Canonicalizing Open Knowledge Bases using Embeddings and Side Information

Shikhar Vashishth, Prince Jain, Partha Talukdar
Indian Institute of Science, India

Knowledge Graphs

- Knowledge in graph form
- **Nodes** represent entities
- **Edges** represent relationships b/w entities
- **Examples**: Freebase, Wikidata ...

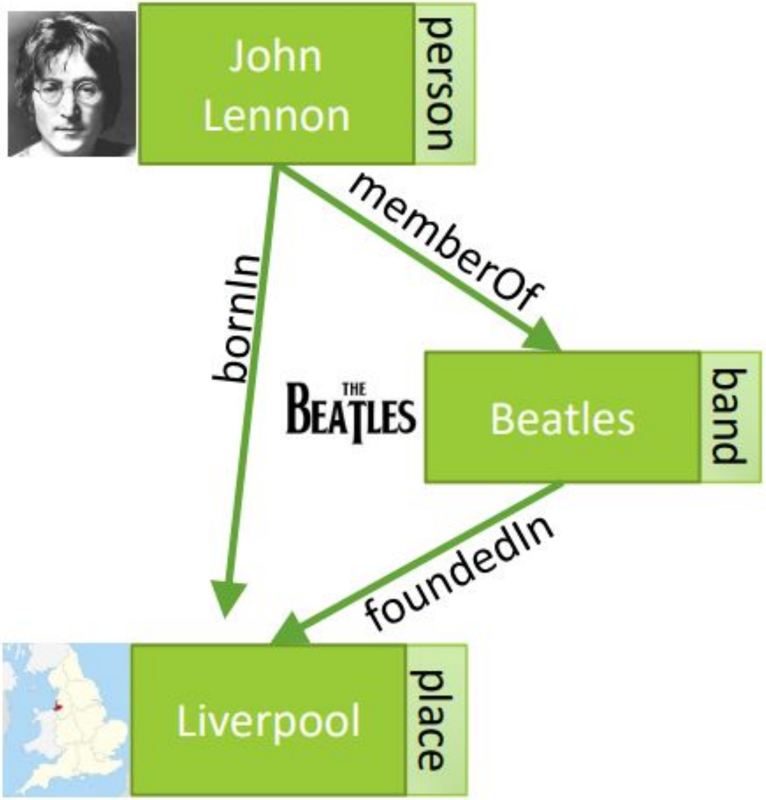


**Figure source: Mining Knowledge Graphs from Text, WSDM '18 tutorial*

What are Open KGs?

- KGs with entities and relations **not restricted to a defined set.**

- **Construction**: Automatically extracting ***(noun-phrase, relation-phrase, noun-phrase)*** from unstructured text.
    - *Obama was the President of US. → (Obama, was president of, US)*
    - Examples: TextRunner, ReVerb, Ollie etc.

- **Use cases:**
    - Extract knowledge from a new domains without supervision.

Challenges with Open KG

- **Problem:** May store redundant and ambiguous facts
  - *(Barack Obama, was president of, US)*
  - *(Obama, born in, Honolulu)*

- Querying for "*Barack Obama*" will not return all extracted facts.

- **Solution:** Need to Canonicalize Open KGs

Canonicalization

Noun Phrases

- Barack Obama; Obama
- George Bush
- New York City; NYC

Relation phrases:

- born_in; took_birth_in
- is_employed_in; works_for
- capital_of

Previous works

- **RESOLVER** system [Yates, 2009] uses string similarity based features to cluster phrases in **TextRunner.**

- **[Galárraga, 2014]** perform noun phrase canonicalization by clustering over **manually-defined feature** spaces which is followed by relation phrase canonicalization using **AMIE** [Galárraga, 2013]

Issues

- **Surface form not sufficient** for disambiguation
  - E.g. (US, America)
- **Manual feature engineering** is expensive and often sub-optimal
- **Sequentially canonicalizing** of noun and relation phrases can lead to error propagation

Contributions

- We propose **CESI**, a novel method for canonicalizing Open KBs using **learned embeddings.**

- CESI **jointly canonicalize** both noun phrase (NP) and relation phrase using relevant side information.

- We build a new data, **ReVerb45K** which has **20x more NPs** than previous biggest dataset for the task.

