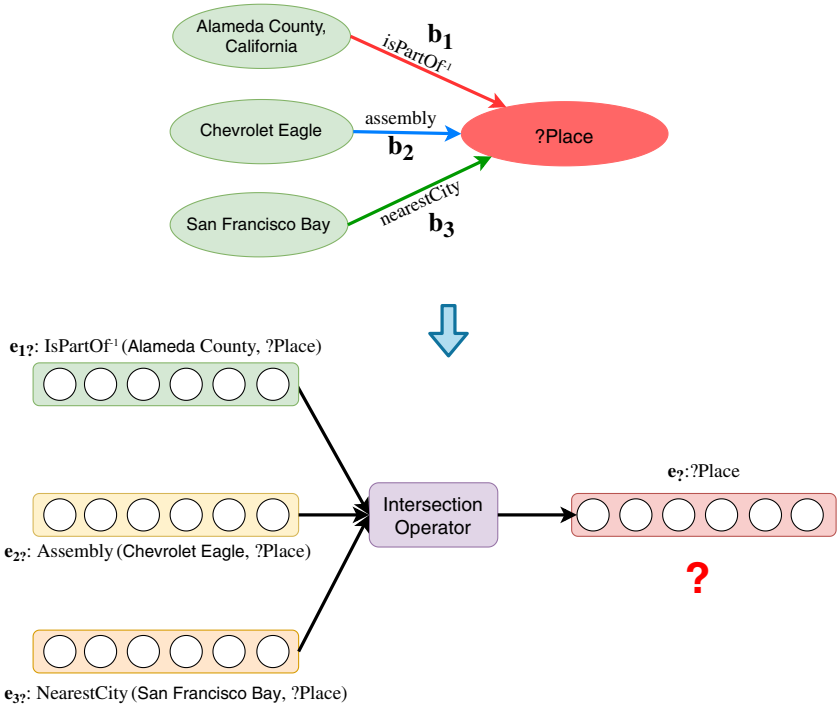


FIGURE 4 Illustration of intersection operator $\mathcal{I}()$

intersection operator which we denote by $\mathcal{I}^{(GQE)}()$. Mai, Janowicz, et al. (2019) showed that their CGA model with a self-attention based intersection operator $\mathcal{I}^{(CGA)}()$ can outperform *GQE*. So in this work, we use $\mathcal{I}^{(CGA)}()$ as the intersection operator $\mathcal{I}()$. Readers who are interested in this technique are invited to check Mai, Janowicz, et al. (2019) for more details.

5.4 | Query embedding computing

Hamilton et al. (2018) proposed a way to compute the query embedding of a CGQ q based on these three components. Given a CGQ q, we can encode all its anchor nodes (entities) into the entity embedding space using Enc(). Then we recursively apply the projection operator $\mathcal{P}()$ and intersection operator $\mathcal{I}()$ by following the DAG of q until we get an embedding for the target node (variable $V_?$), namely, $\mathbf{q} = \mathbf{\Phi}_{\mathcal{G},\theta}(q) = \mathbf{v}_?$. Then we use nearestneighbor search in the entity embedding space to find the *closest* embedding, whose corresponding entity will be the predicted answer to Query q. For details of the query embedding algorithm, see Hamilton et al. (2018).

Figure 5 gives an illustration of the query embedding computation process in the embedding space by using Query q_C as an example. We first use Enc() to get the embeddings of three anchor nodes (see the dashed green box in Figure 5). Then $\mathcal{P}()$ (the three green arrows) is applied to each basic graph pattern to get three embeddings $\mathbf{e}_{1?}$, $\mathbf{e}_{2?}$ and $\mathbf{e}_{3?}$. $\mathcal{I}()$ (red arrows) is used later on to integrate them into one single embedding $\mathbf{e}_?$ or $\mathbf{q}$ for the target variable `?Place`.

In this work, we follow the same query embedding computation process. Furthermore, we extend the current model architecture to perform spatial semantic lifting.

6 | *SE-KGE* MODEL

Since many geographic questions rely on spatial information (e.g., coordinates) and spatial reasoning, a spatially-explicit model is desired for the geographic logic query answering task (Task 1). Moreover, the spatial semantic lifting task (Task 2) is only possible if we have an entity encoder which can encode the spatial information of geographic entities as well as a specially designed projection operator. To solve these problem, we propose a new entity encoder Enc() (see Section 6.1) and a new projection operator (see Section 6.2) for our *SE-KGE* model. Tasks 1 and 2 then require different training processes which will be discussed in Sections 6.3 and 6.4. *SE-KGE* extends the general logic query answering framework of GQE (Hamilton et al., 2018) and CGA (Mai, Janowicz, et al., 2019) with explicit spatial embedding representations.

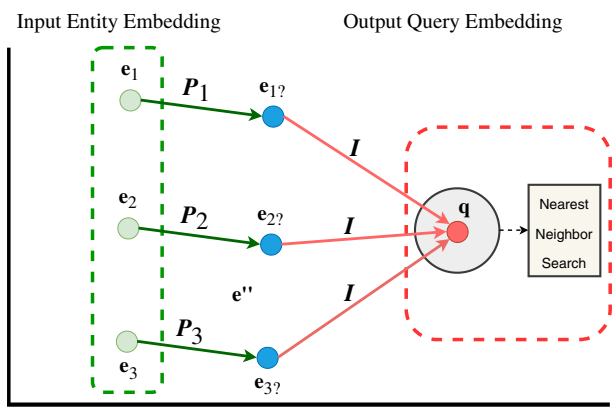


FIGURE 5 Illustration of (geographic) logic query answering in the embedding space

6.1 | Entity encoder

Definition 6 (Entity Encoder Enc())Given a geographic knowledge graph $\mathcal{G}$, the entity encoder $\mathsf{Enc}(): \mathcal{V} \to \mathbb{R}^d$ is defined as a function parameterized by θ_{Enc}, which maps any entity $e_i \in \mathcal{V}$ to a vector representation of d dimension, the so-called entity embedding $\mathbf{e}_i \in \mathbb{R}^d$. Enc() consists of two parts, the entity feature encoder $\mathsf{Enc}^{(c)}(): \mathcal{V} \to \mathbb{R}^{d^{(c)}}$ and the entity space encoder $\mathsf{Enc}^{(x)}(): \mathcal{V} \to \mathbb{R}^{d^{(x)}}$. These two encoders map any entity $e_i \in \mathcal{V}$ to a feature embedding $\mathbf{e}_i^{(c)} \in \mathbb{R}^{d^{(c)}}$ and space embedding $\mathbf{e}_i^{(x)} \in \mathbb{R}^{d^{(x)}}$, respectively. The final entity embedding $\mathbf{e}_i$ is the concatenation of $\mathbf{e}_i^{(c)}$ and $\mathbf{e}_i^{(x)}$:

$$\mathbf{e}_i = \mathsf{Enc}(e_i) = [\mathsf{Enc}^{(c)}(e_i); \mathsf{Enc}^{(x)}(e_i)] = [\mathbf{e}_i^{(c)}; \mathbf{e}_i^{(x)}]. \tag{11}$$

Here $[\cdot;\cdot]$ denotes vector concatenation of two column vectors and $d = d^{(c)} + d^{(x)}$. $\mathsf{Enc}^{(c)}()$ is defined in Definition 4.

6.1.1 | Entity space encoder

In our work, rather than using the terms *location encoder* and *location embedding* (Mac Aodha et al., 2019), we use *space encoder* to refer to the neural network model that encodes the spatial information of an entity and call the encoding results *space embeddings*. While location encoders focus on encoding a single point location, our space encoder $\mathsf{Enc}^{(x)}()$ aims to handle spatial information of geographic entities at different scales:

1. For a small geographic entity $e_i \in \mathcal{V}_{pt} \setminus \mathcal{V}_{pn}$ such as a radio station or restaurant, we use its location $\mathbf{x}_i = \mathcal{PT}(e_i)$ as the input to $\mathsf{Enc}^{(x)}()$.
2. For a geographic entity with a large extent $e_i \in \mathcal{V}_{pn}$ such as a country or state, at each encoding time, we randomly generate a point $\mathbf{x}_i^{(t)}$ as the input for $\mathsf{Enc}^{(x)}()$ based on the two-dimensional uniform distribution defined on its spatial extent (bounding box) $\mathcal{PN}(e_i) = [\mathbf{x}_i^{\min}; \mathbf{x}_i^{\max}]$ (i.e., $\mathbf{x}_i^{(t)} \sim \mathcal{U}(\mathbf{x}_i^{\min}, \mathbf{x}_i^{\max})$). Since during training $\mathsf{Enc}^{(x)}()$ will be called multiple times, it will at the end learn a uniform distribution over e_i's bounding box. In practice, one can sample using any process, such as stratified random sampling, or vary the sampling density by expected variation.
3. For a non-geographic entity $e_i \in \mathcal{V} \setminus \mathcal{V}_{pt}$, we randomly initialize its space embedding. One benefit of this approach is that during the KG embedding training process, these embeddings will be updated based on backpropagation in neural networks so that the spatial information of its connected entities in $\mathcal{G}$ will propagate to this embedding as its pseudo-space footprint. For example, a person's spatial embedding will be close to the embedding of his/her birthplace or hometown.

