IterefinE: Iterative KG Refinement Embeddings using Symbolic Knowledge

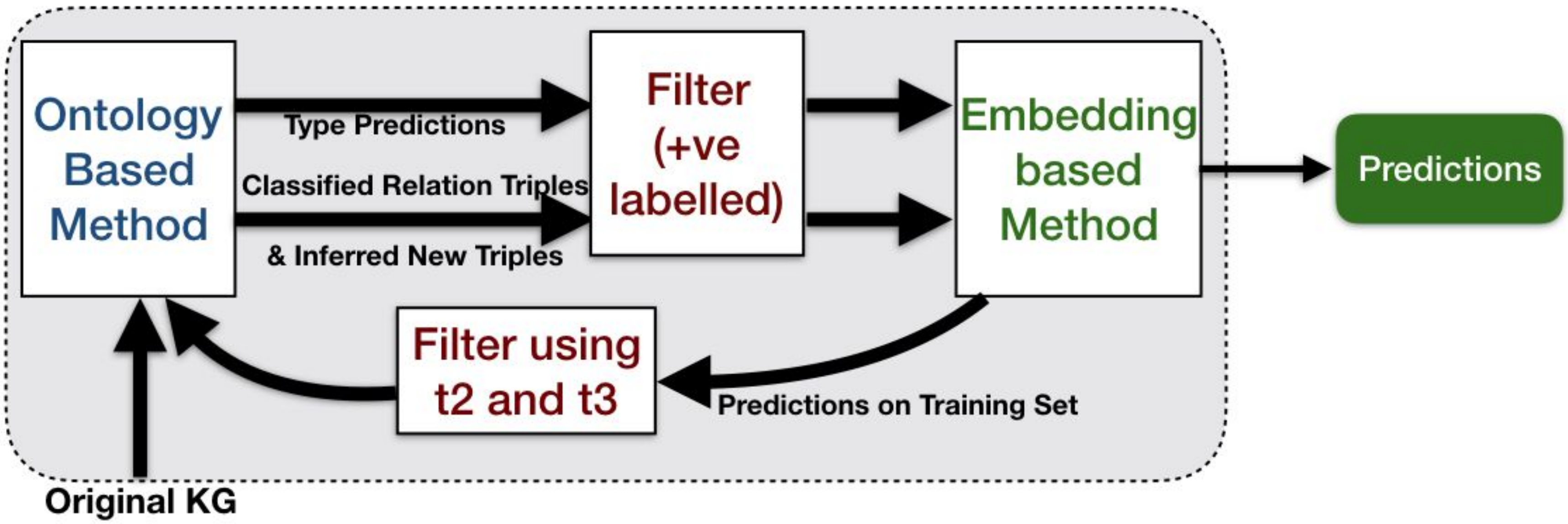
Siddhant Arora, Srikanta Bedathur,
Maya Ramanath and Deepak Sharma
Indian Institute of Technology Delhi

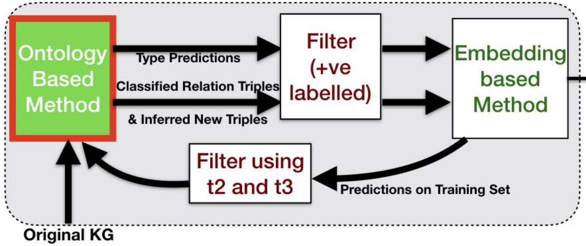
Motivation

- KGs are often noisy and incomplete which decreases performance in downstream task
- Noise refers to various kind of errors in KG like different names for same entity, incorrect relationships and incompatible entity types
- Cleaning up of noise in KGs (KG Refinement) is usually performed using inference rules and reasoning over KGs
- New facts are inferred using KG embeddings
- GOAL : Combine ontology/inference rules with embeddings methods to improve KG refinement

Contributions

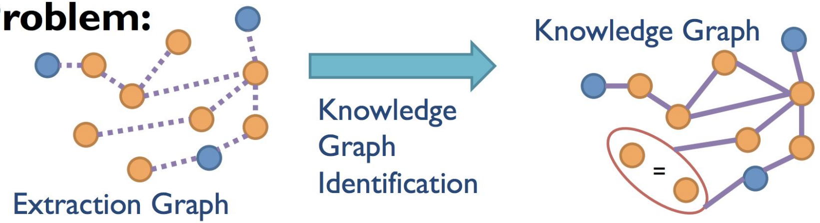
- Propose IterefinE, an iterative method to combine rule-based methods with embeddings-based methods
- Extensive experiments showing improvements upto 9% over baselines





PSL-KGI[1]

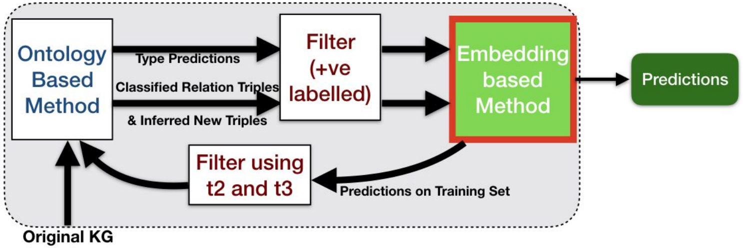
Problem:



Ontological Information
Domain (DOM)
Range (RNG)
Same Entity (SAMEENT)
MUT
Subclass (SUB)
INV
RMUT
SUBPROP (RSUB)

Ontological Rule
$w_{CR-T} : CANDREL_T(E_1, E_2, R) \Rightarrow REL(E_1, E_2, R)$ $w_{CL-T} : CANDLBL_T(E, L) \Rightarrow LBL(E, L)$
$SAMEENT(E_1, E_2) \wedge LBL(E_1, L) \Rightarrow LBL(E_2, L)$ $SAMEENT(E_1, E_2) \wedge REL(E_1, E, R) \Rightarrow REL(E_2, E, R)$ $SAMEENT(E_1, E_2) \wedge REL(E, E_1, R) \Rightarrow REL(E, E_2, R)$

[1] Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. Knowledge graph identification. In International Semantic Web Conference, pages 542–557. Springer, 2013.

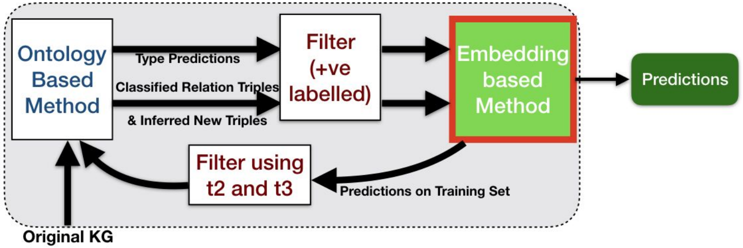


KG Embeddings

$$L(G) = \sum_{(s,r,o,y)\in G} y \log f(s,r,o) + (1-y)\log(1-f(s,r,o))$$

- **ComplEx**[2] - $f(s,r,o) = e_s r_r \overline{e_o}$
- **ConvE**[3] - $f(s,r,o) = f(vec(f([\overline{e_s};\overline{r_r}] * w))W)e_o$
- **Implicit Type Supervision**[4] $f(s,r,o) = \sigma(\mathbf{s_t}\cdot\mathbf{r_h}) * \mathbf{Y}(s,r,o) * \sigma(\mathbf{o_t}\cdot\mathbf{r_t})$,
    - s_t and o_t are implicit type embeddings of s and o,
    - r_h and r_t are implicit embeddings of relation dom and range
    - Y is scoring function

[2] T. Trouillon, J. Welbl, S. Riedel,E. Gaussier, and G. Bouchard.Complex embeddings for simple link prediction. In ICML, 2016
[3] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel. Convolutional 2d knowledge graph embeddings. In AAAI, 2018
[4] P. Jain, P. Kumar, S. Chakrabarti, et al. Type-sensitive knowledge base inference without explicit type supervision. In ACL, 2018

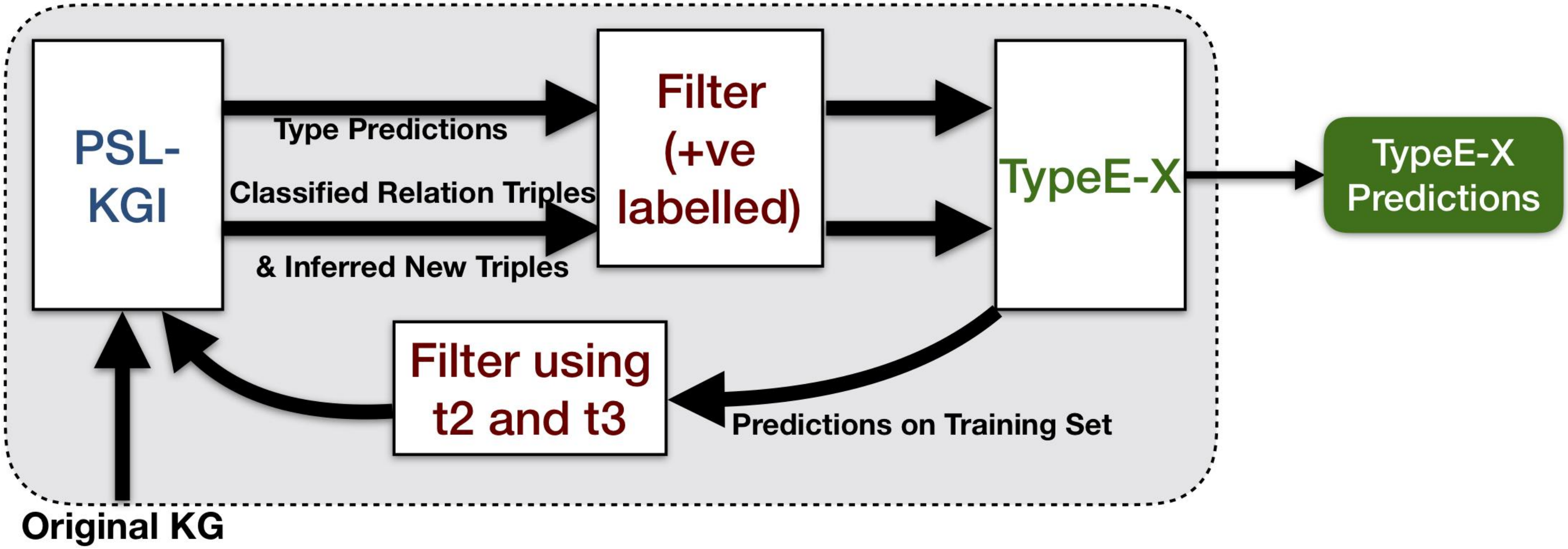


Explicit Type Supervision (TypeE-X)

