Behavioral Testing of Knowledge Graph Embedding Models

Wiem Ben Rim, Carolin Lawrence, Kiril Gashteovski, Mathias Niepert, Naoaki Okazaki
Link Prediction Evaluation

Returns averaged evaluation metrics calculated over a given test set.

Test set
FB15K-237

KGE model

Averaged evaluation metrics:
- Mean Rank: 177
- Mean Reciprocal Rank: 0.338
- Hits@1: 0.241
- Hits@3: 0.375
- Hits@10: 0.533

Results of RotatE model (Zhiqing Sun, et al, 2019)

Typically, if $\text{MRR}_{\text{Model A}} > \text{MRR}_{\text{Model B}}$, then Model A is better than B.
Motivation

Link Prediction Evaluation

What is the model **good** at?

What is the model **bad** at?
Motivation
Behavioral Testing

Tests different capabilities of a system without any knowledge of the internal structure (Beizer, 1995).

- **input 1** → expected output 1
- **input 2** → expected output 2
- **input 3** → unexpected behavior
Proposed Method
Behavioral Testing for KGE models
Proposed Method
Behavioral Testing for KGE models

For KGEs, we can target several different capabilities through different tests. We take the example of symmetric relations:

- Good at memorization
- Bad at recognizing symmetry
A model can have a low average score but be the best at generalizing on symmetry.

**KGEval**: Understand the strengths and weaknesses of your link prediction models, source code at: [https://github.com/nec-research/KGEval](https://github.com/nec-research/KGEval)