Regex Queries over Incomplete Knowledge Bases

Vaibhav Adlakha, Parth Shah, Srikanta Bedathur, Mausam

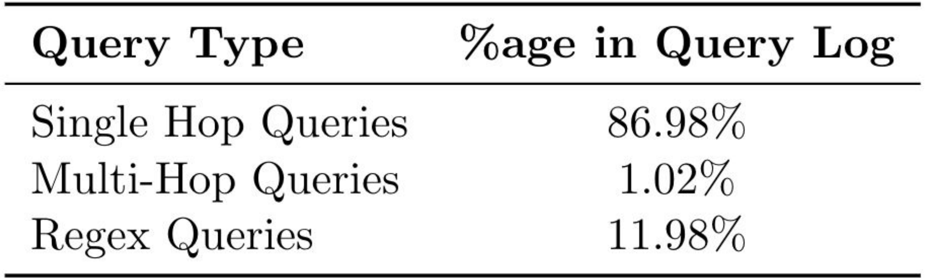
Indian Institute of Technology, Delhi

Types of Knowledge Base Queries

- Single-hop queries
  - Who founded Microsoft?
- Multi-hop queries
  - Where do founders of Apple live?
- First-order logic queries
  - Where did Canadian citizens with Turing Award graduate?

Regex Queries over Knowledge Base

WIKIDATA
SPARQL



Query Type	%age in Query Log
Single Hop Queries	86.98%
Multi-Hop Queries	1.02%
Regex Queries	11.98%

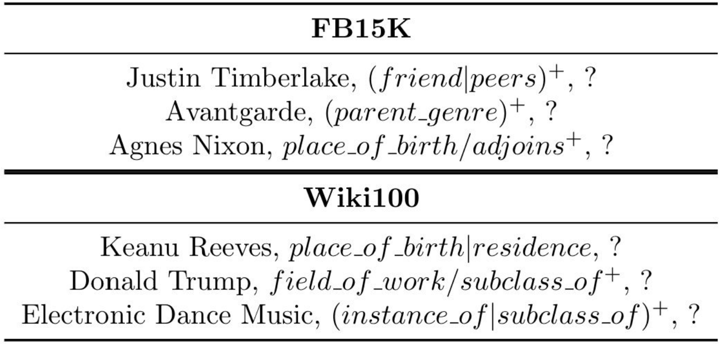
Table 1: User queries in Wikidata logs

Regex queries are characterized by **Kleene plus (+)** and **Disjunction (V)** operators

Datasets for Regex Queries

- **Wiki100-Regex**
  - Queries harvested from **actual query logs**
  - 5 unique query types

- **FB15K-Regex**
  - Queries formed by aggregating random walks
  - 21 unique query types



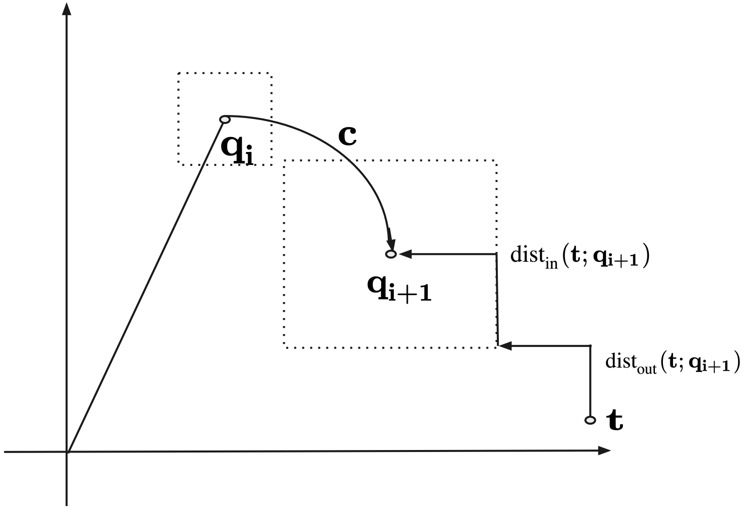
FB15K
Justin Timberlake, $(friend \mid peers)^+$, ?
Avantgarde, $(parent\_genre)^+$, ?
Agnes Nixon, $place\_of\_birth/adjoins^+$, ?
Wiki100
Keanu Reeves, $place\_of\_birth \mid residence$, ?
Donald Trump, $field\_of\_work/subclass\_of^+$, ?
Electronic Dance Music, $(instance\_of \mid subclass\_of)^+$, ?

