NANJING UNIVERSITY

南京大学万维网软件研究组

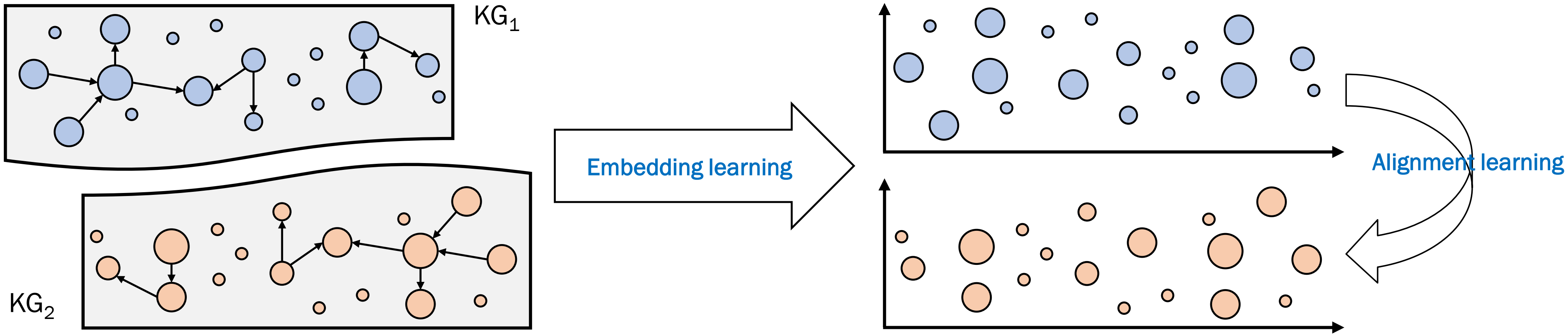
The Websoft Research Group, Nanjing University, China

Part III: Entity Alignment

Zequn Sun

Embedding-based Entity Alignment

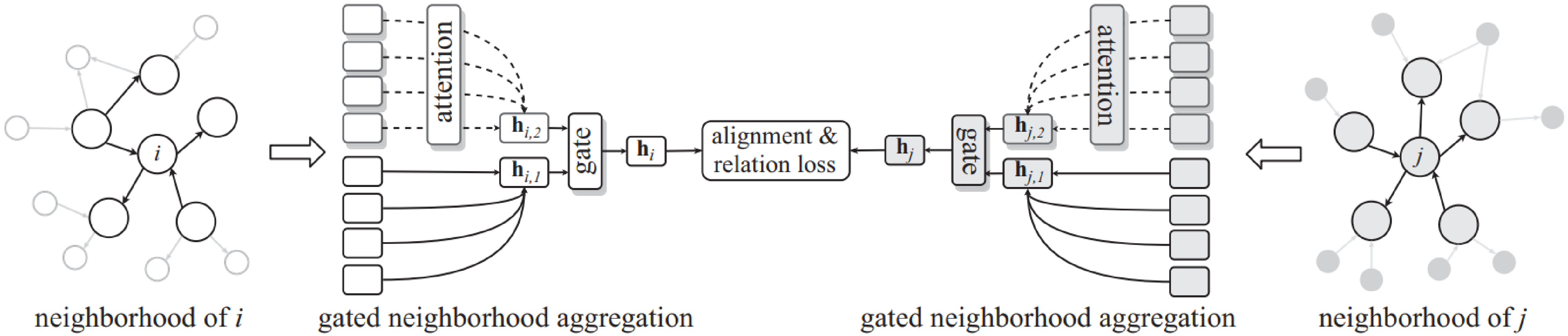
- Two modules for embedding-based entity alignment
  - Embedding learning
    - Structure-based methods, e.g., MTransE (Chen et al., IJCAI-2017)
    - Auxiliary information enhanced methods
  - Alignment learning
    - Supervised methods
    - Semi-supervised methods



Structure-based Methods: AliNet (Sun et al., AAAI-2020)

- Gated multi-hop neighborhood aggregation
- Entities with similar neighborhood subgraphs should have similar embeddings.
  - However, the counterpart entities usually have non-isomorphic neighborhood structures.

  - Gated multi-hop GNN can mitigate the non-isomorphism of neighborhood structures by attentively aggregating distant neighbor information.



