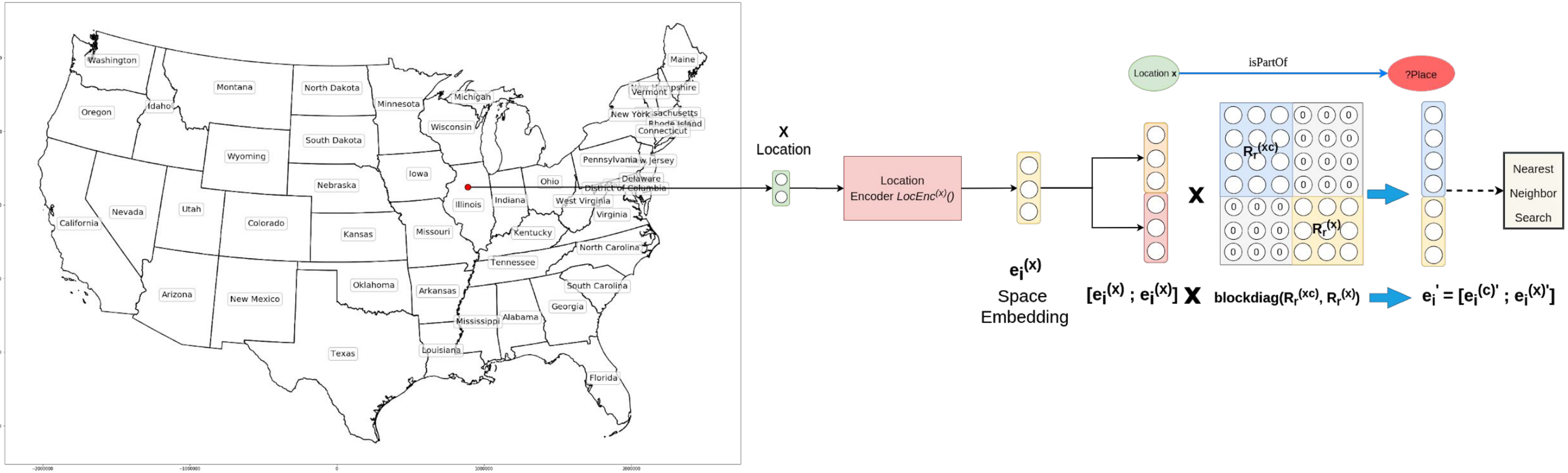
SE-KGE: A Location-Aware Knowledge Graph Embedding Model for Geographic Question Answering and Spatial Semantic Lifting

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]STKO Lab, UC Santa Barbara; [2] LinkedIn Corporation; [3] SayMosaic Inc.



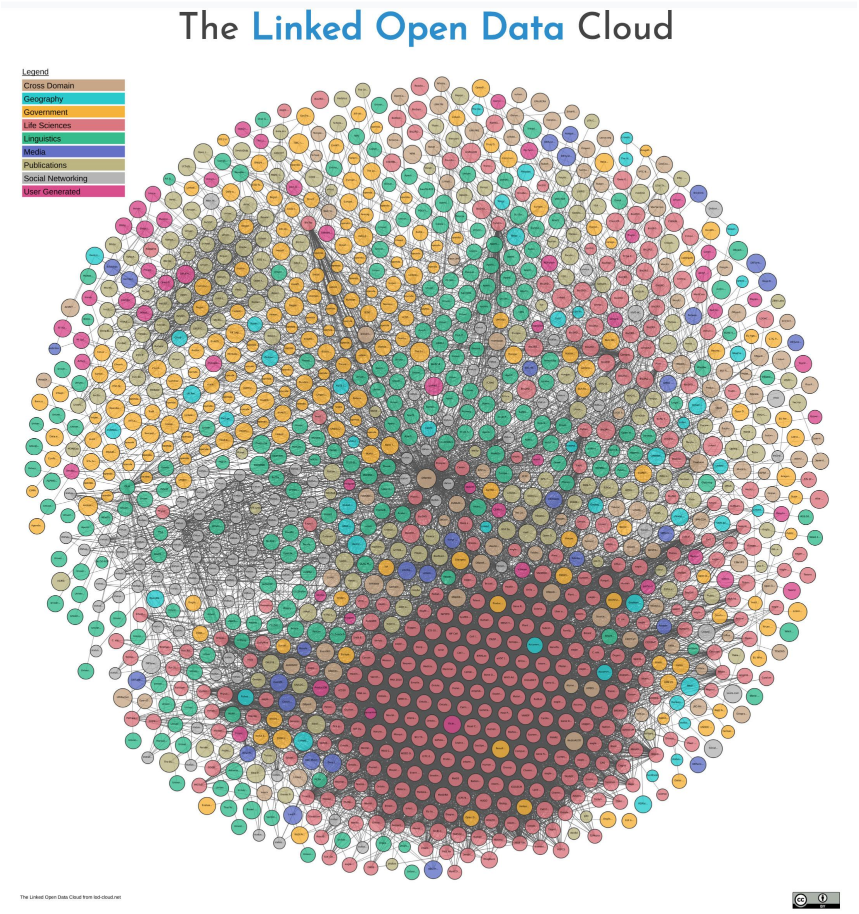
Spatial semantic lifting in the SE-KGE embedding space

