Relaxing Unanswerable Geographic Questions Using A Spatially Explicit Knowledge Graph Embedding Model

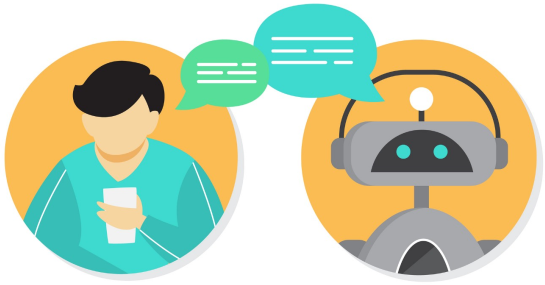
AAG 2019, April 4th 2019

Gengchen Mai Bo Yan Krzysztof Janowicz Rui Zhu

STKO Lab, University of California, Santa Barbara

Introduction

- **Question Answering** (QA): In the field of NLP, **QA** refers to the methods, processes, and systems which allow users to ask questions in the form of **natural language sentences** and receive **one or more answers**, often in the form of sentences.
- **Examples**: Apple Siri and Amazon Alexa.
- Although QA systems have been studied and developed for a long time, **geographic question answering** remained nearly untouched.



Introduction

Geographic questions are fundamentally different from other questions in several ways.

- Many geographic questions are highly **context-dependent** and **subjective**.
- The answers are typically derived from **a sequence of spatial operations** rather than extracted from a piece of unstructured text or retrieved from Knowledge Graphs (KG).
- Geographic questions are often affected by **vagueness** and **uncertainty** at the conceptual level.

Introduction

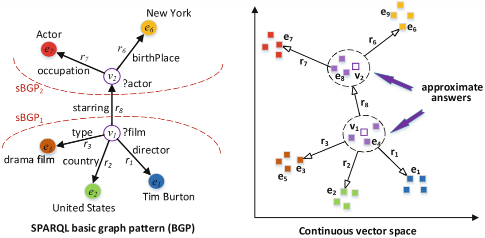
- Due to the above reasons, it is likely to receive **no answer** given a geographic question.
- **Query relaxation and rewriting** for **unanwerable questions** in general QA
- **Hypothesis**: **Geographic questions** will benefit from **spatially-explicit relaxation methods** in which the **spatial adjacency** is taken into account.

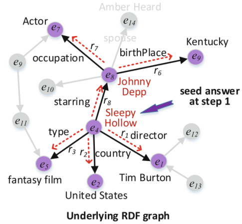
Handling Unanswerable Questions

- **Reason** for unanswerable Questions:
- **Missing Information** from the current KB:
  - **Question A**: *what is the weather like in Creston, California?*
  - **Option A**: **Go up the place hierarchy**. *what is the weather like in San Luis Obispo County?*
  - **Option B**: **Go to sibling nodes**. *what is the weather like in San Luis Obispo (City)?*
- **Logical inconsistencies**:
  - **Question B**: *which city spans Texas and Colorado*
  - **Delete one of the contradictory conditions**. *which city locates in Texas?*
- **Problem**: current relaxation/rewriting techniques **do not consider spatial adjacency** when handling unanswerable questions, and, thus, often return **surprising and counter-intuitive results**.

Introduction

- Get American drama films directed by Tim Burton one of whose star actors was born in New York





M. Wang et al. Towards Empty Answers in SPARQL:Approximating Querying with RDF Embedding. ISWC 2018

Our Contribution

- We propose a **spatially explicit knowledge graph embedding model, TransGeo,** which explicitly models the distance decay effect.
- This spatially explicit embedding model is utilized to **relax/rewrite unanswerable geographic queries**.
- We present a benchmark dataset to evaluate the performance of the unanswerable geographic question handling framework. The evaluation results show that **our spatially explicit embedding model outperforms non-spatial models**.

