Interpreting Embedding Models of Knowledge Bases: Model Agnostic Approaches

2018 ICML Workshop on Human Interpretability in Machine Learning

Arthur C. Gusmão[1], Alvaro H. C. Correia[1], Glauber De Bona[1], and Fabio G. Cozman[1]

[1]Escola Politécnica – Universidade de São Paulo, Brazil

July 14, 2018
Stockholm, Sweden

Outline

Knowledge Bases (KBs): sets of triples

⟨ Jane, child_of, Mom ⟩
⟨ John, child_of, Mom ⟩
⟨ Patti, child_of, Mom ⟩
⟨ Mom, born_in, Miami ⟩
⟨ Jane, born_in, Miami ⟩
⟨ John, born_in, Miami ⟩

(Example adapted from [1])

Knowledge Bases (KBs): sets of triples

⟨ Jane, child_of, Mom ⟩
⟨ John, child_of, Mom ⟩
⟨ Patti, child_of, Mom ⟩
⟨ Mom, born_in, Miami ⟩
⟨ Jane, born_in, Miami ⟩
⟨ John, born_in, Miami ⟩

(Example adapted from [1])

Used in many applications!

- Natural language processing (NLP)
- Semantic web search

Knowledge Bases (KBs): sets of triples

⟨ Jane, child_of, Mom ⟩
⟨ John, child_of, Mom ⟩
⟨ Patti, child_of, Mom ⟩
⟨ Mom, born_in, Miami ⟩
⟨ Jane, born_in, Miami ⟩
⟨ John, born_in, Miami ⟩

(Example adapted from [1])

Used in many applications!

- Natural language processing (NLP)
- Semantic web search

But often incomplete...

Knowledge Base Completion

⟨ Jane, child_of, Mom ⟩
⟨ John, child_of, Mom ⟩
⟨ Patti, child_of, Mom ⟩
⟨ Mom, born_in, Miami ⟩
⟨ Jane, born_in, Miami ⟩
⟨ John, born_in, Miami ⟩
⟨ Patti, ?, Miami ⟩

Knowledge Base Completion

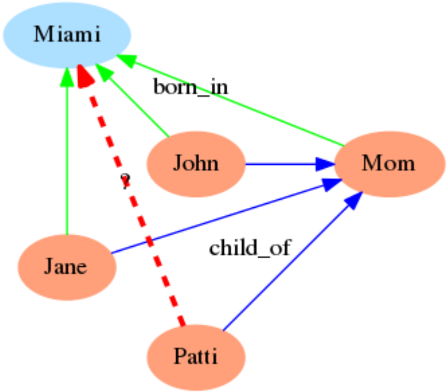
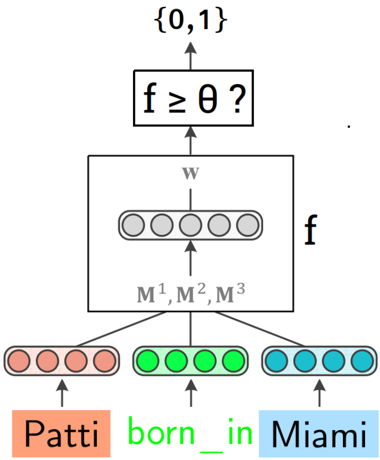


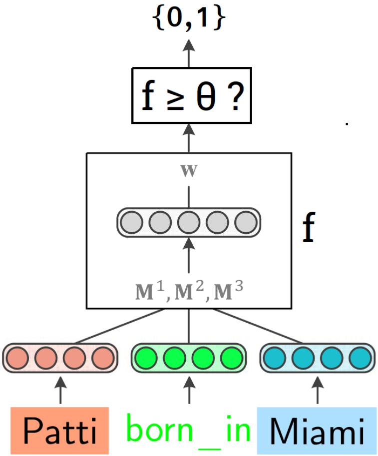
Figure adapted from [1].

Embedding Models for KB Completion



Embedding models map entities and relations into vectors.

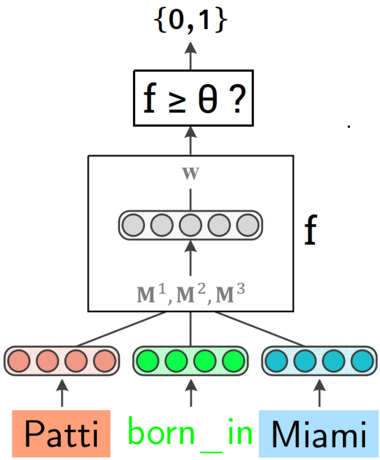
Embedding Models for KB Completion



Embedding models map entities and relations into vectors.

- Achieve state-of-the-art results and are scalable;
- But are poorly interpretable.