Arxiv paper: https://arxiv.org/abs/2004.14171

Knowledge Graph

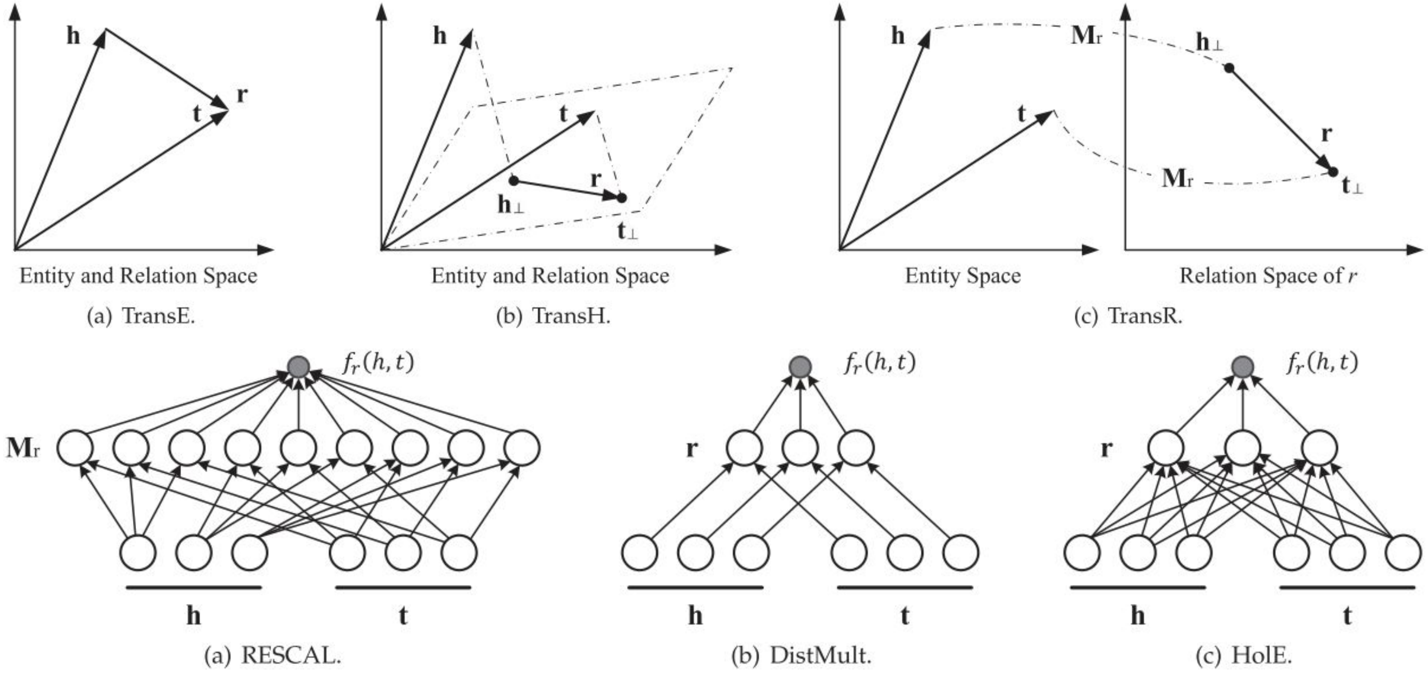
Knowledge Graph (KG): a labeled and directed multi-graph of statements (called triples) about the world

Problem: incompleteness and sparsity



Knowledge Graph Embedding

Knowledge Graph Embedding (KGE): project entities and relations in a KG onto **a continuous vector space** while preserving the inherent structure of the KG



Illustrations for several well-known knowledge graph embedding models (Wang et al., 2017)

Problems of Knowledge Graph Embedding

- Incompleteness and sparsity problems affect the performance of downstream tasks such as question answering (QA) since missing triples result in **certain questions becoming unanswerable**

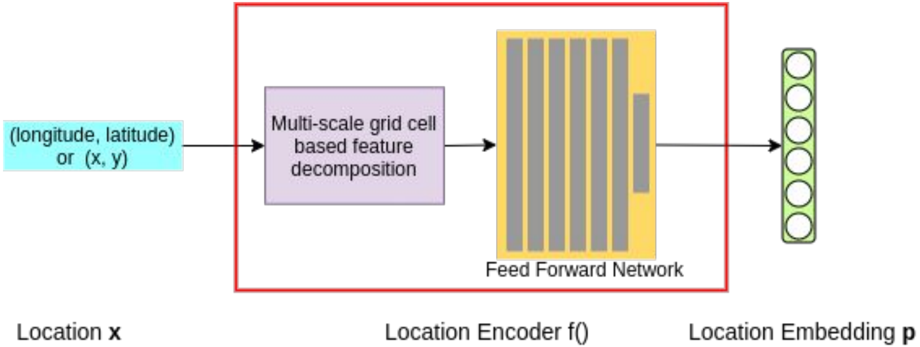
- Neglected spatial aspects, e.g., the spatial footprints of geographic entities despite the fact that they are important for many KG downstream tasks:
  - Geographic knowledge graph completion (Qiu et al., 2019)
  - Geographic ontology alignment (Zhu et al., 2016)
  - Geographic entity alignment (Trisedya et al., 2019)
  - Geographic question answering (Mai et al., 2019b)
  - Geographic knowledge graph summarization (Yan et al., 2019)
  - ….