Workflow

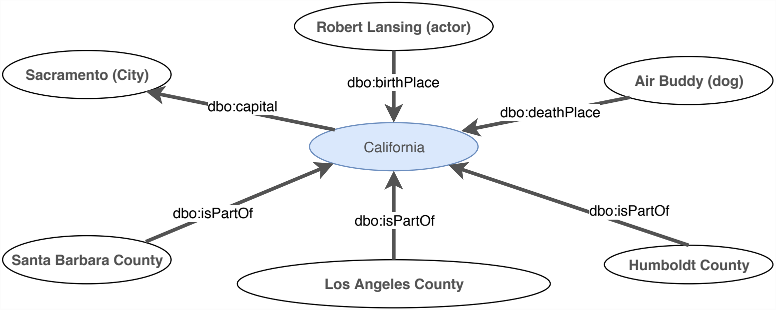
Given an unanswerable SPARQL query Q_j, our goal is:

- Learn a **spatially explicit KG embedding model** for the current KG which takes **distance decay** into account;
- Use the embedding model to **infer a ranked list of approximated answers** to this question;
- Generate a **relaxed/related SPARQL query for each approximate answer** as an explanation for the query relaxation/rewriting process.

Traditional Entity Context Modeling in Semantic Web

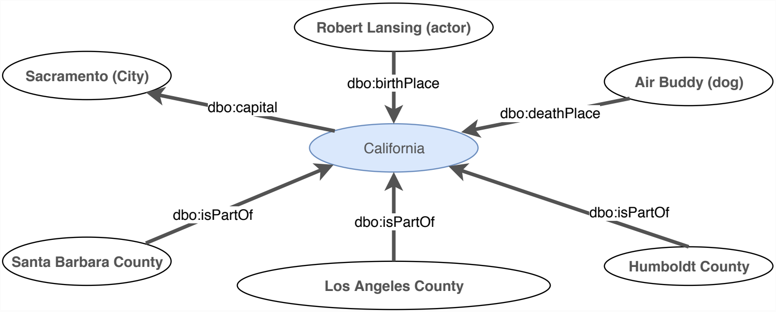
Definition

Entity Context: Given an entity $e \in E$ in the knowledge graph G, the context of e is defined as $C(e) = \{(r_c, e_c) | (e, r_c, e_c) \in G \vee (e_c, r_c, e) \in G\}$.



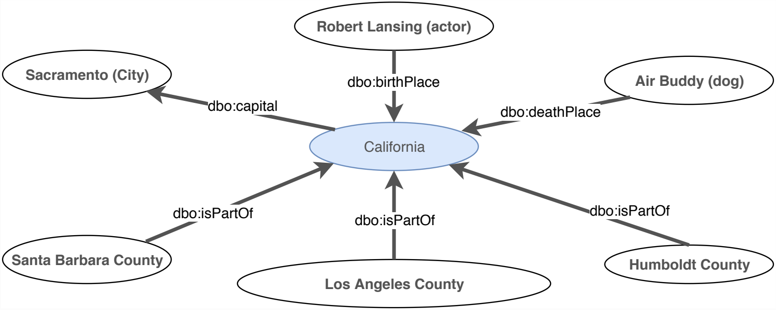
Modeling Geographic Entity Context in Knowledge Graphs

- **The traditional entity context modeling method**: Each 1-d triple in the entity context has **equal weight**.
- This method falls apart when geographic entities are considered in two ways:
  - Does not fully reflect **Tobler's first law of geography**
  - The **canonical predicates** used in the place hierarchy



Modeling Geographic Entity Context in Knowledge Graphs

Major Idea: Assign **larger weights** to **geographic triples** in an entity context where the weights are modeled from **a distance decay function**



Spatially Explicit KG Embedding Model

Learning a **Spatially Explicit KG Embedding Model**: Given a KG $G = \langle E, R \rangle$, a set of geographic entities $P \subseteq E$, and a triple $T_i = (h_i, r_i, t_i) \in G$.

$$w(T_i) = \begin{cases} \max(\ln \frac{D}{dis(h_i,t_i)+\varepsilon}, l) & \text{if } h_i \in P \land t_i \in P \\ l & \text{otherwise} \end{cases} \tag{1}$$

- $dis(h_i, t_i)$ is the geodesic distance between geographic entity h_i and t_i;
- l is the lowest edge weight we allow for each triple;
- D is the longest (simplified) earth surface distance;
- ε is a hyperparameter.