The entity space encoder $\mathsf{Enc}^{(x)}()$ is formally defined as follows:

Definition 7 (Entity space encoder $\mathsf{Enc}^{(x)}()$) Given any entity $e_i \in \mathcal{V}$ from $\mathcal{G}$, $\mathsf{Enc}^{(x)}()$ computes the space embedding $\mathbf{e}_i^{(x)} = \mathsf{Enc}^{(x)}(e_i) \in \mathbb{R}^{d^{(x)}}$ by

$$\mathbf{e}_i^{(x)} = \begin{cases} \mathsf{LocEnc}^{(x)}(\mathbf{x}_i), \text{where } \mathbf{x}_i = \mathcal{PT}(e_i), & \text{if } e_i \in \mathcal{V}_{pt} \setminus \mathcal{V}_{pn}, \\ \mathsf{LocEnc}^{(x)}(\mathbf{x}_i^{(t)}), \text{where } \mathbf{x}_i^{(t)} \sim \mathcal{U}(\mathbf{x}_i^{\min}, \mathbf{x}_i^{\max}), \mathcal{PN}(e_i) = [\mathbf{x}_i^{\min}; \mathbf{x}_i^{\max}], & \text{if } e_i \in \mathcal{V}_{pn} \\ \frac{\mathbf{Z}_x \mathbf{h}_i^{(x)}}{\|\mathbf{Z}_x \mathbf{h}_i^{(x)}\|_{L2}}, & \text{if } e_i \in \mathcal{V} \setminus \mathcal{V}_{pt}. \end{cases} \tag{12}$$

Here $\mathbf{Z}_x$ and $\mathbf{h}_i^{(x)}$ are the embedding matrix and one-hot vector for non-geographic entities in the entity space encoder $\mathsf{Enc}^{(x)}()$, similarly to Equation(7). $\mathsf{LocEnc}^{(x)}()$ denotes a location encoder module (see Equation 2). Figure 6

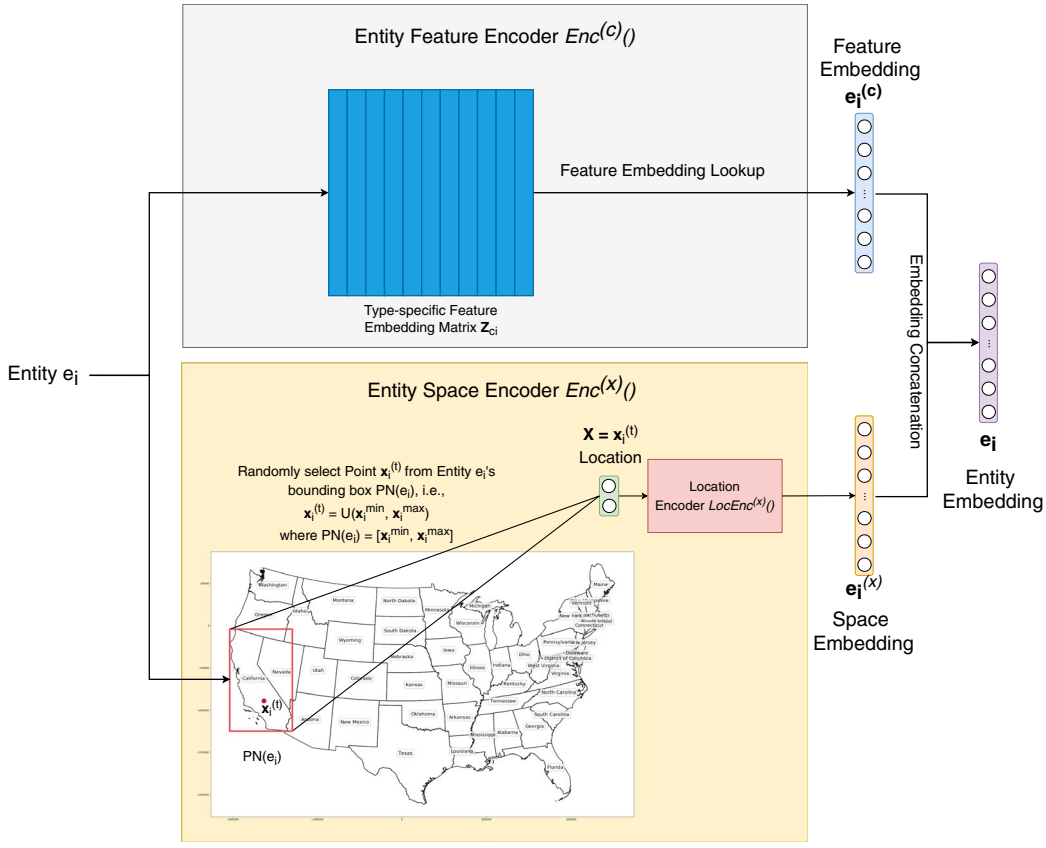


FIGURE 6 Entity encoder Enc() of *SE-KGE*. Compared with previous work (Figure 2), an entity space encoder component $\text{Enc}^{(x)}()$ is added to capture the spatial information of geographic entities

illustrates the architecture of the entity encoder Enc(). Compared to the *GQE* entity encoder $\text{Enc}^{(GQE)}()$ shown in Figure 2, the proposed entity encoder of *SE-KGE* adds the entity space encoder $\text{Enc}^{(x)}()$ which leverages a multi-scale grid cell representation to capture the spatial information of geographic entities.

As far as using a bounding box as an approximation is concerned, one reason to use bounding boxes instead of real geometries is that performing the point-in-polygon operation in real time during machine learning model training is very expensive and not efficient. Many spatial databases use bounding boxes as approximations of real geometries to avoid intensive computation. We adopt the same strategy here. Moreover, the detailed spatial footprint of e_i is expected to be captured through the training process of the entity embedding. For example, even if the model is only aware of the bounding box of California, by using the `dbo:isPartOf` relations between California and its subdivisions, the model will be informed of all the spatial extents of its subdivisions.

6.2 | Projection operator

Definition 8 (Projection operator $\mathcal{P}()$)Given a geographic knowledge graph $\mathcal{G}$, a projection operator $\mathcal{P}():\mathcal{V}\cup\mathcal{A}\times\mathcal{R}\rightarrow\mathbb{R}^d$ maps a pair (e_i,r), (V_i,r), or $(\mathbf{x}_i,r)$ to an embedding $\mathbf{e}'_i$. According to the input, $\mathcal{P}()$ can be treated as one of the following:

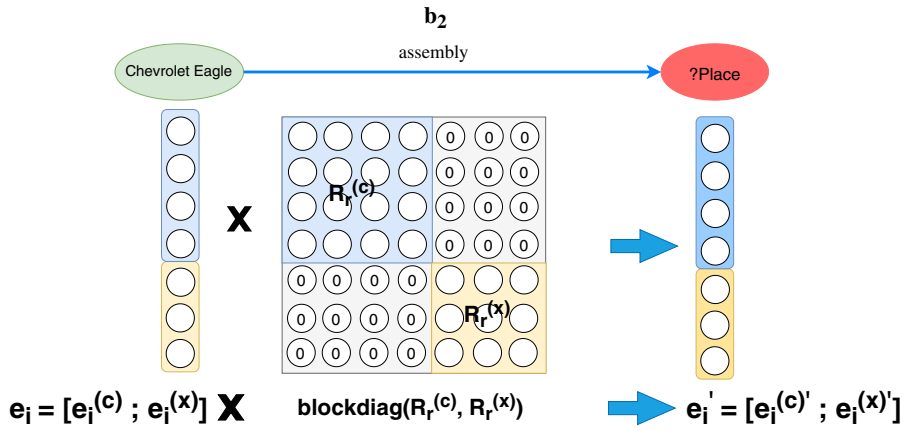


FIGURE 7 Illustration of projection operator $\mathcal{P}^{(e)}()$ of *SE-KGE* with input (e_i,r)

1. *link prediction* $\mathcal{P}^{(e)}(e_i,r)$—given a triple's head entity e_i and relation r, predicting the tail;
2. *link prediction*$\mathcal{P}^{(e)}(V_i,r)$—given a basic graph pattern $b = r(V_i,V_j)$ and $\mathbf{v}_i$, which is the computed embedding for the existentially quantified bound variable V_i, predicting the embedding for variable V_j; and
3. *spatial semantic lifting*$\mathcal{P}^{(x)}(\mathbf{x}_i,r)$—given an arbitrary location $\mathbf{x}_i$ and relation r, predicting the most probable linked entity.

Formally, $\mathcal{P}()$ is defined as:

$$\mathbf{e}'_i = \begin{cases} \mathcal{P}^{(e)}(e_i,r) = \text{diag}(\mathbf{R}_r^{(c)},(\mathbf{R}_r^{(x)}(\mathbf{R}_r^{(x)})\text{Enc}(e_i) = \text{diag}(\mathbf{R}_r^{(c)},(\mathbf{R}_r^{(x)})\mathbf{e}_i, & \text{if input } = (e_i,r) \\ \mathcal{P}^{(e)}(V_i,r) = \text{diag}(\mathbf{R}_r^{(c)},(\mathbf{R}_r^{(x)})\mathbf{v}_i, & \text{if input } = (V_i,r) \\ \mathcal{P}^{(x)}(\mathbf{x}_i,r) = \text{diag}(\mathbf{R}_r^{(xc)},(\mathbf{R}_r^{(x)})[\text{LocEnc}^{(x)}(\mathbf{x}_i);\text{LocEnc}^{(x)}(\mathbf{x}_i)], & \text{if input } = (\mathbf{x}_i,r) \end{cases} \tag{13}$$

where $\mathbf{R}_r^{(c)} \in \mathbb{R}^{d^{(c)}\times d^{(c)}}$, $\mathbf{R}_r^{(x)} \in \mathbb{R}^{d^{(x)}\times d^{(x)}}$ and $\mathbf{R}_r^{(xc)} \in \mathbb{R}^{d^{(c)}\times d^{(x)}}$ are three trainable and relation-specific matrices. $\mathbf{R}_r^{(c)}$ and $\mathbf{R}_r^{(x)}$ focus on the feature embedding and space embedding. $\mathbf{R}_r^{(xc)}$ transforms the space embedding $\mathbf{e}_i^{(x)}$ to its correspondence in feature embedding space. $\text{diag}(\mathbf{R}_r^{(c)},\mathbf{R}_r^{(x)}) \in \mathbb{R}^{d\times d}$ and $\text{diag}(\mathbf{R}_r^{(xc)},\mathbf{R}_r^{(x)}) \in \mathbb{R}^{d\times 2d^{(x)}}$ denote two block-diagonal matrices based on $\mathbf{R}_r^{(c)}$, $\mathbf{R}_r^{(x)}$ and . $[\text{LocEnc}^{(x)}(\mathbf{x}_i);\text{LocEnc}^{(x)}(\mathbf{x}_i)]$ indicates the concatenation of two identical space embeddings $\text{LocEnc}^{(x)}(\mathbf{x}_i)$. Here, we use the same $\mathcal{P}^{(e)}()$ for the first two cases to indicate they share the same neural network architecture. This is because both of them are link prediction tasks with different inputs.

Figure 7 illustrates the idea of projection operator $\mathcal{P}^{(e)}()$ by using the basic graph pattern b_2 in q_C (see Figure 1) as an example of link prediction (case (1)). Given the embedding of `dbr:Chevrolet_Eagle` and the relation-specific matrix $\text{diag}(\mathbf{R}_r^{(c)},\mathbf{R}_r^{(x)})$ for relation `dbo:assembly`, we can predict the embedding of the variable `?Place`, $\mathbf{e}_{?2}$.

Figure 8 shows how to use $\mathcal{P}^{(x)}()$ in the semantic lifting task (case (3); see Section 6.4. for a detailed description). Note that "×" in Figures 7 and 8 indicates $\text{diag}(\mathbf{R}_r^{(c)},\mathbf{R}_r^{(x)})\mathbf{e}_i$ and $\text{diag}(\mathbf{R}_r^{(xc)},\mathbf{R}_r^{(x)})[\text{LocEnc}^{(x)}(\mathbf{x}_i);\text{LocEnc}^{(x)}(\mathbf{x}_i)]$, respectively.

