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

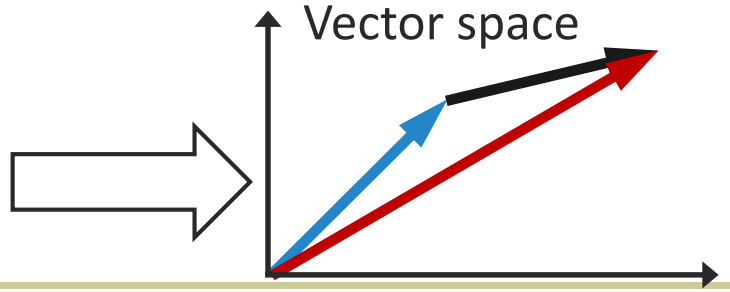
- Introduction
- **Preliminaries**
- JAPE – Joint Attribute-Preserving Embedding
- Evaluation
- Conclusion and Future Work

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), $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ is expected.

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

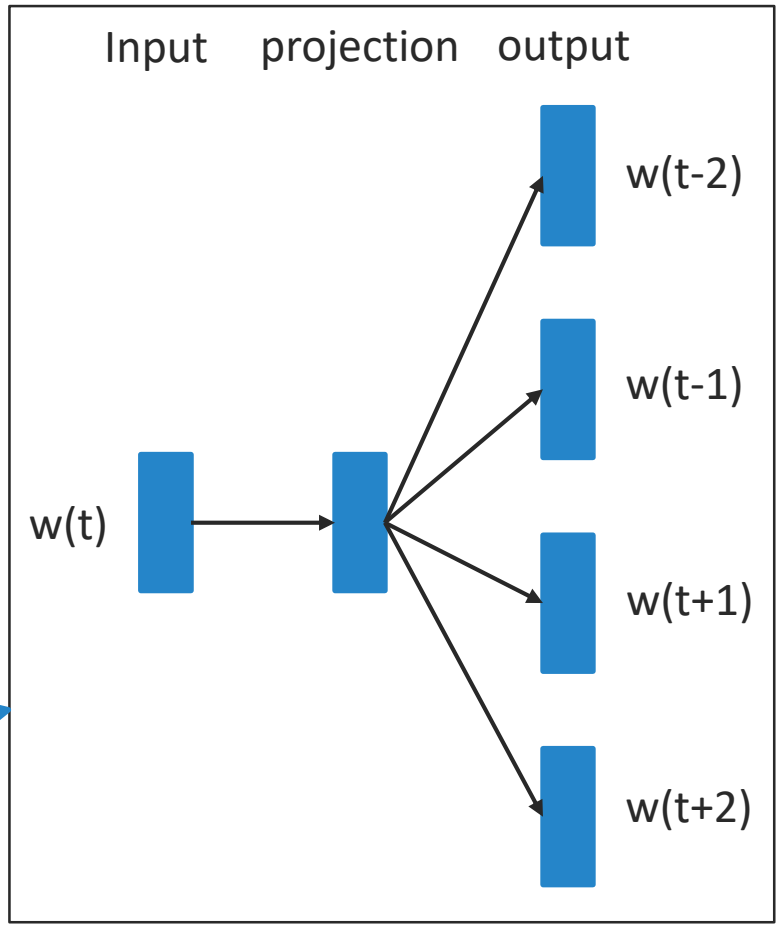
(***Washington***, ***capitalOf***, ***America***)



Bordes, Antoine, et al. "Translating embeddings for modeling multi-relational data." Advances in neural information processing systems. 2013.

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.



w(t-2), w(t-1), w(t+1), w(t+2) are the contextual words of w(t).

