Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding

Zequn Sun, Wei Hu*
Nanjing University, China

Chengkai Li
University of Texas at Arlington, USA
Outline

- Introduction
- Preliminaries
- JAPE – Joint Attribute-Preserving Embedding
- Evaluation
- Conclusion and Future Work
**Background**

- Knowledge bases (KBs) store rich structured real-world facts
  - Often suffer from two issues: low coverage and multi-linguality gap
  - It is both necessary and beneficial to integrate cross-lingual KBs

- **Cross-lingual entity alignment**
  - Find entities in two KBs that refer to the same real-world object
    - Each KB is labeled in a different natural language
      - e.g., “Vienna” in English vs. “维也纳” in Chinese
  - Why important?
    - Play a vital role in automatically integrating multiple KBs
    - Help construct a coherent KB
    - Enable different expressions of knowledge across diverse natural languages
Challenges for Existing Methods

- **Traditional methods**
  - rely on machine translation to eliminate the language barrier
  - costly
  - error-prone

- **Embedding-based Methods**
  - based solely on relationship triples
    - ignore attributes
  - use existing alignment as supervision
    - usually account for a small proportion
Motivation of Our Approach

- Leverage KB embedding for cross-lingual entity alignment
  - Independent of diverse natural languages

Methodology

- Map two KBs into a unified vector space
- **Structure Embedding (SE)**
  - two alignable KBs are likely to have many aligned relationship triples
- **Attribute Embedding (AE)**
  - aligned entities have high similarity in attributes

Make full use of seed alignment

- Each pair of seed alignment shares the same embedding
Outline

- Introduction
- Preliminaries
- JAPE – Joint Attribute-Preserving Embedding
- Evaluation
- Conclusion and Future Work
Knowledge Base Embedding

- Knowledge Base Embedding
  - Encode entities and relations as vectors

- Translation-based KB Embedding [Bordes et al., 2013]
  - Interpret a relationship as the translation from its head entity to its tail entity
  - Given a relationship triple \((h, r, t)\), \(h + r \approx t\) is expected.

\(h, r, t\) denote head entity, relationship and tail entity, respectively. **Boldface** denotes the corresponding vector.

\((\text{Washington}, \text{capitalOf}, \text{America})\)

Word2vec

- Learn word embeddings that capture precise syntactic and semantic word relationships [Mikolov et al., 2013].

- Skip-gram model
  - Learn word embeddings that are good at predicting the nearby words.

Outline

- Introduction
- Preliminaries
- JAPE – Joint Attribute-Preserving Embedding
- Evaluation
- Conclusion and Future Work
Framework of JAPE

- **Input**: two KBs, and seed alignment
- **Structure Embedding (SE)** models relationship structures of KBs
- **Attribute Embedding (AE)** models correlations of attributes
- **Joint**: refine SE representations by clustering entities with AE-based similarities
- **Output**: search the nearest neighbors in the embedding space
Structure Embedding

- We use the score function of TransE to measure the plausibility of relationship triples:

\[ f(tr) = \| h + r - t \|^2 \]

where \( tr \) denotes a relationship triple \((h, r, t)\).

- We minimize the following objective function:

\[ \mathcal{O}_{SE} = \sum_{tr \in T^+} \sum_{tr' \in T_{tr}^-} f(tr) - \alpha \cdot f(tr') \]
Attribute Embedding

We call a set of attributes correlated if they are commonly used together to describe an entity.

- **AE captures the correlations of attributes**
  - Given an attribute, AE predicts its correlated attributes
  - AE assigns higher correlations to the attributes that have the same range type

\[
O_{AE} = - \sum_{(a,c) \in H} w_{a,c} \cdot \log p(c|a)
\]

- \(p(c|a)\) denotes the probability that \(c\) is a correlated attribute of \(a\)
- \(w_{a,c}\) is the weight for the attribute pair \((a, c)\)
Joint Embedding

We want similar entities to be clustered to refine their embeddings:

\[ O_s = \left\| \mathbf{E}_{SE}^{(1)} - \mathbf{S}^{(1,2)} \mathbf{E}_{SE}^{(2)} \right\|_F^2 + \beta \cdot \left( \left\| \mathbf{E}_{SE}^{(1)} - \mathbf{S}^{(1)} \mathbf{E}_{SE}^{(1)} \right\|_F^2 + \left\| \mathbf{E}_{SE}^{(2)} - \mathbf{S}^{(2)} \mathbf{E}_{SE}^{(2)} \right\|_F^2 \right) \]

To preserve both the structure and attribute information of two KBs, we jointly minimize the combined objective function:

\[ O_{joint} = O_{SE} + \delta \cdot O_s \]
Outline

- Introduction
- Preliminaries
- JAPE – Joint Attribute-Preserving Embedding
- Evaluation
- Conclusion and Future Work
Evaluation

