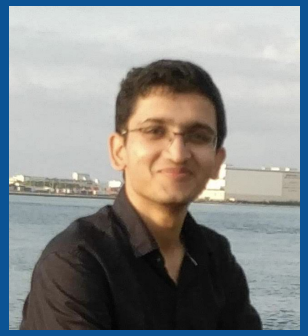
A Re-evaluation of Knowledge Graph Completion Methods

Zhiqing Sun[1*]
zhiqings@cs.cmu.edu



Shikhar Vashishth[1,2*]
svashish@cs.cmu.edu

Soumya Sanyal[2*]
soumyasanyal@iisc.ac.in

Partha Talukdar[2]
ppt@iisc.ac.in

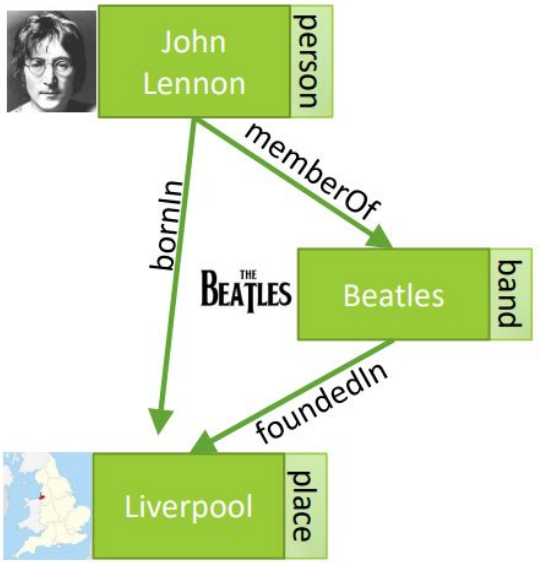
Yiming Yang[1]
yiming@cs.cmu.edu

[1]Carnegie Mellon University, [2]Indian Institute of Science

Language Technologies Institute

Knowledge Graphs

- Knowledge in graph form
- Nodes represent entities
- Edges represent relationships
- Examples: Freebase, Wikidata …
- **Use cases:**
  - Question Answering
  - Dialog systems
  - Web Search



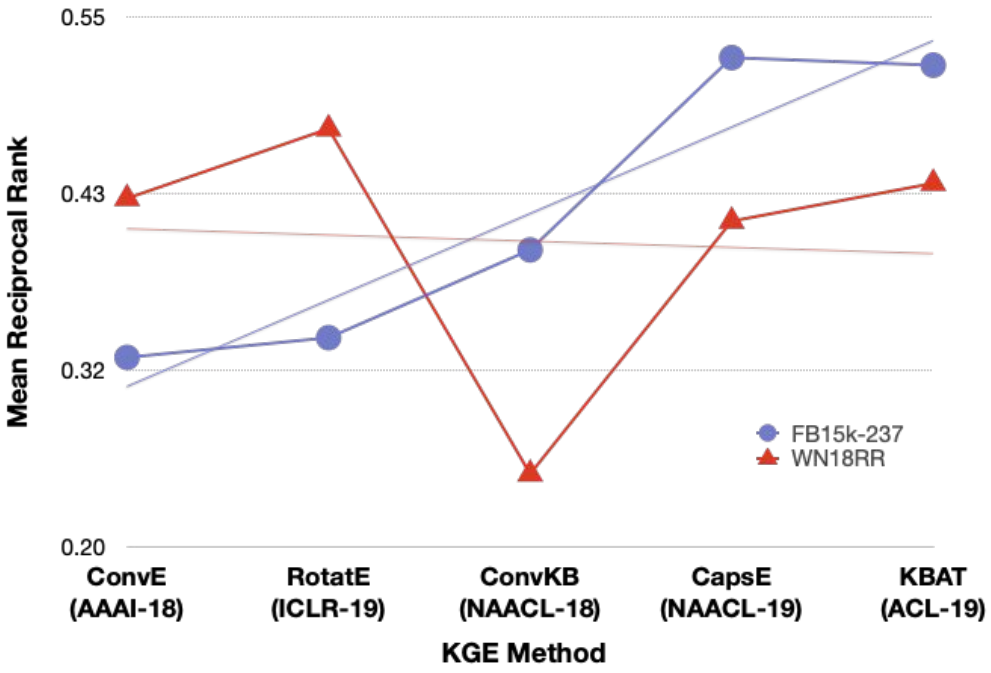
Link Prediction (KG Completion)

- **Definition:**
  Task of inferring missing facts based on known ones.

- **General technique** involves learning a representation for all entities and relations in KG.

Why Re-evaluation?