6.3 | Geographic logic query answering $\Phi_{\mathcal{G},\theta}(q)$ model training

We train the *SE-KGE* model on both the original KG structure with an unsupervised objective $\mathcal{L}_{\text{KG}}$ (Section 6.3.1) and the query–answer pairs with a supervised objective $\mathcal{L}_{\text{QA}}$ (Section 6.3.2):

$$\mathcal{L}^{(\text{QA})} = \mathcal{L}_{\text{KG}} + \mathcal{L}_{\text{QA}} \tag{14}$$

FIGURE 8 Spatial semantic lifting in the embedding space by using Enc() and $\mathcal{P}^{(x)}()$

6.3.1 | Unsupervised KG training phase

In this phase, we train *SE-KGE* components based on the local KG structure. In $\mathcal{G} = (\mathcal{V},\mathcal{E})$, for every entity $e_i \in \mathcal{V}$, we first obtain its 1-degree neighborhood $N(e_i) = \{(r_{ui}, e_{ui}) | r_{ui}(e_{ui}, e_i) \in \mathcal{G}\} \cup \{(r_{oi}^{-1}, e_{oi}) | r_{oi}(e_i, e_{oi}) \in \mathcal{G}\}$. We need to sample n different tuples from $N(e_i)$ to form a sampled neighborhood $N_n(e_i) \subseteq N(e_i)$ and $|N_n(e_i)| = n$. We treat this subgraph as a CGQ with n basic graph patterns, in which entity e_i holds the target variable position. The model predicts the embedding of e_i such that the correct embedding $\mathbf{e}_i$ is the closest one to the predicted embedding $\mathbf{e}_i''$ against all embeddings $\mathbf{e}_i^-$ in the negative sample set $\text{Neg}(e_i)$:

$$\mathcal{L}_{\text{KG}} = \sum_{e_i \in \mathcal{V}} \sum_{e_i^- \in \text{Neg}(e_i)} \max(0, \Delta - \Omega(\mathbf{H}_{\text{KG}}(e_i), \mathbf{e}_i) + \Omega(\mathbf{H}_{\text{KG}}(e_i), \mathbf{e}_i^-)) \tag{15}$$

where

$$\mathbf{e}_i'' = \mathbf{H}_{\text{KG}}(e_i) = \mathcal{I}(\{\mathcal{P}^{(e)}(e_{ci}, r_{ci}) | (r_{ci}, e_{ci}) \in N_n(e_i)\}) \tag{16}$$

Here $\mathcal{L}_{\text{KG}}$ is a max-margin loss and Δ is the margin.

6.3.2 | Supervised query–answer pair training phase

We train *SE-KGE* by using conjunctive query–answer pairs. We first sample X different conjunctive graph query (logical query)–answer pairs $S = \{(q_i, a_i)\}$ from $\mathcal{G}$. We treat each entity as the target variable of a CGQ and sample K queries for each DAG structure. All DAG structures we considered in this work are shown in Figure 9. The way to do query sampling is to sort the nodes in a DAG in topological order and sample one basic graph pattern at a time by following this order and navigating on the $\mathcal{G}$ (Hamilton et al., 2018). In order to generate a GCGQ, we have the restriction $e_i \in \mathcal{V}_{pt}$.

The training objective is to make the correct answer entity embedding $\mathbf{a}_i$ the closest one to the predicted query embedding $\mathbf{q}_i = \mathbf{\Phi}_{\mathcal{G},\theta}(q_i)$ against all the negative answers' embeddings $\mathbf{a}_i^-$ in negative answer set $\text{Neg}(r_i, a_i)$. We also use a max-margin loss:

$$\mathcal{L}_{\text{QA}} = \sum_{(q_i, a_i) \in S} \sum_{a_i^- \in \text{Neg}(q_i, a_i)} \max(0, \Delta - \Omega(\mathbf{q}_i, \mathbf{a}_i) + \Omega(\mathbf{q}_i, \mathbf{a}_i^-)) \tag{17}$$

For $\text{Neg}(q_i, a_i)$ we compared two negative sampling strategies: 1) *negative sampling*: $\text{Neg}(r_i, a_i) \subseteq \mathcal{V}$ is a fixed-size set of entities such that, for *all* $e_i^- \in \text{Neg}(r_i, a_i), \Gamma(e_i^-) = \Gamma(e_i)$ and $e_i^- \neq e_i$; 2) *hard negative sampling*: $\text{Neg}(q_i, a_i)$ is a fixed-size set of entities which satisfy some of the basic graph patterns b_{ij} (see Definition 2) in q_i but not all.

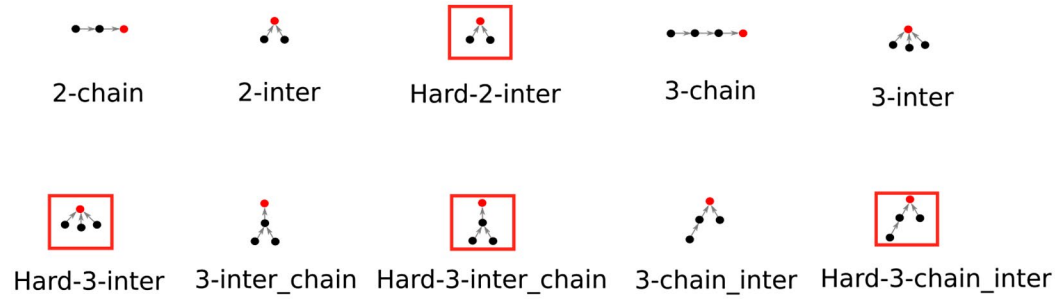


FIGURE 9 DAG structures of the conjunctive graph queries we sampled from $\mathcal{G}$. Nodes indicate entities or variables and edges indicate basic graph patterns. The red node is the target variable of the corresponding query. The DAG structures surrounded by red boxes indicate queries sampled with the hard negative sampling method

6.4 | Spatial semantic lifting $\Psi_{\mathcal{G},\theta_{ssl}}(\mathbf{x}, r)$ model training

We randomly select a point $\mathbf{x}_i \in \mathcal{A} \subseteq \mathbb{R}^2$ from the study area, and use location encoder $\text{LocEnc}^{(x)}$ to encode its location embedding $\mathbf{e}_i^{(x)} \in \mathbb{R}^{d^{(x)}}$. Since we do not have the feature embedding for this location, to make the whole model an inductive learning one, we use $\mathcal{P}^{(x)}()$ to predict the tail embedding $\mathbf{e}' = \Psi_{\mathcal{G},\theta_{ssl}}(\mathbf{x}_i, r)$ of this virtual triple $r(\mathbf{x}_i, e')$. This is equivalent to posing a query $r(\mathbf{x}_i, ?e)$ to $\mathcal{G}$. A nearest-neighbor search in the entity embedding space will produce the predicted entity which can link to location $\mathbf{x}_i$ with relation r. Since given any location $\mathbf{x}_i$ from the study area, $\Psi_{\mathcal{G},\theta_{ssl}}(\mathbf{x}_i, r)$ can predict the entity embedding that $\mathbf{x}_i$ can link to given relation r, this is a fully inductive learning-based model. This model does not require location $\mathbf{x}_i$ to be selected from a predefined set of locations which is a requirement for transductive learning-based models such as Kejriwal and Szekely (2017). Figure 8 shows the idea of spatial semantic lifting.

We train the spatial semantic lifting model $SE-KGE_{ssl}$ with Enc(), $\mathcal{P}^{(e)}()$, and $\mathcal{P}^{(x)}()$ by using two objectives: a link prediction objective $\mathcal{L}_{\text{LP}}$ (Section 6.4.1) and a spatial semantic lifting objective $\mathcal{L}_{\text{SSL}}$ (Section 6.4.2):

$$\mathcal{L}^{(SSL)} = \mathcal{L}_{\text{LP}} + \mathcal{L}_{\text{SSL}} \tag{18}$$

6.4.1 | Link prediction training phase

The link prediction training phase aims to train the feature embeddings of each entity. For each triple $s_i = (h_i, r_i, t_i) \in \mathcal{T}$, we can use Enc() and $\mathcal{P}^{(e)}()$ to predict the tail entity embedding given the head and relation ($\mathcal{P}^{(e)}(h_i, r_i)$), or predict the head entity embedding given the tail and relation ($\mathcal{P}^{(e)}(t_i, r_i^{-1})$). Note that we have two separate $\mathcal{P}^{(e)}()$ for r_i and r_i^{-1}. The loss function is given by:

$$\begin{aligned}\mathcal{L}_{\text{LP}} = &\sum_{s_i=(h_i,r_i,t_i)\in\mathcal{T}} \sum_{t_i^-\in \text{Neg}_t(t_i)} \max\left(0, \Delta - \Omega(\mathcal{P}^{(e)}(h_i,r_i),\mathbf{t}_i) + \Omega(\mathcal{P}^{(e)}(h_i,r_i),\mathbf{t}_i^-)\right) \\ &+ \sum_{s_i=(h_i,r_i,t_i)\in\mathcal{T}} \sum_{h_i^-\in \text{Neg}_t(h_i)} \max\left(0, \Delta - \Omega(\mathcal{P}^{(e)}(t_i,r_i^{-1}),\mathbf{h}_i) + \Omega(\mathcal{P}^{(e)}(t_i,r_i^{-1}),\mathbf{h}_i^-)\right)\end{aligned}, \tag{19}$$

where $\text{Neg}_t(e_i)$ is the set of negative entities which share the same type with entity e_i.

6.4.2 | Spatial semantic lifting training phase

We also directly optimize our model on the spatial semantic lifting objective. We denote by $\mathcal{T}_s$ and $\mathcal{T}_o$ sets of triples whose head (or tail) entities are geographic entities (i.e., $\mathcal{T}_s = \{s_i | s_i = (h_i, r_i, t_i) \in \mathcal{T} \wedge h_i \in \mathcal{V}_{pt}\}$ and $\mathcal{T}_o = \{s_i | s_i = (h_i, r_i, t_i) \in \mathcal{T} \wedge t_i \in \mathcal{V}_{pt}\}$). The training objective is to make the tail entity embedding $\mathbf{t}_i$ the closest one to the predicted embedding $\mathcal{P}^{(x)}(\mathcal{X}(h_i), r_i)$ against all negative entity embeddings $\mathbf{t}_i^-$. We do the same for the inverse triple (t_i, r_i^{-1}, h_i). The loss function is given by:

$$\begin{aligned}\mathcal{L}_{\text{SSL}} = &\sum_{s_i=(h_i,r_i,t_i)\in\mathcal{T}_s}\sum_{t_i^-\in\text{Neg}_t(t_i)} \max\left(0, \Delta - \Omega(\mathcal{P}^{(x)}(\mathcal{X}(h_i), r_i), \mathbf{t}_i) + \Omega(\mathcal{P}^{(x)}(\mathcal{X}(h_i), r_i), \mathbf{t}_i^-)\right) \\ &+ \sum_{s_i=(h_i,r_i,t_i)\in\mathcal{T}_o}\sum_{h_i^-\in\text{Neg}_t(h_i)} \max\left(0, \Delta - \Omega(\mathcal{P}^{(x)}(\mathcal{X}(t_i), r_i^{-1}), \mathbf{h}_i) + \Omega(\mathcal{P}^{(x)}(\mathcal{X}(t_i), r_i^{-1}), \mathbf{h}_i^-)\right),\end{aligned} \tag{20}$$

where

$$\mathcal{X}(e_i) = \begin{cases} \mathbf{x}_i = \mathcal{PT}(e_i), & \text{if } e_i \in \mathcal{V}_{pt} \setminus \mathcal{V}_{pn} \\ \mathbf{x}_i^{(t)} \sim \mathcal{U}(\mathbf{x}_i^{\min}, \mathbf{x}_i^{\max}), \mathcal{PN}(e_i) = [\mathbf{x}_i^{\min}; \mathbf{x}_i^{\max}], & \text{if } e_i \in \mathcal{V}_{pn} \end{cases} \tag{21}$$

7 | EXPERIMENT

To demonstrate how *SE-KGE* incorporates spatial information of geographic entities such as locations and spatial extents we experimented with two tasks: geographic logic query answering and spatial semantic lifting. To demonstrate the effectiveness of spatially-explicit models and the importance of considering the scale effect in location encoding we select multiple baselines on the geographic logic query answering task. To show that *SE-KGE* is able to link a randomly selected location to entities in the existing KG with some relation, which none of the existing KG embedding models can solve, we proposed a new task - spatial semantic lifting.

