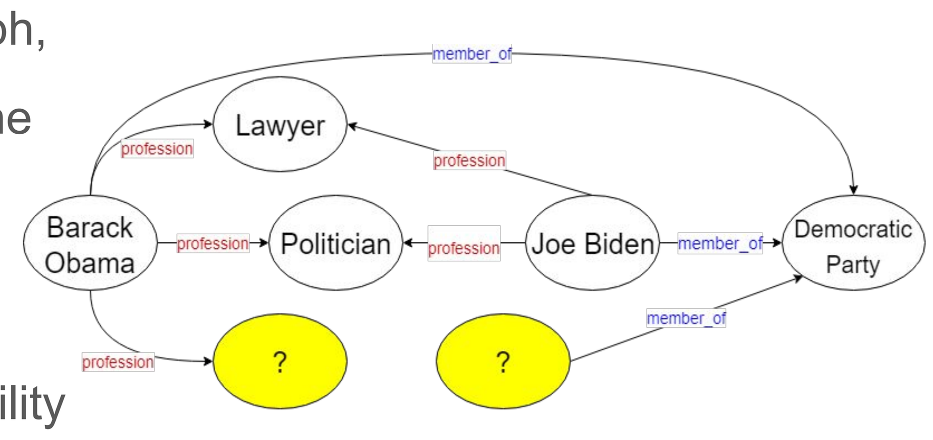
Link Prediction for Knowledge Graphs

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction

Knowledge Graphs

In KGs data is in represented as a graph, where the nodes refer to *entities* and the edges refer to *relations*.



Link Prediction

Predicting the existence (or the probability of existence) of edges in the graph.

Goal and Task

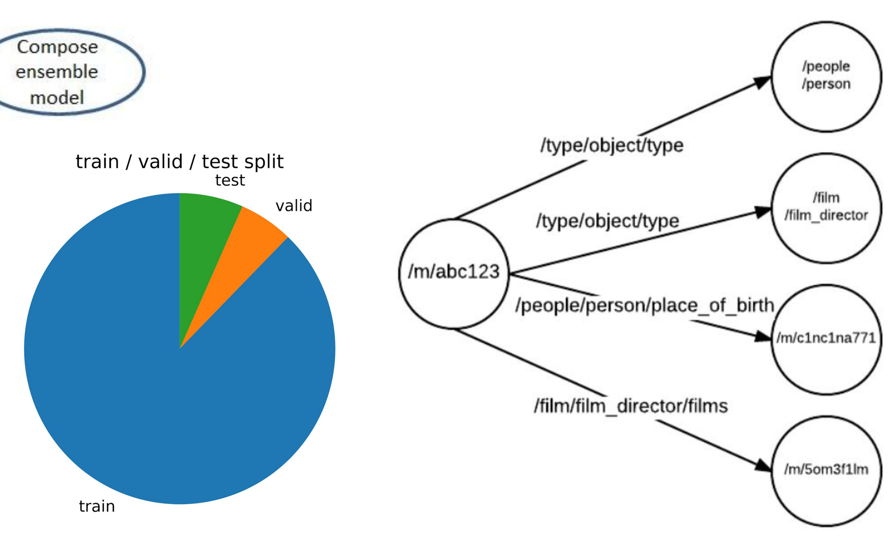
Given (s, p, _) ⟶ what are the most likely entities for an object ?

Given (_, p, o) ⟶ what are the most likely entities for a subject ?

Dataset

FB15k-237: Subset of Freebase

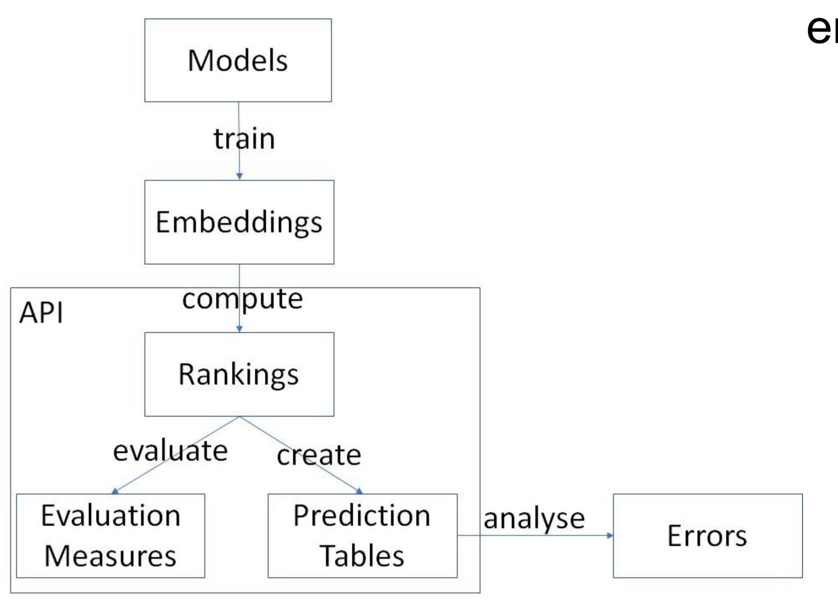
- Inverse relations are removed
- 3 Splits: Train, Test and Valid