Spatially Explicit KG Embedding Model

- The Knowledge Graph $G = \langle E, R \rangle$ becomes a **weighted multigraph** (MG)
- **An edge-weighted PageRank** is applied to MG.
- Compute **entity score** $w(e_i)$ based on PageRank result
- $w(e_i)$ encodes the **structural information of the original KG** and the **distance decay effect on interaction** among geographic entities.

$$w(e_i) = N \cdot \frac{\frac{1}{-\ln PR(e_i)}}{\sum_i \frac{1}{-\ln PR(e_i)}} \tag{2}$$

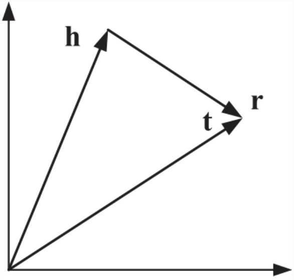
  - $PR(e_i)$ be the PageRank score for each entity e_i in the knowledge graph;
  - N is the number of entities in G

Spatially Explicit KG Embedding Model

- **Spatially Explicit KG Embedding Model**: Translation-based KG embedding model based on **TransE** which utilizes $w(e_i)$
- **TransE** embeds entities into **low-dimensional vector spaces** while relations are treated as **translation operations** in the entity embedding space
- Let $(h, r, t) \in G$ is a triple:

$$f_r(h, t) = \| \mathbf{h} + \mathbf{r} - \mathbf{t} \| \quad (3)$$

- In a perfect situation, if $(h, r, t) \in G$, $\| \mathbf{h} + \mathbf{r} - \mathbf{t} \| = 0$
- The original **TransE** is not an **entity context preserving model** which is required by **query relaxation/rewriting** process.



(a) TransE.

Spatially Explicit KG Embedding Model

- For each entity e_i in G, we sample an entity context $C_{samp}(e_i) \subseteq C(e_i)$ where the **sampling probability** $P(r_{ci}, e_{ci})$ of each context item $(r_{ci}, e_{ci}) \in C(e_i)$ is based on **entity score** $w(e_{ci})$

$$P(r_{ci}, e_{ci}) = \frac{w(e_{ci})}{\sum_{(r_{cj}, e_{cj}) \in C(e_i)} w(e_{cj})}, \text{ where } (e_i, r_{ci}, e_{ci}) \in G \vee (e_{ci}, r_{ci}, e_i) \in G \tag{4}$$

- A **compatibility score** between $C_{samp}(e_i)$ and an arbitrary entity e_k can be computed:

$$f(e_k, C_{samp}(e_i)) = \frac{1}{|C_{samp}(e_i)|} \cdot \sum_{(r_{cj}, e_{cj}) \in C_{samp}(e_i)} \phi(e_k, r_{cj}, e_{cj}) \tag{5}$$

$$\phi(e_k, r_{cj}, e_{cj}) = \begin{cases} \|\mathbf{e_k} + \mathbf{r_{cj}} - \mathbf{e_{cj}}\| & \text{if } (e_i, r_{cj}, e_{cj}) \in G \\ \|\mathbf{e_{cj}} + \mathbf{r_{cj}} - \mathbf{e_k}\| & \text{if } (e_{cj}, r_{cj}, e_i) \in G \end{cases} \tag{6}$$

Spatially Explicit KG Embedding Model

- **Pairwise ranking loss function**:

$$\mathcal{L} = \sum_{e_i \in G} \sum_{e'_i \in Neg(e_i)} max\Big(\gamma + f(e_i, C_{samp}(e_i)) - f(e'_i, C_{samp}(e_i)), 0\Big) \tag{7}$$