7.1 | *DBGeo* data set generation

In order to evaluate our proposed location-aware KG embedding model *SE-KGE*, we first build a geographic KG which is a subgraph of DBpedia by following the common practice in KG embedding research (Bordes et al., 2013; Mai, Yan, et al., 2019; Wang et al., 2014). We select the U.S. mainland as the study area $\mathcal{A}$ since previous research (Janowicz et al., 2016) has shown that DBpedia has relatively richer geographic coverage in the U.S. The KG construction process is as follows:

1. We collect all the geographic entities within the U.S. mainland as the seed entity set $\mathcal{V}_{\text{seed}}$, which accounts for 18,780 geographic entities,5 We then collect their 1- and 2-degree object property triples with `dbo:` prefix predicates/relations (http://dbpedia.org/sparql?help=nsdecl).
2. We compute the degree of each entity in the collected KG and delete any entity, together with its corresponding triples, if its node degree is less than a threshold η. We use η = 10 for non-geographic entities and η = 5 for geographic entities, because many geographic entities, such as radio stations, have fewer object-type property triples and a smaller threshold ensures that a relatively large number of geographic entities can be extracted from the KG.
3. We further filter out those geographic entities that are newly added from Step 2 and are outside of the U.S. mainland. The resulting triples form our KG, and we denote the geographic entity set as $\mathcal{V}_{pt}$.

4. We split $\mathcal{G}$ into training, validation and testing triples in the ratio of 90:1:9, so that every entity and every relation appears in the training set. We denote the KG formed by the training triples as $\mathcal{G}_{\text{train}}$, while denoting the whole KG as $\mathcal{G}$.
5. We generate K conjunctive graph query–answer pairs from $\mathcal{G}$ for each DAG structure shown in Figure 9 based on the query–answer generation process described in Section 6.3. $Q(\mathcal{G})$ and $Q(\mathcal{G})_{\text{geo}}$ denote the resulting query–answer (QA) set, while $Q_{\text{geo}}(\mathcal{G})$ denotes the geographic QA set. For each query q_i in the training QA set, we make sure that each query is answerable based on $\mathcal{G}_{\text{train}}$ (i.e., $\varphi(\mathcal{G}_{\text{train}}, q_i) \neq \emptyset$). As for query q_i in the validation and testing QA set, we make sure each query q_i satisfies $\varphi(\mathcal{G}_{\text{train}}, q_i) = \emptyset$ and $\varphi(\mathcal{G}, q_i) \neq \emptyset$.
6. For each geographic entity $e \in \mathcal{V}_{pt}$, we obtain its location/coordinates by extracting its `geo:geometry` triple from DBpedia. We project the locations of geographic entities into the U.S. National Atlas Equal Area projection coordinate system (epsg:2163) $\mathcal{XY}$. $\mathcal{PT}(e) = \mathbf{x}$ indicates the location of e in $\mathcal{XY}$.
7. For each geographic entity $e \in \mathcal{V}_{pt}$, we get its spatial extent (bounding box) $\mathcal{PN}(e)$ in $\mathcal{XY}$ by using the ArcGIS Geocoding API (https://geocode.arcgis.com/arcgis/rest/services/World/GeocodeServer/find) and OpenStreetMap API. We obtain 80.6% of geographic entities; we denote them by $\mathcal{V}_{pn}$.
8. For each entity $e_i \in \mathcal{V}$, we obtain its types by using `rdf:type` triples. Note that there are entities having multiple types. We look up the DBpedia Ontology (class hierachy) to get their level-1 superclass. We find out that every entity in $\mathcal{G}$ has only one level-1 superclass type. Table 1 shows statistics of entities in different types.
9. To build the training/validation/testing data sets for spatial semantic lifting, we obtain $\mathcal{T}_s, \mathcal{T}_o \subseteq \mathcal{T}$ (see Section 6.4), each triple of which is composed of geographic entities as its head or tail. We denote $\mathcal{R}_{\text{ssl}} = \{r_i | s_i = (h_i, r_i, t_i) \in \mathcal{T}_s \cap \mathcal{T}_o\}$.

We denote $Q^{(2)}(\mathcal{G}), Q^{(3)}(\mathcal{G})$ as the general QA sets which respectively contain two and three basic graph patterns, and similarly for $Q^{(2)}_{\text{geo}}(\mathcal{G}), Q^{(3)}_{\text{geo}}(\mathcal{G})$. Table 2 shows the statistics of the constructed $\mathcal{G}$, the generated QA sets, and the spatial semantic lifting data set in *DBGeo*. Figure 10 shows the spatial distribution of all geographic entities $\mathcal{V}_{pt}$ in $\mathcal{G}$.

7.2 | Evaluation on the geographic logic query answering task

7.2.1 | Baselines

In order to quantitatively evaluate *SE-KGE* on the geographic QA task, we train $SE-KGE_{\text{full}}$ and multiple baselines on $\mathcal{G}$ in *DBGeo*. Compared to previous work (Hamilton et al., 2018; Mai, Janowicz, et al., 2019), the most important contribution of this work is the entity space encoder $\text{Enc}^{(x)}()$ which makes our model spatially explicit. So we carefully select four baselines to test the contribution of $\text{Enc}^{(x)}()$ on the geographic logic QA task. We have selected four baselines:

TABLE 1 Number of entities of each entity type in *DBGeo*

Entity type	Number of entities
`dbo:Place`	16,527
`dbo:Agent`	8,371
`dbo:Work`	594
`dbo:Thing`	179
`dbo:TopicalConcept`	134
`dbo:MeanOfTransportation`	104
`dbo:Event`	71

TABLE 2 Statistics for our data set in *DBGeo* (Section 7.1)

		DBGeo		
		Training	**Validation**	**Testing**
Knowledge Graph	$\lvert\mathcal{T}\rvert$	214,064	2,378	21,406
	$\lvert\mathcal{R}\rvert$	318	–	–
	$\lvert\mathcal{V}\rvert$	25,980	–	–
	$\lvert\mathcal{V}_{pt}\rvert$	18,323	–	–
	$\lvert\mathcal{V}_{pn}\rvert$	14,769	–	–
Geographic Question Answering	$\lvert Q^{(2)}(\mathcal{G})\rvert$	1,000,000	-	-
	$\lvert Q^{(3)}(\mathcal{G})\rvert$	1,000,000	–	–
	$\lvert Q^{(2)}_{\text{geo}}(\mathcal{G})\rvert$	1,000,000	1,000/QT	10,000/QT
	$\lvert Q^{(3)}_{\text{geo}}(\mathcal{G})\rvert$	1,000,000	1,000/QT	10,000/QT
Spatial Semantic Lifting	$\lvert\mathcal{T}_s \cap \mathcal{T}_o\rvert$	138,193	1,884	17,152
	$\lvert\mathcal{R}_{\text{ssl}}\rvert$	227	71	135

"X/QT" indicates the number of QA pairs per query type.

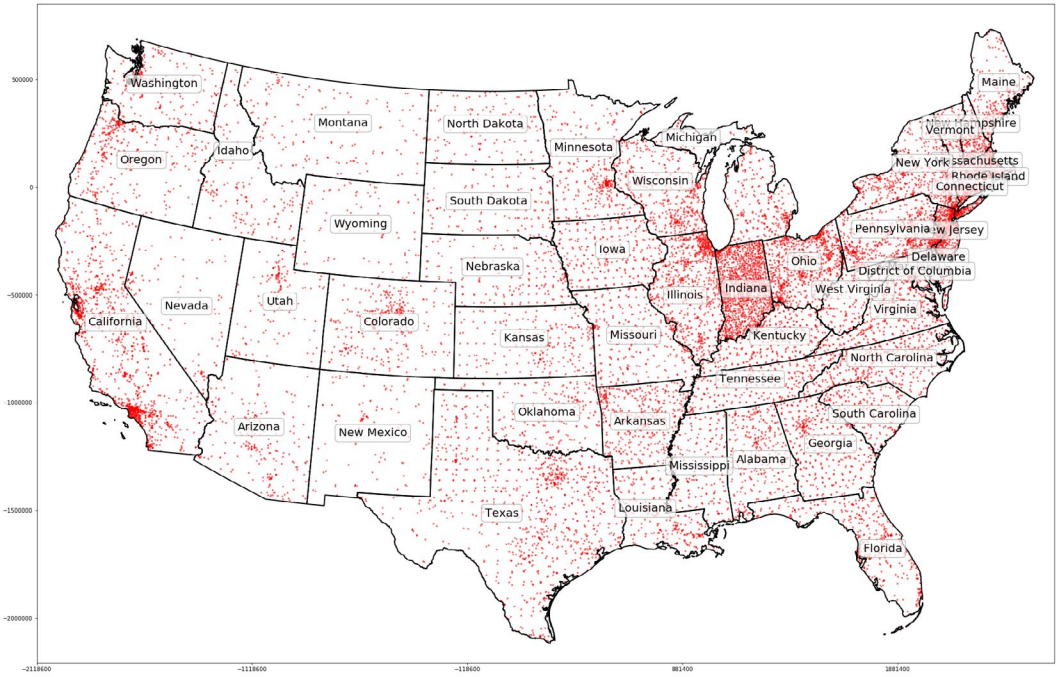


FIGURE 10 Spatial distribution of all geographic entities in $\mathcal{G}$

1. GQE_{diag} and GQE: two versions of the logic query answering model proposed by Hamilton et al. (2018), discussed in detail in Section 5. The main difference between GQE_{diag} and GQE is the projection operator they use: $\mathcal{P}^{(\text{GQE}_{\text{diag}})}$ and $\mathcal{P}^{(\text{GQE})}$, respectively. Compared with $SE-KGE_{\text{full}}$, both GQE_{diag} and GQE only use entity feature encoder $\text{Enc}^{(c)}()$ as the entity encoder and $\mathcal{I}^{(\text{GQE})}$ as the intersection operator. Both methods only use $\mathcal{L}_{\text{QA}}$ in Equation (14) as the training objective. Their two baselines are implemented based on the original code repository (https://github.com/williamleif/graphqembed) of Hamilton et al. (2018).