Structure-based Methods: HyperKA (Sun et al., EMNLP-2020)

- Hyperbolic relational GNN
  - Basic operations of hyperbolic geometry
    - Hyperbolic distance $d_{\mathbb{D}}(\mathbf{u}, \mathbf{v})$
    - Vector translation $\mathbf{u} \oplus \mathbf{v}$
    - Transformation $\mathbf{M} \otimes \mathbf{u}$

  - HyperKA is a GNN-based model.
  - Hyperbolic relation translation at the input layer $\mathbf{M} \otimes \mathbf{u}$.
  - Hyperbolic neighborhood aggregation with highlighting input features.

Auxiliary Information Enhanced Methods: JAPE

- Attribute based entity clustering (Sun et al., ISWC-2017)
  - Aligned entities have high similarity in attributes.
  - Use a Skip-gram model to train attribute embeddings. Correlated attributes have similar embeddings.
  - The attribute-view embedding of an entity is the average of embeddings of its attributes.
  - Expect the **entities** with **similar attribute embeddings** to be **clustered**.

A set of attributes correlated if they are commonly used together to describe an entity.

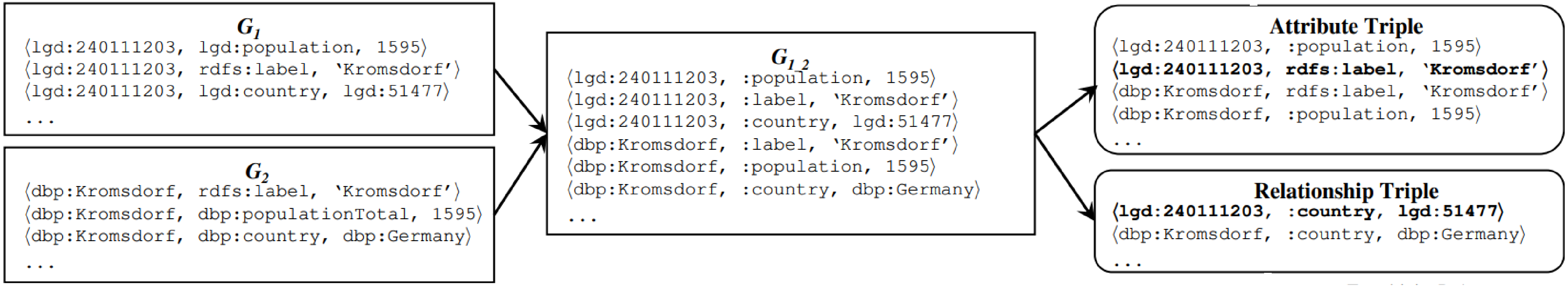
Two group of correlated attributes

Vector space

Refine entity embeddings by clustering entities with attribute-based similarities.

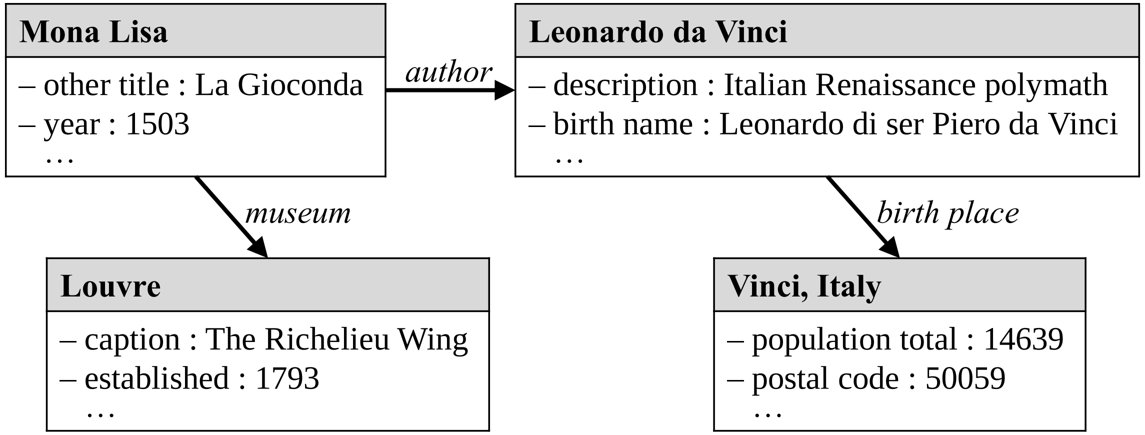
Auxiliary Information Enhanced Methods: AttrE

- Joint structure and attribute embeddings (Trisedya et al., AAAI-2019)
  - Entity embeddings learned from attribute triples also contribute to entity alignment.
  - Attribute values can be represented by pre-trained word and character embeddings using LSTM.
  - Model attribute triples through the same way of modeling relation triples.