Analyse Model Errors

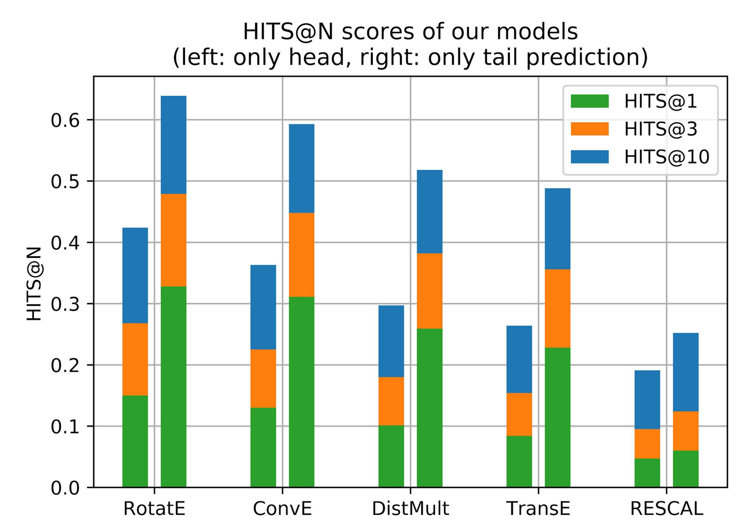
Motivation

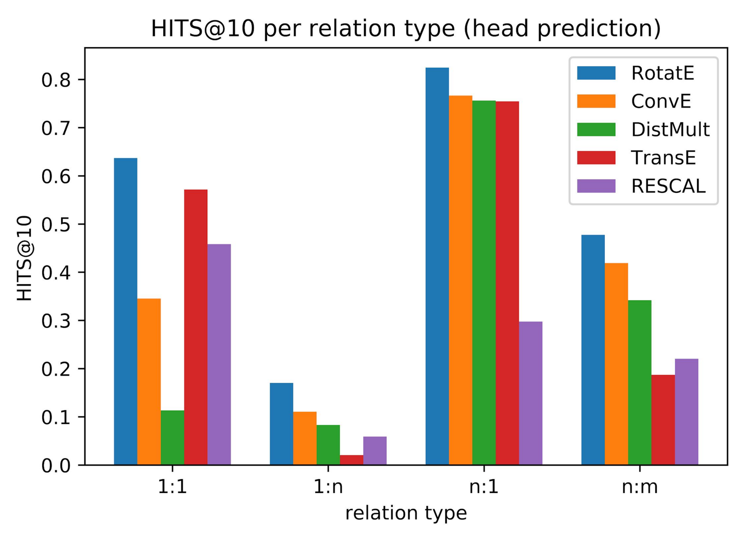
Explore the strengths and weaknesses of each model, find possible reasons for errors