- Calming SOTA hype in Knowledge Graph Link Prediction
- Recently, a large number of papers have reported inconsistent high-performance gains



Contributions

- Identify that inflated performance is because of inappropriate evaluation protocol
- Propose RANDOM, a novel evaluation protocol which addresses the concern and detects inflated performance
- Perform extensive re-examination on recent neural network based KGC techniques

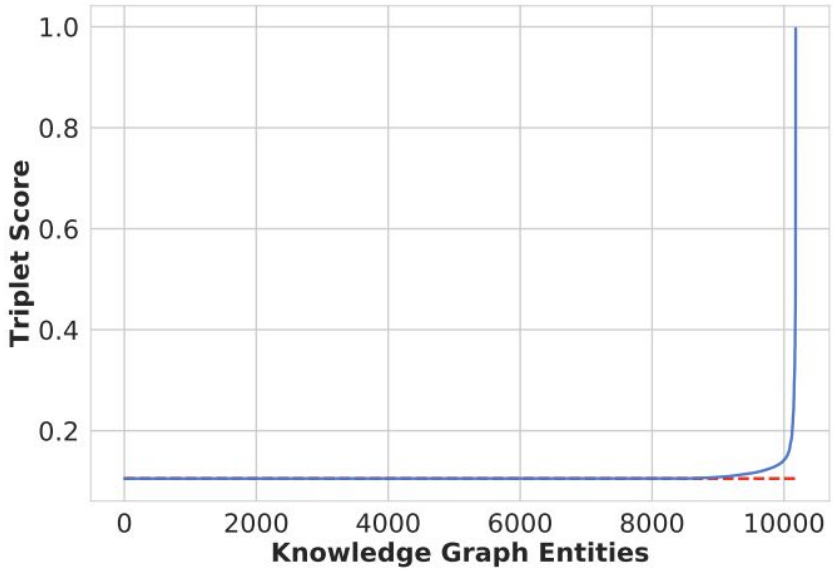
KGC Evaluation

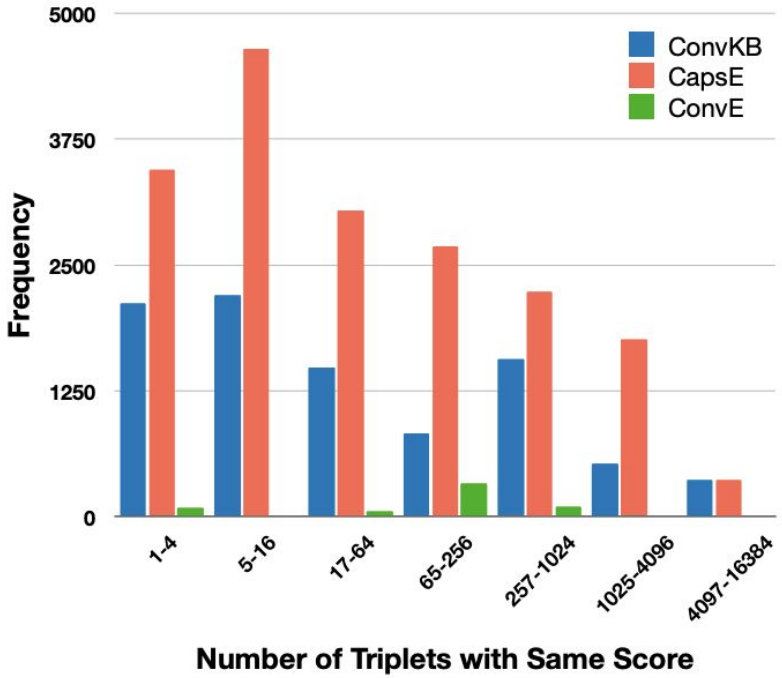
- For a triple *(h, r, t)* in the evaluation set:
  - We predict *t* given *(h, r)* by scoring all $T' = \{(h, r, t') \mid t' \in E\}$
  - Triplets are sorted based on the score, and rank of the valid triplet (h,r,t) is used as an evaluation metric.
  - Similarly, we predict h given (r, t). Report average across both

- Filtered Setting (Bordes et al., 2013)
  - All known correct triplets are removed from *T'* except one being evaluated.

Issues with Existing Methods

- 58.5% negative sampled triplets obtain the exact same score as the valid triplet with ConvKB on FB15k-237.
- On average, ConvKB and CapsE have 125 and 197 such entities, whereas ConvE has around 0.002 over the entire evaluation dataset of FB15k-237