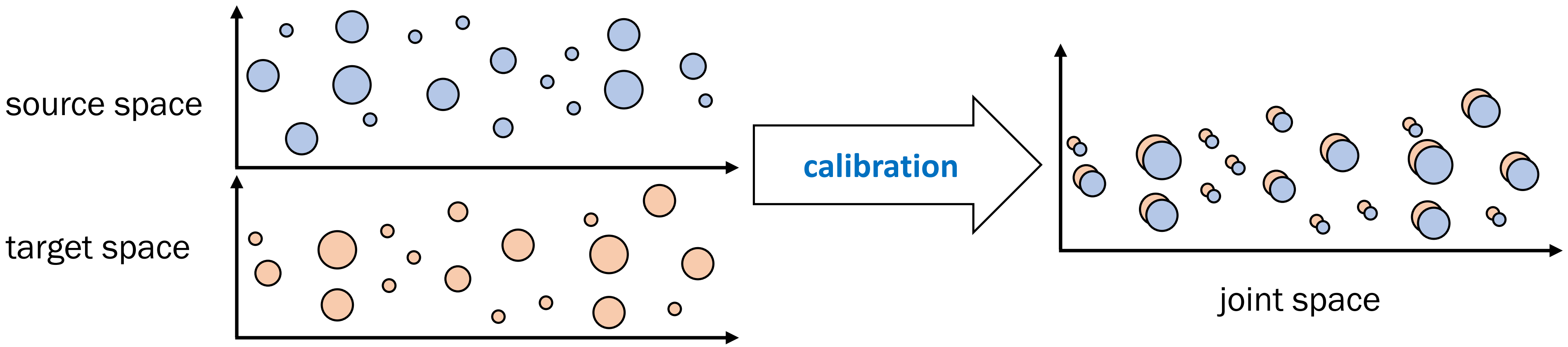
Auxiliary Information Enhanced Methods: MultiKE

- Multi-view KG embedding (Zhang et al., IJCAI-2019)
  - Entities have multi-view features, such as names, relation triples, attribute triples, etc.
  - Learn view-specific embeddings for entities.
    - Represent names using pre-trained word embeddings.
    - Encode attribute triples with CNNs.
    - Encode relation triples with TransE.
  - View combination strategies
    - Weighted view averaging
    - Shared space learning
    - In-training combination

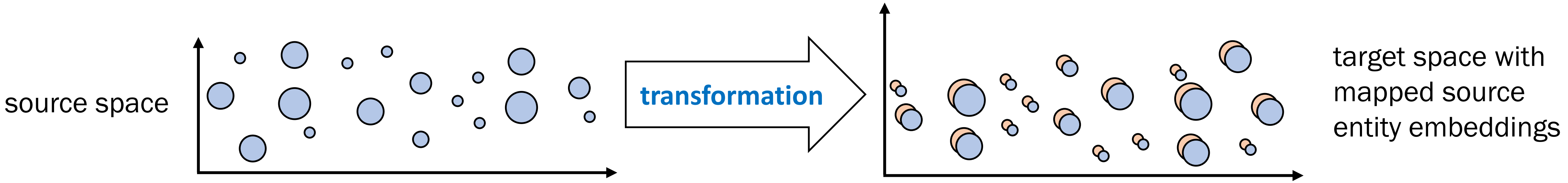


Alignment Learning: Supervised Methods

- Embedding space calibration
  - Minimize the embedding distance of pre-aligned entities



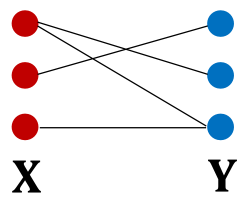
- Embedding space transformation
  - Map source entity embeddings to the target embedding space to match their counterparts.



