Representation Learning for Entity Alignment

from benchmarks to real-world data and backwards

October 2024

SESAME seminars

Konstantin Todorov

University of Montpellier

LIRMM - CNRS

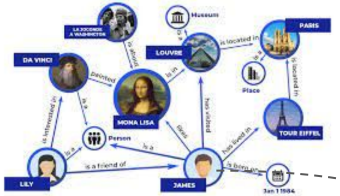
Joint work with




Ensiyeh Raoufi　Pierre Larmande　François Scharffe　Happi Bill Gates

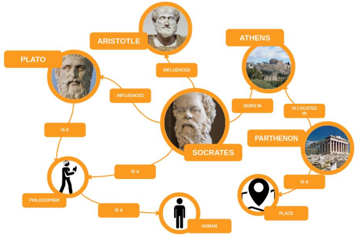
Entity Alignment (EA) across Knowledge Graphs (KG)

Establishing identity links between resources / entities / across two KGs.



source KG

owl:sameAs

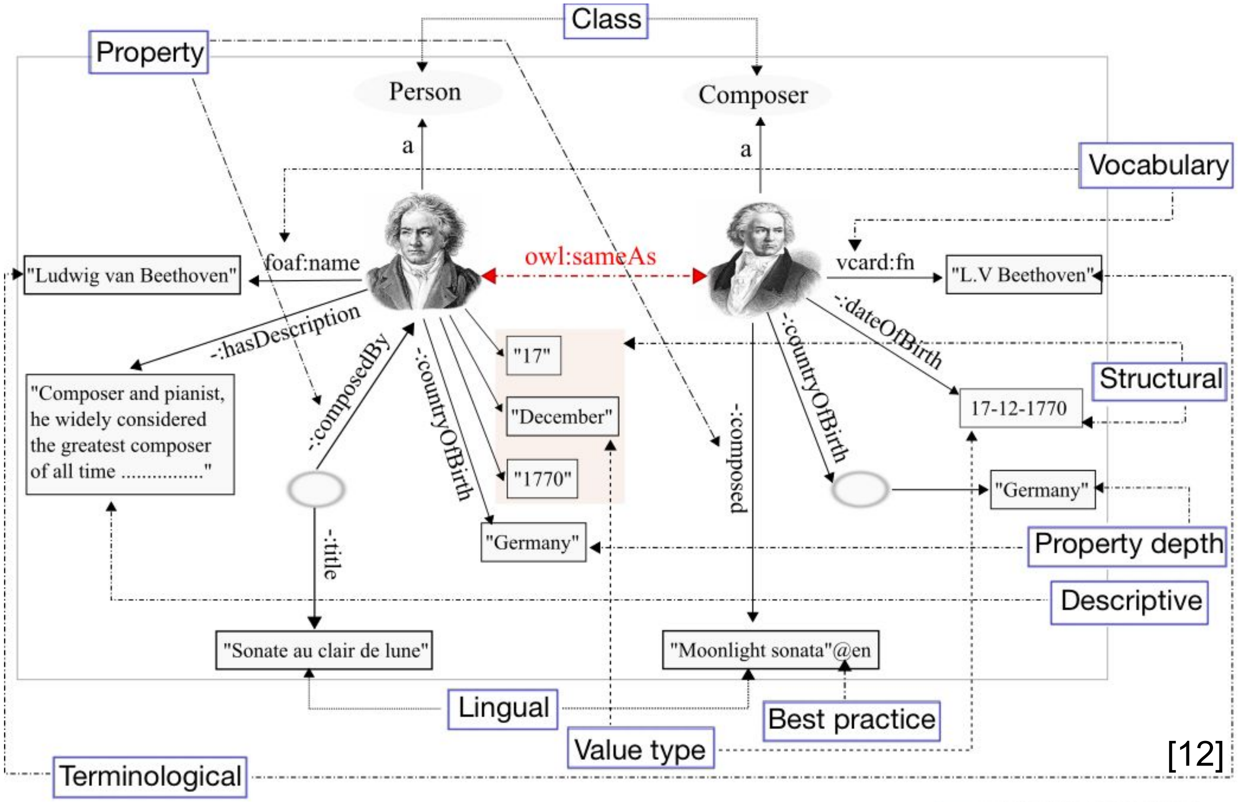


target KG

Example:

http://yago-knowledge.org/resource/Ludwig_van_Beethoven,
owl:sameAs, http://dbpedia.org/resource/Ludwig_van_Beethoven

Entity Alignment (EA) across KGs



Not an easy task for a machine…

- Datasets are **heterogenous**

- The **human effort** for system tuning and link validation is considerable

- There is a multitude of very **specific use-cases**: a challenge for generic approaches

ANR DOREMUS

Spoiler: in 2024 we still have a problem

Two SotA methods

Dataset	BERT-INT		RDGCN	
	Hit@1	Hit@10	Hit@1	Hit@10
DBP15K$_{FR-EN}$	98.8	99.7	80.9	92.6
SPIMBENCH	78.1	78.5	28.8	47.9
DOREMUS	37.1	46.7	0.00	0.3
AGROLD	18.5	28.6	0.00	0.00

Spoiler: in 2024 we still have a problem

Two SotA methods

Benchmark datasets

Real-world datasets

Dataset	BERT-INT Hit@1	BERT-INT Hit@10	RDGCN Hit@1	RDGCN Hit@10
DBP15K_{FR-EN}	98.8	99.7	80.9	92.6
SPIMBENCH	78.1	78.5	28.8	47.9
DOREMUS	37.1	46.7	0.00	0.3
AGROLD	18.5	28.6	0.00	0.00

