

SARA: Semantic-Assisted Reinforced Active Learning for Entity Alignment

Dr. Chuan-Ju Wang

Research Fellow

CITI, Academia Sinica

Joint work with Ching-Hsuan Liu, Dr. Chih-Ming Chen, Dr. Jing-Kai Lou, Prof. Ming-Feng Tsai, Prof. Jiun-Lang Huang

Outline

Introduction

- Entity Alignment (EA)
- Nowadays EA Challenges
- Contributions

Proposed Method

- Model Framework
- Model Training Procedures

Experiments

- Datasets and Settings
- Fine-tuning Results
- Semantic-assisted Loss
- Ablation Study

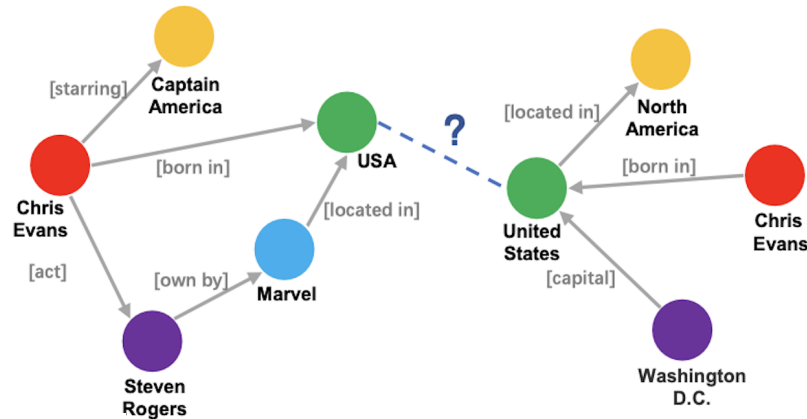
Conclusions

Introduction

- Entity Alignment (EA)
- Nowadays EA challenges
- Contributions

Entity Alignment

- Entity Alignment (EA) is a crucial task in knowledge fusion, aimed at **matching equivalent entities across different knowledge graphs**.
- Linking entities from multiple databases or knowledge graphs enhances the scope and representation of the entire knowledge graph, creating **a more comprehensive and unified knowledge representation**.
- Goal: **Matching equivalent entities across different knowledge graphs**.



Although the current graph-based EA methods have significantly progressed, they still encounter challenges when applied to real-world scenarios.

Challenges and Opportunities

Heavy reliance on labeled data

- State-of-the-art models often **require large amounts of pre-aligned seed alignments**, posing practical limitations.

Underutilization of advanced language models

- Existing methods primarily use **shallow neural-network-based techniques** (e.g., GloVe, fastText) and could benefit from deep contextualized language models like BERT, T5, and GPT.

→ Currently, there are no methods to overcome the two above challenges simultaneously.

→ Opportunities for improvement lie in addressing these challenges to enhance the quality and performance of EA.

Contributions

To overcome challenges in EA, we present the [Semantic-Assisted Reinforced Active Learning \(SARA\)](#) framework. SARA enhances EA [under limited supervision scenarios](#) by combining [reinforced active learning](#) with [the utilization of semantic information](#).

- We propose SARA, an EA framework that combines [structural and semantic information](#) to improve alignment performance with limited labeled data.
- We investigate using [deep contextualized language models](#) for learning embeddings that capture the semantic aspects of entity names.
- We introduce [a semantic-assisted alignment loss](#) that integrates structural and semantic information for improved model learning.
- Extensive experiments on benchmark datasets and a real-world dataset demonstrate the superiority of our method over state-of-the-art approaches in EA.

