A Re-evaluation of Knowledge Graph Completion Methods

Zhiqing Sun\textsuperscript{1*}  
zhiqings@cs.cmu.edu

Shikhar Vashishth\textsuperscript{1,2*}  
svashish@cs.cmu.edu

Soumya Sanyal\textsuperscript{2*}  
soumyasanyal@iisc.ac.in

Partha Talukdar\textsuperscript{2}  
ppt@iisc.ac.in

Soumya Sanyal\textsuperscript{2*}  
soumyasanyal@iisc.ac.in

Yiming Yang\textsuperscript{1}  
yiming@cs.cmu.edu

\textsuperscript{1}Carnegie Mellon University,  \textsuperscript{2}Indian Institute of Science
Knowledge Graphs

- **Knowledge** in graph form
- **Nodes** represent entities
- **Edges** represent relationships
- Examples: Freebase, Wikidata ...

**Use cases:**
- Question Answering
- Dialog systems
- Web Search
**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

![Graph showing performance of different KGE methods](image-url)
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 \(T' = \{(h, r, t') | t' \in E\}\)
  - 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 \( f(h,r,t) = c, \forall h,r,t \) to all triples achieves 100% performance.

Given: \((h,r)\)

Predict: \(t\)

\[ T' = \{(h, r, t') | t' \in E\} \]

- **Naive KGE Method**
- **Sorting/Ranking**
- **Ranks**

\[ \begin{align*}
\text{Ranks} & \quad 1 \quad 2 \quad 3 \\
1 & \quad 2 & \quad 3 \\
n & \quad \text{n}
\end{align*} \]
**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.
- 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

<table>
<thead>
<tr>
<th></th>
<th>Reported</th>
<th>RANDOM</th>
<th>Top</th>
<th>BOTTOM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR ↑</td>
<td>MR ↓</td>
<td>H@10 ↑</td>
<td>MRR ↑</td>
</tr>
<tr>
<td>ConvE</td>
<td>.325</td>
<td>244</td>
<td>.501</td>
<td>.324 ± .0</td>
</tr>
<tr>
<td>RotatE</td>
<td>.338</td>
<td>177</td>
<td>.533</td>
<td>.336 ± .0</td>
</tr>
<tr>
<td>TuckER</td>
<td>.358</td>
<td>-</td>
<td>.544</td>
<td>.353 ± .0</td>
</tr>
<tr>
<td>ConvKB</td>
<td>.396</td>
<td>257</td>
<td>.517</td>
<td>.243 ± .0</td>
</tr>
<tr>
<td></td>
<td>(+.164)</td>
<td>(-63)</td>
<td>(+.106)</td>
<td>.130</td>
</tr>
<tr>
<td>CapsE</td>
<td>.523</td>
<td>303</td>
<td>.593</td>
<td>.150 ± .0</td>
</tr>
<tr>
<td></td>
<td>(+.361)</td>
<td>(-99)</td>
<td>(+.229)</td>
<td>.134</td>
</tr>
<tr>
<td>KBAT</td>
<td>.518†</td>
<td>210†</td>
<td>.626†</td>
<td>.157 ± .0</td>
</tr>
</tbody>
</table>
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