CESI Overview

1. **Side Information Acquisition:**
    - Gathers various noun and relation phrase side Information
2. **Embeddings Noun and relation phrases:**
    - Learns specialized vector embeddings
3. **Clustering Embeddings and Canonicalization:**
    - Clusters embeddings based on distance
    - Assigns a representative to each noun and relation cluster

Side Information Acquisition

- Involves identifying equivalence relations of form:
    - $\mathbf{e}_1 \equiv \mathbf{e}_2$ and $\mathbf{r}_1 \equiv \mathbf{r}_2$

- **Entity Linking:**
    - Identify entity mention and link to KBs like Wikipedia
    - US → United_States, America → United_States

- **Paraphrase database (PPDB):**
    - Large collection of paraphrases in English
    - management ≡ administration, head of ≡ chief of

Side Information Acquisition

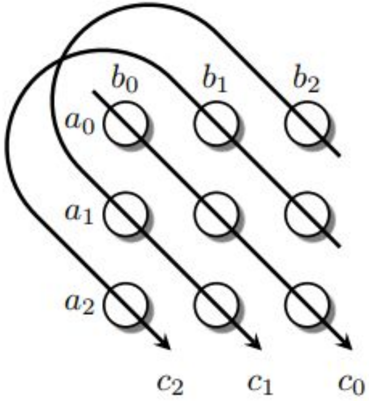
- **WordNet with Word-sense disambiguation:**
  - Identify synsets of NPs
  - picture ≡ image, plant ≡ industry
- **IDF Token Overlap:**
  - NPs and relations sharing infrequent terms
  - Warren Buffett ≡ Mr. Buffett, Mr.Gates ≡ Bill Gates
- Used 9 types of side info, refer paper for more.
- Side information used as **soft constraints**

Embeddings Noun and Relation phrases

- Several KG embedding algorithms available, we use of **HolE (Holographic Embeddings)**

- HolE assigns a **score η** to each triple **(v, r, v')** in KB:

$$\eta = e_r^T (e_v \star e_{v'})$$



$$\boldsymbol{c} = \boldsymbol{a} \star \boldsymbol{b}$$

$$c_0 = a_0 b_0 + a_1 b_1 + a_2 b_2$$
$$c_1 = a_0 b_2 + a_1 b_0 + a_2 b_1$$
$$c_2 = a_0 b_1 + a_1 b_2 + a_2 b_0$$

- Learns embedding by **optimizing**:

$$\sum_{i \in D_+} \sum_{j \in D_-} \max(0, \gamma + \sigma(\eta_j) - \sigma(\eta_i))$$

CESI Optimization Objective

$$\min_{\Theta} \boxed{\lambda_{str} \sum_{i \in D} \sum_{j \in D_-} \max(0, \gamma\, \sigma(\eta_j) - \sigma(\eta_i))}$$

HolE Objective

$$\boxed{\sum_{\theta \in \mathscr{C}_{\text{ent}}} \frac{\lambda_{\text{ent},\theta}}{|\mathcal{Z}_{\text{ent},\theta}|} \sum_{v,v' \in \mathcal{Z}_{\text{ent},\theta}} \|e_v - e_{v'}\|^2}$$

Noun phrase Side Information

$$\boxed{\sum_{\phi \in \mathscr{C}_{\text{rel}}} \frac{\lambda_{\text{rel},\phi}}{|\mathcal{Z}_{\text{rel},\phi}|} \sum_{u,u' \in \mathcal{Z}_{\text{rel},\phi}} \|r_u - r_{u'}\|^2}$$