Embedding Models for KB Completion

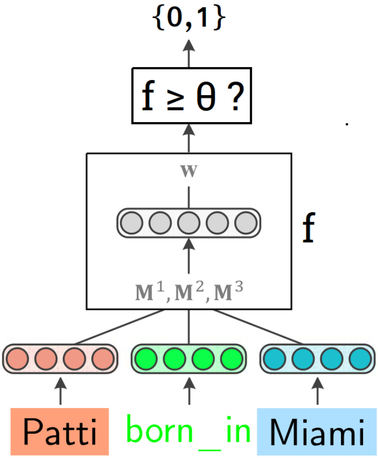


Embedding models map entities and relations into vectors.

- Achieve state-of-the-art results and are scalable;
- But are poorly interpretable.

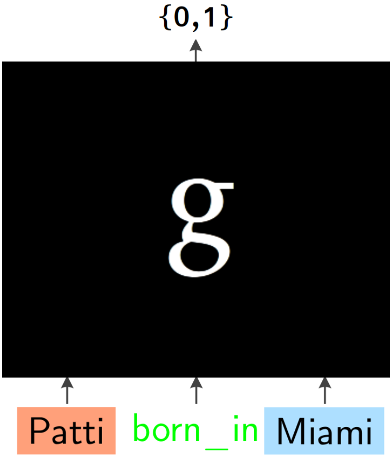
Embeddings turn a semantically rich input into numeric representations where each dimension bears little meaning.

Interpreting Embedding Models of KBs



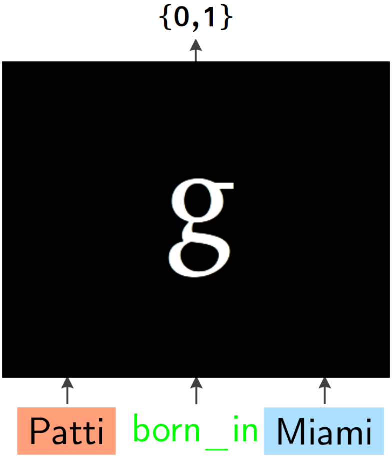
In this work we propose methods to interpret embedding models of KBs.

Interpreting Embedding Models of KBs



- See the embedding model as a **black box**;
- Learn an interpretable model from inputs and outputs.

Interpreting Embedding Models of KBs



- See the embedding model as a **black box**;
- Learn an interpretable model from inputs and outputs.

Model agnostic!

Interpreting Embedding Models of KBs

We propose two methods:

XKE-PRED
Explaining knowledge embedding models with predicted features

XKE-TRUE
Explaining knowledge embedding models with ground truth features

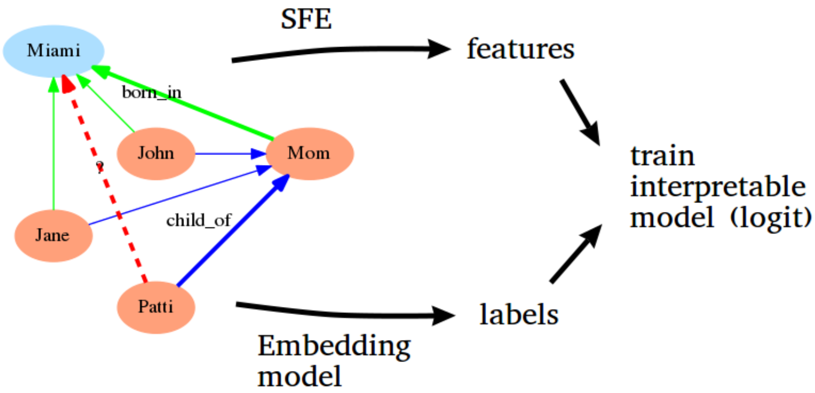
Interpreting Embedding Models of KBs

We propose two methods:

XKE-PRED
Explaining knowledge embedding models
with predicted features

XKE-TRUE
Explaining knowledge embedding models
with ground truth features

XKE-TRUE


SFE
features
Miami
born_in
John
Mom
?
Jane
child_of
Patti
train
interpretable
model (logit)
labels
Embedding
model

Subgraph Feature Extraction

Subgraph Feature Extraction (SFE):

- Binary features;
- Each feature indicates the existence of a path π (a sequence of edges) between two entities;

Advantages:

- Features can be understood as bodies of weighted rules [2];
- Usually regarded as "easily interpretable";
- Can be used with any classification model.

Subgraph Feature Extraction

The only feature with value 1 between Patti and Miami is the path $\pi = (\texttt{child\_of}, \texttt{born\_in})$.

