Cross-lingual Entity Alignment with Incidental Supervision

Muhao Chen$^{1,2}$, Weijia Shi$^3$, Ben Zhou$^2$, Dan Roth$^2$

$^1$Viterbi School of Engineering, USC
$^2$Department of Computer and Information Science, UPenn
$^3$Department of Computer Science, UCLA
Understanding Relations Is Prominent In Practice

QA and Semantic Search

About 34,600,000 results (1.04 seconds)

787B

(?car, produced by, Mazda)
(?car, won, 24 Hours of Le Mans)
Knowledge Graphs: Precise But **Expensive** Knowledge Representation

Obtaining the structural knowledge

- Is expensive (Avg $5.71 per triple [Paulheim+, ISWC-18] in open domain; higher cost in scientific domains).
- Has relied on massive human efforts.
- Has never been close to complete.
Knowledge Is Not Isolated

Different knowledge graphs can possess complementary knowledge.

(The Tale of Genji, Genre, ?e)

DBpedia: Novel

Monogatari (story)
Love story
Royal family story
Realistic novel
Ancient literature
Problem definition

- Given two (multilingual) KGs, identifying the same entity across them

Why important?

- Allows knowledge to be combined and synchronized in different KGs
- Helps with identifying trustworthy facts in KGs
What’s New in This Work

Previous methods rely on (costly) direct supervision that is internal to KGs

- Seed alignment labels
- Entity profiles: entity descriptions, attributes, etc.

This work leverages (cheap) incidental supervision from external free text

- Connecting entities with any available mentions in free text
- Contextual similarity and induced lexical alignment serve as indirect supervision for entity alignment
- Without the need of any additional labeled data

Incidental Supervision From Free Text

Three steps
1. (Noisy) grounding: connecting KGs and text corpora
2. Embedding learning: embedding lexemes based on structures and text
3. Alignment induction: self-learning for both entity and lexical alignment
Noisy Grounding

Combining two modalities of the same language
• KG and Free text

Two choices of techniques (without additional training labels)
• Off-the-shelf EDL models [Khashabi+ 2018]: NER + entity linking
• Surface form matching: longest prefix matching with a Completion Trie [Hsu+ 2013]

High recall and noise-tolerant grounding
Embedding Learning

Jointly training two model components

\[ S^E_L = S^K_L + S^T_L \]

**KG Embedding**
- *l*-layers of GCNs
- A translational learning-to-rank model

\[
S^K_L = - \sum_{T \in G_L} \log \frac{\exp(b - |\mathbf{h} + \mathbf{r} - \mathbf{t}|)}{\sum_{\hat{T} \in G_L} \exp(b - |\hat{\mathbf{h}} + \hat{\mathbf{r}} - \hat{\mathbf{t}}|)}
\]

**Text Embedding**
- A Skip-Gram language model

\[
S^T_L = - \sum_{x \in E_L \cup W_L} \sum_{x_c \in C_x, D_L} \log \frac{\exp(d(x, x_c))}{\sum_{x_n} \exp(d(x, x_n))}
\]

Embedding based on both structural and textual contexts
Alignment Induction

Iteratively inducing alignment

In each iteration

• Obtaining the closed-form Procrustes solution

\[ S^A_{L_iL_j} = \sum_{(x_i,x_j) \in I(L_i,L_j)} \| M_{ij} x_i - x_j \|_2 \]

• Propose new alignment pairs that are **mutual nearest neighbors (NN)**

• Continue until no mutual NNs are found

Lexical alignment serves as incidental supervision signals for entity alignment
Experiments

Datasets
- **DBP15k**: alignment between KGs of 4 languages (EN, FR, JA, ZH); ~30% seed alignment in training
- **WK3I**: alignment between KGs of 3 languages (DE, EN, FR); ~20% seed alignment in training

Metrics
- Ranking metrics including MRR, Hits@k (k=1, 10)

Baselines
- 10 supervised methods (**AliNet** [Sun+ 2020] is the best performing one)
- 3 based on auxiliary information (**HMAN** [Yang+ 2019] is the best performing one with entity descriptions)
- 5 semi-supervised methods (**BootEA** [Sun+ 2018] is the representative method, and **NAEA** [Zhu+ 2019] is the best performing one)
Observations are consistent on all experimental settings

- Incidental supervision from free text effectively improve entity alignment on KGs
- Using pre-trained EDL or simple surface form matching (SFM) as grounding does not affect much the performance
Ablation Study

- Self-learning brings the most contribution
- Structural information from KGs is important
- Text information is a good addition
Conclusion

Contributions of this work
• An incidentally supervised method for entity alignment on KGs
• Instead of using (expensive) direct supervision from internal information of KGs, this work retrieves (cheap) supervision signals from external, unlabeled text
• New SOTA on benchmarks

Future directions
• Low-resource language KG construction and verification
• Application to low-resource scientific domains, e.g. pharmacy and genomics
References in the Slides

3. Hsu and Ottaviano. Space-efficient data structures for top-k completion. WWW 2013
5. Sun, et al. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. AAAI 2020
Thank You