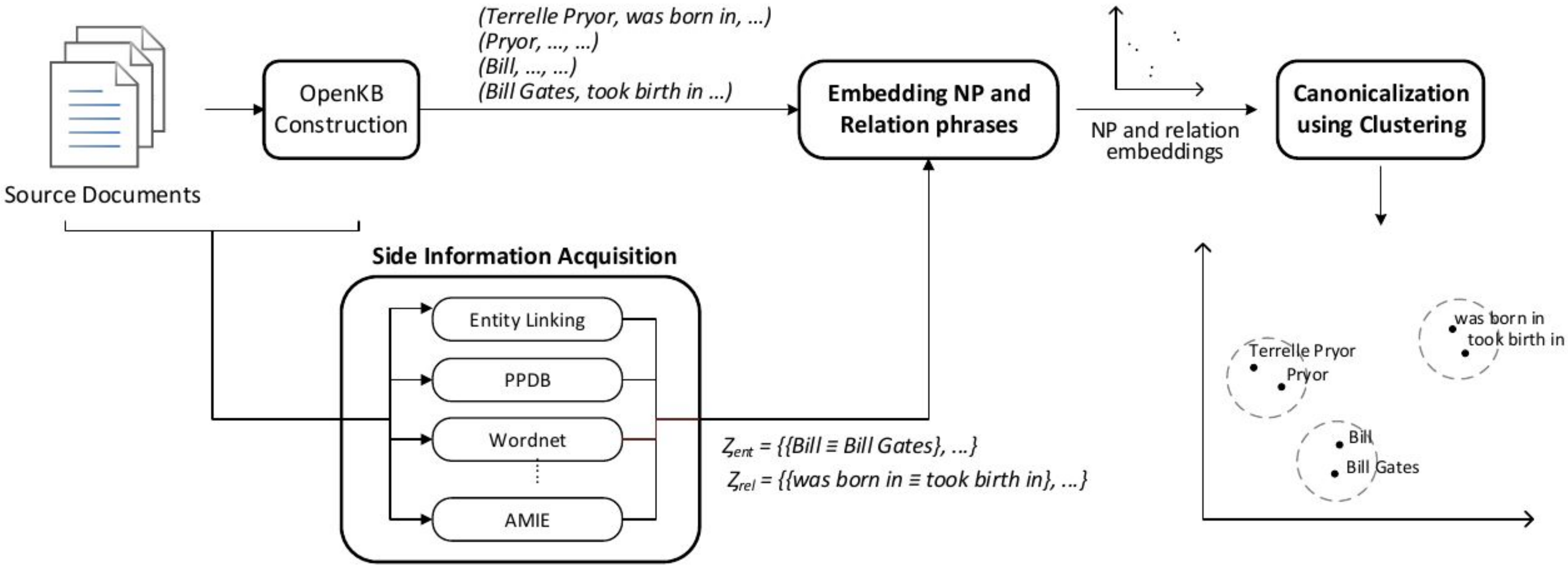
Relation phrase Side Information

$$\boxed{\lambda_{\text{reg}} \Big(\sum_{v \in V} \|e_v\|^2 \sum_{r \in R} \|e_r\|^2 \Big).}$$