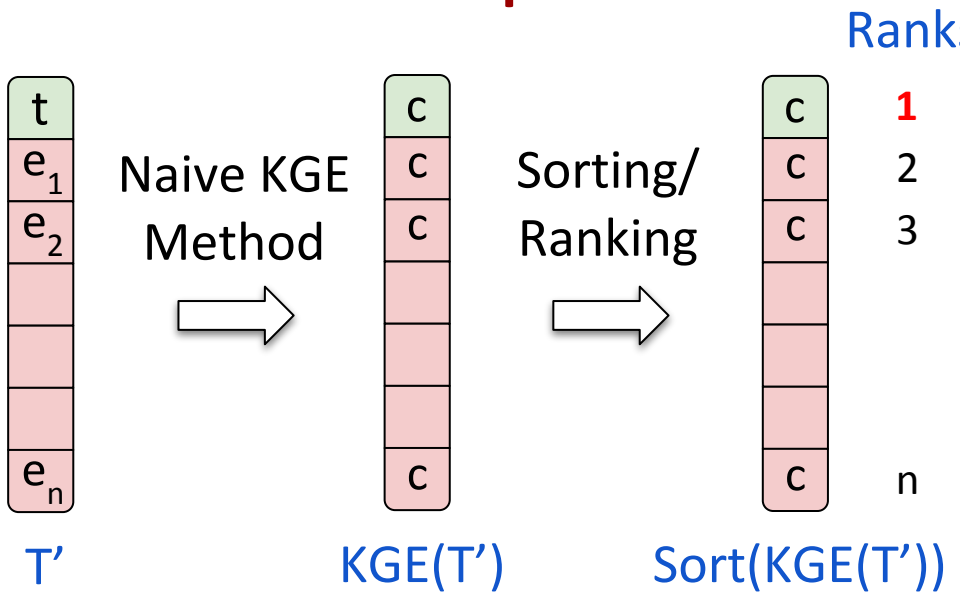


Current Evaluation Protocol (TOP)

- Place the valid entity at the beginning among all entities with the same score ‘c’.
- **Problem:** A naive baseline which gives an identical score (f(h,r,t) = c, ∀ h,r,t) to all triples **achieves 100% performance**.

Given: (h,r)
Predict: t

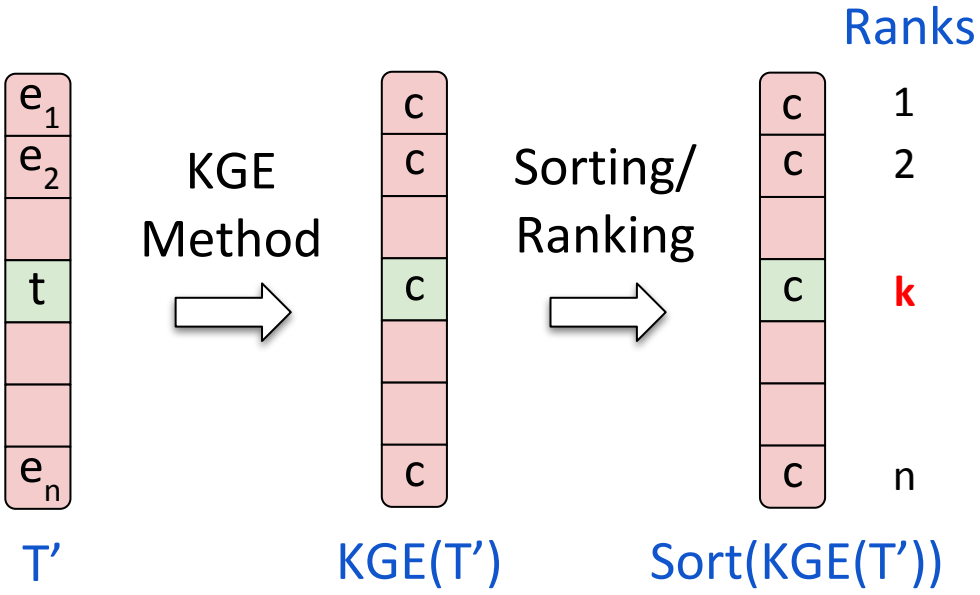
$T' = \{(h, r, t') \mid t' \in E\}$



Proposed Evaluation Protocol

- RANDOM:

  Place the valid entity at a random position among entities with the same score ‘c’



RANDOM protocol eliminates the bias in evaluation.

Results

- We observe a drastic change in results on FB15k-237 on switching the evaluation protocol
- BOTTOM: Place valid entity at the end among entities with the same score

	Reported			Random			Top			Bottom		
	MRR ↑	MR ↓	H@10 ↑	MRR ↑	MR ↓	H@10 ↑	MRR ↑	MR ↓	H@10 ↑	MRR ↑	MR ↓	H@10 ↑
ConvE	.325	244	.501	.324 ± .0	285 ± 0	.501 ± .0	.324	285	.501	.324	285	.501
RotatE	.338	177	.533	.336 ± .0	178 ± 0	.530 ± .0	.336	178	.530	.336	178	.530
TuckER	.358	-	.544	.353 ± .0	162 ± 0	.536 ± .0	.353	162	.536	.353	162	.536
ConvKB	.396	257	.517	.243 ± .0	309 ± 2	.421 ± .0	.407 (+.164)	246 (-63)	.527 (+.106)	.130 (-.113)	373 (+64)	.383 (-.038)
CapsE	.523	303	.593	.150 ± .0	403 ± 2	.356 ± .0	.511 (+.361)	305 (-99)	.586 (+.229)	.134 (-.016)	502 (+99)	.297 (-.059)
KBAT	.518†	210†	.626†	.157 ± .0	270 ± 0	.331 ± .0	.157	270	.331	.157	270	.331

Conclusion

- Along with making progress on KG embedding techniques, it is equally important to use the right evaluation.
- Experimentally demonstrate that many recent KGE methods suffer from using biased evaluation protocols.
- Strongly recommend using the RANDOM evaluation strategy for evaluating the task of Link Prediction.

Paper Link:
A Re-evaluation of Knowledge Graph Completion Methods

Thank you!

Research Supported by:

Source Code:

github.com/svjan5/kg-reeval