Alignment Learning: Self-training Methods

- Bootstrapping entity alignment (Sun et al., IJCAI-2018)
  - The accessible pre-aligned entity pairs usually accounts for a small proportion.

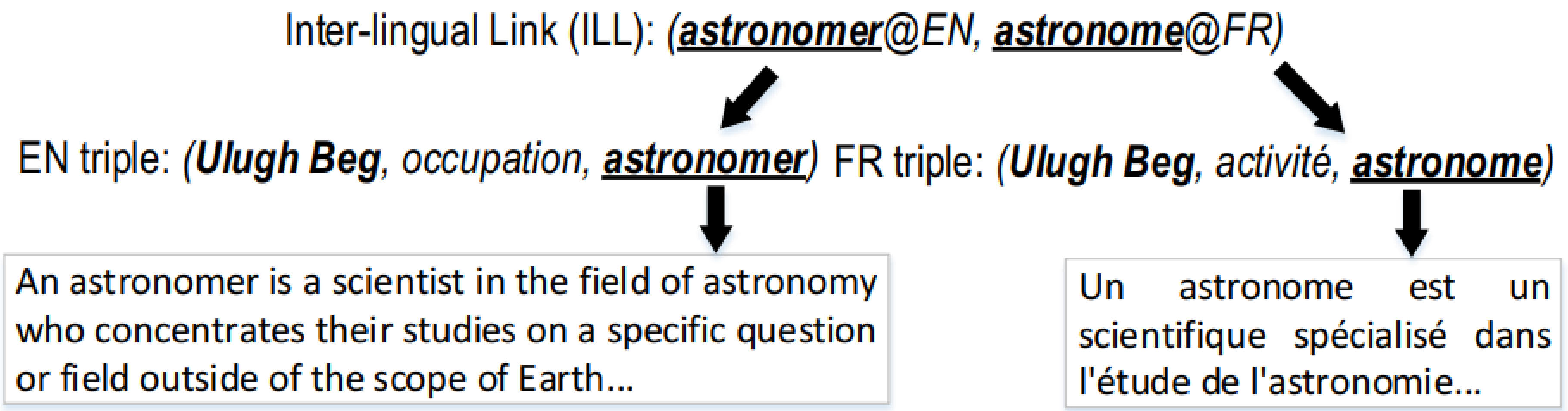
  - Iteratively label likely entity alignment as training data.
  - Bootstrapping strategies to reduce error accumulation
    - Select new alignment by solving the max-weighted matching on bipartite graphs.



    - Detect labeling conflicts when accumulating the newly-labeled alignment of different iterations.
    - Edit the new alignment using a greedy strategy
      - Choose the label with more alignment likelihood as the final label.

Alignment Learning: Co-training Methods

- Co-training embeddings of structures and descriptions (Chen et al., IJCAI-2018)
  - Learn structure embeddings by TransE.
  - Learn description embeddings by GRU with pre-trained word embeddings.
  - Alternately propose new entity alignment based on structure embeddings and description embeddings.



Entity Alignment Datasets

- DBP15K (Sun et al., ISWC-2017)
  - Three cross-lingual datasets built from the multilingual versions of DBpedia: $DBP_{ZH\text{-}EN}$ (Chinese to English), $DBP_{JA\text{-}EN}$ (Japanese to English) and $DBP_{FR\text{-}EN}$ (French to English). Each dataset contains 15 thousand reference entity alignment.

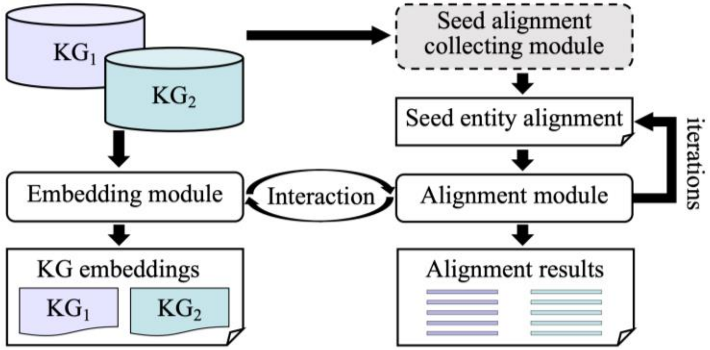
Datasets		Entities	Relationships	Attributes	Rel. triples	Attr. triples
$DBP15K_{ZH\text{-}EN}$	Chinese	66,469	2,830	8,113	153,929	379,684
	English	98,125	2,317	7,173	237,674	567,755
$DBP15K_{JA\text{-}EN}$	Japanese	65,744	2,043	5,882	164,373	354,619
	English	95,680	2,096	6,066	233,319	497,230
$DBP15K_{FR\text{-}EN}$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

