CESI: Canonicalizing Open Knowledge Bases using Embeddings and Side Information

{ Shikhar Vashishth, Prince Jain, Partha Talukdar} @ IISc

Abstract

- Open IE methods extract triples from text, resulting in the construction of large Open KBs. Non-canonicalized noun and relation phrases in such Open KBs leads to storage of redundant and ambiguous facts.
- To overcome this challenge, we propose CESI – a novel approach which performs canonicalization over learned embeddings of Open KBs. CESI extends recent advances in KB embedding by incorporating relevant side information in a principled manner.

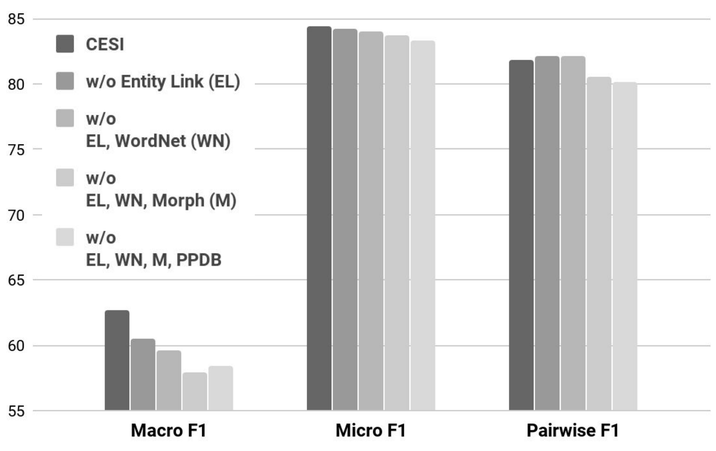
Contributions

1. Novel method for canonicalizing Open KBs using learned embeddings.
2. CESI performs **joint canonicalization** of noun and relation phrases using relevant side information in a principled manner.
3. Propose a new dataset for Open KB canonicalization, **ReVerb45K** with 20x more NPs than the previous biggest dataset for this task.

T-SNE Visualization

t-SNE visualization of embeddings

Abaltion Results



Performance on removing Side Info

References

[1] Galárraga, Heitz, Murphy, and Suchanek Canonicalizing Open Knowledge Bases In *CIKM '14*

[2] M. Nickel, L. Rosasco, and T. A. Poggio HolE: Holographic Embeddings of KG In *AAAI '16*

Overview

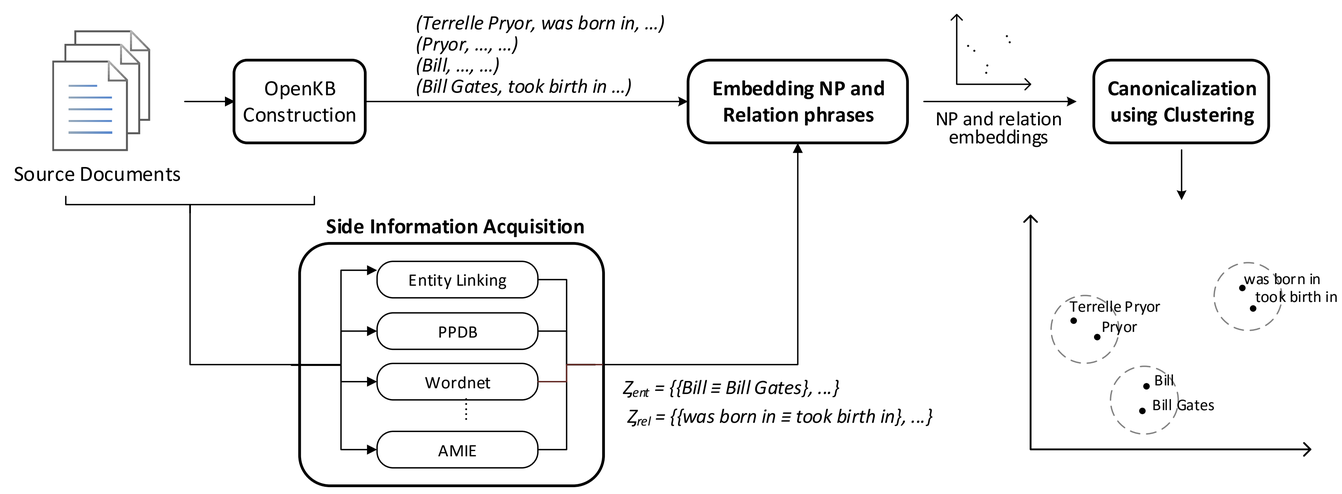


Figure: Overview of CESI

CESI performs canonicalization through its three step procedure:

1. **Side Information Acquisition:** Gather various NP and relation phrase side information for each triple.
2. **Embedding NP and Relation Phrases:** Learn specialized vector embeddings for all NPs and relation phrases.
3. **Clustering Embeddings and Canonicalization:** Cluster the NPs and relation phrases based on their distance in the embedding space.

Method

We aim to optimize the following objective function for learning embeddings for noun and relation phrases:

$$\min_{\Theta} \lambda_{str} \sum_{i \in D_+} \sum_{j \in D_-} \max(0, \gamma + \sigma(\eta_j) - \sigma(\eta_i)) + \sum_{\theta \in \mathscr{C}_{\text{ent}}} \frac{\lambda_{\text{ent},\theta}}{|\mathcal{Z}_{\text{ent},\theta}|} \sum_{v,v' \in \mathcal{Z}_{\text{ent},\theta}} \|e_v - e_{v'}\|^2 + \sum_{\phi \in \mathscr{C}_{\text{rel}}} \frac{\lambda_{\text{rel},\phi}}{|\mathcal{Z}_{\text{rel},\phi}|} \sum_{u,u' \in \mathcal{Z}_{\text{rel},\phi}} \|r_u - r_{u'}\|^2 + \lambda_{\text{reg}} \left(\sum_{v \in V} \|e_v\|^2 + \sum_{r \in R} \|e_r\|^2 \right).$$

Side Information

Noun Phrase Side Information:

- *Entity Linking:* Identify entity mentions and link to KBs like Wikipedia.
- *PPDB:* Large collection of paraphrases.
- *WordNet with sense disambiguation:* Identifying synsets of NPs
- *IDF Token Overlap:* Identifying NPs sharing infrequent terms.
- *Morphological normalization:* Tense removal, pluralization, capitalization etc.

Relation Phrase Side Information:

- *AMIE* for identifying equivalent relations
- *Knowledge Base Population* systems for detecting relations between entities

Quantitative Evaluation

Method	Base Dataset			Ambiguous Dataset			ReVerb45K		
	Macro	Micro	Pair.	Macro	Micro	Pair.	Macro	Micro	Pair.
Morph Norm	58.3	88.3	83.5	49.1	57.2	70.9	1.4	77.7	75.1
PPDB	42.4	46.9	32.2	37.3	60.2	69.3	46.0	45.4	64.2
EntLinker	54.9	65.1	75.2	49.7	83.2	68.8	62.8	81.8	80.4
Galárraga-StrSim	88.2	96.5	97.7	66.6	85.3	82.2	69.9	51.7	0.5
Galárraga-IDF	94.8	97.9	98.3	67.9	82.9	79.3	71.6	50.8	0.5
Galárraga-Attr	76.1	51.4	18.1	**82.9**	27.7	8.4	**75.1**	20.1	0.2
GloVe	95.7	97.2	91.1	65.9	89.9	90.1	56.5	82.9	75.3
HolE (Random)	69.5	91.3	86.6	53.3	85.0	75.1	5.4	74.6	50.9
HolE (GloVe)	75.2	93.6	89.3	53.9	85.4	76.7	33.5	75.8	51.0
CESI	**98.2**	**99.8**	**99.9**	66.2	**92.4**	**91.9**	62.7	**84.4**	**81.9**

Future Direction

In future, we plan to employ soft clustering for canonicalizing NPs and RPs. Also, other types of side information can be incorporated for further improving the performance.

Source Code

The source code and dataset are available at:
`https://github.com/mallabiisc/cesi`

Acknowledgement

This work was supported by MHRD, Govt. of India, and by gifts from Google Research and Accenture