EA: some terminology

EA: some terminology

Dataset: a pair of a source and a target KG to be interlinked, together with a reference alignment

Reference (or seed) alignment: a manually curated set of correspondences across the two KGs

Unmatchable entities: pairs of entities from the source and target KGs that refer to separate real-world entities

EA: some terminology

Synthetic benchmark dataset: generated artificially

- generated from scratch (statistical methods)
- sampling entities from existing KGs under some conditions (being sparse or dense, retaining a similar degree distribution as the KGs they are sampled from); often under the 1-to-1 assumption

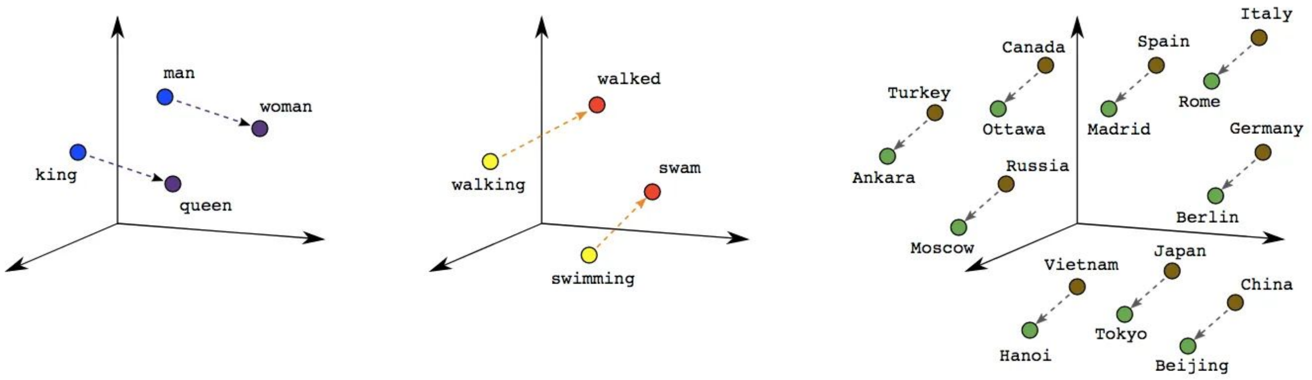
The 1:1 assumption: each source entity has exactly one match in the target graph

Real-world dataset: unchanged KGs from a real-world scenario; not sub-sampled from larger KGs

EA: some terminology

Heterogeneity: any difference in the expression of a given piece of knowledge across two KGs (be it structural, syntactical, terminological, or other) [1]

Embeddings: vector representations of data that capture relationships and similarities of things in a lower-dimensional space, learnt *by* and *for* ML.



Male-Female Verb Tense Country-Capital

EA methods

EA methods

“Traditional” methods [2]:

- user-crafted representations of entities and relations
- alignment via similarity measures or logic axioms.
- prioritizes symbolic reasoning, logical inferences and linking specifications defined by domain experts to guide the alignment process
- Examples: LogMap, DLinker

Embedding-based methods [6,7]:

- automatically learnt representations of entities and relations
- predicted alignments based on training
- ensures that corresponding entities have vectors that are close in the embedding space
- prioritizes full automation
- Examples: BERT-INT, RDGCN

EA methods

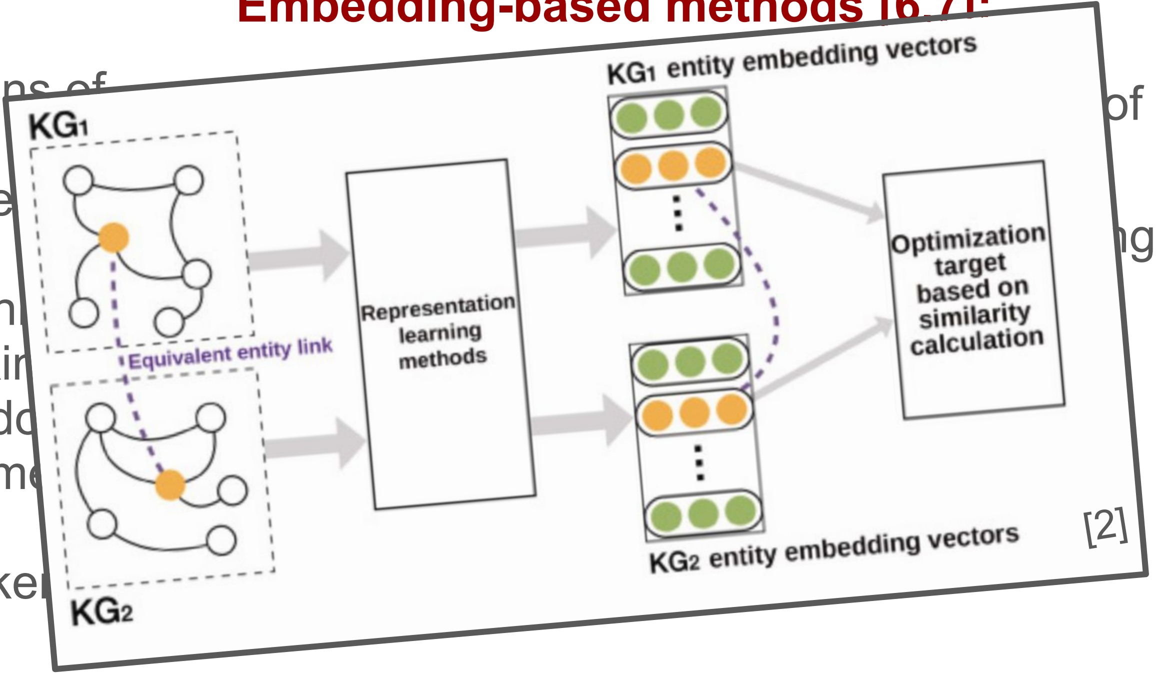
“Traditional” methods [2]:

- user-crafted representations of entities and relations
- alignment via similarity me… or logic axioms.
- prioritizes symbolic reason… logical inferences and linki… specifications defined by do… experts to guide the alignme… process
- Examples: LogMap, DLinke…

Embedding-based methods [6,7]:

… of

… ng

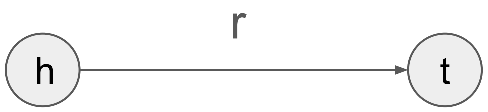


EA embedding-based methods

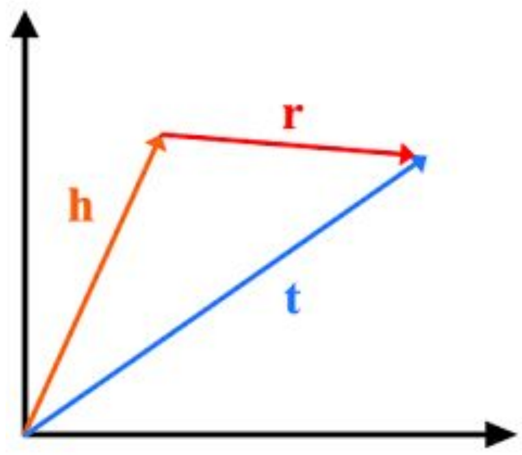
Tanslational GNN-based Graph Transformers Graph Co-training

EA embedding-based methods

Tanslational | GNN-based | Graph Transformers | Interaction models



head (h), relation (r), tail (t)



Embed a relation predicate as a translation vector from a head to a tail entity.

head + relation = tail

TransE, TransH and many variants [3]

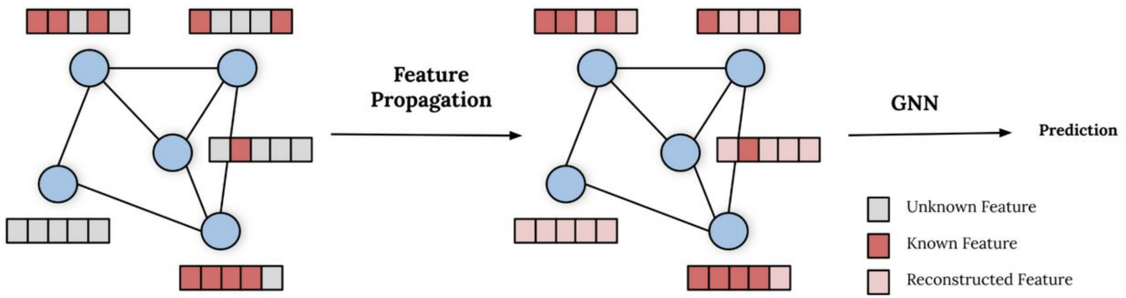
For names: literal embeddings
For attributes: convolutional neural networks.

Our pick for EA: **MultiKE [11]**

EA embedding-based methods

Tanslational GNN-based Graph Transformers Interaction models

Message passing (or feature propagation)



Graph convolution

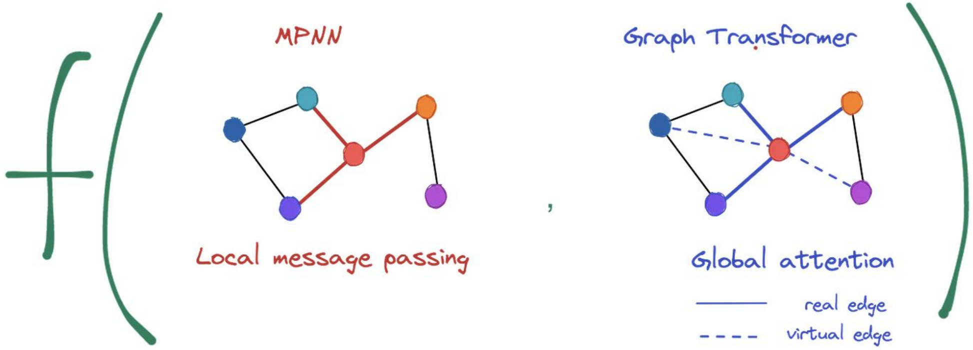
- inspired by CNNs
- run over all nodes and their neighbors
- at the end everyone knows something about everyone else

Our pick for EA: **RDGCN [8]**

EA embedding-based methods

Tanslational GNN-based **Graph Transformers** Interaction models

Enhancing GNNs' message passing by using global attention framework



Meeting point of GNNs and Transformers

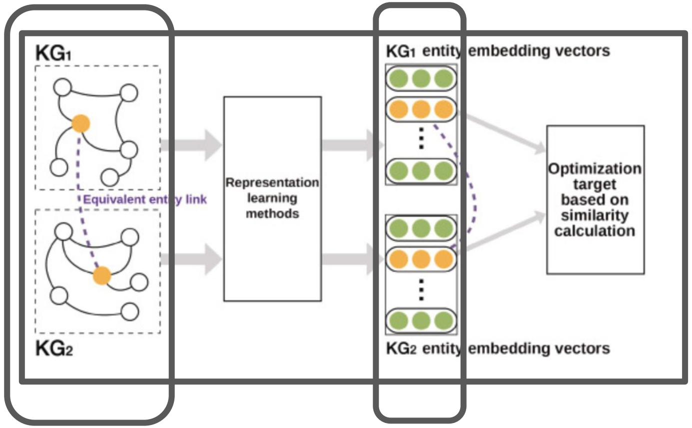
GNNs: oversmoothing, poor capturing of long-range dependencies

GT: a node's update is a function of **all nodes** in a graph, thanks to the self-attention mechanism in the Transformer layer; textual attributes are also used

Our pick for EA: **i-Align [9]**

EA embedding-based methods

Tanslational GNN-based Graph Transformers Interaction models



- BERT model: generates embeddings for entity names, descriptions, and attributes.
- Interaction model: comparisons between corresponding features across KGs (names, descriptions, neighbors, and attributes)
- Generate an interaction vector between entities, which is then used through neural networks or other techniques.