Table 2: Example queries from FB15K and Wiki100

RotatE-Box

Based on:

- RotatE (Sun et al. 2019)
- Query2Box (Ren et al. 2020)


$\mathbf{q_i}$
$\mathbf{c}$
$\mathbf{q_{i+1}}$
$\text{dist}_{\text{in}}(\mathbf{t};\mathbf{q_{i+1}})$
$\text{dist}_{\text{out}}(\mathbf{t};\mathbf{q_{i+1}})$
$\mathbf{t}$

Handling Regex Operators – Kleene Plus

- **Projection**

  $kp(\mathbf{c}) = \mathbf{c}' = (e^{i\boldsymbol{\theta}_{c'}}, \mathbf{K}_{\text{off}}\text{Off}(\mathbf{c}))$, where $\boldsymbol{\theta}_{c'} = \mathbf{K}_{\text{cen}}\boldsymbol{\theta}_{c}$

- **Free parameter**

  $\mathbf{r}^{+}$ embedding for each relation *r*

Handling Regex Operators – Disjunction

- **Aggregation**

  Minimum distance to the closest query box

  $$\mathrm{dist}(\mathbf{e};\mathbf{q}) = \mathrm{Min}(\{\mathrm{dist}(\mathbf{e};\mathbf{q}_1), \mathrm{dist}(\mathbf{e};\mathbf{q}_2), \ldots, \mathrm{dist}(\mathbf{e};\mathbf{q}_N)\})$$