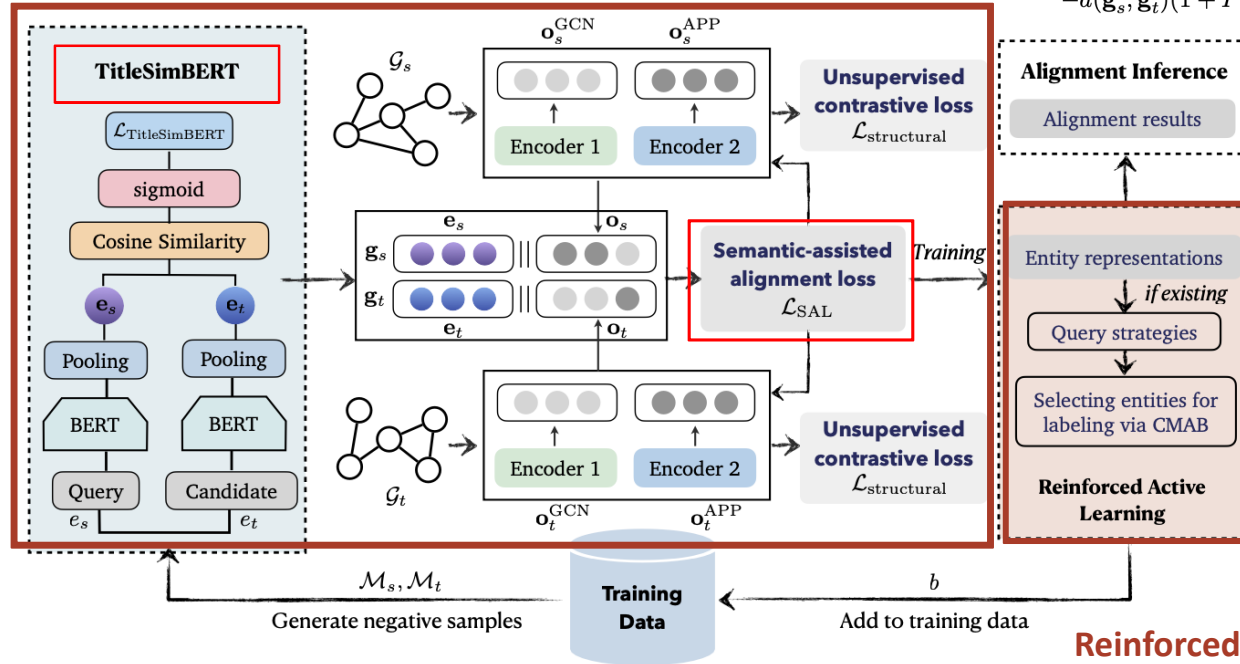
Proposed Method

- Model framework
- Model training procedures

Model Framework

Semantic-assisted Model Training

$$\mathcal{L}_{SAL} = \sum_{(s,t) \in L} \sum_{(s',t') \in L'_{(s,t)}} \max(d(\mathbf{g}_s, \mathbf{g}_t) + \gamma - d(\mathbf{g}'_s, \mathbf{g}'_t)(1 + P(\mathbf{e}'_s, \mathbf{e}'_t)), 0)$$



Reinforced Active Learning

Figure 4.1: The framework of our proposed SARA

$$\mathcal{L}_{\text{TitleSimBERT}} = -\frac{1}{M} \sum_{i=1}^M y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i), \quad (2)$$

$$\mathcal{L}_{\text{SAL}} = \sum_{(s,t) \in L} \sum_{(s',t') \in L_{(s,t)}} \max(d(\mathbf{g}_s, \mathbf{g}_t) + \gamma, -d(\mathbf{g}'_s, \mathbf{g}'_t)(1 + P(e'_s, e'_t)), 0), \quad (4)$$

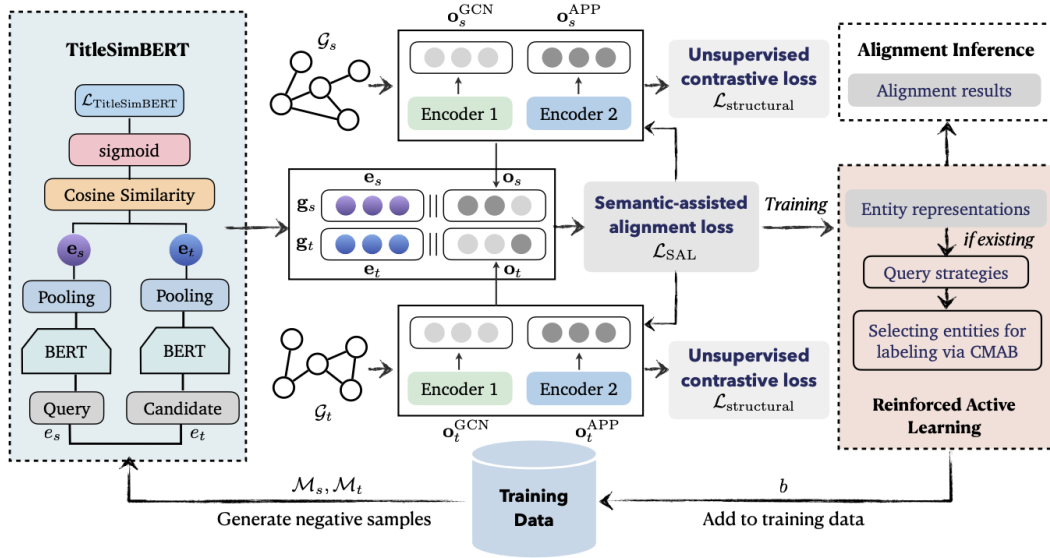


Figure 4.1: The framework of our proposed SARA

$$\mathcal{L} = \lambda_a \mathcal{L}_{\text{SAL}} + \lambda_b \mathcal{L}_{\text{structural}}, \quad (6)$$