- do not need to embed entire KGs, more adaptable inference with unseen data
- insights into the correlation of features between entities across two KGs

Our pick for EA: **BERT-INT [10]**

EA embedding-based methods

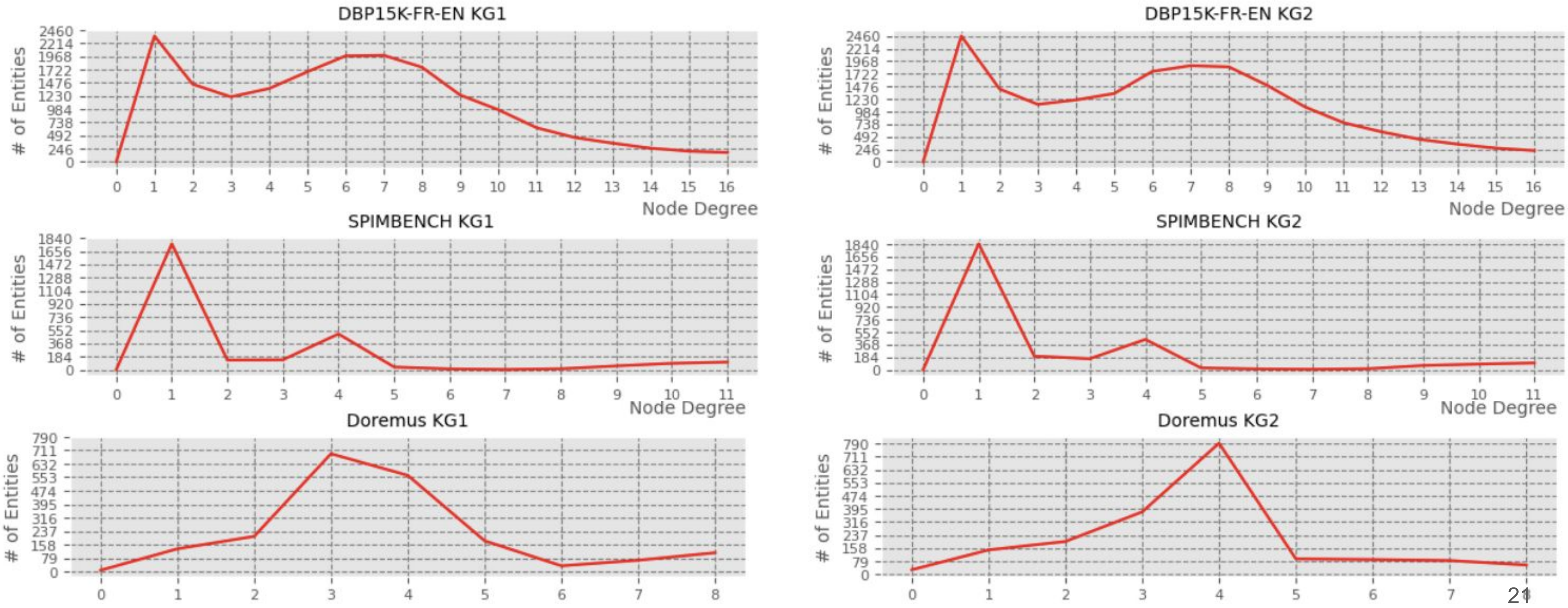
Tanslational GNN-based Graph Transformers Interaction models

Method	KG embedding approach	Best-evaluated benchmark dataset	Hit@1 on benchmark dataset	Input features		
				Relation predicate	Attribute predicate	Entity name
MultiKE	Translational	DBP_WD_100K	0.918	Relation name	Attr name/value	Entity name
RDGCN	GNN	DBP15K	0.886 (on FR-EN)	-	-	Entity name
i-Align	Graph Transformer	DBP_YG_15K	0.912	-	Attr name/value	Entity name
BERT-INT	KGs co-training	DBP15K	0.992 (on FR-EN)	Relation name	Attr name/value	Entity name/descripti

EA datasets: benchmark vs. real-world

EA datasets: benchmark vs. real-world

node degree distributions: some examples



EA datasets: benchmark vs. real-world

benchmark

benchmarks often present idealized scenarios with a limited set of relationships, controlled noise, and specific characteristics;
1:1 assumption is often the rule

real-world

contain real-world graphs with all their challenges:
degree distribution & scale differences, noise, etc.;
no 1:1 assumption

EA datasets: benchmark vs. real-world

all numbers indicate percentages except for KG Sizes which indicates the number of entities