SE-KGE: A Location-Aware KG Embedding Model

A novel KGE model which **directly encodes spatial footprints**, namely **point coordinates** and **bounding boxes**, thereby making them available while learning knowledge graph embeddings.

Encoding spatial footprints of geographic entities:

- **Location encoder** (Mai et al., 2020): the neural network models which encode a pair of coordinates into a high dimensional embedding which can be used in multi downstream tasks



Challenges of SE-KGE

1. Location encoding can handle **point-wise metric relations** (e.g., dbo:nearestCity) and **directional relations** (e.g., dbp:north) in KGs, but it is not easy to encode containment relations (e.g., dbo:isPartOf).
    - Represent geographic entities as **regions** instead of points in the embedding space

2. How to seamlessly handle **geographic** and **non-geographic entities**?

3. How to capture the **spatial** and **other semantic aspects** at the same time?

4. **Spatial Semantic Lifting**: How to design a KGE model so that it can be used to infer new relations between entities in a KG and any arbitrary location in the study area?

Method: GeoKG Definition

Given a geographic knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- V : the set of entities/nodes
- E : the set of directed edges
- $\mathcal{V}_{pt} \subseteq \mathcal{V}$: the geographic entity set
- $\mathcal{PT}(\cdot)$: entity $e \in \mathcal{V}_{pt}$ => $\mathcal{PT}(e) = \mathbf{x}$ where $\mathbf{x} \in \mathcal{A} \subseteq \mathbb{R}^2$
- $\mathcal{V}_{pn} \subseteq \mathcal{V}_{pt}$: the set of large-scale geographic entity
- $\mathcal{PN}(\cdot)$: entity $e \in \mathcal{V}_{pn}$ => $\mathcal{PN}(e) = [\mathbf{x}^{min}; \mathbf{x}^{max}] \in \mathbb{R}^4$ where $\mathbf{x}^{min}, \ \mathbf{x}^{max} \in \mathcal{A} \subseteq \mathbb{R}^2$

Method: CQG Definition

Definition 2 (Conjunctive Graph Query (CGQ)). *A query $q \in Q(\mathcal{G})$ that can be written as follows:*

$$
\begin{aligned}
& q = V_?.\exists V_1, V_2, .., V_m : b_1 \wedge b_2 \wedge ... \wedge b_n \\
\textit{where} \quad & b_i = r_i(e_k, V_l), V_l \in \{V_?, V_1, V_2, .., V_m\}, e_k \in \mathcal{V}, r \in \mathcal{R} \\
\textit{or} \quad & b_i = r_i(V_k, V_l), V_k, V_l \in \{V_?, V_1, V_2, .., V_m\}, k \neq l, r \in \mathcal{R}
\end{aligned}
$$

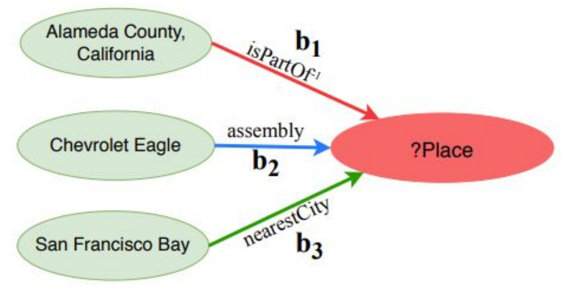
- $Q(\mathcal{G})$: a set of all conjunctive graph queries that can be asked over G
- $V_?$: the target variable of query q (target node)
- $V_1, V_2, .., V_m$: existentially quantified bound variables (bound nodes)
- b_i : a basic graph pattern in this CGQ
- e_k : the entity node appeared in the question (anchor node)

The dependency graph of Query q is a **directed acyclic graph** (DAG)

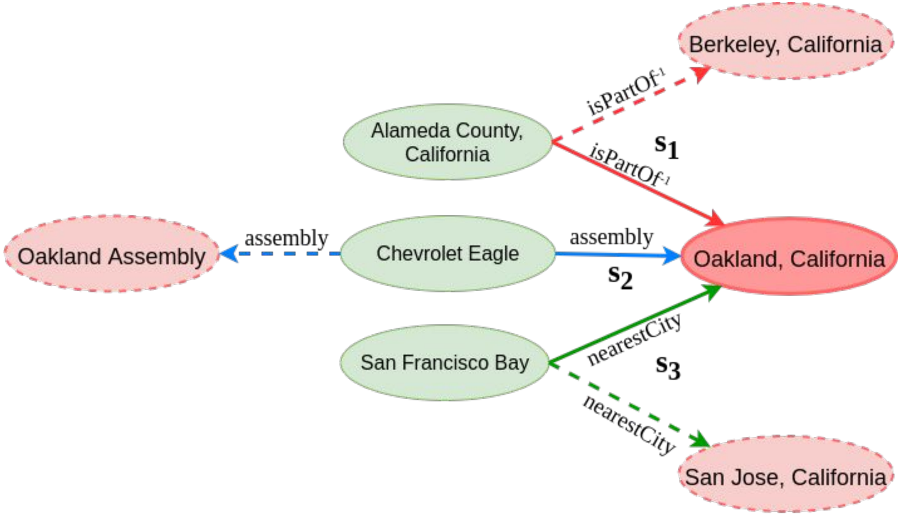
Geographic CGQ: the answer entity is a geographic entity

Method: CQG Example

Which city in Alameda County, California is the assembly place of Chevrolet Eagle and the nearest city to San Francisco Bay?