Regularization

Optimized using SGD

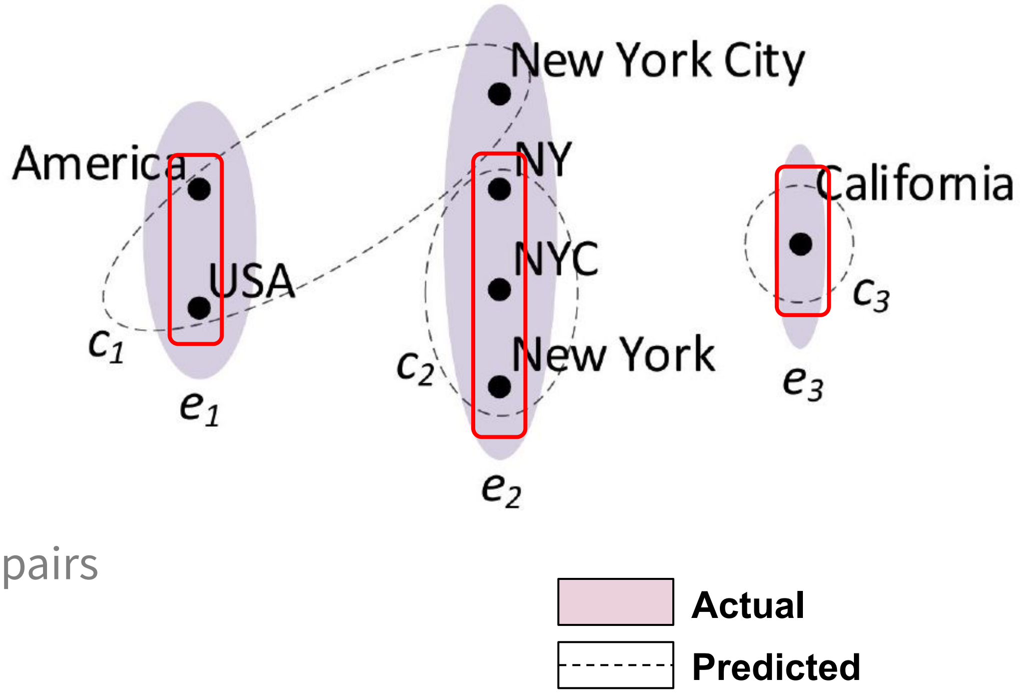
CESI Architecture



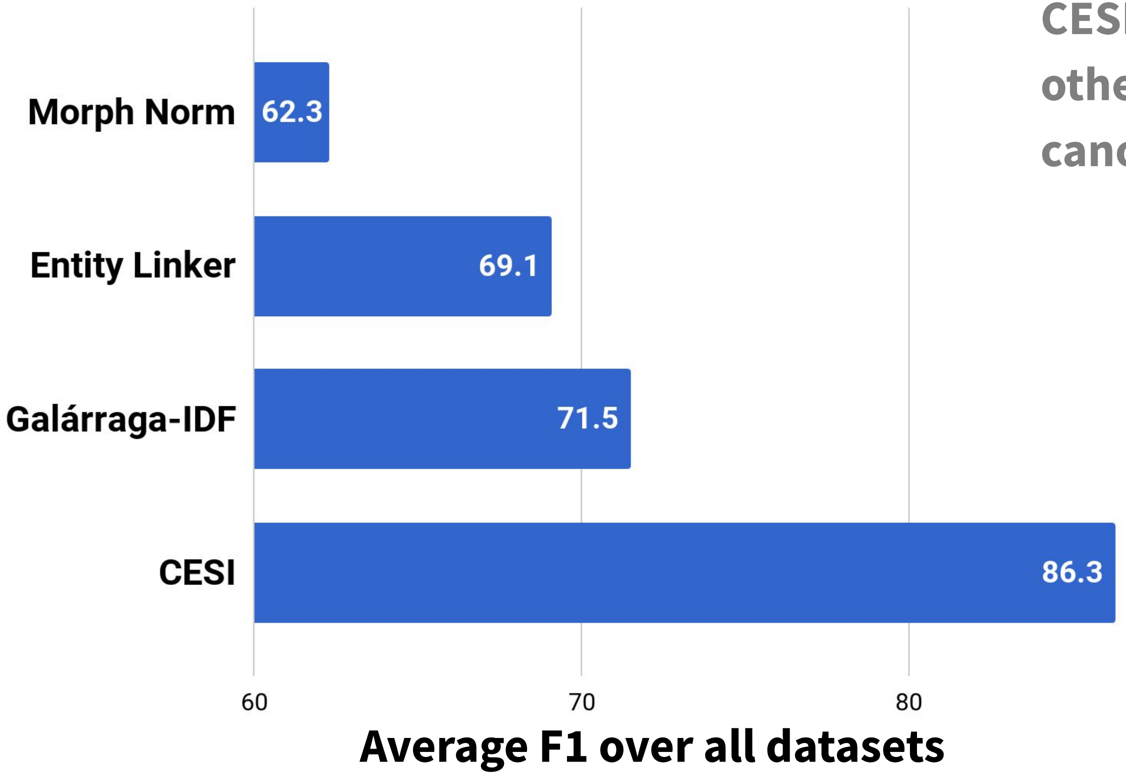
Experiments

Evaluation Metrics

- **Macro:**
  - Fraction of pure clusters
  - Precision = 2/3
- **Micro:**
  - Purity of clusters
  - Precision = 6/7
- **Pairwise:**
  - Ratio of hits to all possible pairs
  - Precision = 4/6