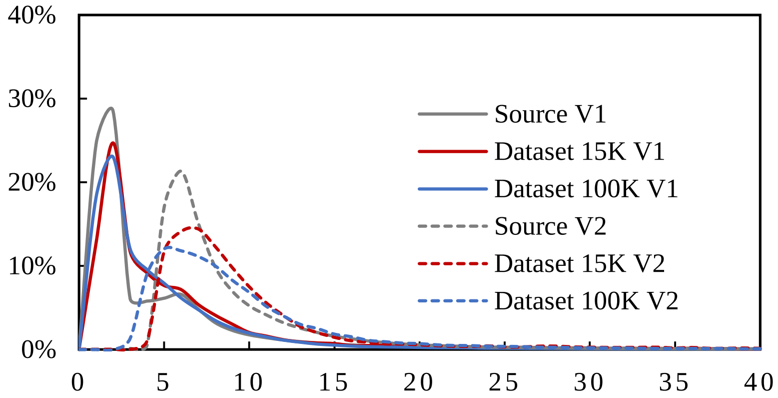
- DWY100K (Sun et al., IJCAI-2018)
  - Two large-scale datasets extracted from DBpedia, Wikidata and YAGO3, denoted by DBP-WD and DBP-YG. Each dataset has 100 thousand reference entity alignment.

Datasets		# Ent.	# Rel.	# Attr.	# Rel tr.	# Attr tr.
DBP-WD	DBpedia	100,000	330	351	463,294	381,166
	Wikidata	100,000	220	729	448,774	789,815
DBP-YG	DBpedia	100,000	302	334	428,952	451,646
	YAGO3	100,000	31	23	502,563	118,376

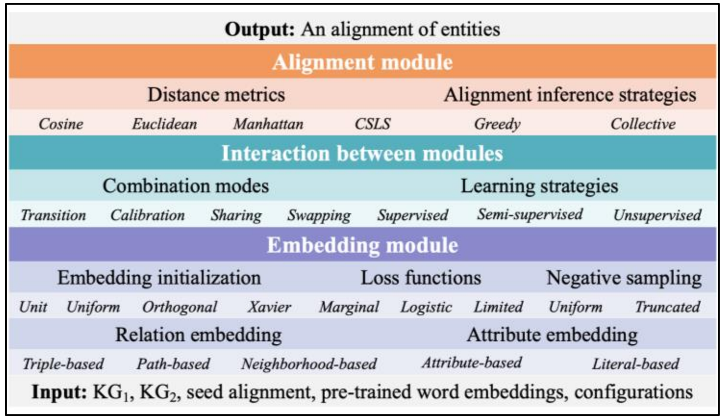
Benchmarking Study of Entity Alignment (Sun et al., VLDB-2020)

- Survey the field of embedding-based entity alignment.
- Create benchmark datasets (2.0 version is coming!).
- Conduct an experimental study of the representative approaches.
- Perform exploratory experiments for future studies.