- **Gold standard:** 15K matched entity pairs from **DBpedia inter-language links**

<table>
<thead>
<tr>
<th>KBs</th>
<th>Entities</th>
<th>Relationships</th>
<th>Rel. triples</th>
<th>Attributes</th>
<th>Att. triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBP15K_{EN-ZH}</td>
<td>English</td>
<td>98,125</td>
<td>2,317</td>
<td>237,674</td>
<td>7,173</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>66,469</td>
<td>2,830</td>
<td>153,929</td>
<td>8,113</td>
</tr>
<tr>
<td>DBP15K_{EN-JA}</td>
<td>English</td>
<td>95,680</td>
<td>2,096</td>
<td>233,319</td>
<td>6,066</td>
</tr>
<tr>
<td></td>
<td>Japanese</td>
<td>65,744</td>
<td>2,043</td>
<td>164,373</td>
<td>5,882</td>
</tr>
<tr>
<td>DBP15K_{EN-FR}</td>
<td>English</td>
<td>105,889</td>
<td>2,209</td>
<td>278,590</td>
<td>6,422</td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>66,858</td>
<td>1,379</td>
<td>192,191</td>
<td>4,547</td>
</tr>
</tbody>
</table>

- **Hits@K:** the proportion of correct entities ranked in the top-k
- **Mean rank:** the mean of these ranks
- **Higher Hits@K and lower mean rank** indicate better performance
Results

- DBP15K\textsubscript{EN-ZH} with 30% supervising data
  - JAPE > MTransE > JE
  - Negative examples and attribute embedding are all useful

<table>
<thead>
<tr>
<th>Approaches</th>
<th>( \text{ZH} \rightarrow \text{EN} )</th>
<th>( \text{EN} \rightarrow \text{ZH} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hits@1</td>
<td>Hits@10</td>
</tr>
<tr>
<td>JE</td>
<td>21.27</td>
<td>42.77</td>
</tr>
<tr>
<td>MTransE</td>
<td>30.83</td>
<td>61.41</td>
</tr>
<tr>
<td>SE w/o neg.</td>
<td>38.34</td>
<td>68.86</td>
</tr>
<tr>
<td>JAPE SE</td>
<td>39.78</td>
<td>72.35</td>
</tr>
<tr>
<td>SE + AE</td>
<td>\textbf{41.18}</td>
<td>\textbf{74.46}</td>
</tr>
</tbody>
</table>

- Same conclusions on DBP15K\textsubscript{EN-JA} and DBP15K\textsubscript{EN-FR}
Results

- Sensitivity to Proportion of Seed Alignment
  - Results become better with the increase of the proportion
  - Still achieved promising results even with a very small proportion of seed alignment like 10%

(A) ZH → EN

(B) JA → EN

(C) FR → EN
Example visualization

- Aligned entities are embedded closely
- Correlated attributes are embedded closely

(A) Entity alignment

(B) Attribute correlations
Results

- Comparison between JAPE and the MT-based method
  - Employs Google Translate to translate the labels
  - Consider the lower rank of the two results as the combined rank

<table>
<thead>
<tr>
<th>Approaches</th>
<th>ZH → EN</th>
<th></th>
<th>EN → ZH</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hits@1</td>
<td>Hits@10</td>
<td>Hits@50</td>
<td>Mean</td>
</tr>
<tr>
<td>MT-based</td>
<td>55.76</td>
<td>67.61</td>
<td>74.30</td>
<td>820</td>
</tr>
<tr>
<td>JAPE</td>
<td>41.18</td>
<td>74.46</td>
<td>88.90</td>
<td>64</td>
</tr>
<tr>
<td>MT-based &amp; JAPE</td>
<td><strong>73.09</strong></td>
<td><strong>90.43</strong></td>
<td><strong>96.61</strong></td>
<td><strong>11</strong></td>
</tr>
</tbody>
</table>

- Machine translation achieves satisfying results due to the high accuracy of Google Translate
- The combined results are significantly better, which reveals the mutual complementarity between JAPE and machine translation
Outline

- Introduction
- Preliminaries
- JAPE – Joint Attribute-Preserving Embedding
- Evaluation
- Conclusion and Future Work
Conclusion and Future Work

■ Our contributions

- We propose an embedding-based approach to cross-lingual entity alignment, which does not depend on machine translation between cross-lingual KBs
- To the best of our knowledge, we are among the first to learn embeddings of cross-lingual KBs while preserving their attribute information

■ Future work

- Introduce attribute values
- Extend it for holistic alignment of entities, relations and attributes or for cross-lingual KB completion
Thank you for your time!

- This work is supported by the National Natural Science Foundation of China (Nos. 61370019, 61572247 and 61321491)
- Codes and datasets of JAPE are now available at https://github.com/nju-websoft/JAPE