Type-Sensitive Knowledge Base Inference Without Explicit Type Supervision

Prachi Jain*[1], Pankaj Kumar*[1], Mausam[1] and **Soumen Chakrabarti[2]**
[1]Indian Institute of Technology, Delhi and [2]Indian Institute of Technology, Bombay

Conventional Knowledge Base Completion Models

- A knowledge base consists of triplets $T \subseteq E \times R \times E$
- E is the set of entities and R is the set of relations
- Models learn embeddings for each element in E and R
- A scoring function $f(s,r,o)\colon \mathrm{T} \to \mathbb{R}$ is defined in terms of the above embeddings, representing it's confidence

Model	Embeddings	Scoring Function
E	$a_e, b_r, c_r \in \mathbb{R}^n$	$b_r \cdot a_s + c_r \cdot a_o$
DistMult	$a_e, b_r \in \mathbb{R}^n$	$\langle a_s \vert b_r \vert a_o \rangle$
Complex	$a_e, b_r \in \mathbb{C}^n$	$\Re(\langle a_s \vert b_r \vert a_o \rangle)$

Problems with conventional Models

Subject	Relation	Gold Object	Prediction 1	Prediction 2
Howard Leslie Shore	follows-religion	Jewism(religion)	Walk Hard (film)	21 Jump Street (film)
Spyglass Entertainment	headquarterd-in	El lay(location)	The Real World (tv)	Contraband (film)
Les Franklin	born-in-location	New York(location)	Federico Fellini(person)	Louie De palma (person)
Eugene Alden Hackman	studied	Rural Journalism(education)	L Snowden Wainwright III(person)	The Bourne Legacy (film)
Chief Phillips (film)	released-in-region	Yankee land(location)	Akira Isida (person)	Presidential Medal of Freedom (award)

Samples of top two DistMult predictions (having *inconsistent types*) on FB15K

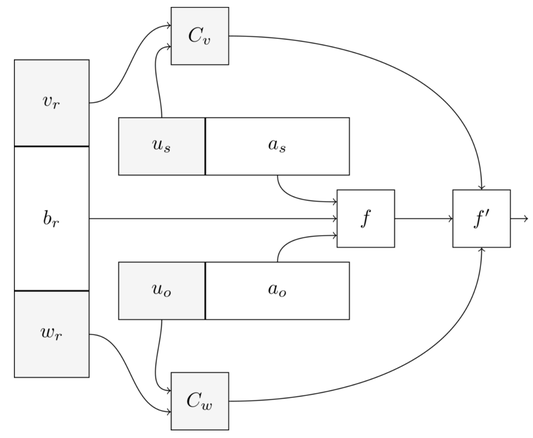
TypeDM and TypeComplex

- Typed Models extend a conventional model (base model)
- Compatibility functions $C_v(s,r)$ and $C_w(o,r)$ represent subject and object type compatibility between entity and relations
- u_e denotes the type embedding of entity e
- v_r and w_r denote the type of head and tail entity of relation r
- Trained without any additional type catalogue or other information
- Final Score of a typed model, in terms of base model score f'

$$f(s,r,o) = f'(s,r,o)C_v(s,r)C_w(o,r)$$
$$C_v(s,r) = \sigma(v_r \cdot u_s)$$
$$C_w(o,r) = \sigma(w_r \cdot u_o)$$

Model	FB15K MRR	FB15K HITS@1	FB15K HITS@10	FB15K237 MRR	FB15K237 HITS@1	FB15K237 HITS@10	YAGO3-10 MRR	YAGO3-10 HITS@1	YAGO3-10 HITS@10
E	23.40	17.39	35.29	21.30	14.51	36.38	7.87	6.22	10.00
DM+E	60.84	49.53	79.70	38.15	28.06	58.02	52.48	38.72	77.40
DistMult	67.47	56.52	84.86	37.21	27.43	56.12	55.31	46.80	70.76
TypeDM	75.01	66.07	87.92	38.70	29.30	57.36	58.16	51.36	70.08
Complex	70.50	61.00	86.09	37.58	26.97	55.98	54.86	46.90	69.08
TypeComplex	75.44	66.32	88.51	38.93	29.57	57.50	58.65	51.62	70.42