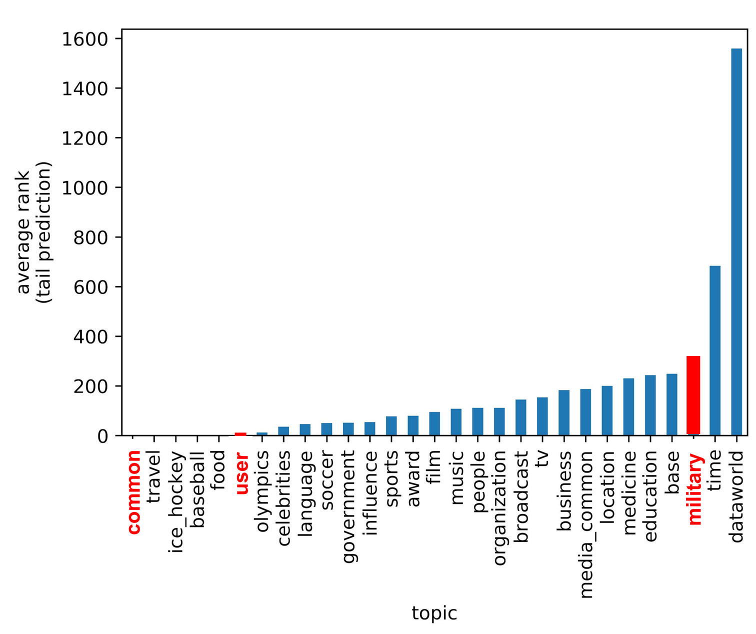
Results from Analysis

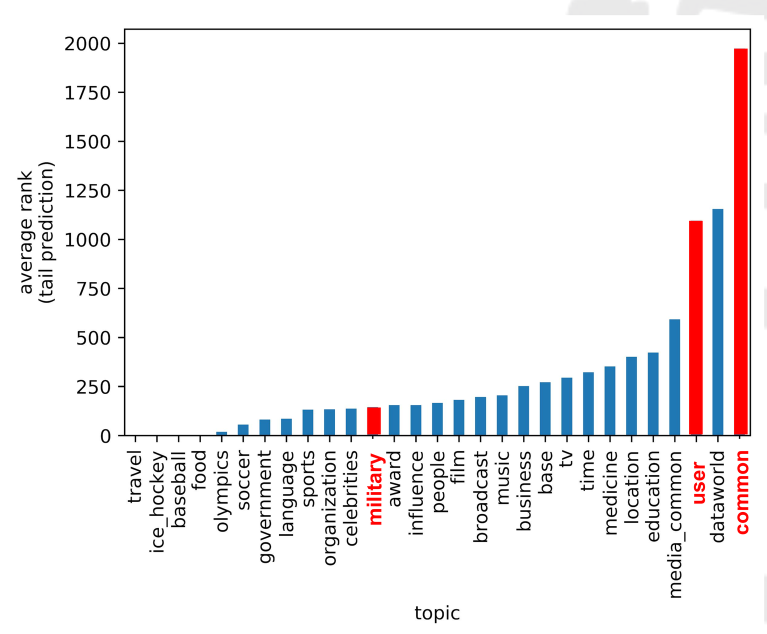
- In some aspects the models don't differ much from each other
  - All models are better at predicting tails
  - They don't vary widely on how good they handle different relation types





- In some aspects they differ from each other:
  - In comparison to other models RESCAL and ConvE predict a wrong named entity type less frequently
  - Our models handle different topics differently well, for example RotatE and DistMult exhibit different capabilities when handling topics such as "common", "user" and "military":

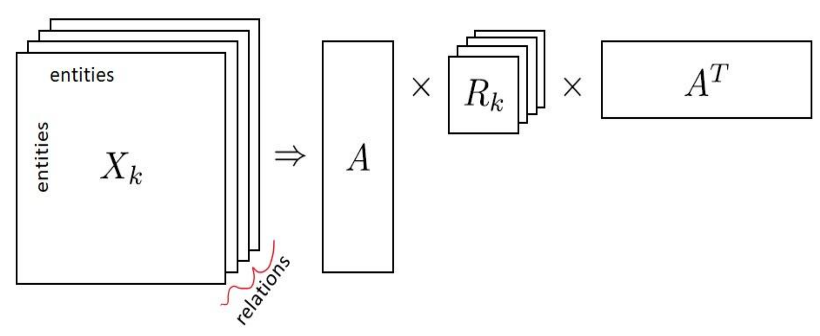




- We distinguish between types of errors (e.g. wrong named entity type) and reasons for errors (e.g. entity/relation is unknown, some models can't handle specific types of relationships, model is biased)