		15K (V1) Hits@1	15K (V1) Hits@5	15K (V1) MRR	15K (V2) Hits@1	15K (V2) Hits@5	15K (V2) MRR	100K (V1) Hits@1	100K (V1) Hits@5	100K (V1) MRR	100K (V2) Hits@1	100K (V2) Hits@5	100K (V2) MRR
EN-FR	MTransE	.247 ± .006	.467 ± .009	.351 ± .007	.240 ± .005	.436 ± .007	.336 ± .005	.138 ± .002	.261 ± .004	.202 ± .002	.090 ± .003	.174 ± .003	.135 ± .003
	IPTransE	.169 ± .013	.320 ± .025	.243 ± .019	.236 ± .012	.449 ± .021	.339 ± .016	.158 ± .004	.277 ± .008	.219 ± .006	.234 ± .007	.431 ± .015	.329 ± .010
	JAPE	.262 ± .006	.497 ± .010	.372 ± .007	.292 ± .009	.524 ± .006	.402 ± .007	.165 ± .002	.310 ± .002	.240 ± .002	.125 ± .003	.239 ± .005	.183 ± .004
	KDCoE	.581 ± .004	.680 ± .004	.628 ± .003	.730 ± .007	.837 ± .006	.778 ± .005	.482 ± .005	.515 ± .006	.499 ± .005	.611 ± .012	.653 ± .015	.632 ± .014
	BootEA	.507 ± .010	.718 ± .012	.603 ± .011	.660 ± .006	.850 ± .005	.745 ± .005	.389 ± .004	.561 ± .004	.474 ± .004	.640 ± .001	.806 ± .001	.716 ± .000
	GCNAlign	.338 ± .002	.589 ± .009	.451 ± .005	.414 ± .005	.698 ± .007	.542 ± .005	.230 ± .002	.412 ± .004	.319 ± .003	.257 ± .002	.455 ± .003	.351 ± .002
	AttrE	.481 ± .010	.671 ± .009	.569 ± .010	.535 ± .015	.746 ± .014	.631 ± .014	.403 ± .019	.572 ± .019	.483 ± .019	.466 ± .011	.644 ± .012	.549 ± .011
	IMUSE	.569 ± .006	.717 ± .010	.638 ± .008	.607 ± .013	.760 ± .014	.678 ± .013	.439 ± .002	.546 ± .004	.492 ± .003	.461 ± .003	.605 ± .005	.529 ± .004
	SEA	.280 ± .015	.530 ± .026	.397 ± .019	.360 ± .018	.651 ± .018	.494 ± .017	.225 ± .011	.399 ± .013	.314 ± .012	.297 ± .002	.500 ± .002	.395 ± .002
	RSN4EA	.393 ± .007	.595 ± .012	.487 ± .009	.579 ± .006	.759 ± .006	.662 ± .006	.293 ± .004	.452 ± .006	.371 ± .004	.495 ± .003	.672 ± .005	.578 ± .004
	MultiKE	.749 ± .004	.819 ± .005	.782 ± .004	.864 ± .007	.909 ± .005	.885 ± .006	.629 ± .002	.680 ± .002	.655 ± .002	.642 ± .003	.696 ± .003	.670 ± .003
	RDGCN	.755 ± .004	.854 ± .003	.800 ± .003	.847 ± .006	.919 ± .004	.880 ± .005	.640 ± .004	.732 ± .004	.683 ± .004	.715 ± .003	.787 ± .002	.748 ± .002
EN-DE	MTransE	.307 ± .007	.518 ± .004	.407 ± .006	.193 ± .016	.352 ± .023	.274 ± .018	.140 ± .003	.264 ± .004	.204 ± .004	.115 ± .003	.215 ± .004	.168 ± .003
	IPTransE	.350 ± .009	.515 ± .012	.43 ± .011	.476 ± .012	.678 ± .011	.571 ± .010	.226 ± .014	.357 ± .019	.292 ± .017	.346 ± .013	.535 ± .016	.437 ± .014
	JAPE	.288 ± .016	.512 ± .018	.394 ± .016	.167 ± .011	.329 ± .015	.250 ± .013	.152 ± .006	.291 ± .009	.223 ± .007	.11 ± .004	.218 ± .006	.167 ± .005
	KDCoE	.529 ± .014	.629 ± .015	.580 ± .014	.649 ± .017	.788 ± .017	.715 ± .016	.506 ± .014	.591 ± .019	.549 ± .016	.651 ± .011	.756 ± .010	.701 ± .011
	BootEA	.675 ± .004	.820 ± .004	.740 ± .004	.833 ± .015	.912 ± .008	.869 ± .012	.518 ± .003	.673 ± .003	.592 ± .003	.739 ± .004	.851 ± .003	.791 ± .004
	GCNAlign	.481 ± .003	.679 ± .005	.571 ± .003	.534 ± .005	.717 ± .005	.618 ± .005	.317 ± .007	.485 ± .008	.399 ± .007	.375 ± .005	.549 ± .006	.457 ± .005
	AttrE	.517 ± .011	.687 ± .013	.597 ± .011	.650 ± .015	.816 ± .008	.726 ± .012	.399 ± .010	.554 ± .012	.473 ± .011	.464 ± .011	.637 ± .010	.546 ± .011
	IMUSE	.580 ± .017	.720 ± .014	.647 ± .015	.674 ± .011	.803 ± .008	.734 ± .010	.421 ± .005	.516 ± .005	.469 ± .005	.457 ± .005	.588 ± .007	.521 ± .006
	SEA	.530 ± .027	.718 ± .026	.617 ± .025	.606 ± .024	.779 ± .018	.687 ± .020	.341 ± .016	.502 ± .017	.421 ± .016	.447 ± .006	.625 ± .006	.532 ± .006
	RSN4EA	.587 ± .001	.752 ± .003	.662 ± .001	.791 ± .009	.890 ± .006	.837 ± .008	.430 ± .002	.57 ± .001	.497 ± .001	.639 ± .001	.763 ± .001	.697 ± .001
	MultiKE	.756 ± .004	.809 ± .003	.782 ± .003	.755 ± .008	.813 ± .008	.784 ± .007	.668 ± .002	.712 ± .002	.690 ± .001	.661 ± .004	.709 ± .004	.686 ± .004
	RDGCN	.830 ± .006	.895 ± .004	.859 ± .005	.833 ± .007	.891 ± .005	.860 ± .006	.722 ± .002	.794 ± .002	.756 ± .002	.766 ± .002	.829 ± .002	.796 ± .002