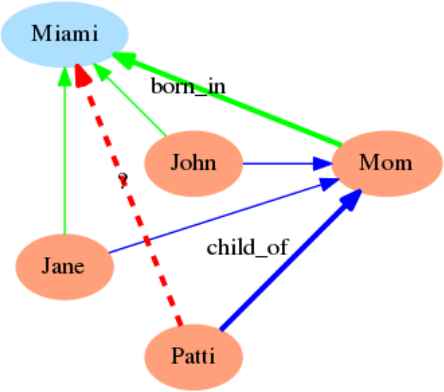


Figure adapted from [1].

XKE-TRUE

More formally:

XKE-TRUE

- Construct a set of examples $\mathbb{D}$ of arbitrary size n in which, for each triple $x_{h,r,t} = \langle e_h, r_r, e_t \rangle$,
  - Features $F(x_{h,r,t} \mid \mathcal{G})$ are extracted using SFE from a ground truth knowledge graph $\mathcal{G}$;
  - The label corresponds to the embedding model's prediction.

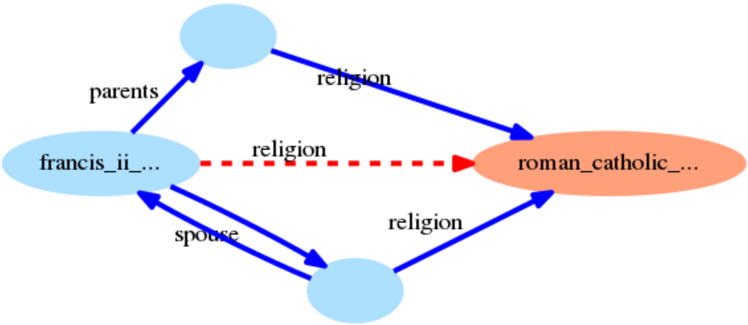
$$\mathbb{D} = \{(F(x_{h,r,t} \mid \mathcal{G}), g(x_{h,r,t}))\}^n$$

- Train an interpretable classifier (logit) using $\mathbb{D}$;
- Draw explanations from the interpretable classifier.

Experiments & Results

Dataset	FB13				NELL186			
XKE variant	TRUE	PRED$_3$	PRED$_5$	PRED$_7$	TRUE	PRED$_3$	PRED$_5$	PRED$_7$
Embedding Accuracy	82.55				86.40			
# Positive triples in $\mathcal{G}$ (XKE-TRUE) or $\hat{\mathcal{G}}$ (XKE-PRED)	322k	830k	1,668k	2,658k	36k	196k	524k	987k
$\hat{\mathcal{G}}$ positive over predicted ratio	-	0.286	0.207	0.168	-	0.604	0.581	0.558
# Features per example	2.91	0.91	1.34	1.79	70.66	159.54	249.86	337.41
% Examples with # features > 0	54.73	33.83	37.88	41.81	50.01	39.39	45.57	51.87
Explanation Mean # Rules (for explanations with size > 0)	2.29	2.19	2.70	2.57	105.30	51.33	159.02	158.87
Explanation Mean Rule Length	3.09	3.00	2.87	2.82	3.86	3.78	3.89	3.89
Fidelity	73.26	66.65	74.36	69.99	86.55	77.00	74.94	75.64
Fidelity (filtered for examples with # features > 0)	80.52	84.30	85.74	83.28	87.02	85.00	83.07	84.47
Fidelity (weighted by the # features)	75.21	82.67	84.58	84.80	85.66	88.09	86.24	88.22
Accuracy	73.43	64.58	71.78	68.11	89.10	75.79	76.18	76.44
Accuracy (filtered for examples with # features > 0)	80.78	81.00	82.02	80.34	91.19	84.08	84.30	85.11
Accuracy (weighted by the # features)	71.68	78.42	81.28	82.19	82.12	86.56	89.11	89.41
F1 (Fidelity)	76.66	50.11	71.14	61.13	83.19	61.41	68.07	68.03
F1 (Accuracy)	77.35	49.07	69.16	59.69	86.89	62.66	71.14	70.68

ID	#1 (XKE-TRUE)
Triple	⟨ francis_ii_of_the_two_sicilies, religion, roman_catholic_church ⟩
Reason #1	(2.456) parents,religion
Reason #2	(1.946) spouse^{-1},religion
Reason #3	(1.913) spouse,religion
Bias	(1.017)
XKE	0.999346
Embedding	1



Conclusion

- We presented techniques to explain KB embeddings models, where features can be understood as weighted Horn clauses.
- Future work: fidelity is a point for improvement.
- We expect this initial work to serve as a basis of comparison and inspiration for the development of novel methods for explaining embedding models in KB completion.

Code available: https://github.com/arthurcgusmao/xke

Thank you!
Questions?

References I

Antoine Bordes and Jason Weston.
Embedding Methods for Natural Language Processing, October 2014.
Tutorial.

Matt Gardner, Partha Talukdar, and Tom Mitchell.
Combining Vector Space Embeddings with Symbolic Logical Inference over Open-Domain Text.
page 5, 2015.