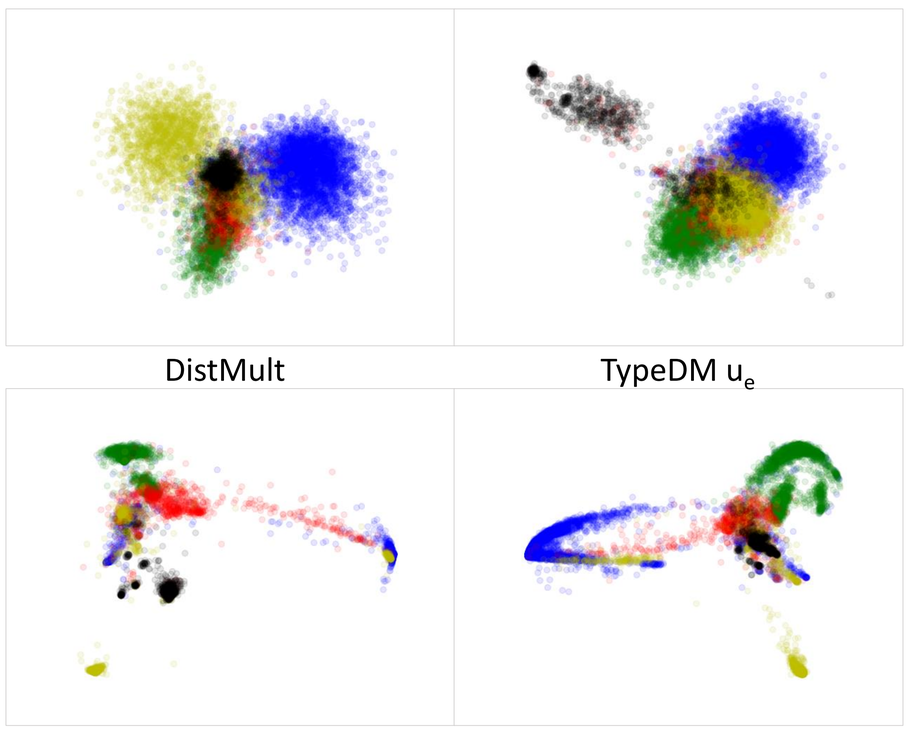
Comparing Conventional Models and Typed Models of equivalent sizes for Knowledge Base Completion



Analysis of Trained Embedding

Method	Embedding	Size	Homogeneity	Completeness	Prediction F1
TypeDM	u_e	19	66.72	66.29	81.77
TypeDM	a_e	180	57.89	59.67	75.96
TypeDM	Both	199	66.75	66.29	82.57
DM	a_e	200	51.40	48.12	81.34
TypeComplex	u_e	19	65.90	62.97	82.70
TypeComplex	a_e	360	50.76	48.57	74.75
TypeComplex	Both	379	66.03	63.09	84.14
Complex	a_e	400	51.56	47.20	81.58
DM+E	u_e	19	0.48	2.05	74.66
DM+E	a_e	180	49.62	47.24	82.72
DM+E	Both	199	49.66	47.26	82.68
E	a_e	200	39.83	37.62	74.23

Interpreting Embeddings with respect to Supervised Type Classification (7 clusters of people, location, organisation, film, sports and others for Homogeneity and Completeness) on FB15K



Entities plotted by the PCA Projection of the embeddings learnt for Knowledge base completion, colored by their types in FB15K (People, Location, Organisation, Film, Sports)

Conclusion

- Without supervision from a type catalogue, typed models outperform base models for knowledge base completion across datasets
- u_e embeddings learnt by typed models are better correlated with the entity types

Code: http://github.com/dair-iitd/kbi
*Equal contribution