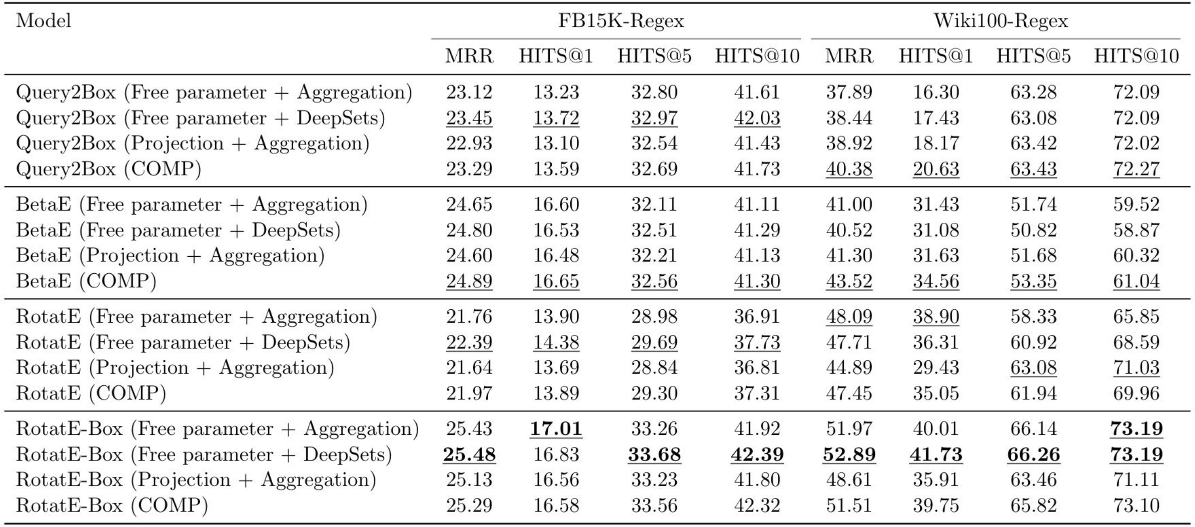
- **DeepSets (Zaheer et al. 2017)**

  Learnable permutation-invariant functions

  $$\begin{aligned}\boldsymbol{\theta}_c &= \mathbf{W}_{\mathrm{cen}} \cdot \Psi(\mathrm{MLP}_{\mathrm{cen}}(\boldsymbol{\theta}_{c_1}), \mathrm{MLP}_{\mathrm{cen}}(\boldsymbol{\theta}_{c_2}), \ldots, \mathrm{MLP}_{\mathrm{cen}}(\boldsymbol{\theta}_{c_N}))\\ \mathrm{Off}(\mathbf{c}) &= \mathbf{W}_{\mathrm{off}} \cdot \Psi(\mathrm{MLP}_{\mathrm{off}}(\mathrm{Off}(\mathbf{c}_1)), \mathrm{MLP}_{\mathrm{off}}(\mathrm{Off}(\mathbf{c}_2)), \ldots, \mathrm{MLP}_{\mathrm{off}}(\mathrm{Off}(\mathbf{c}_N)))\end{aligned}$$

Results

RotatE-box variants outperform other models



Model	FB15K-Regex				Wiki100-Regex			
	MRR	HITS@1	HITS@5	HITS@10	MRR	HITS@1	HITS@5	HITS@10
Query2Box (Free parameter + Aggregation)	23.12	13.23	32.80	41.61	37.89	16.30	63.28	72.09
Query2Box (Free parameter + DeepSets)	23.45	13.72	32.97	42.03	38.44	17.43	63.08	72.09
Query2Box (Projection + Aggregation)	22.93	13.10	32.54	41.43	38.92	18.17	63.42	72.02
Query2Box (COMP)	23.29	13.59	32.69	41.73	40.38	20.63	63.43	72.27
BetaE (Free parameter + Aggregation)	24.65	16.60	32.11	41.11	41.00	31.43	51.74	59.52
BetaE (Free parameter + DeepSets)	24.80	16.53	32.51	41.29	40.52	31.08	50.82	58.87
BetaE (Projection + Aggregation)	24.60	16.48	32.21	41.13	41.30	31.63	51.68	60.32
BetaE (COMP)	24.89	16.65	32.56	41.30	43.52	34.56	53.35	61.04
RotatE (Free parameter + Aggregation)	21.76	13.90	28.98	36.91	48.09	38.90	58.33	65.85
RotatE (Free parameter + DeepSets)	22.39	14.38	29.69	37.73	47.71	36.31	60.92	68.59
RotatE (Projection + Aggregation)	21.64	13.69	28.84	36.81	44.89	29.43	63.08	71.03
RotatE (COMP)	21.97	13.89	29.30	37.31	47.45	35.05	61.94	69.96
RotatE-Box (Free parameter + Aggregation)	25.43	**17.01**	33.26	41.92	51.97	40.01	66.14	**73.19**
RotatE-Box (Free parameter + DeepSets)	**25.48**	16.83	**33.68**	**42.39**	**52.89**	**41.73**	**66.26**	**73.19**
RotatE-Box (Projection + Aggregation)	25.13	16.56	33.23	41.80	48.61	35.91	63.46	71.11
RotatE-Box (COMP)	25.29	16.58	33.56	42.32	51.51	39.75	65.82	73.10

Table 6: Performance on subset of regex query types answerable by all variants. Best overall score is in bold. Best score amongst variants of the same model is underlined.

Modeling Challenges – Kleene Plus

- Kleene Plus is an idempotent unary operator

  $$(r^+)^+ = r^+$$

- Kleene plus is an infinite union of path queries

  $$r^+ = r \vee (r/r) \vee (r/r/r) \ldots$$