2. CGA: a logic query answering model proposed by Mai, Janowicz, et al. (2019) (see Section 5 above). Compared with $SE-KGE_{\text{full}}$, CGA uses a different entity encoder ($\text{Enc}^{(\text{CGA})}$) and projection operator ($\mathcal{P}^{(\text{CGA})}$) such that the spatial information of each geographic entity is not considered. This baseline is used to test whether designing a spatially-explicit logic query answering model can outperform general models on the geographic query answering task.
3. $SE-KGE_{\text{direct}}$: a simpler version of $SE-KGE_{\text{full}}$ which uses a *single scale location encoder* in the entity encoder instead of the multi-scale periodic location encoder as shown in Equation (2) in Section 2.3. Instead of first decomposing input $\mathbf{x}$ into a multi-scale periodic representation by using sinusoidal functions with different frequencies (Mai et al., 2020), the location encoder of $SE-KGE_{\text{direct}}$ directly inputs $\mathbf{x}$ into a feed-forward network. This single-scale location encoder is proposed in Mai et al. (2020) as one baseline model—direct. Moreover, its entity space encoder does not consider the spatial extent of each geographic entity either and just uses its coordinates to do location encoding. This baseline is used to test the effectiveness of using multi-scale periodical representation learning in our *SE-KGE* framework.
4. $SE-KGE_{pt}$: a simpler version of $SE-KGE_{\text{full}}$ whose entity space encoder does not consider the spatial extents of geographic entities. The only different between $SE-KGE_{pt}$ and $SE-KGE_{\text{direct}}$ is that $SE-KGE_{pt}$ uses *Space2Vec* (Mai et al., 2020) as the location encoder while $SE-KGE_{\text{direct}}$ utilizes the single scale direct model as the location encoder. This baseline is used to test the necessity to consider the spatial extent of geographic entities in our *SE-KGE* framework. In other words, it uses the following equation for its space encoder:

$$\mathbf{e}_i^{(x)} = \begin{cases} \text{LocEnc}^{(x)}(\mathbf{x}_i), \text{where } \mathbf{x}_i = \mathcal{PT}(e_i) & \text{if } e_i \in \mathcal{V}_{pt} \\ \frac{\mathbf{z}_x \mathbf{h}_i^{(x)}}{\|\mathbf{z}_x \mathbf{h}_i^{(x)}\|_{L2}} & \text{if } e_i \in \mathcal{V} \setminus \mathcal{V}_{pt} \end{cases} \tag{22}$$

5. $SE-KGE_{\text{space}}$: a simpler version of $SE-KGE_{\text{full}}$ whose entity encoder does not have the feature encoder component. This baseline is used to understand how the space encoder $\text{Enc}^{(x)}()$ captures the connectivity information of $\mathcal{G}$.

7.2.2 | Training details

We train our model $SE-KGE_{\text{full}}$ and six baselines on the *DBGeo* data set. GQE_{diag} and GQE are trained on the general QA pairs and geographic QA pairs as in Hamilton et al. (2018). The other models are additionally trained on the original KG structure. Gird search is used for hyperparameter tuning: $d = [32, 64, 128]$, $d^{(c)} = [16, 32, 64]$, $d^{(x)} = [16, 32, 64]$, $S = [8, 16, 32, 64]$, $\lambda_{\text{min}} = [10, 50, 200, 1{,}000]$. The best performance is obtained when $d = 128$, $d^{(c)} = 64$, $d^{(x)} = 64$, $S = 16$, $\lambda_{\text{min}} = 50$. $\lambda_{\text{max}} = 5{,}400{,}000$ is determined by the study area $\mathcal{A}$. We also try different activation functions (i.e., sigmoid, ReLU, LeakyReLU) for the full connected layers $\mathbf{NN}()$ of location encoder $\text{LocEnc}^{(x)}()$. We find that $SE-KGE_{\text{space}}$ performs best with LeakyReLU as the activation function together with L^2 normalization on the location embedding. $SE-KGE_{\text{direct}}$, $SE-KGE_{pt}$ and $SE-KGE_{\text{full}}$ perform best with the sigmoid activation function without L^2 normalization on the location embedding. We implement all models in PyTorch and train/evaluate each model on an Ubuntu machine with two GeForce GTX Nvidia GPU cores, each of which has 10 GB memory. The *DBGeo* data set and related code are open-source (https://github.com/gengchenmai/se-kge).

7.2.3 | Evaluation results

We evaluate $SE-KGE_{\text{full}}$ and six baselines on the validation and testing QA data sets of *DBGeo*. Each model produces a cosine similarity score between the predicted query embedding $\mathbf{q}$ and the correct answer embedding $\mathbf{a}$

(as well as the embedding of negative answers). The objective is to rank the correct answer top 1 among itself and all negative answers given their cosine similarity to **q.** Two evaluation metrics are computed: the area under the receiver operating characteristic (ROC) curve (AUC) and the average percentile rank (APR). AUC compares the correct answer with one randomly sampled negative answer for each query. An ROC curve is computed based on model performance on all queries and the area under this curve is obtained. As for the APR, the percentile rank of the correct answer among all negative answers is obtained for each query based on the prediction of a QA model. Then the APR is computed as the average of the percentile ranks of all queries. Since AUC only uses one negative sample per query while APR uses all negative samples for each query, we consider APR as a more robust evaluation metric.

Table 3 shows the evaluation results of $SE-KGE_{\text{full}}$ as well as six baselines on the validation and testing QA data set of *DBGeo*. We split each data set into different categories based on their DAG structures (see Figure 9). Note that logic query answering is a very challenging task. As for the two works which share a similar setup to ours, Hamilton et al. (2018) show that their GQE model outperforms the TransE baseline by 1.6% of APR on the Bio data set. Similarly, Mai, Janowicz, et al. (2019) demonstrate that their CGA model outperforms the GQE model by 1.39% and 1.65% of APR on the DB18 and WikiGeo19 data sets. In this work we show that our $SE-KGE_{\text{full}}$ model outperforms the current state-of-the-art CGA model by 2.17% and 1.31% in terms of APR on the validation and testing data set of *DBGeo*, respectively. We regard this as a sufficient signal to show the effectiveness of $SE-KGE_{\text{full}}$ on the geographic QA task. Some interesting conclusions can be drawn from Table 3:

1. CGA has a significant performance improvement over GQE_{diag} and GQE on *DBGeo*. This result is consistent with that of Mai, Janowicz, et al. (2019), which demonstrates the advantage of the self-attention mechanism in $\mathcal{I}^{(\text{CGA})}$.
2. The performance of $SE-KGE_{\text{direct}}$ and CGA are similar, which shows that a simple single-scale location encoder ($SE-KGE_{\text{direct}}$) is not sufficient to capture the spatial information of geographic entities.
3. $SE-KGE_{\text{full}}$ performs better than $SE-KGE_{pt}$, which only considers the location information of geographic entities. This illustrates that the scale effect is beneficial for the geographic logic QA task.
4. The performance of $SE-KGE_{\text{space}}$ is the worst among all models. This indicates that it is not enough to only consider spatial information as the input features for entity encoder Enc(). This makes sense because each entity in $\mathcal{G}$ has a lot of semantic information other than its spatial information, and only using spatial information for entity embedding learning is insufficient. However, $SE-KGE_{\text{space}}$ is a fully inductive learning model which enables us to do spatial semantic lifting.
5. Comparing $SE-KGE_{\text{full}}$ with CGA, we can see that $SE-KGE_{\text{full}}$ outperforms CGA for almost all DAG structures on the testing data set except "Hard-3-chain_inter" (−0.58%), while the top 2 DAG structures with the largest margin are "3-inter_chain" (2.15%) and "3-chain_inter" (2.08%). On the validation data set, $SE-KGE_{\text{full}}$ gets a higher ΔAPR than CGA on "Hard-3-inter_chain" (7.42%) and "3-inter_chain" (6.08%). GQE_{diag} shows the best performance on the "Hard-3-chain_inter" query structure.

In order to demonstrate how the intersection operator $\mathcal{I}()$ helps to improve the model performance on the geographic QA task, we show $SE-KGE_{\text{full}}$'s predicted ranking list of entities on Query q_C as well as its three basic graph patterns in Table 4. The 12 entities in this table represent the hard negative sampling set of Query q_C. `dbr:Oakland, _ California` is the correct answer for query q_C. We can see that the four top ranked entities of b_1: IsPartOf^{-1}(Alameda County,?Place) are all subdivisions of Alameda County. The five top ranked entities of b_2 : *Assembly(Chevrolet Eagle,?Place)* are all assembly places of Chevrolet Eagle. Similarly, the top ranked entities of b_3: *NearestCity(San Francisco Bay, ?Place)* are close to San Francisco Bay. The full query q_C yield the best rank of the correct answer. This indicates that each basic graph pattern contributes to the query embedding prediction of $SE-KGE_{\text{full}}$. Moreover, to compare the performance of different models on query q_C, the percentile ranks given by CGA, $SE-KGE_{pt}$, and $SE-KGE_{\text{full}}$ are 53.9%, 61.5%, and 77.0%, respectively.