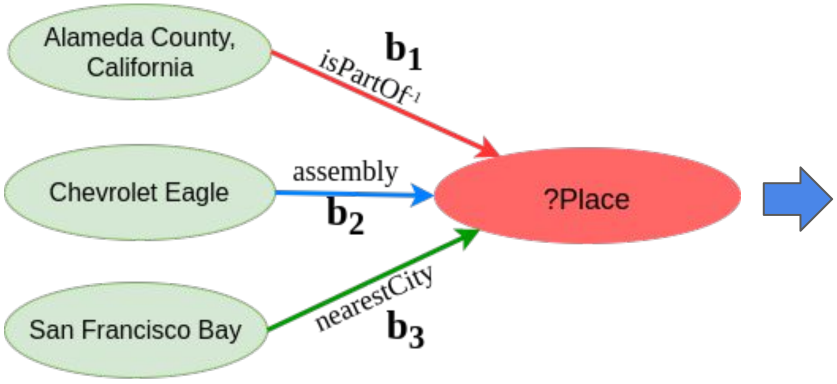
$$?Place : IsPartOf^{-1}(Alameda\ County, ?Place) \wedge$$
$$Assembly(Chevrolet\ Eagle, ?Place) \wedge$$
$$NearestCity(San\ Francisco\ Bay, ?Place)$$

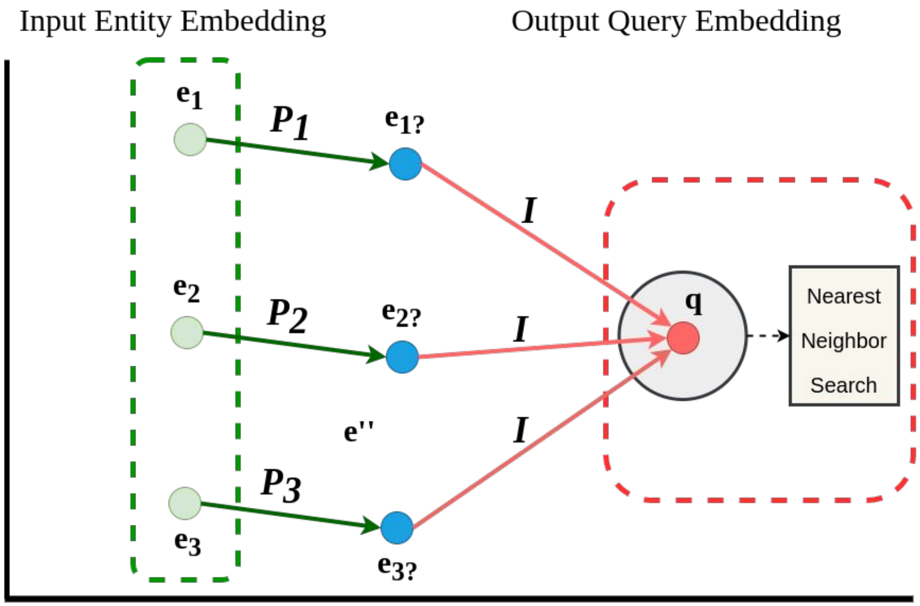


Method: Three Components for GeoQA

There major components of SE-KGE:

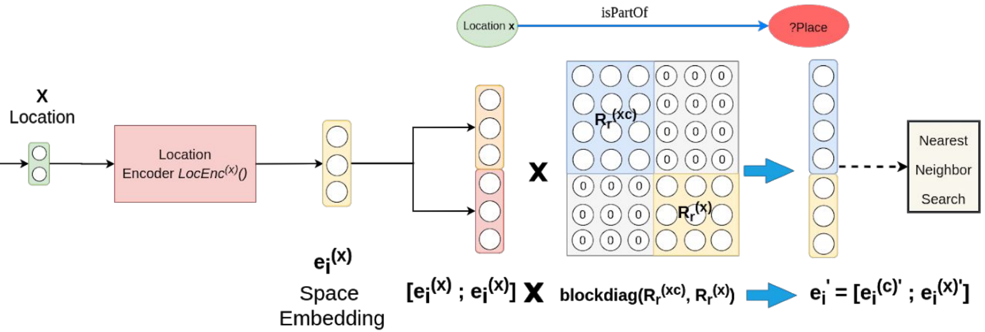
- **Entity encoder** $Enc()$
- **Projection operator** $\mathcal{P}()$
- **Intersection operator** $\mathcal{I}()$





Method: Space Semantic Lifting

Use entity encoder $Enc()$ and projection operator $\mathcal{P}()$ for spatial semantic lifting:



Note that location encoder is one component of entity encoder

Method: Location-Aware Entity Encoder

- Semantic Aspect:

Definition 4 (Entity Feature Encoder: $Enc^{(c)}()$). *Given any entity $e_i \in \mathcal{V}$ with type $c_i = \Gamma(e_i) \in \mathcal{C}$ from $\mathcal{G}$, entity feature encoder $Enc^{(c)}()$ 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:*

$$\mathbf{e}_i^{(c)} = Enc^{(c)}(e_i) = \frac{\mathbf{Z}_{c_i}\mathbf{h}_i^{(c)}}{\| \mathbf{Z}_{c_i}\mathbf{h}_i^{(c)} \|_{L2}} \tag{5}$$

- Space Aspect:

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

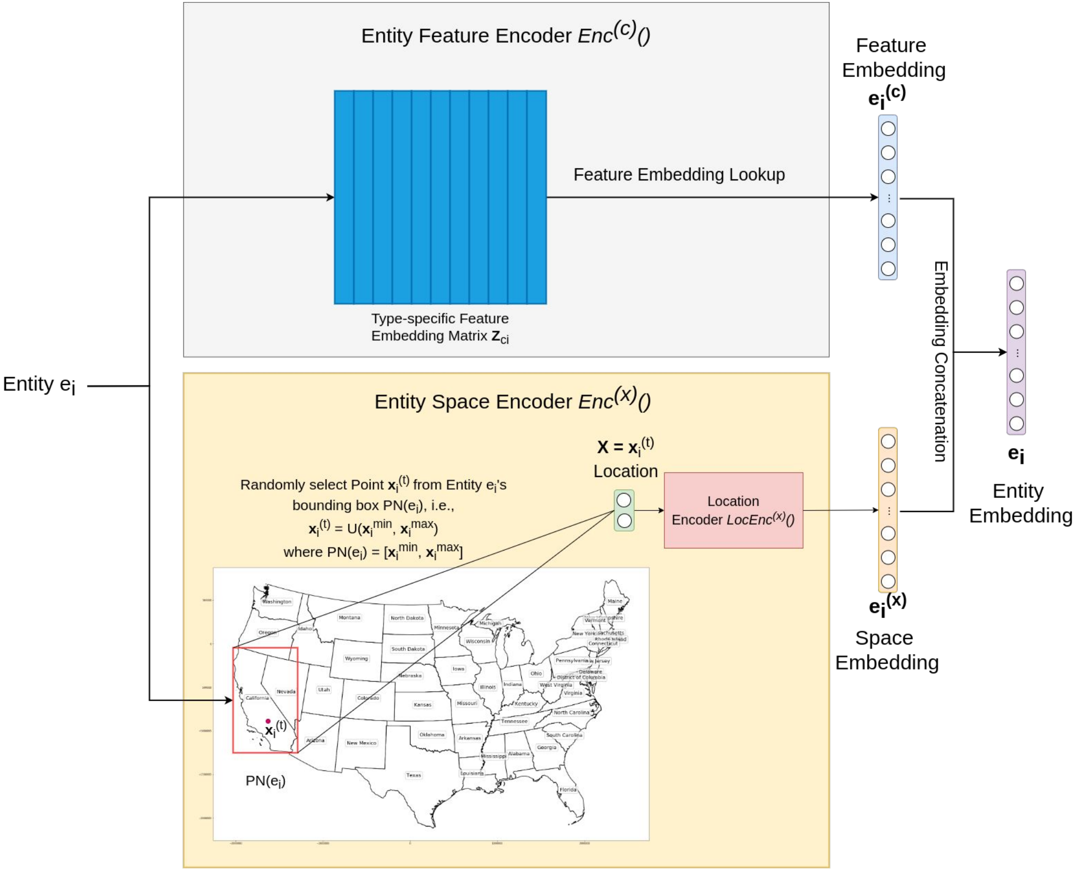
$$\mathbf{e}_i^{(x)} = \begin{cases} LocEnc^{(x)}(\mathbf{x}_i)\ , where\ \mathbf{x}_i = \mathcal{PT}(e_i)\ , & if\ e_i \in \mathcal{V}_{pt} \backslash \mathcal{V}_{pn} \\ LocEnc^{(x)}(\mathbf{x}_i^{(t)})\ , where\ \mathbf{x}_i^{(t)} \sim \mathcal{U}(\mathbf{x}_i^{min}, \mathbf{x}_i^{max})\ ,\ \mathcal{PN}(e_i)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