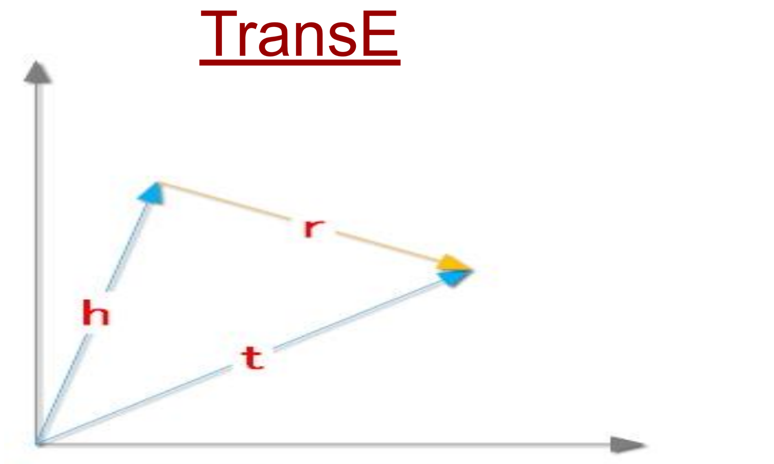
Models and Evaluation

RESCAL



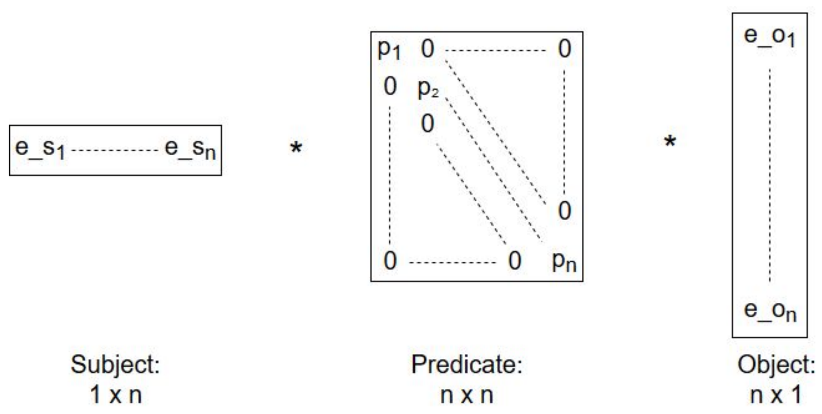
Models binary relations as a tensor

TransE



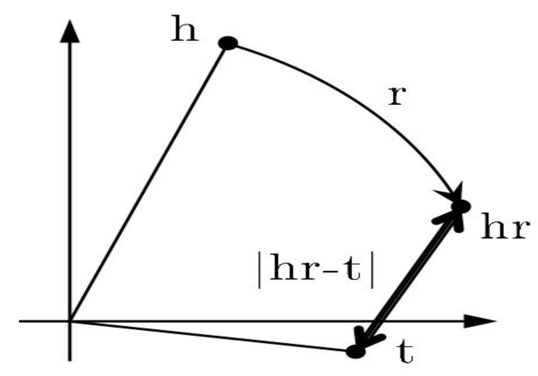
Models relationships as translations

DistMult



Models relations as diagonal matrix

RotatE



Each relation is defined as a rotation from the source to the target entity

ConvE

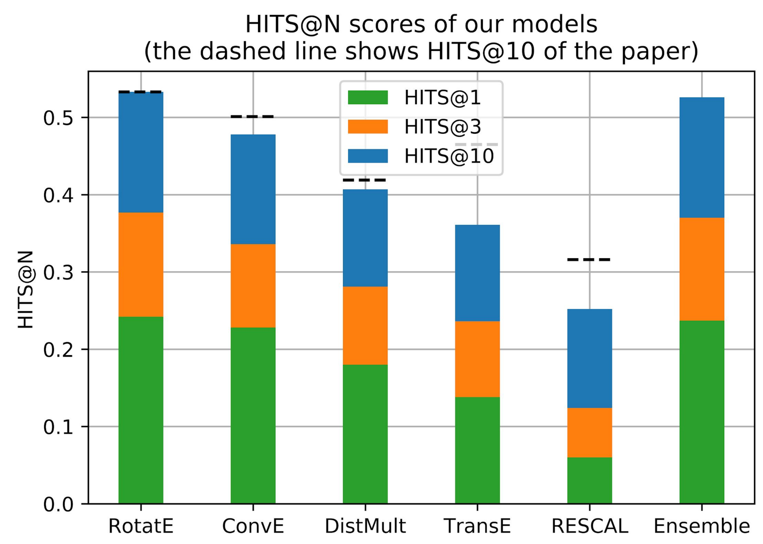
$$\left(\begin{bmatrix} a & a & a\\ a & a & a\end{bmatrix};\begin{bmatrix} b & b & b\\ b & b & b\end{bmatrix}\right)=\begin{bmatrix} a & a & a\\ a & a & a\\ b & b & b\\ b & b & b\end{bmatrix}.$$