TABLE 3 Evaluation of geographic logic query answering on *DBGeo*, using AUC (%) and APR (%) as evaluation metrics

	DAG type	GQE_{diag}		GQE		CGA		$SE-KGE_{direct}$		$SE-KGE_{pt}$		$SE-KGE_{space}$		$SE-KGE_{full}$	
		AUC	APR	AUC	APR	AUC	APR	AUC	APR	AUC	APR	AUC	APR	AUC	APR
Valid	2-chain	63.37	64.89	84.23	**88.68**	84.56	86.8	83.12	84.79	**85.97**	84.90	76.81	67.07	85.26	87.25
	2-inter	97.23	97.86	96.00	97.02	98.87	98.58	98.98	98.28	98.95	98.52	85.51	87.13	**99.04**	**98.95**
	Hard-2-inter	70.99	73.55	66.04	73.83	73.43	79.98	73.27	76.36	**74.38**	82.16	63.15	62.91	73.42	**82.52**
	3-chain	61.42	67.94	79.65	79.45	79.11	80.93	77.92	79.26	79.38	83.97	70.09	60.80	**80.90**	**85.02**
	3-inter	98.01	99.21	96.24	98.17	99.18	**99.62**	**99.28**	99.41	99.10	99.56	87.62	89.00	99.27	99.59
	Hard-3-inter	78.29	85.00	68.26	77.55	79.59	86.06	79.50	84.28	**80.48**	**87.40**	63.37	67.17	78.86	85.20
	3-inter_chain	90.56	94.08	93.39	91.52	94.59	90.71	95.99	95.11	95.86	94.41	81.16	83.01	**96.70**	**96.79**
	Hard-3-inter_chain	74.19	**83.79**	70.64	74.54	73.97	76.28	74.81	78.90	**76.45**	75.95	65.54	68.21	76.33	83.70
	3-chain_inter	**98.01**	97.45	92.69	93.31	96.72	97.61	97.31	98.67	97.79	**98.76**	83.70	84.42	97.7	98.65
	Hard-3-chain_inter	**83.59**	**88.12**	66.86	74.06	72.12	77.53	73.23	79.24	74.74	80.47	65.13	69.29	74.72	78.11
	Full Valid	81.57	85.19	81.40	84.81	85.21	87.41	85.34	87.43	**86.31**	88.61	74.21	73.90	86.22	**89.58**
Test	2-chain	64.88	65.61	85.00	87.41	84.91	86.74	83.61	85.97	86.08	88.08	75.46	73.38	**86.35**	**88.12**
	2-inter	96.98	97.99	95.86	97.18	98.79	98.71	98.98	98.94	**98.98**	**99.08**	87.01	85.78	98.93	99.01
	Hard-2-inter	70.39	76.19	64.50	71.86	72.15	79.26	72.04	79.11	**73.72**	**81.78**	61.22	62.97	72.62	81.04
	3-chain	62.30	62.29	79.19	80.19	78.93	80.17	77.53	78.86	79.43	**81.28**	70.55	68.04	**80.49**	80.63
	3-inter	98.09	99.12	96.54	97.94	99.33	99.56	**99.45**	99.47	99.41	**99.63**	88.05	87.63	99.39	99.59
	Hard-3-inter	77.27	83.92	68.69	75.42	78.93	83.52	78.58	84.14	**80.11**	84.87	64.44	64.53	78.76	**84.89**
	3-inter_chain	90.39	91.96	92.54	93.13	93.46	94.36	95.23	95.92	95.02	95.78	81.52	79.61	**95.92**	**96.51**
	Hard-3-inter_chain	72.89	79.12	70.67	75.55	73.47	79.61	73.93	80.21	74.88	79.36	64.99	65.52	**75.36**	**80.72**
	3-chain_inter	97.35	98.27	92.22	94.08	96.55	96.67	97.29	98.39	**97.79**	98.68	85.28	84.08	97.64	**98.75**
	Hard-3-chain_inter	**83.33**	**86.24**	66.77	72.10	72.31	77.89	73.55	77.08	75.19	77.42	65.07	65.41	74.62	77.31
	Full test	81.39	84.07	81.20	84.49	84.88	87.65	85.02	87.81	**86.06**	88.20	74.36	73.70	86.01	**88.96**

Note: The bold is used to highlight the difference.

TABLE 4 Rank of entities in the hard negative sample set of query q_C based on the $SE-KGE_{full}$ prediction for different queries: b_1, IsPartOf^{-1}(Alameda County,?Place); b_2, *Assembly(Chevrolet Eagle,?Place)*; b_3, *NearestCity(San Francisco Bay,?Place)*; the full query q_C. The correct answer is highlighted in bold

	b_1	b_2	b_3	Query q_C
1	dbr:Emeryville,_California	dbr:Flint_Truck_Assembly	dbr:Alameda,_California	dbr:San_Jose,_California
2	dbr:Castro_Valley,_California	dbr:Norwood,_Ohio	dbr:San_Jose,_California	**dbr:Oakland,_California**
3	dbr:Alameda,_California	dbr:Flint,_Michigan	dbr:Berkeley,_California	dbr:Berkeley,_California
4	dbr:Berkeley,_California	dbr:Tarrytown,_New_York	**dbr:Oakland,_California**	dbr:Alameda,_California
5	dbr:San_Jose,_California	**dbr:Oakland,_California**	dbr:Emeryville,_California	dbr:San_Francisco
6	dbr:Fremont,_California	dbr:Alameda,_California	dbr:Fremont,_California	dbr:Fremont,_California
7	**dbr:Oakland,_California**	dbr:San_Jose,_California	dbr:Norwood,_Ohio	dbr:Emeryville,_California
8	dbr:San_Francisco	dbr:San_Francisco	dbr:San_Francisco	dbr:Castro_Valley,_California
9	dbr:Norwood,_Ohio	dbr:Berkeley,_California	dbr:Flint_Truck_Assembly	dbr:Flint,_Michigan
10	dbr:Flint_Truck_Assembly	dbr:Fremont,_California	dbr:Flint,_Michigan	dbr:Norwood,_Ohio
11	dbr:Tarrytown,_New_York	dbr:Emeryville,_California	dbr:Castro_Valley,_California	dbr:Flint_Truck_Assembly
12	dbr:Flint,_Michigan	dbr:Castro_Valley,_California	dbr:Tarrytown,_New_York	dbr:Tarrytown,_New_York

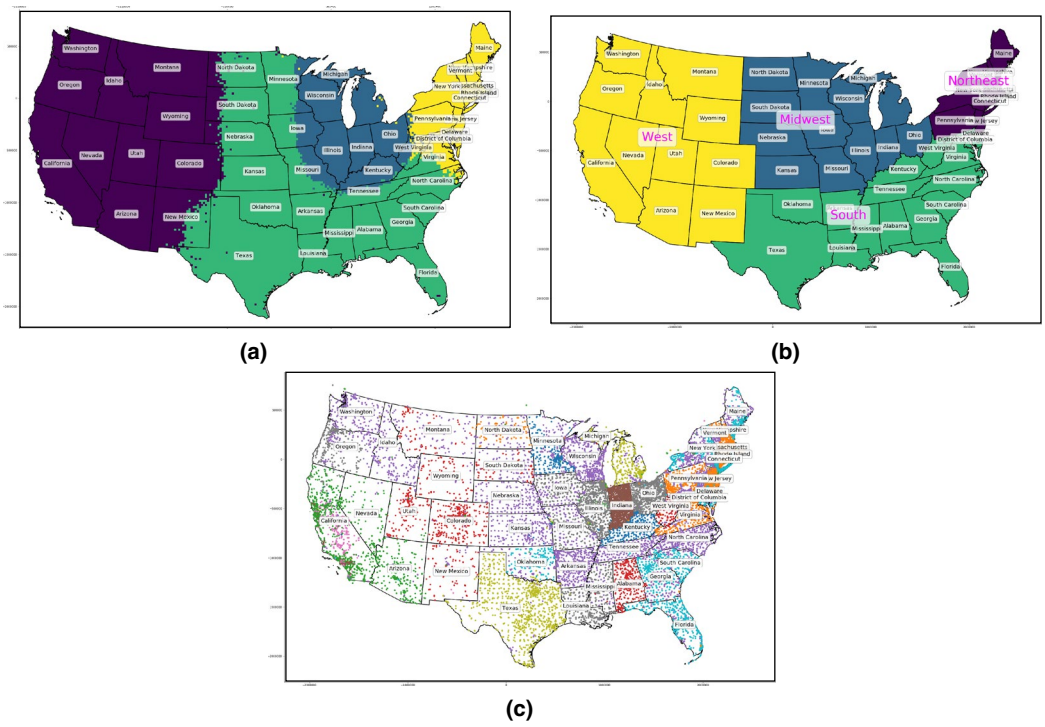


FIGURE 11 (a) Clustering result of location embeddings produced by the location encoder $\text{LocEnc}^{(x)}()$ in $SE-KGE_{\text{space}}$. It illustrates spatial coherence and semantics; (b) Census Bureau-designated regions of the U.S.; and (c) the community detection (shuffled Louvain) results of knowledge graph $\mathcal{G}$ by treating $\mathcal{G}$ as a undirected unlabeled multi-graph. It lacks spatial coherence

We also test how well the location encoder $\text{LocEnc}^{(x)}()$ in *SE-KGE* can capture the global position information and how $\text{LocEnc}^{(x)}()$ interacts with other components of *SE-KGE*. We use $SE-KGE_{\text{space}}$ as an example. Since $\text{LocEnc}^{(x)}()$ is an inductive learning model, we divide the study area $\mathcal{A}$ into 20 km by 20 km grids and take the location of each grid center as the input of $\text{LocEnc}^{(x)}()$. Each grid will get a $d^{(x)}$-dimensional location embedding after location encoding. We apply hierarchical clustering on these embeddings. Figure 11a shows the clustering result. We compare it with the widely used US Census Bureau-designated regions (https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States) (see Figure 11b). We can see that Figure 11a and b look very similar to each other. We use two clustering evaluation metrics—normalized mutual information (NMI) and Rand index—to measure the degree of similarity, yielding .62 on NMI and .63 on the Rand index. Taking a closer look at Figure 11a, we can also see that the clusters are divided on the state borders. We hypothesize that this is because $\text{LocEnc}^{(x)}()$ is informed of the connectivity of different geographic entities in $\mathcal{G}$ during model training, resulting in locations which are connected in the original $\mathcal{G}$ also being clustered after training.

To validate this hypothesis, we apply the Louvain community detection algorithm with a shuffled node sequence (https://github.com/tsakim/Shuffled_Louvain) on the original $\mathcal{G}$ by treating $\mathcal{G}$ as an undirected and unlabeled graph. Figure 11c shows the community structure with the best modularity, which contains 32 communities. Some interesting observations can be made by comparing these three figures:

1. Most communities in Figure 11c are separated at state borders, which is evidence for our hypothesis.
2. Some communities contain locations in different states, which are far away from each other. For example, the red community contains locations in Utah, Colorado, and Alabama. This indicates that some locations are very

similar purely based on the graph structure of $\mathcal{G}$. As $\text{LocEnc}^{(x)}()$ imposes spatial constraints on entities, spatially coherent clusters in Figure 11a are presented.

One hypothesis why Figure 11a and b look similar is that in the KG, the number of connections between entities within one Bureau-designated region is more than the number of connections among entities in different regions. This may be due to the fact that DBpedia uses census data as one of the data sources while census data are organized in a way which reflects Bureau-designated regions of the U.S. More research is needed to validate this hypothesis in the future.