Method: Location-Aware Entity Encoder

- Entity Feature Encoder

- Entity Space Encoder

Encoding results are concatenated as the final output

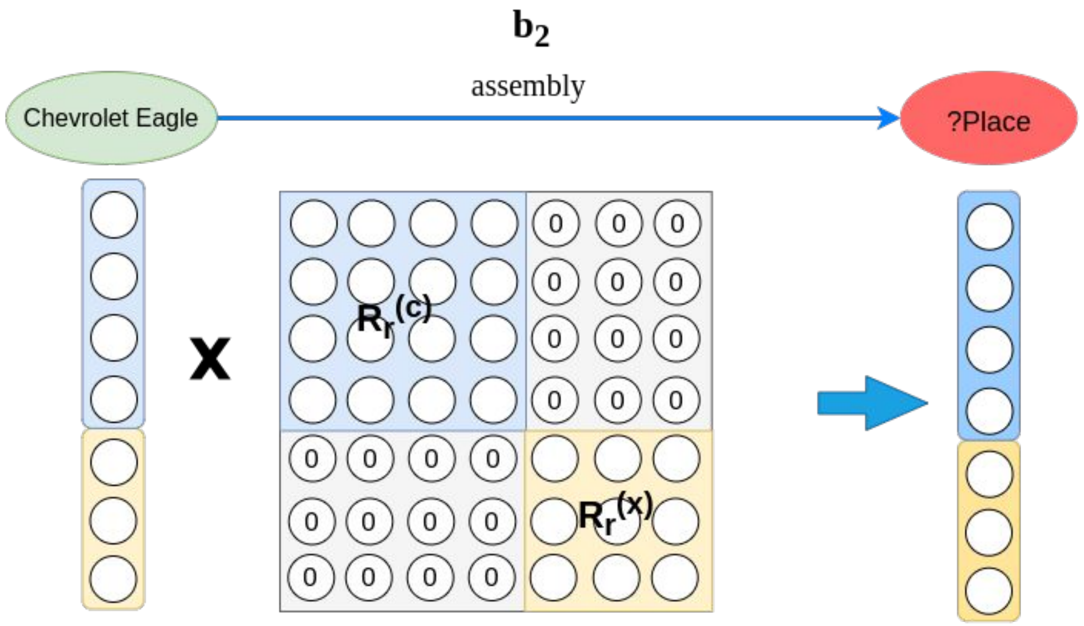


Method: Location-Aware Projection Operator

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 of (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:* **(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;* **(2) 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:*

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

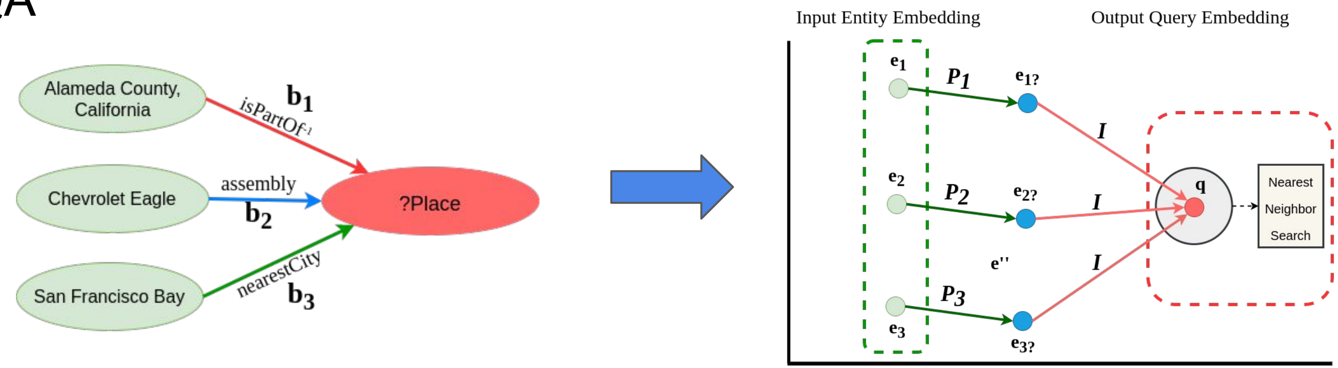
Method: Location-Aware Projection Operator



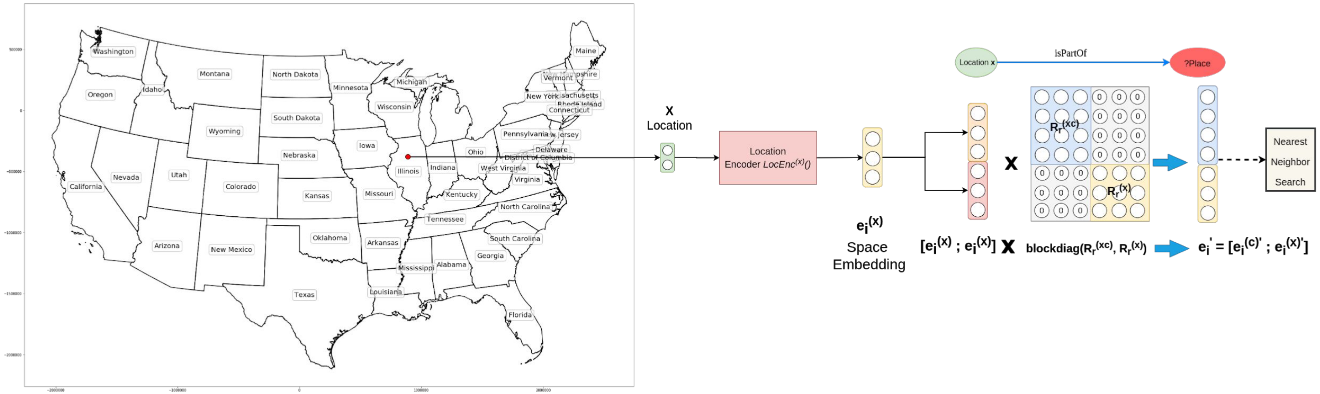
$$e_i = [e_i^{(c)} ; e_i^{(x)}] \times \quad \mathrm{blockdiag}(R_r^{(c)}, R_r^{(x)}) \quad \rightarrow \quad e_i' = [e_i^{(c)\prime} ; e_i^{(x)\prime}]$$