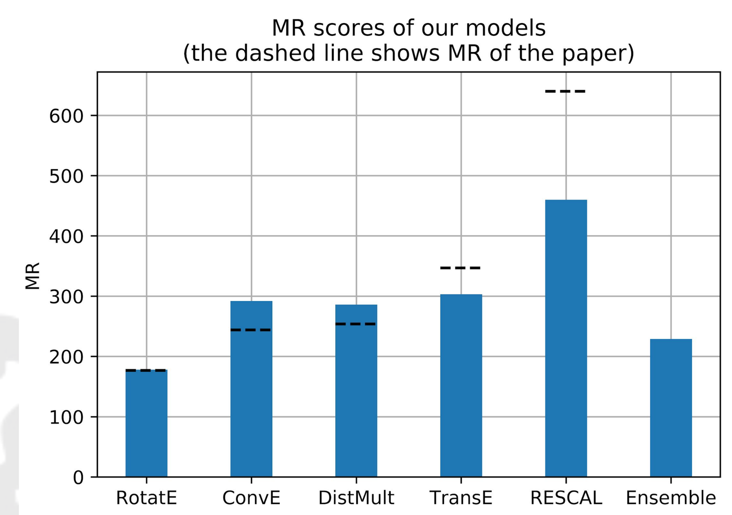
A deep neural network to model the statistical relationships between the entities

References

- Rescal: A Three-Way Model for Collective Learning on Multi-Relational Data, M.Nickel, V.Tresp, H.Kriegel, 2011
- Distmult, arXiv:1412.6575
- TransE: Translating Embeddings for Modeling Multi-relational Data, A.Bordes, N.Usunier, A.Garcia-Duran, 2013
- RotatE: arXiv: 1902.10197
- ConvE: arXiv: 1707.01476

Evaluation Results

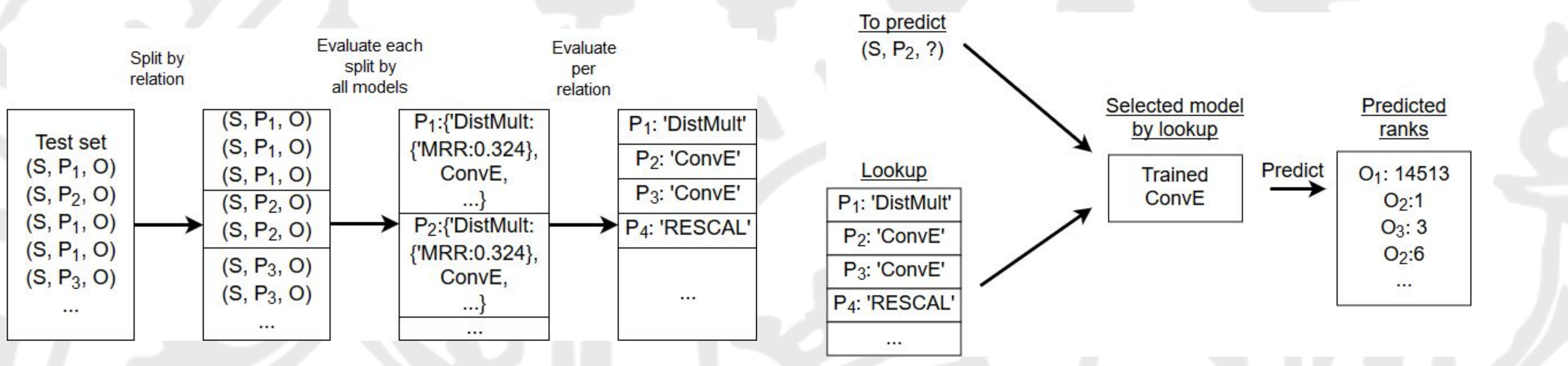




Ensemble Model

Motivation

Achieving better performance by using different models together. Because all models behave differently for different relations, the ensemble model selects the best model for prediction based on the relation:



Conclusion

- RotatE achieved the best and RESCAL the worst results on FB15k-237
- An in-depth error analysis shows significant strengths and weaknesses of each model motivating an ensemble of these models
- Our ensemble model achieved comparable results to RotatE. The individual model scores had different value ranges, which made ensembling difficult.
- Each model should further be evaluated on different and potentially larger datasets to have more stable results.
- Future work could focus on relation-types and topics that were particularly difficult for all models

Practical Course "Big Data Science"

M.Özdemir, S.Bosch, D.Kim, Y.Wang, H.Dang (Group ADL-AI)

Lehrstuhl für Datenbanksysteme und Data Mining, Ludwig-Maximilians Universität München

muammer.oezdemir, sarah.bosch, do.kim, wang.yuhao, h.dang@campus.lmu.de