- For each entity e_i, we randomly sample K entities as the negative sampling set $Neg(e_i)$ for e_i

KG Embedding Model Based Query Relaxation and Rewriting

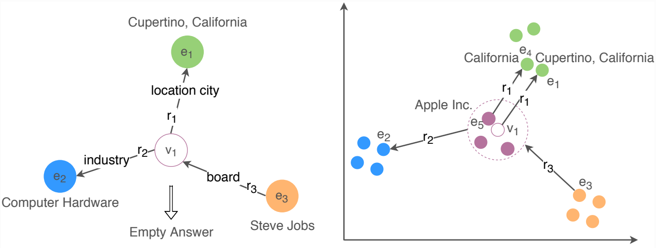
- *In which computer hardware company located in Cupertino is/was Steve Jobs a board member*

```
SELECT ?v
WHERE {
?v dbo:locationCity dbr:Cupertino,_California .
?v dbo:industry dbr:Computer_hardware .
dbr:Steve_Jobs dbo:board ?v .}
```

Listing 1: An example SPARQL query generated by a semantic parser.

KG Embedding Model Based Query Relaxation and Rewriting

- *In which computer hardware company located in Cupertino is/was Steve Jobs a board member*



Incomplete knowledge graph yields an empty answer — Relaxation in the embedding space

- Use $\mathbf{v_i} = \mathbf{e_i} + \mathbf{r_i}$ to **predict** variable embedding $\mathbf{v_i}$ from each triple path;
- Computed the final variable embedding $\mathbf{v}$ as weighted average of $\mathbf{v_i}$;
- Use **nearest neighbor search** in entity embedding space to get the **approximate answer**;
- Use the **approximate answer** to **relax** the original query.

DB18 DataSet

- We collect a new KG embedding training dataset, *DB18*[1], which is a subgraph of DBpedia.

Summary statistic for *DB18*

DB18	Total	Training	Testing
# of triples	139155	138155	1000
# of entities	22061	-	-
# of relations	281	-	-
# of geographic entities	1681 (7.62%)	-	-

[1] https://github.com/gengchenmai/TransGeo

GeoUQ DataSet

- We construct an evaluation dataset, *GeoUQ*, which is composed of **20 unanswerable geographic questions** base on *DB18*.
- These queries satisfy 2 conditions:
  - each query *Q* will yield **empty answer set** when executing *Q* on **training KG**;
  - *Q* will return **only one answer** when executing *Q* on **the whole KG**.

Evaluation

- **Link Prediction Task**: Given h, r, Predict the correct t against the negative samples.
- **Answer prediction by relaxation/rewriting task**: The rank of the correct answer in the predicted answer ranking list

Two evaluation tasks for different KG embedding models

	Link Prediction				SPARQL Relaxation	
	MRR		HIT@10		MRR	HIT@10
	Raw	Filter	Raw	Filter		
TransE Model	**0.122**	0.149	30.00%	34.00%	0.008	5% (1 out of 20)
Wang et al. (2018)	0.113	0.154	27.20%	30.50%	0.000	0% (0 out of 20)
$TransGeo_{regular}$	0.094	0.129	28.50%	33.40%	0.098	25% (5 out of 20)
$TransGeo_{unweighted}$	0.108	0.152	30.80%	37.80%	0.043	15% (3 out of 20)
TransGeo	0.104	**0.159**	**32.40%**	**42.10%**	**0.109**	**30% (6 out of 20)**

Query Relaxation Example

Original SPARQL Query:

Query:

SELECT ?v
WHERE {
?v dbo:locationCity **dbr:Cupertino, _California** .
?v dbo:industry dbr:Computer_hardware .
dbr:Steve_Jobs **dbo:board** ?v .
}

Answer: dbr:Apple_Inc

Relaxed Query by TransGeo:

Query:

SELECT ?v
WHERE {
?v dbo:locationCity **dbr:California** .
?v dbo:industry dbr:Computer_hardware .
dbr:Steve_Jobs **dbo:board** ?v .
}

Answer: dbr:Apple_Inc

Conclusion

- We propose a **spatially explicit KG embedding models**, TransGeo, which include the **distance decay effect** into the KG embedding model training.
- We show how to use **TransGeo** to do **spatially explicit query relaxation**.
- The evaluation of two evaluation tasks - **link prediction** and **answer prediction by relaxation/rewriting** - shows that our spatially explicit embedding model, TransGeo, can **outperform** all the other 4 baseline methods on both tasks

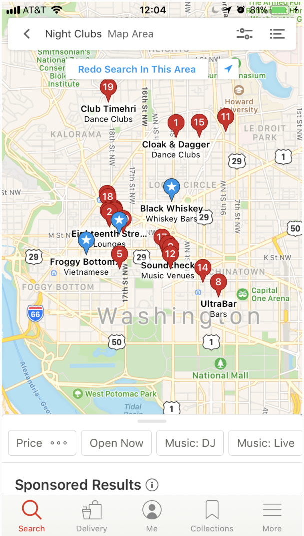
Future Work

- We want to explore ways to **only consider distance decay during query relaxation** rather than the model training step.
- We used point geometries to compute distance between geographic entities. In the future, **complex geometries** and **topology** should be considered.

Reference: **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.

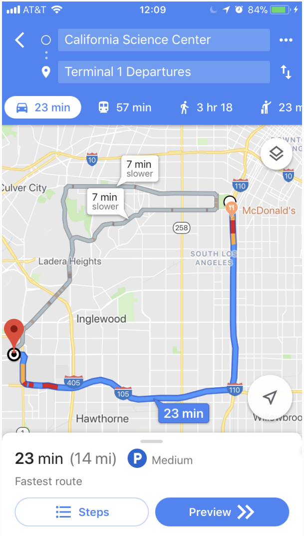
Uniqueness of Geographic Questions

- Many geographic questions are highly **context-dependent** and **subjective**.
- The answers to many geographic questions vary according to **when** and **where** these questions are asked, and **who** asks them.
  - What is the location of the California Science Center? (**context independently**)
  - V.S.
  - Nightclubs near me that are 18+ (**location-dependent**)
  - How expensive is a ride from Stanford University to Googleplex? (**time-dependent**)
  - How safe is Isla Vista? (**subjective**)



Uniqueness of Geographic Questions

- The answers are typically derived from **a sequence of spatial operations** rather than extracted from a piece of unstructured text or retrieved from Knowledge Graphs (KG).
  - What is the shortest route from California Science Center to LAX?
  - A shortest path algorithm on a route dataset rather than searching in a text corpus



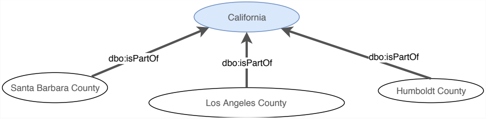
Uniqueness of Geographic Questions

- Geographic questions are often affected by **vagueness** and **uncertainty** at the conceptual level.
  - How many lakes are there in Michigan?
  - The answer can vary between 63,000 and 10 depending on the conceptualization of Lake

Modeling Geographic Entity Context in Knowledge Graphs

Tobler's first law of geography:

- **Place hierarchy** is far too coarse to model **distance decay**.



Modeling Geographic Entity Context in Knowledge Graphs

Canonical predicates used in the place hierarchy:

- For any given populated place, even if **no other triples are known about a small settlement**, the KG will still contain **at least a triple about a higher-order unit** the place belongs to, e.g., a county
- Example:
  - all populated places in **Coconino County, Arizona** are parts of `dbr:Coconino_County,_Arizona`
  - **Tiny deserted settlements**: nearly 100% of all triples about **Two Guns, AZ**
  - **Major cities**: a small percentage of all triples about **Flagstaff**
- This will result in **places about which not much is known to have an artificially increased similarity**.