7.3 | Evaluation on spatial semantic lifting task

7.3.1 | Baselines

The spatial semantic lifting model is composed of Enc(), $\mathcal{P}^{(e)}()$ and $\mathcal{P}^{(x)}()$, and is denoted by $SE-KGE_{ssl}$. In order to study the contribution of the feature and location encoders, we create a baseline $SE-KGE'_{\text{space}}$ whose entity encoder does not have the feature encoder component, similar to $SE-KGE_{\text{space}}$. The difference is that they are trained on different objectives. These are the only two models that can perform the spatial semantic lifting task, since they are fully inductive learning models directly using locations as the only input features.

7.3.2 | Training details

We train $SE-KGE_{ssl}$ and $SE-KGE'_{\text{space}}$ based on $\mathcal{L}^{(SSL)}$. To quantitatively evaluate them on the spatial semantic lifting task, we use $\mathcal{T}_s \cap \mathcal{T}_o$ in the validation and testing data set with different relations (see Table 2). For each triple $s_i = (h_i, r_i, t_i) \in \mathcal{T}_s$, given the head entity's location and r_i, we use $\mathcal{P}^{(x)}(\mathcal{X}(h_i), r_i)$ (see Equation 21) to predict the tail entity embedding. A similar process can be carried out for $s_j = (h_j, r_j, t_j) \in \mathcal{T}_o$ but from the reverse direction. We also use AUC and APR as the evaluation metrics. Note that since $\mathcal{X}(h_i) = \mathbf{x}_i^{(t)} \sim \mathcal{U}(\mathbf{x}_i^{\min}, \mathbf{x}_i^{\max})$, $\mathcal{PN}(h_i) = [\mathbf{x}_i^{\min}; \mathbf{x}_i^{\max}]$ if $h_i \in \mathcal{V}_{pn}$, the location of head entity is randomly generated, which can be treated as unseen in the training process. We use the same hyperparameter configuration as $SE-KGE_{\text{full}}$.

7.3.3 | Evaluation results

Table 5 shows the overall evaluation results. We can see that $SE-KGE_{ssl}$ outperforms $SE-KGE'_{\text{space}}$ by a significant margin (ΔAUC = 9.86% and ΔAPR = 9.59% on the testing data set), which clearly shows the strength of considering both feature embedding and space embedding in the spatial semantic lifting task.

TABLE 5 Evaluation of spatial semantic lifting on *DBGeo* over all validation/testing triples

	$SE-KGE_{\text{space}}$		$SE-KGE_{ssl}$		$SE-KGE_{ssl}$ - $SE-KGE_{\text{space}}$	
	AUC	APR	AUC	APR	ΔAUC	ΔAPR
Valid	72.85	75.49	**82.74**	**85.51**	9.89	10.02
Test	73.41	75.77	**83.27**	**85.36**	9.86	9.59

Note: The bold is used to highlight the difference.

Next, among all validation and testing triples with different relations, we select a few relations and report the APR of two models on these triples with specific relations. The results are shown in Table 6. These relations are selected since they are interesting from a spatial reasoning perspective. We can see that $SE-KGE_{ssl}$ outperforms $SE-KGE'_{space}$ on all these triple sets with different relations.

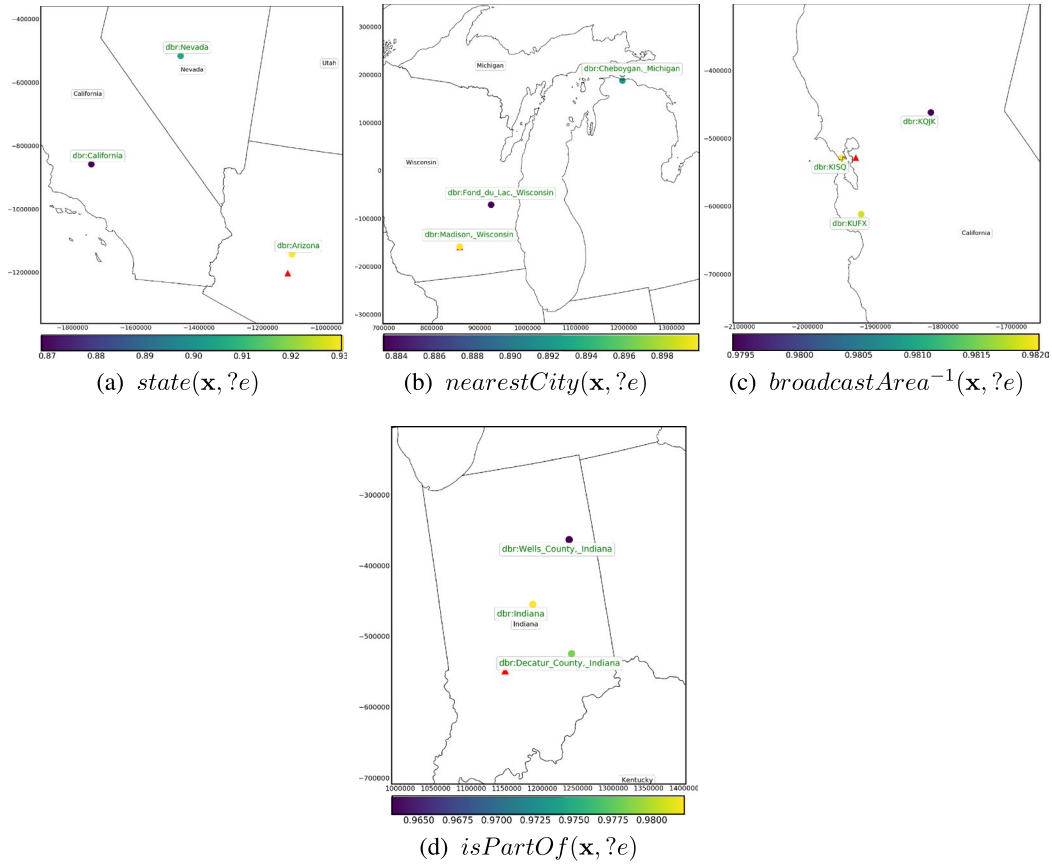
In order to know how well $SE-KGE_{ssl}$ understands the semantics of different types of (spatial) relations, we visualize the spatial semantic lifting results in Figure 12 for four spatial relations: `dbo:state`, `dbo:nearestCity`, `dbo:broadcastArea-1` and `dbo:isPartOf`. `dbo:state`, `dbo:isPartOf` and `dbo:broadcastArea-1` are about partonomy relations, while `dbo:nearestCity` is an example of pointwise metric spatial relations. Some interesting observations can be made:

1. $SE-KGE_{ssl}$ is capable of capturing the spatial proximity such that the top ranked geographic entity (yellow point) in each case is the closest to location **x** (red triangle). We also treat this as an indicator of the capability of $SE-KGE_{ssl}$ to handle partonomy relations and pointwise metric spatial relations.
2. $SE-KGE_{ssl}$ can capture the semantics of relations (e.g., the domain and range of each relation/predicate). All top-ranked entities are within the range of the corresponding relation. For example, in Figure 12a with Query *state*(**x**,?*e*), the top three entities are all states spatially close to **x**. In Figure 12b with Query broadcastArea^{-1}(**x**, ?*e*), all top three entities are nearby radio stations. In Figure 12d with Query isPartOf(**x**, ?*e*), all top three entities are states (`dbo:Indiana`) and counties.
3. We notice that the result of query nearestCity(**x**, ?*e*) in Figure 12b is not good enough since the second result, `dbo:Cheboygan, _Michigan`, is outside of Wisconsin. After investigating the triples with `dbo:nearestCity` as the relation, we find that `dbo:nearestCity` usually links a natural resource entity (e.g., lakes, national parks) to a city. These natural resource entities usually cover large areas and complex geometries. So `dbo:nearestCity` is not a purely pointwise distance-based relation but a complex distance-based relation based on real geometries. Since our model only takes the bounding box of each entity and there are usually no subdivisions of these nature resource entities, it is hard for our model to learn the semantics of `dbo:nearestCity`.

TABLE 6 Evaluation of $SE-KGE_{ssl}$ and $SE-KGE'_{space}$ on *DBGeo* for a few selected relations *r*, using APR (%) as the evaluation metric

	Query type	$SE-KGE'_{space}$	$SE-KGE_{ssl}$	ΔAPR
Valid	*state*(**x**,?*e*)	92.00	**99.94**	7.94
	nearestCity(**x**,?*e*)	84.00	**94.00**	10.00
	broadcastArea^{-1}(**x**,?*e*)	91.60	**95.60**	4.00
	isPartOf(**x**,?*e*)	88.56	**98.88**	10.32
	locationCity(**x**,?*e*)	83.50	**99.00**	15.50
	residence^{-1}(**x**,?*e*)	90.50	**93.50**	3.00
	hometown^{-1}(**x**,?*e*)	61.14	**74.86**	13.71
Test	*state*(**x**,?*e*)	89.06	**99.97**	10.91
	nearestCity(**x**,?*e*)	87.60	**99.80**	12.20
	broadcastArea^{-1}(**x**,?*e*)	90.81	**96.63**	5.82
	isPartOf(**x**,?*e*)	87.66	**98.87**	11.21
	locationCity(**x**,?*e*)	84.80	**99.10**	14.30
	residence^{-1}(**x**,?*e*)	61.21	**77.68**	16.47
	hometown^{-1}(**x**,?*e*)	61.44	**76.83**	15.39

Note: The bold is used to highlight the difference.



(a) $state(\mathbf{x}, ?e)$ (b) $nearestCity(\mathbf{x}, ?e)$ (c) $broadcastArea^{-1}(\mathbf{x}, ?e)$

(d) $isPartOf(\mathbf{x}, ?e)$

FIGURE 12 Visualization of spatial semantic lifting of $SE-KGE_{ssl}$. (a)–(d) show the top three geographic entities which can answer query $r(\mathbf{x}, ?e)$ where r is the relation we pick. Red triangle: the selected location $\mathbf{x}$. Circles: top three geographic entities ranked by our model, their colors indicating cosine similarity between the geographic entities and the predicted query embedding

Based on the evaluation results and model analysis, we can see that, given a relation r, $SE-KGE_{ssl}$ is able to link a location $\mathbf{x}$ to an entity e in $\mathcal{G}$ by considering the semantics of r and spatial proximity.

8 | CONCLUSIONS

In this work we propose a location-aware knowledge graph embedding model called *SE-KGE* which enables spatial reasoning in the embedding space for its three major components: the entity embedding encoder Enc(), the projection operator $\mathcal{P}()$, and the intersection operator $\mathcal{I}()$. We demonstrate how to incorporate spatial information of geographic entities such as locations and spatial extents into Enc() such that *SE-KGE* can handle different types of spatial relations such as pointwise metric spatial relations and partonomy relations. To the best of our knowledge, this is the first KG embedding model which incorporates location encoding into the model architecture instead of relying on some form of distance measure among entities while capturing the scale effect of different geographic entities. Two tasks have been used to evaluate the performance of *SE–KGE*: geographic logic query answering and spatial semantic lifting. Results show that $SE-KGE_{full}$ can outperform multiple baselines on the geographic logic query answering task, which indicates the effectiveness of spatially-explicit models. It also demonstrates

the importance of considering the scale effect in location encoding. We also proposed a new task, spatial semantic lifting, with the aim of linking a randomly selected location to entities in the existing KG with some relation. None of the existing KG embedding models can solve this task except our model. We have shown that $SE-KGE_{ssl}$ can significantly outperform the baseline $SE-KGE'_{space}$ (ΔAUC = 9.86% and ΔAPR = 9.59% on the testing data set). Visualizations show that $SE-KGE_{ssl}$ can successfully capture the spatial proximity information as well as the semantics of relations. In the future, we hope to explore a more concise way to encode the spatial footprints of geographic entities in a KG. Moreover, we want to explore more varieties of the spatial semantic lifting task.