Method: GeoQA and Spatial Semantic Lifting

- GeoQA



- Spatial Semantic Lifting

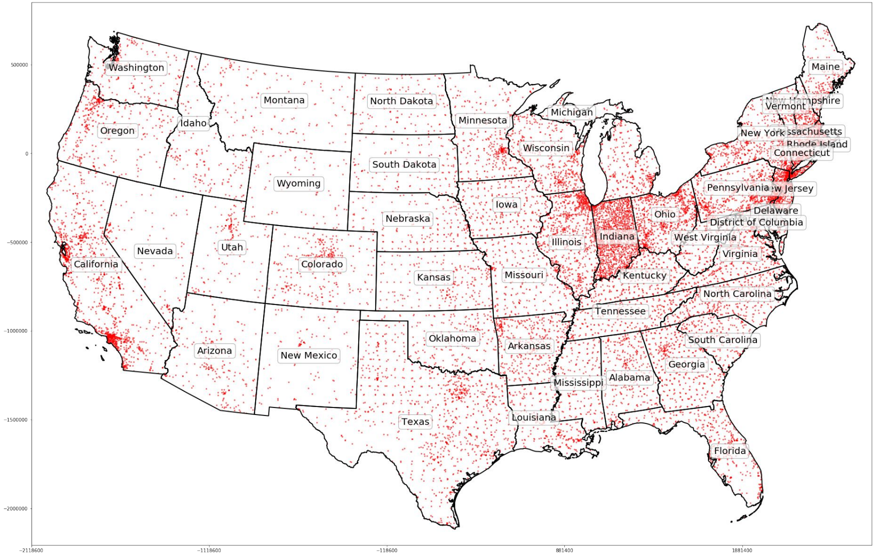


Experiment

Evaluate SE-KGE using the DBGeo dataset which is built based on a subgraph of DBpedia

Table 1: Statistics for our dataset in *DBGeo* (Section 7.1). “XXXX/QT” indicates the number of QA pairs per query type.

		DBGeo Training	*DBGeo* Validation	*DBGeo* 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	-	-
	$\lvert Q^{(2)}_{geo}(\mathcal{G})\rvert$	1,000,000	1000/QT	10000/QT
	$\lvert Q^{(3)}_{geo}(\mathcal{G})\rvert$	1,000,000	1000/QT	10000/QT
Spatial Semantic Lifting	$\lvert\mathcal{T}_s \cap \mathcal{T}_o\rvert$	138,193	1,884	17,152
	$\lvert\mathcal{R}_{ssl}\rvert$	227	71	135

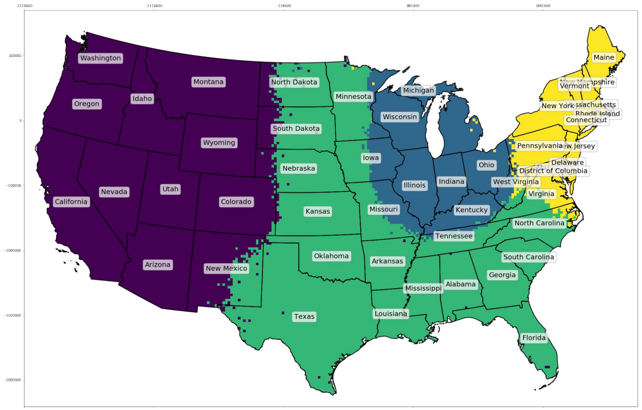


Geographic Question Answering

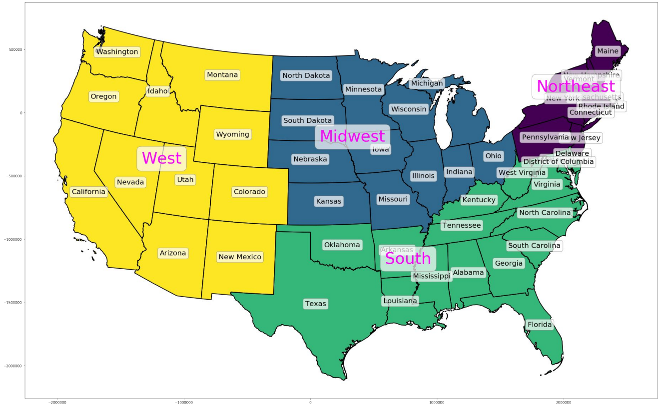
Table 3: The evaluation of geographic logic query answering on *DBGeo* (using AUC (%) and APR (%) as evaluation metric)