	Dataset	JS divergence	Max difference in percentage of nodes	KG Sizes	Size similarity	Reference alignment: Levenshtein normalized similarity	Reference alignment: EA semantic similarity
benchmark	DBP15K$_{FR-EN}$	5.55	1.87	#S 19661 #T 19993	98.3	60.1	90.5
benchmark	SPIMBENCH	4.41	2.45	#S 2966 #T 3082	96.2	36.6	66.2
benchmark	ICEWS-WIKI	36.1	6.21	#S 11047 #T 15831	69.78	-	-
benchmark	ICEWS-YAGO	43.0	10.37	#S 26863 #T 22555	83.9	-	-
real-world	DOREMUS	16.8	14.0	#S 2057 #T 1889	92.8	30.3	53.4
real-world	AgroLD	6.84	6.22	#S 96117 #T 51488	53.6	19.6	43.4

EA datasets: benchmark vs. real-world

all numbers indicate percentages except for KG Sizes which indicates the number of entities

Dataset	JS divergence	Max difference in percentage of nodes	KG Sizes	Size similarity	Reference alignment	
					Levenshtein normalized similarity	EA semantic similarity
$DBP15K_{FR-EN}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	90.5
SPIMBENCH						66.2
ICEWS-WIKI						-
ICEWS-YAGO						-
DOREMUS						53.4
AgroLD			#T 51488			43.4

benchmark

real-world

The datasets show to be

- diverse
- highly heterogeneous

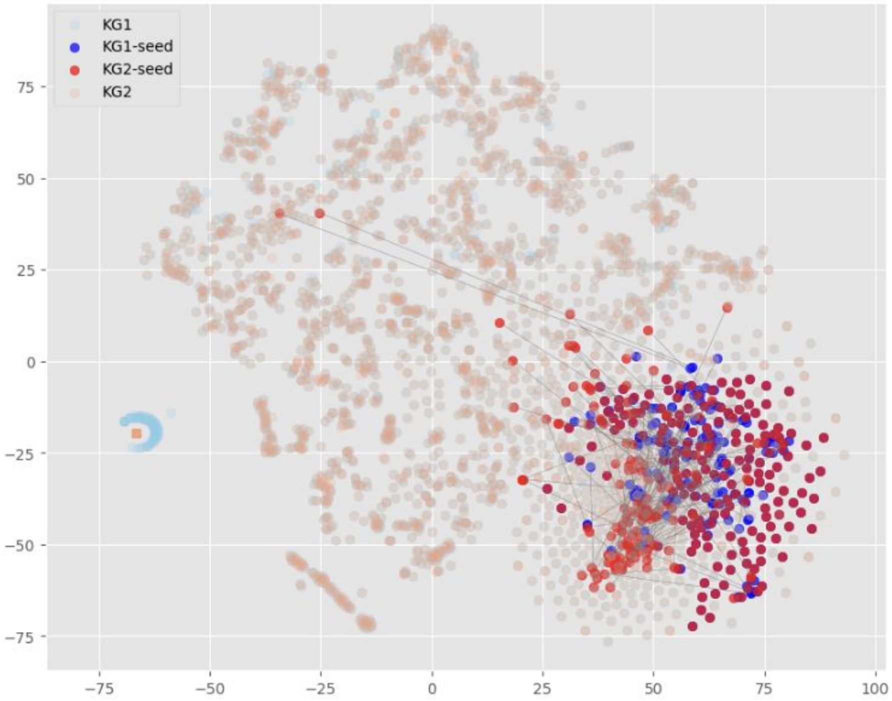
⇒ adequate insights beyond the specific choice of datasets

⇒ better understanding the challenges for the EA task when dealing with real-world datasets

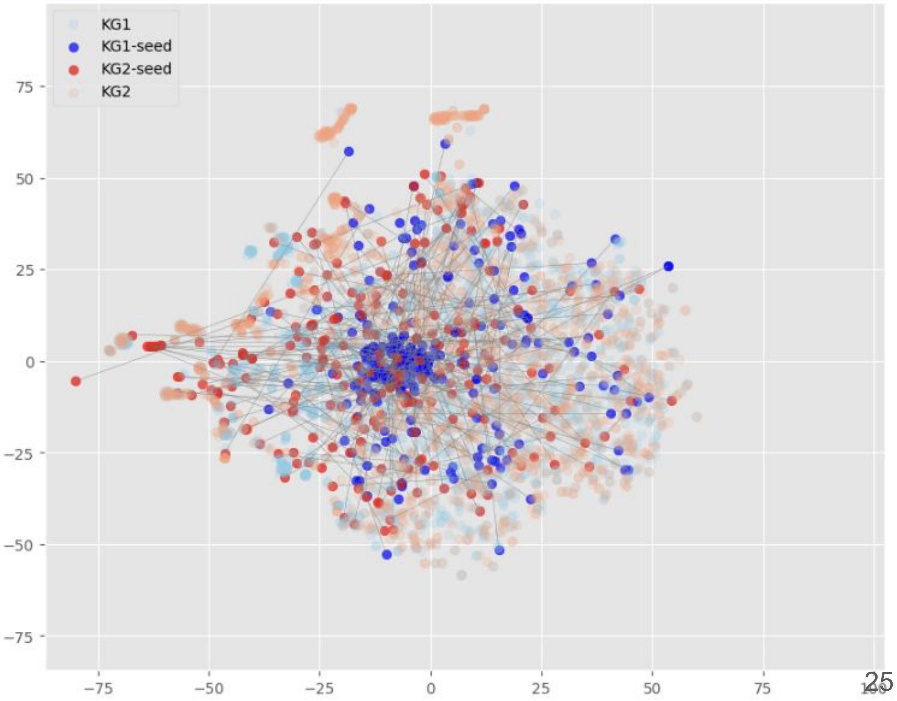
EA datasets: benchmark vs. real-world

Reduced-dimension BERT-based initial entity embeddings of SPIMBENCH (left) and DOREMUS (right).