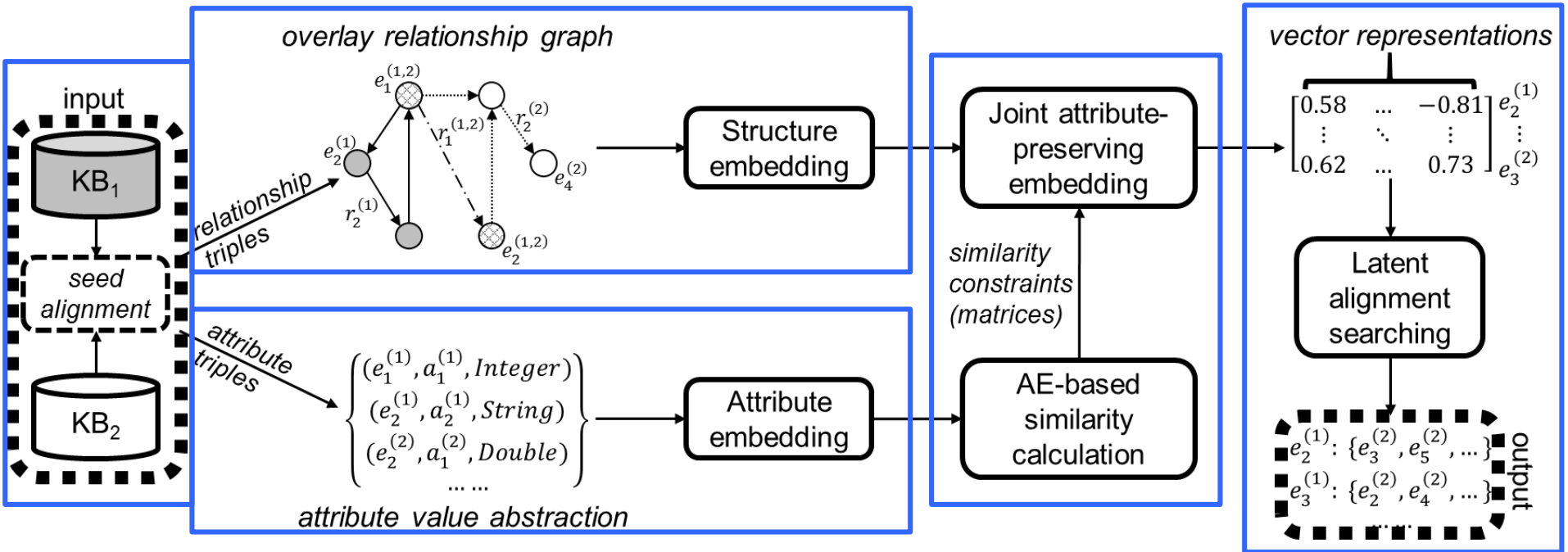
Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

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

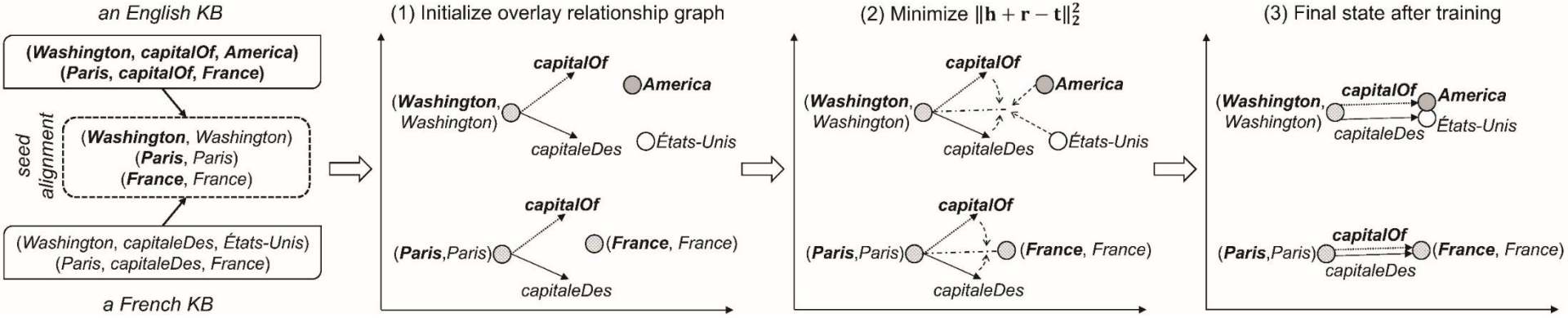
- We use the score function of TransE to measure the plausibility of relationship triples:

$$f(tr) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2$$

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

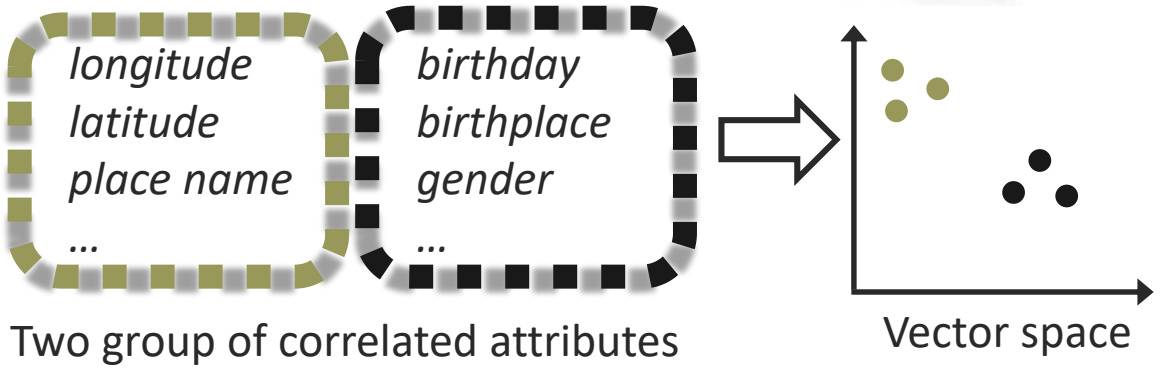
- We minimizes the following objective function:

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



Attribute Embedding

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



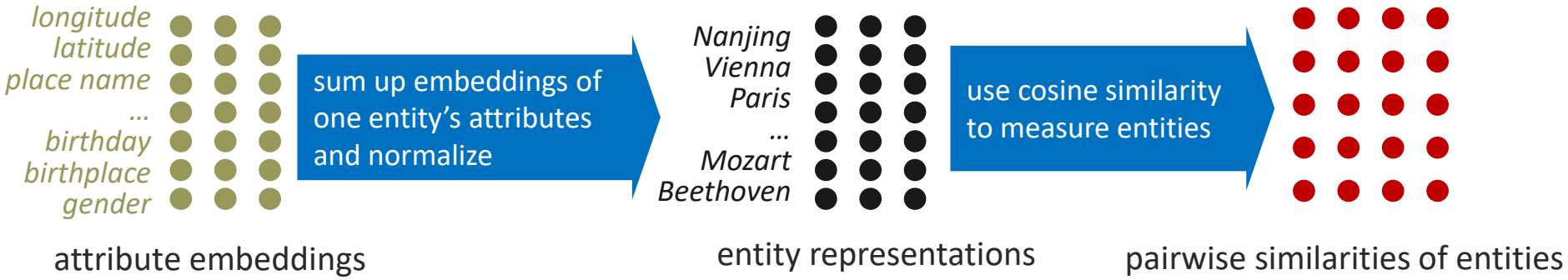
Two group of correlated attributes

- 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

KBs		Entities	Relationships	Rel. triples	Attributes	Att. triples
$DBP15K_{EN-ZH}$	English	98,125	2,317	237,674	7,173	567,755
	Chinese	66,469	2,830	153,929	8,113	379,684
$DBP15K_{EN-JA}$	English	95,680	2,096	233,319	6,066	497,230
	Japanese	65,744	2,043	164,373	5,882	354,619
$DBP15K_{EN-FR}$	English	105,889	2,209	278,590	6,422	576,543
	French	66,858	1,379	192,191	4,547	528,665

- 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

- $DBP15K_{EN-ZH}$ with 30% supervising data
  - JAPE > MTransE > JE
  - Negative examples and attribute embedding are all useful

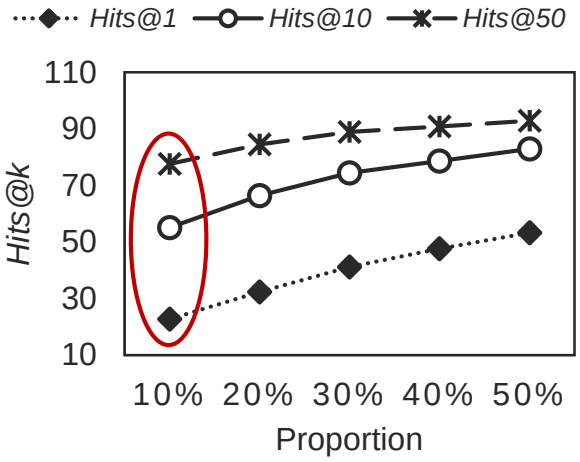
Approaches		ZH → EN				EN → ZH			
		Hits@1	Hits@10	Hits@50	Mean	Hits@1	Hits@10	Hits@50	Mean
JE		21.27	42.77	56.74	776	19.52	39.36	53.25	841
MTransE		30.83	61.41	79.12	154	24.78	52.42	70.45	208
JAPE	SE w/o neg.	38.34	68.86	84.07	103	31.66	59.37	76.33	147
	SE	39.78	72.35	87.12	84	32.29	62.79	80.55	109
	SE + AE	**41.18**	**74.46**	**88.90**	**64**	**40.15**	**71.05**	**86.18**	**73**

