Knowledge base embedding

Daniel Trizna

What is knowledge base embedding?

- Knowledge Graph representation of knowledge base
- n-dimensional vector space representation
  - each entity e is associated with a vector $e \in R^n$
  - each relation name r is associated with a scoring function $s_r : R^n \times R^n \rightarrow R$

Why knowledge base embedding?

- Find underlying connections in our dataset
- Can be used for explainable malware detection
- Interesting field of science

Different approaches

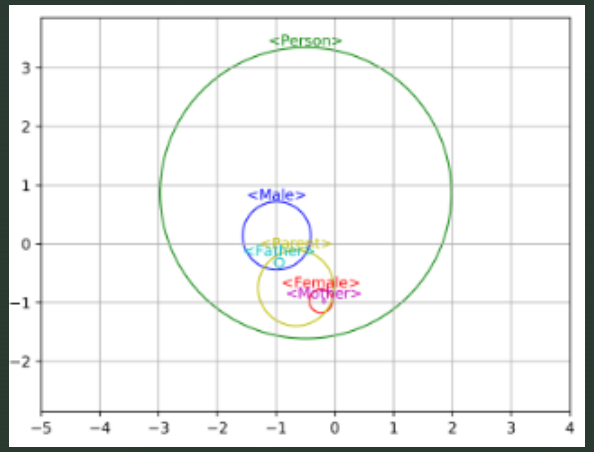
- TransE
- EL-Embeding
- Al-Cone model

TransE

- No convex objects
- Requires dataset in triples
- Tail is nearest object to head + relationship
- Minimize margin-based ranking $\mathcal{L} = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} [\gamma + d(\boldsymbol{h}+\boldsymbol{\ell},\boldsymbol{t}) - d(\boldsymbol{h}'+\boldsymbol{\ell},\boldsymbol{t}')]_+$
  - $[x]_+$ - positive part of x
  - γ - margin hyperparameter
  - d - dissimilarity (L1 or L2-norm)

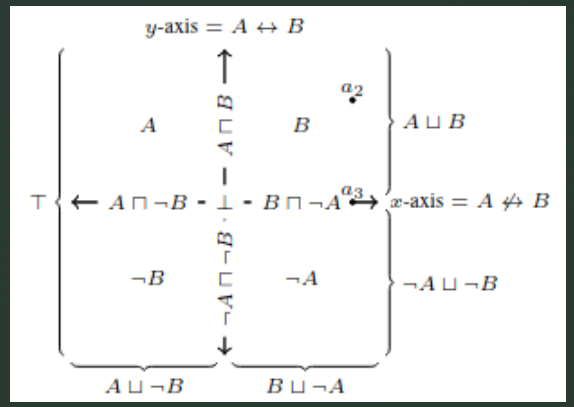
EL-Embedding

- n-ball
- Two functions:
  - $f_\eta : C \cup R \rightarrow R^n$
  - $r_\eta : C \rightarrow R$



Al-Cone embedding

- Axis-aligned cones
- Enable usage of concept negation – polar Al-Cone
- Easy to link to ML
  - conic optimization



Experiments

- TransE embedding
- PyKeen library (https://pykeen.readthedocs.io)
- Ontology provided by Peter Švec
- Dataset:
  - 25000 positive examples
  - 25343 negative examples
  - Testing data:
    - 4968 positive examples
    - 5101 negative examples

Results

- Accuracy :
  - Top_1: 69.89%
  - Top_5: 50.25%
  - Top_10: 50.11%

TOP_1	True positives	True negatives
Predicted positives	2776	840
Predicted negatives	2192	4261

TOP_5	True positives	True negatives
Predicted positives	4869	4910
Predicted negatives	99	191

TOP_10	True positives	True negatives
Predicted positives	4928	4983
Predicted negatives	40	118

Thank you for your attention