SPIMBENCH



DOREMUS



EA models: performance analyses

EA models: performance analyses

Is there a difference in performance on benchmark data and real-world data and, if so—why?

What are the real inference capacities of embeddings-based models?

How to evaluate EA tasks correctly?

EA models: performance analyses

Evaluation metrics: two families of measures

In pre-embeddings EA

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}; \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \quad F_1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Embeddings-based methods
inspired by link prediction

$$\text{Hit@1} = \frac{\text{Number of times the top-ranked prediction is correct}}{\text{Total number of predictions}}$$

EA models: performance analyses

Evaluation metrics: two families of measures

In pre-embeddings EA

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}; \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \quad F_1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Embeddings-based methods
inspired by link prediction

$$\text{Hit@1} = \frac{\text{Number of times the top-ranked prediction is correct}}{\text{Total number of predictions}}$$

Benchmark datasets often rely on the 1:1 assumption (each source entity has exactly one match in the target graph). Under that assumption

Hit@1 is equivalent to Pr, Re and F1-score.

⚠ This is not the case when this assumption doesn’t hold (often in real-world scenarios).

EA models: performance analyses

Benchmark vs. real-world datasets

		Methods								
		BERT-INT		**RDGCN**		**MultiKE**		**i-Align**		**DLinker**
		Hit@1	Hit@10	**Hit@1**	Hit@10	**Hit@1**	Hit@10	**Hit@1**	Hit@10	F_1-score
Datasets	$DBP15K_{FR-EN}$	**99.3**	99.8	88.6*	95.7*	37.5	43.6	26.6	43.2	-
	SPIMBENCH	**82.4**	82.4	77.7	94.7	57.1	57.1	75.0	86.5	70.2
	ICEWS-WIKI	-	-	**75.1**	84.2	-	-	-	-	-
	ICEWS-YAGO	-	-	**68.3**	80.8	-	-	-	-	-
	DOREMUS	47.9	64.1	1.33	5.92	2.70	8.70	53.1	68.0	**95.6**
	AgroLD	21.1	33.2	0.02	0.3	2.30	5.7	4.4	12.1	**59.0**

EA models: performance analyses: BERT-INT

Benchmark vs. real-world datasets

		BERT-INT	
		Hit@1	Hit@10
Datasets	$DBP15K_{FR-EN}$	**99.3**	99.8
	SPIMBENCH	**82.4**	82.4
	ICEWS-WIKI	-	-
	ICEWS-YAGO	-	-
	DOREMUS	47.9	64.1
	AgroLD	21.1	33.2

BERT-INT relies heavily on the quality and amount of textual information (entity descriptions)

DOREMUS and AgroLD are datasets with less textual and semantic similarity and fewer descriptive features ⇒ decrease in performance.

This emphasizes the importance of high-quality data descriptions for BERT-INT's success.

EA models: performance analyses: RDGCN

Benchmark vs. real-world datasets

	Datasets	RDGCN Hit@1	RDGCN Hit@10
Datasets	$DBP15K_{FR-EN}$	88.6*	95.7*
	SPIMBENCH	77.7	94.7
	ICEWS-WIKI	**75.1**	84.2
	ICEWS-YAGO	**68.3**	80.8
	DOREMUS	1.33	5.92
	AgroLD	0.02	0.3

RDGCN relies on graph structure, while the real-world dataset are heterogeneous in structure.

RDGCN uses word embedding on entity names, looking up the URIs suffixes — bad idea when it comes to real-world dataset where we simply have IDs and no meaningful words.

AgroLD manifests a long-tail issue (many nodes having few neighbours and a few having many), and its graphs are bi-pirtite — both issues for GNNs [5].

EA models: performance analyses: MultiKE

Benchmark vs. real-world datasets

		MultiKE	
		Hit@1	Hit@10
Datasets	DBP15K$_{FR-EN}$	37.5	43.6
	SPIMBENCH	57.1	57.1
	ICEWS-WIKI	-	-
	ICEWS-YAGO	-	-
	DOREMUS	2.70	8.70
	AgroLD	2.30	5.7

MultiKE's performance is the weakest.

Higher level of structural and qualitative heterogeneities in DOREMUS and AgroLD than in the benchmark datasets.

MultiKE relies on both the graph structure and textual information of entities and their attributes.

EA models: performance analyses: i-Align

Benchmark vs. real-world datasets

Datasets		i-Align Hit@1	i-Align Hit@10
Datasets	$DBP15K_{FR-EN}$	26.6	43.2
	SPIMBENCH	75.0	86.5
	ICEWS-WIKI	-	-
	ICEWS-YAGO	-	-
	DOREMUS	53.1	68.0
	AgroLD	4.4	12.1

Performs better on SPIMBENCH and DOREMUS as compared to DBP15K and its performance drops significantly for AgroLD.

Only the first ten characters of the attribute values are considered by the textual transformer-based encoder

⇒ Again illustrates the importance of retaining the informative attribute descriptions included in the values.

Curse of multilinguality [4] affecting DBP15K results.

EA models: performance analyses: Baseline comparison

Benchmark vs. real-world datasets

		DLinker
		F_1-score
Datasets	DBP15K_{FR-EN}	-
	SPIMBENCH	70.2
	ICEWS-WIKI	-
	ICEWS-YAGO	-
	DOREMUS	**95.6**
	AgroLD	**59.0**

DLinker [13] does not support entity alignment on the multilingual dataset of DBP15K.

DLinker outperforms embedding-based

- using a greedy strategy that focuses on finding the longest common subsequence
- ignoring other structural or literal data that can introduce noise, especially in real-world data.