  - Same conclusions on $DBP15K_{EN-JA}$ and $DBP15K_{EN-FR}$

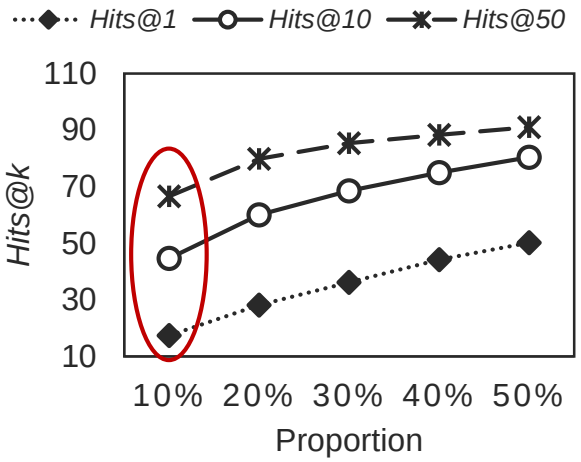
Results

- 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%

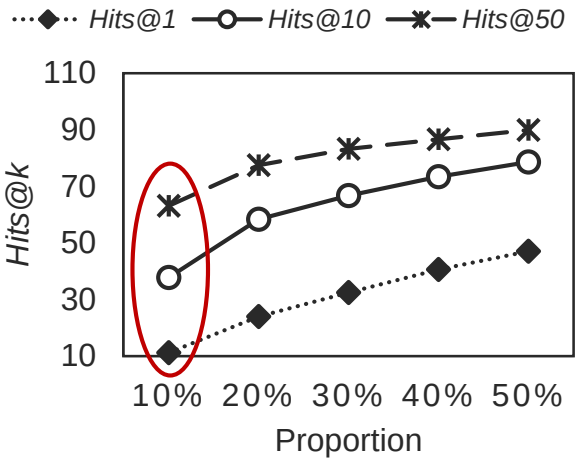
(A) ZH → EN



(B) JA → EN



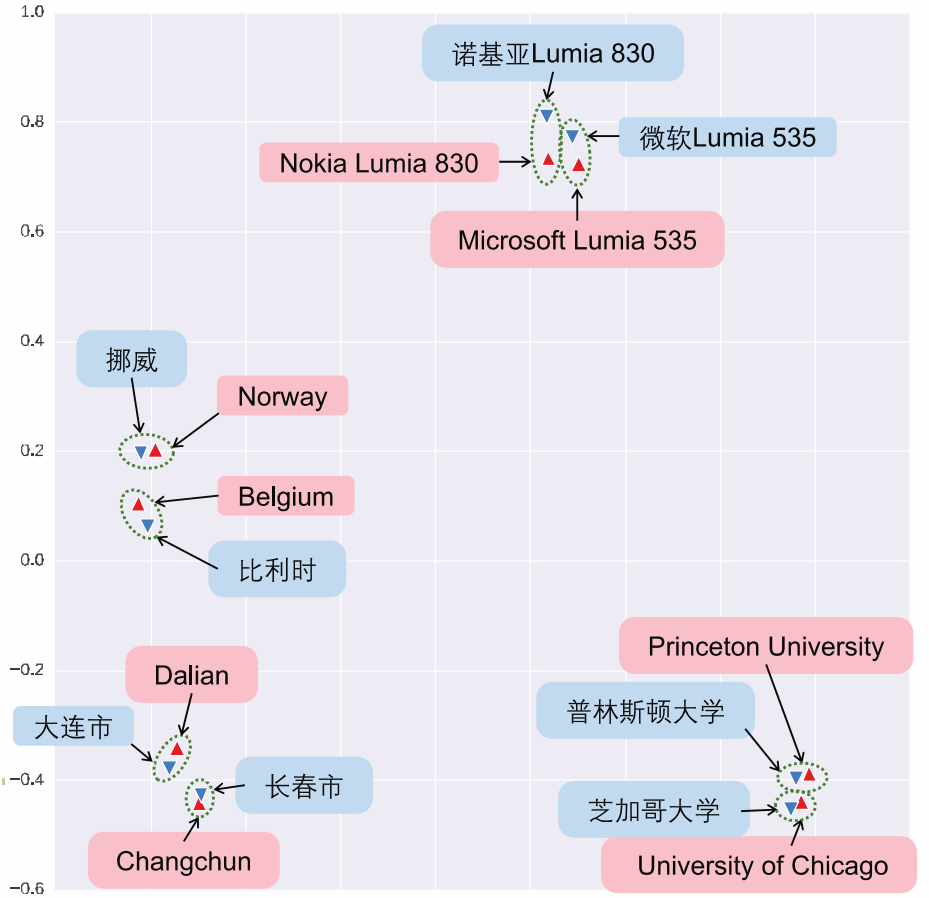
(C) FR → EN



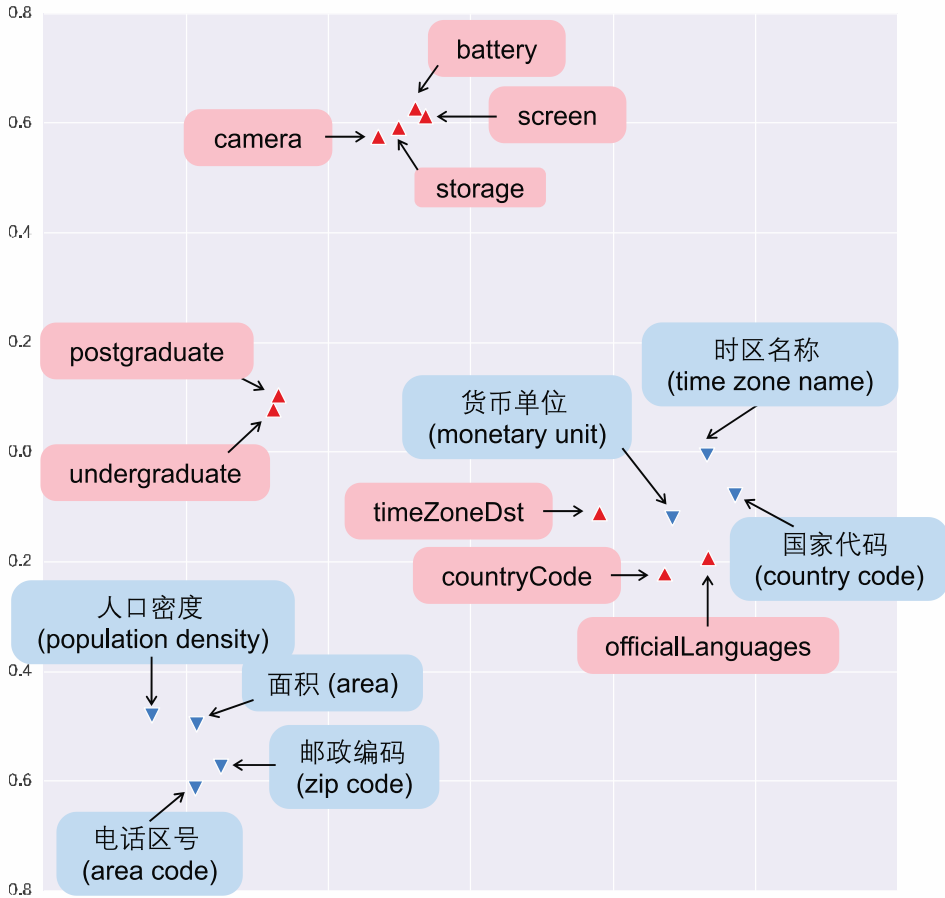
Example visualization

- 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

Approaches	ZH → EN				EN → ZH			
	Hits@1	Hits@10	Hits@50	Mean	Hits@1	Hits@10	Hits@50	Mean
MT-based	55.76	67.61	74.30	820	40.38	54.27	62.27	1,551
JAPE	41.18	74.46	88.90	64	40.15	71.05	86.18	73
MT-based & JAPE	**73.09**	**90.43**	**96.61**	**11**	**62.70**	**85.21**	**94.25**	**26**

  - 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