	DAG Type	GQE_{diag}		GQE		CGA		$SE\text{-}KGE_{direct}$		$SE\text{-}KGE_{pt}$		$SE\text{-}KGE_{space}$		$SE\text{-}KGE_{full}$	
		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**	84.9	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	60.8	**80.9**	**85.02**
	3-inter	98.01	99.21	96.24	98.17	99.18	**99.62**	**99.28**	99.41	99.1	99.56	87.62	89	99.27	99.59
	Hard-3-inter	78.29	85	68.26	77.55	79.59	86.06	79.5	84.28	**80.48**	**87.4**	63.37	67.17	78.86	85.2
	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	**96.7**	**96.79**
	Hard-3-inter_chain	74.19	**83.79**	70.64	74.54	73.97	76.28	74.81	78.9	**76.45**	75.95	65.54	68.21	76.33	83.7
	3-chain_inter	**98.01**	97.45	92.69	93.31	96.72	97.61	97.31	98.67	97.79	**98.76**	83.7	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	81.4	84.81	85.21	87.41	85.34	87.43	**86.31**	88.61	74.21	73.9	86.22	**89.58**
Test	2-chain	64.88	65.61	85	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	64.5	71.86	72.15	79.26	72.04	79.11	**73.72**	**81.78**	61.22	62.97	72.62	81.04
	3-chain	62.3	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	72.1	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	81.2	84.49	84.88	87.65	85.02	87.81	**86.06**	88.2	74.36	73.7	86.01	**88.96**

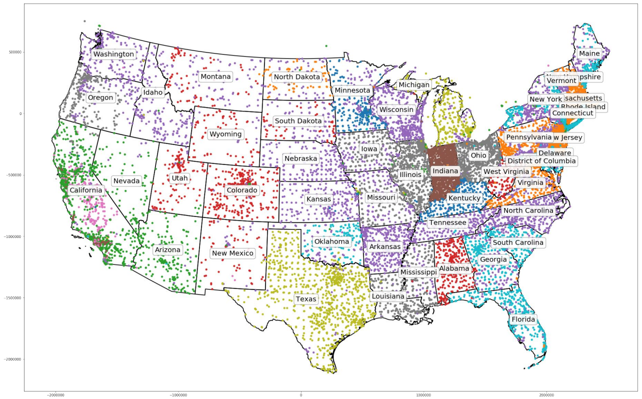
Geographic Question Answering



(a) Clustering result of location embeddings produced by the location encoder in SE-KGE$_{space}$



(b) Census Bureau-designated regions of United States



(c) The community detection (Shuffled Louvain) results of KG

Spatial Semantic Lifting

Table 5: The evaluation of spatial semantic lifting on *DBGeo* over all validation/testing triples

	SE-KGE$_{space}$		*SE-KGE*$_{ssl}$		*SE-KGE*$_{ssl}$ - *SE-KGE*$_{space}$	
	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

Table 6: The evaluation of *SE-KGE*$_{ssl}$ and *SE-KGE*$'_{space}$ on *DBGeo* for a few selected relation r (using APR (%) as evaluation metric).

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

Conclusion

- We develop a spatially-explicit knowledge graph embedding model, SE-KGE, which applies a location encoder to incorporate spatial information (coordinates and spatial extents) of geographic entities.
- SE-KGE is extended as end-to-end models for two tasks: geographic question answering and spatial semantic lifting (a new task).
- Evaluation results show that SE-KGE can outperform multiple baselines on two tasks.
- Visualization shows that SE-KGE can successfully capture the spatial proximity information as well as the semantics of relations.

Future work:

- We want to explore a more concise way to encode the spatial footprints of geographic entities in a KG

Reference

1. **Gengchen Mai**, Krzysztof Janowicz, Ling Cai, Rui Zhu, Blake Regalia, Bo Yan, Meilin Shi, Ni Lao. SE-KGE: A Location-Aware Knowledge Graph Embedding Model for Geographic Question Answering and Spatial Semantic Lifting. *Transactions in GIS*. DOI:10.1111/TGIS.12629 [arxiv paper]
2. **Gengchen Mai**, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, Ni Lao. Multi-Scale Representation Learning for Spatial Feature Distributions using Grid Cells, In: *Proceedings of International Conference on Learning Representations (ICLR) 2020*, Apr. 26 - 30, 2020, Addis Ababa, ETHIOPIA . [OpenReview paper] [arxiv paper] [code] [video] [slides] *** Spotlight Paper (Acceptance Rate 6%, 156 out of 2594 submissions)**
3. **Gengchen Mai**, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, Ni Lao. Contextual Graph Attention for Answering Logical Queries over Incomplete Knowledge Graphs, In: *Proceedings of K-CAP 2019*, Nov. 19 - 21, 2019, Marina del Rey, CA, USA. [arxiv]
4. **Gengchen Mai**, Bo Yan, Krzysztof Janowicz, Rui Zhu. Relaxing Unanswerable Geographic Questions Using A Spatially Explicit Knowledge Graph Embedding Model, In: *Proceedings of AGILE 2019*, June 17 - 20, 2019, Limassol, Cyprus. *** 1st Best Full Paper Award**
5. Bo Yan, Krzysztof Janowicz, **Gengchen Mai**, Rui Zhu. A Spatially-Explicit Reinforcement Learning Model for Geographic Knowledge Graph Summarization. *Transactions in GIS*, 23(2019), 620-640. DOI:10.1111/tgis.12547
6. Will Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. Embedding logical queries on knowledge graphs. In Advances in Neural Information Processing Systems, pp. 2026-2037. 2018.
7. Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge Graph Embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering* 29, no. 12 (2017): 2724-2743.