EA models: performance analyses

Benchmark vs. real-world datasets

- All embedding-based methods face the noise issue
  - worse for Translational and GNN-based methods that rely heavily on graph structures
- A local comparison of entity properties in two KGs, rather than treating them as parts of larger KGs, results in higher quality EA predictions
  - i-Align, which uses a graph transformer for embedding local subgraphs, propagates less noise compared to GNN and Translational systems
- i-Align also outperforms RDGCN and MultiKE in real-world datasets
  - focuses more on literals and textual properties
- BERT-INT and methods using extensive textual data are best for handling structurally and semantically diverse large-scale knowledge graphs
- However: difficult to find a structure-related meta-feature which justifies the performance of all methods, because each method embeds the structure from a different aspect.

EA models: performance analyses

Inference capacities: extending the candidate search space

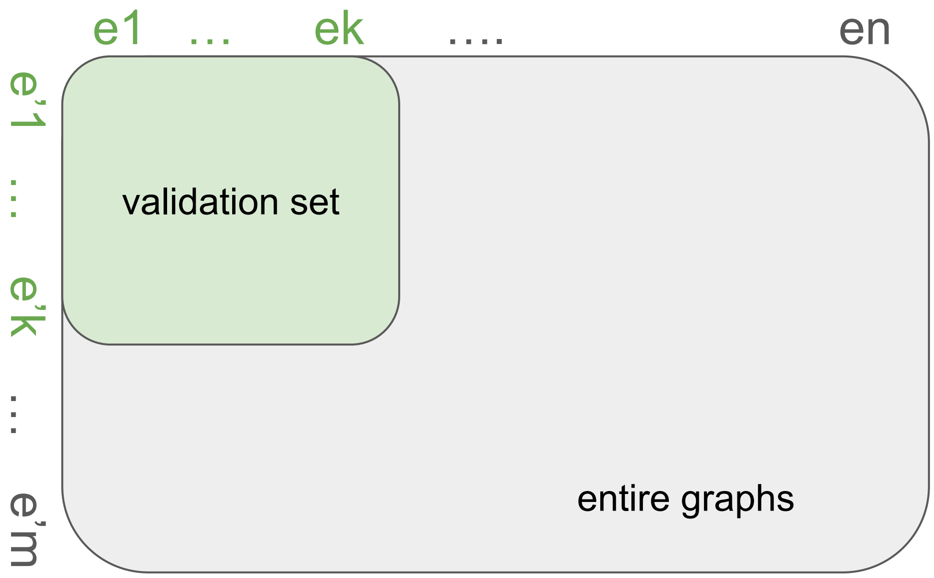
- Under the 1-to-1 assumption, models like RDGCN and BERT-INT focus only on a subset of reference alignments during evaluation
    - This ignores much of the search space, which limits the models' ability to predict correct alignments beyond the validation set
- Many EA models still focus only on ground truth data, even for training, and ignore non-matchable entities added to the dataset.

The study assesses model performance in two scenarios:

- Limited validation set (traditional approach)
- An extended scenario: all entities from the target KG are included in the candidate search space

EA models: performance analyses

Inference capacities: extending the candidate search space



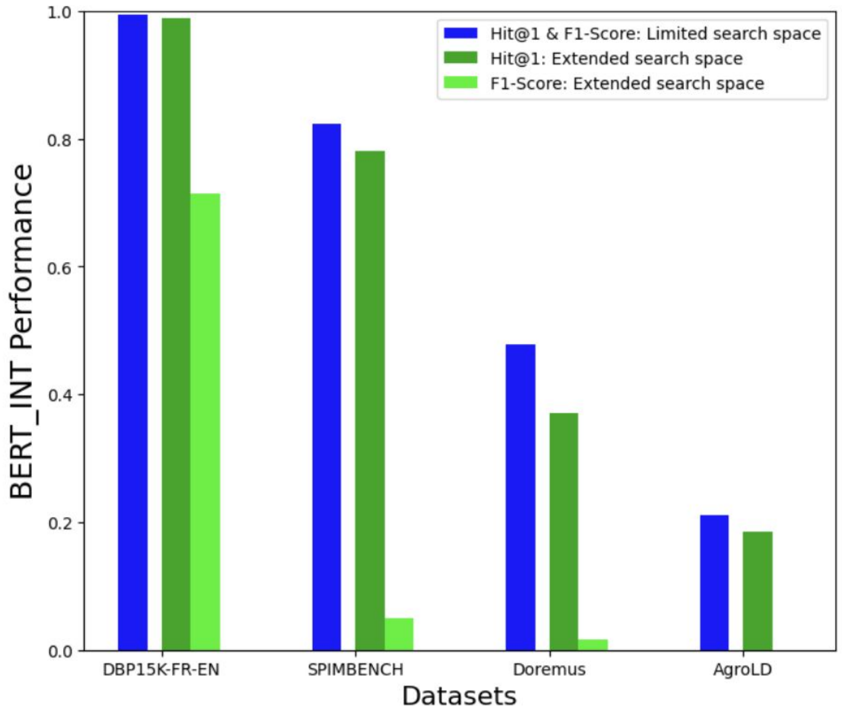
Limited case: all candidates are within the green square matrix
Extended case: candidates include the entire graphs