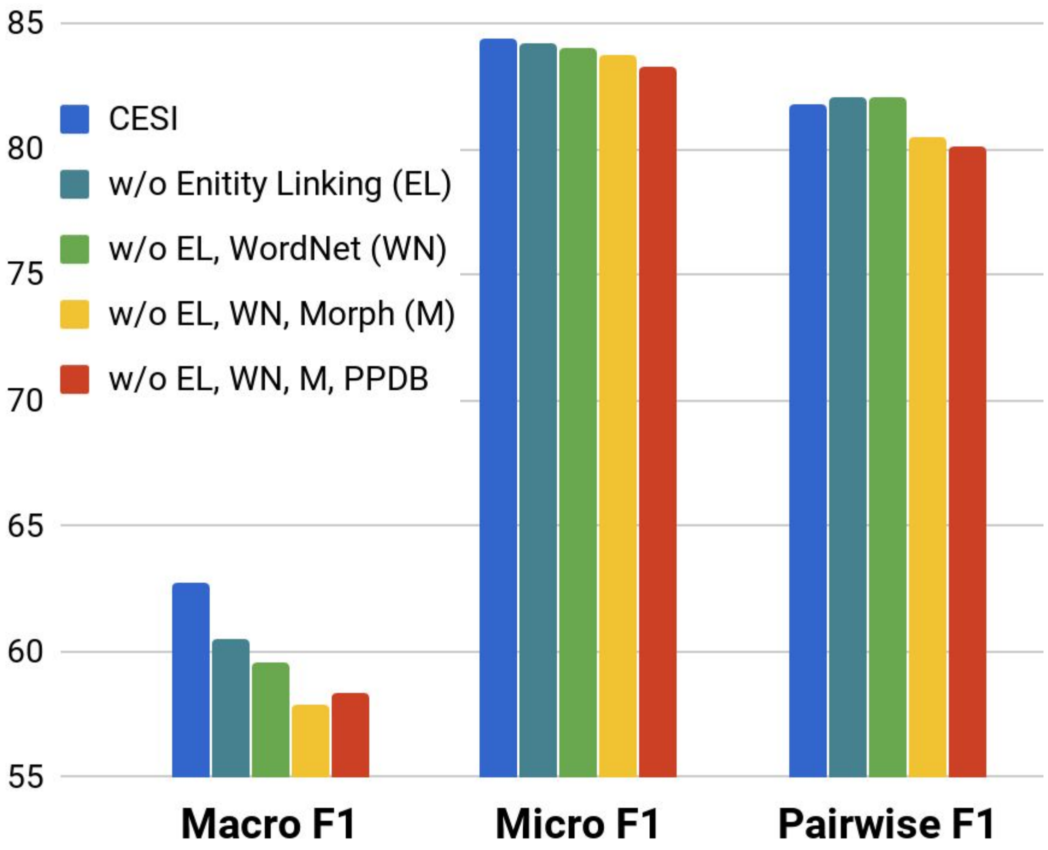


NP Canonicalization



CESI out-performs others in noun phrase canonicalization

Effect of Side Information



Side information improves performance

Relation Canonicalization

	Macro Precision	Micro Precision	Pairwise Precision	Induced Relation Clusters
		Base Dataset		
AMIE	42.8	63.6	43.0	7
CESI	**88.0**	**93.1**	**88.1**	**210**
		Ambiguous Dataset		
AMIE	55.8	64.6	23.4	46
CESI	**76.0**	**91.9**	**80.9**	**952**
		ReVerb45K		
AMIE	69.3	84.2	66.2	51
CESI	**77.3**	**87.8**	**72.6**	**2116**

CESI produces more and better relation canonicalized clusters

Qualitative Evaluation (t-sne)

Conclusion

- Canonicalization is necessary for Open KG
- Existing approaches are based on manually feature engineering which can be sub-optimal
- CESI, presents an embedding based joint noun and relation phrase canonicalization
  - Utilizes several types of side information
  - Obtains state-of-the-art results for the problem

Questions?

Contact email:
shikhar@iisc.ac.in

Source code and data are available
github.com/malllabiisc/cesi

- **References:**

  - Vashishth, Shikhar, Prince Jain, and Partha Talukdar. "CESI: Canonicalizing Open Knowledge Bases using Embeddings and Side Information." *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2018. https://arxiv.org/abs/1902.00172
  - Galárraga, Luis, et al. "Canonicalizing open knowledge bases." *Proceedings of the 23rd acm international conference on conference on information and knowledge management*. ACM, 2014. https://dl.acm.org/citation.cfm?id=2662073