$$f(s,r,o) = \sigma((\mathbf{s_t} \| \mathbf{s_l}) \cdot (\mathbf{r_h} \| \mathbf{r_{dom}})) * \mathbf{Y}(s,r,o) * \sigma((\mathbf{o_t} \| \mathbf{o_l}) \cdot (\mathbf{r_t} \| \mathbf{r_{range}}))$$

- Here s_1 and o_1 are explicit entity type embeddings,
- r_{dom} and r_{range} are explicit embedding of domain and range of relation.
- The entity types, domain and range type of relation are transferred from PSL-KGI

Algorithm Workflow



Dataset Preparation

Dataset	$\|E\|$	$\|R\|$	#triples in train / valid / test
NELL	820K	222	1.02M / 4K / 4K
FB15K-237	14K	238	246K / 27K / 30K
YAGO3-10	123K	38	1.13M / 10K / 10K
WN18RR	40K	12	116K / 6K / 6K

Table 3: Number of entities, relation types and triples in each dataset.

NELL already has noisy labels whereas for other datasets-

- Randomly sample 25% and corrupt them.
- Make 50% of the noise is **type compatible** and the rest is type non compatible

Ontology Information

Dataset	DOM	RNG	SUB	RSUB	MUT	RMUT	INV	SAMEENT
NELL	418	418	288	461	17K	48K	418	8K
FB15K-237	237	237	44K	0	147K	53K	44	20K
YAGO3-10	37	37	828	2	30	870	8	20K
WN18RR	11	11	13	0	0	66	0	20K

Table 4: Number of instances of each ontological component in datasets considered.

- NELL and YAGO come with rich ontology
- Type Labels are obtained for FB15k-237[5] and for WN18RR[6]. All other rules are automatically mined for both datasets

[5] Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Representation learning of knowledge graphs with hierarchical types. In IJCAI, pages 2965–2971, 2016. - Check citation
[6] Johannes Villmow. Transforming wn18 / wn18rr back to text., 2018.

Results

PSL KGI is hard to beat on NELL

Slightly worse on WN18RR because of very limited ontology

Method	NELL			YAGO3-10			FB15K-237			WN18RR		
	+ve F1	-ve F1	wF1	+ve F1	-ve F1	wF1	+ve F1	-ve F1	wF1	+ve F1	-ve F1	wF1
ComplEx	0.82	0.58	0.73	0.94	0.43	0.88	0.96	0.4	0.92	0.93	0.26	0.86
ConvE	0.74	0.55	0.67	0.94	0.37	0.87	0.95	0.37	0.90	0.93	0.07	0.84
PSL-KGI	0.85	**0.68**	**0.79**	0.91	0.39	0.85	0.92	0.39	0.88	0.91	**0.37**	0.85
ConvE +ComplEx	0.82	0.58	0.73	0.95	0.43	0.89	0.96	0.39	0.92	0.93	0.15	0.85
α - ComplEx	0.85	**0.68**	**0.79**	0.94	0.50	0.89	0.96	0.58	0.93	**0.94**	0.24	0.87
α - ConvE	0.85	**0.68**	**0.79**	0.94	0.41	0.88	0.95	0.47	0.92	0.92	0.34	0.85
TypeE-ComplEx	**0.86**	**0.68**	**0.79**	**0.95**	**0.56**	**0.91**	**0.98**	**0.82**	**0.97**	0.93	0.24	0.85
TypeE-ConvE	**0.86**	0.67	**0.79**	**0.95**	0.47	0.89	**0.98**	0.77	0.96	**0.94**	0.31	**0.87**

[1] Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. Knowledge graph identification. In International Semantic Web Conference, pages 542–557. Springer, 2013.
[2] T. Trouillon, J. Welbl, S. Riedel,E. Gaussier, and G. Bouchard.Complex embeddings for simple link prediction. In ICML, 2016
[3] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel. Convolutional 2d knowledge graph embeddings. In AAAI, 2018

Additional Results

- Accuracy of TypeE-X methods do not vary very much with additional iterations for rich and good quality ontology
- Adding type inferences from PSL-KGI boost performance over implicit type embeddings
- Subclass, Domain and Range constraints are the most important however none of the individual ontological components alone show performance comparable to using all the component
- Datasets with high quality ontology more stable in KG sizes with increasing iterations
- Type compatible noise are harder to remove than type non compatible noise

Thank You

Contact: siddhantarora1806@gmail.com