ACKNOWLEDGMENTS

This work was partially supported by the NSF award 1936677, Spatially-Explicit Models, Methods, and Services for Open Knowledge Networks as well as Esri Inc.

ORCID

Gengchen Mai http://orcid.org/0000-0002-7818-7309
Ling Cai http://orcid.org/0000-0001-7106-4907
Blake Regalia http://orcid.org/0000-0001-9303-1591
Bo Yan http://orcid.org/0000-0002-4248-7203

ENDNOTES

1 Note that, in many KGs (e.g., DBpedia, Wikidata), an entity can belong to multiple types. We use this definition to be in line with many existing works (Hamilton et al., 2018; Mai, Janowicz, et al., 2019) so that we can compare our results. It is easy to relax this requirement, which we will discuss in Section 6.1.

2 It is worth mentioning that most KGs to date merely store point geometries even for features such as the U.S.

3 For a detailed comparison between CGQs and SPARQL 1.1 queries, see Section 2.1 of Mai, Janowicz, et al. (2019).

4 `dbo:nearestCity` triples in DBpedia are triplified from the "nearest major city" row of the info box in each entity's corresponding Wikipedia page which may contain several cities. See http://dbpedia.org/resource/San_Francisco_Bay.

5 We treat an entity as a geographic entity if its has a `geo:geometry` triple in DBpedia.

REFERENCES

Abbott, A., & Callaway, E. (2014). Prize for place cells: Discoverers of brain's navigation system get physiology Nobel. *Nature*, *514*(7521), 153.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski. M., ⋯ Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. Preprint, arXiv:1806.01261.

Berant, J., Chou, A., Frostig, R., & Liang, P. (2013). Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, WA (pp. 1533–1544). Stroudsburg, PA: ACL.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. *26*, pp. 2787–2795). Red Hook, NY: Curran Associates.

Cai, L., Yan, B., Mai, G., Janowicz, K., & Zhu, R. (2019). TransGCN: Coupling transformation assumptions with graph convolutional networks for link prediction. In *Proceedings of the 10th International Conference on Knowledge Capture*, Marina del Ray, CA (pp. 131–138). New York, NY: ACM.

Chu, G., Potetz, B., Wang, W., Howard, A., Song, Y., Brucher, F., ⋯ Adam, H. (2019). Geo-aware networks for fine-grained recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, Long Beach, CA. Piscataway, NJ: IEEE.

De Nicola, A., DiMascio, T., Lezoche, M., & Tagliano, F. (2008). Semantic lifting of business process models. In *Proceedings of the 12th Enterprise Distributed Object Computing Conference Workshops*, Munich, Germany (pp. 120–126). New York, NY: ACM.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., ⋯ Zhang, W. (2014). Knowledge Vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY (pp. 601–610). New York, NY: ACM.

Hamilton, W., Bajaj, P., Zitnik, M., Jurafsky, D., & Leskovec, J. (2018). Embedding logical queries on knowledge graphs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. *32*, pp. 2026–2037). Red Hook, NY: Curran Associates.

Hamilton, W., Ying, Z., & Leskovec, J. (2017a). Inductive representation learning on large graphs. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. *31*, pp. 1024–1034). Red Hook, NY: Curran Associates.

Hamilton, W. L., Ying, R., & Leskovec, J. (2017b). Representation learning on graphs: Methods and applications. Preprint, arXiv:1709.05584.

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, *194*, 28–61.

Janowicz, K., Hu, Y., McKenzie, G., Gao, S., Regalia, B., Mai, G., ⋯ Taylor, K. (2016). Moon landing or safari? A study of systematic errors and their causes in geographic linked data. In J. Miller, D. O'Sullivan, & N. Wiegand (Eds.), *Geographic Information Science*, *9th International Conference*, *GIScience 2016* (pp. 275–290). Berlin, Germany: Springer.

Janowicz, K., Scheider, S., Pehle, T., & Hart, G. (2012). Geospatial semantics and linked spatiotemporal data: Past, present, and future. *Semantic Web*, *3*(4), 321–332.

Kejriwal, M., & Szekely, P. (2017). Neural embeddings for populated geonames locations. In C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. CudrT-Mauroux, J. Sequeda, ⋯ J. Heflin (Eds.), *The semantic web—ISWC 2017* (Lecture Notes in Computer Science, Vol. *10588*, pp. 139–146). Cham, Switzerland: Springer.

Kristiadi, A., Khan, M. A., Lukovnikov, D., Lehmann, J., & Fischer, A. (2019). Incorporating literals into knowledge graph embeddings. In C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, L. Cruz, A. Hogan, ⋯ F. Gandon (Eds.), *The semantic web—ISWC 2019* (Lecture Notes in Computer Science, Vol. *11778*, pp. 347–363). Cham, Switzerland: Springer.

Lao, N., Mitchell, T., Cohen, W. W. (2011). Random walk inference and learning in a large scale knowledge base. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK (pp. 529–539). Stroudsburg, PA: ACL.

Liang, C., Berant, J., Le, Q., Forbus, K., & Lao, N. (2017). Neural symbolic machines: Learning semantic parsers on Freebase with weak supervision. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, BC, Canada (Vol. *1*, pp. 23–33). Stroudsburg, PA: ACL.

Mac Aodha, O., Cole, E., Perona, P. (2019). Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, South Korea (pp. 9596–9606). Piscataway, NJ: IEEE.

Mai, G., Janowicz, K., & Yan, B. (2018). Support and centrality: Learning weights for knowledge graph embedding models. In C. Faron Zucker, C. Ghidini, A. Napoli, & Y. Toussaint (Eds.), *Knowledge engineering and knowledge management - EKAW 2018* (Lecture Notes in Computer Science, Vol. *11313*, pp. 212–227). Cham, Switzerland: Springer.

Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L., & Lao, N. (2019). Contextual graph attention for answering logical queries over incomplete knowledge graphs. In *Proceedings of the 10th International Conference on Knowledge Capture*, Marina del Ray, CA (pp. 171–178). New York, NY: ACM.

Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L., & Lao, N. (2020). Multi-scale representation learning for spatial feature distributions using grid cells. In *Proceedings of the Eighth International Conference on Learning Representations*, Addis Ababa, Ethiopia. openreview.

Mai, G., Yan, B., Janowicz, K., & Zhu, R. (2019). Relaxing unanswerable geographic questions using a spatially explicit knowledge graph embedding model. In P. Kyriakidis, D. Hadjimitsis, D. Skarlatos, & A. Mansourian (Eds.), *Geospatial technologies for local and regional development: Proceedings of the 22nd AGILE Conference on Geographic Information Science* (pp. 21–39). Berlin, Germany: Springer.

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2015). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, *104*(1), 11–33.

Nickel, M., Rosasco, L., & Poggio, T. A. (2016). Holographic embeddings of knowledge graphs. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Phoenix, AR (pp. 1955–1961). Menlo Park, CA: AAAI.

Nickel, M., Tresp, V., & Kriegel, H.-P. (2012). Factorizing YAGO: Scalable machine learning for linked data. In *Proceedings of the 21st World Wide Web Conference*, Lyon, France (pp. 271–280). New York, NY: ACM.

Pennington, J., Socher, R., Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar (pp. 1532–1543). Stroudsburg, PA: ACL.

Pezeshkpour, P., Chen, L., & Singh, S. (2018). Embedding multimodal relational data for knowledge base completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium (pp. 3208–3218). Stroudsburg, PA: ACL.

Qiu, P., Gao, J., Yu, L., & Lu, F. (2019). Knowledge embedding with geospatial distance restriction for geographic knowledge graph completion. *ISPRS International Journal of Geo-Information*, *8*(6), 254–277.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia (Vol. *2*, pp. 784–789). Stroudsburg, PA: ACL.

Regalia, B., Janowicz, K., & McKenzie, G. (2019). Computing and querying strict, approximate, and metrically refined topological relations in linked geographic data. *Transactions in GIS*, *23*(3), 601–619.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. In A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, ⋯ M. Alam (Eds.), *The semantic web—ESWC 2018* (Lecture Notes in Computer Science, Vol. *10843*, pp. 593–607). Cham, Switzerland: Springer.

Schölkopf, B. (2001). The kernel trick for distances. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems* (Vol. *13*, pp. 301–307). Cambridge, MA: MIT Press.

Schölkopf, B., Sung, K.-K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., & Vapnik, V. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, *45*(11), 2758–2765.

Trisedya, B. D., Qi, J., & Zhang, R. (2019). Entity alignment between knowledge graphs using attribute embeddings. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Honolulu, HI (pp. 297–304). Menlo Park, CA: AAAI.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ⋯ Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. *30*, pp. 5998–6008). Red Hook, NY: Curran Associates.

Wang, M., Wang, R., Liu, J., Chen, Y., Zhang, L., & Qi, G. (2018). Towards empty answers in SPARQL: Approximating querying with RDF embedding. In D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, ⋯ E. Simperl (Eds.), *The semantic web—ISWC 2018* (Lecture Notes in Computer Science, Vol. *11136*, pp. 513–529). Berlin, Germany: Springer.

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge & Data Engineering*, *29*(12), 2724–2743.

Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014). Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Québec City, Québec, Canada (pp. 1112–1119). Menlo Park, CA: AAAI.

Yan, B., Janowicz, K., Mai, G., & Zhu, R. (2019). A spatially explicit reinforcement learning model for geographic knowledge graph summarization. *Transactions in GIS*, *23*(3), 620–640.

Yang, B., Yih, W.-T., He, X., Gao, J., & Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the Third International Conference on Learning Representations*, San Diego, CA. Baixis, France: ISCA.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., & Smola, A. J. (2017). Deep sets. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. *30*, pp. 3391–3401). Red Hook, NY: Curran Associates.

Zhu, R., Hu, Y., Janowicz, K., & McKenzie, G. (2016). Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Transactions in GIS*, *20*(3), 333–355.

How to cite this article: Mai G, Janowicz K, Cai L, et al. *SE-KGE*: A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting. *Transactions in GIS*. 2020;00:1–33. https://doi.org/10.1111/tgis.12629