https://github.com/nju-websoft/OpenEA

Benchmarking Study of Entity Alignment (Sun et al., VLDB-2020)

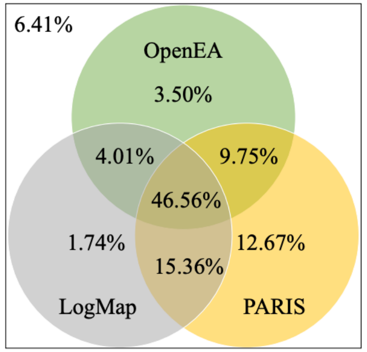
- Some conclusions from main results
  - All the relation-based approaches run better in aligning entities with rich relation triples while their results decline on long-tail entities.
  - Attribute heterogeneity has a strong effect on capturing attribute correlations, and literal embeddings facilitate entity alignment.
  - The quantity and quality of the augmented entity alignment have great impact on the semi-supervised approaches.
  - Using auxiliary information or techniques to boost performance usually increases training time and GPU memory cost.
  - Not all KG embedding models are suitable for entity alignment, and non-Euclidean embeddings are still worth further exploration.

Benchmarking Study of Entity Alignment (Sun et al., VLDB-2020)

- Comparison to conventional approaches
  - Conventional approaches better support the scenario with attribute information.
  - Embedding-based approaches cover most of the typical scenarios with either relation information, attribute information or both.
  - We find that they can produce complementary entity alignment.

	Using relation triples only			Using attribute triples only		
	Precision	Recall	F1-score	Precision	Recall	F1-score
LogMap	-	-	-	$.816_{\pm.003}$	$.723_{\pm.002}$	$.767_{\pm.001}$
PARIS	-	-	-	$.917_{\pm.000}$	$.769_{\pm.000}$	$.837_{\pm.000}$
BootEA	$.507_{\pm.010}$	$.507_{\pm.010}$	$.507_{\pm.010}$	-	-	-
MultiKE	$.337_{\pm.005}$	$.337_{\pm.005}$	$.337_{\pm.005}$	$.719_{\pm.005}$	$.719_{\pm.005}$	$.719_{\pm.005}$
RDGCN	$.255_{\pm.004}$	$.255_{\pm.004}$	$.255_{\pm.004}$	-	-	-

Comparison with conventional approaches using different features (only relation or attribute triples)



Proportions of the correct alignment found. OpenEA additionally finds 13.25% (3.50% + 9.75%) and 7.51% (3.50% + 4.01%) of the alignment that LogMap and PARIS do not

End of Part III