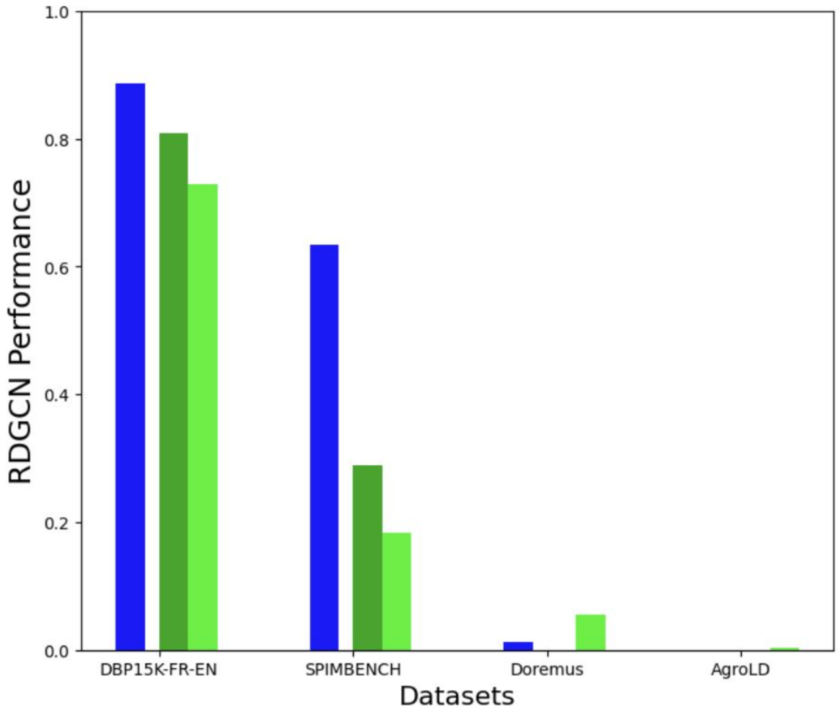
Consequence: in the extended case, if the best predicted match (or even the 10th best for Hit@10) is *not* in the validation set, then no Hit@k is recorded, i.e. best predicted match in the extended case != best predicted match in the limited case

⇒ this would come to show that the embeddings fail to discover the correct alignment on a large scale under real-world conditions

EA models: performance analyses

Inference properties: extending the search space





Key takeaways

- Focus on the challenges posed by different types of datasets and the nature of the evaluation process, highlighting the need for more robust models that can handle real-world data complexities.

- An in-depth analysis of real-world datasets, compared to popular benchmark datasets: performance drop in EA models, such as BERT-INT and RDGCN, when applied to heterogeneous real-world data.

- Benchmark overfitting, where models struggle with generalization to unseen, real-world data.

- Semantic similarity over reference alignments is correlated with the performance of EA models using language models, which helps explain performance variations.

- Interaction models are identified as a better fit for EA tasks, especially in large-scale, real-world scenarios, due to their ability to handle data heterogeneity more effectively.

References

[1] K. Todorov, Datasets First! A Bottom-up Data Linking Paradigm., in: ISWC (Satellites), 2019, pp. 338–342
[2] K. Zeng, C. Li, L. Hou, J. Li and L. Feng, A comprehensive survey of entity alignment for knowledge graphs, AI Open 2 (2021), 1–13.
[3] Bordes, N. Usunier, A. Garcia-Duran, J. Weston and O. Yakhnenko, Translating embeddings for modeling multi-relational data, Advances in neural information processing systems 26 (2013).
[4] Wang, D. Wang, H. Liu, B. Hu, Y. Yan, Q. Zhang and Z. Zhang, Optimizing Long-tailed Link Prediction in Graph Neural Networks through Structure Representation Enhancement, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD ’24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 3222–3232
[5] Giamphy, J.-L. Guillaume, A. Doucet and K. Sanchis, A survey on bipartite graphs embedding, Social Network Analysis and Mining 13 (2023). doi:10.1007/s13278-023-01058-z.
[6] Zhang, B.D. Trisedya, M. Li, Y. Jiang and J. Qi, A benchmark and comprehensive survey on knowledge graph entity alignment via representation learning, The VLDB Journal 31(5) (2022), 1143–1168.
[7] N. Fanourakis, V. Efthymiou, D. Kotzinos and V. Christophides, Knowledge Graph Embedding Methods for Entity Alignment: An Experimental Review, Data Mining and Knowledge Discovery 37(5) (2023), 2070–2137
[8] Wu, Y., Liu, X., Feng, Y., Wang, Z., Yan, R., & Zhao, D. (2019). Relation-aware entity alignment for heterogeneous knowledge graphs. *arXiv preprint arXiv:1908.08210*.
[9] Trisedya, B. D., Salim, F. D., Chan, J., Spina, D., Scholer, F., & Sanderson, M. (2023). i-Align: an interpretable knowledge graph alignment model. *Data Mining and Knowledge Discovery*, *37*(6), 2494-2516.
[10] Tang, X., Zhang, J., Chen, B., Yang, Y., Chen, H., & Li, C. (2020). BERT-INT: A BERT-based interaction model for knowledge graph alignment. *interactions*, *100*, e1.
[11] Zhang, Q., Sun, Z., Hu, W., Chen, M., Guo, L., & Qu, Y. (2019). Multi-view knowledge graph embedding for entity alignment. *arXiv preprint arXiv:1906.02390*.
[12] Achichi, M., Bellahsene, Z., Ellefi, M. B., & Todorov, K. (2019). Linking and disambiguating entities across heterogeneous RDF graphs. *Journal of Web Semantics*, *55*, 108-121.
[13] B.G. Happi Happi, G. Fokou Pelap, D. Symeonidou and P. Larmande, DLinker Results for OAEI 2022, in: 17th International Workshop on Ontology Matching, OM 2022, CEUR Workshop Proceedings, Vol. 3324, CEUR-WS, 2022, pp. 166–173.

Thank you for listening.

ANR **DACE-DL**
https://dace-dl.github.io/