Input: \mathcal{G}_s and \mathcal{G}_t : source and target knowledge graphs;
 L_0 : the set of initially labeled pairs (seed aligned pairs); B : labeling budget

Output: \mathcal{A} : the set of aligned entity pairs

- 1 $L = L_0$; **while** $|L| - |L_0| < B$ **do**
- /* (1) Reinforced active learning */
- 2 Generate query strategies (using entity representations); Use MAB to select b entities and use them to generate a set of labeled entity pairs L_b ; $L \leftarrow L \cup L_b$;
- /* (2) Fine-tuning the TitleSimBERT module for learning name embedding */
- 3 **for** $(s, t) \in L$ **do**
- $\mathcal{M}_s \leftarrow$ Generate a set of negative samples for s ;
- $\mathcal{M}_t \leftarrow$ Generate a set of negative samples for t ; Build the training data $L_{\text{TitleSimBERT}}$ from \mathcal{M}_s , \mathcal{M}_t , and L (i.e., for each positive sample $(s, t) \in L$, we have (s, t') , (s', t) for all $s' \in \mathcal{M}_s$ and $t' \in \mathcal{M}_t$ as negative samples;
- 6 Use Eq. (2) to update the parameters in TitleSimBERT with $L_{\text{TitleSimBERT}}$;
- 7 $\mathbf{e}_{x_i} \leftarrow$ Generate name embedding for each entity $x_i \in \mathcal{E}_s \cup \mathcal{E}_t$ with TitleSimBERT;
- /* (3) Semantic-assisted model training */
- 8 $\mathbf{o}_{x_i} \leftarrow$ Generate structural embedding for each entity $x_i \in \mathcal{E}_s \cup \mathcal{E}_t$ using graph encoders;
- 9 Calculate \mathcal{L}_{SAL} using L via Eq. (4);
- 10 Calculate $\mathcal{L}_{\text{structural}}$ via Eq. (3);
- 11 Calculate \mathcal{L} via Eq. (6);
- 12 Infer the results \mathcal{A} using the learned embeddings;
- 13 **return** \mathcal{A} ;

Experiments

- Datasets and Settings
- Fine-tuning Results
- Semantic-assisted Loss
- Ablation Study

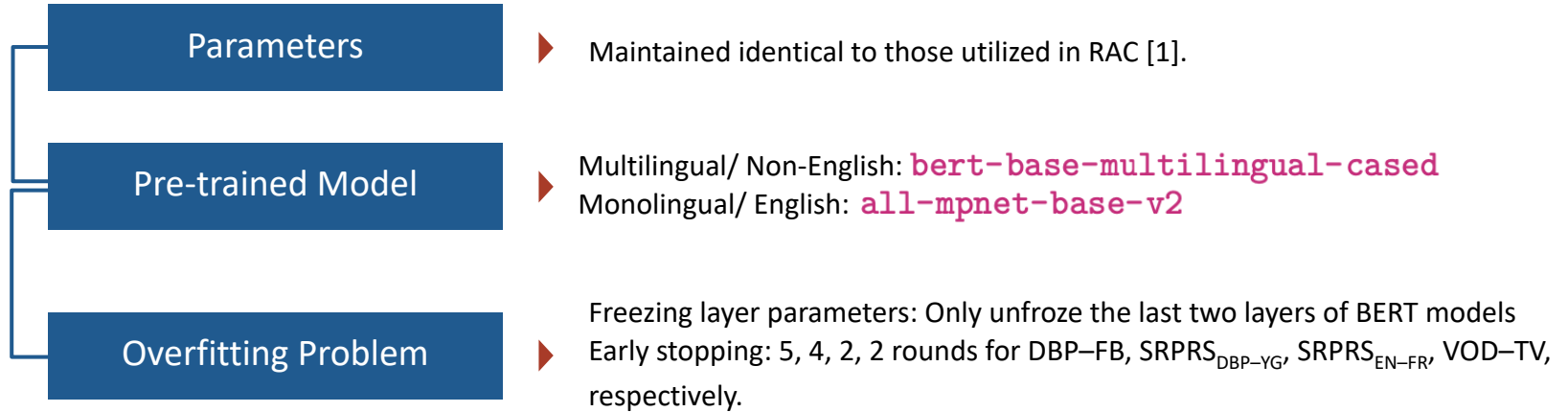
Datasets

Table 5.1: Dataset statistics

Dataset	#Triples	#Ents	#Rels	#Aligns
DBP--FB	208,388	55,403	1,289	25,542
SRPRS _{DBP--YG}	70,317	30,000	253	15,000
SRPRS _{EN--FR}	70,040	30,000	398	15,000
VOD--TV	2,091,334	336,594	4	6,541

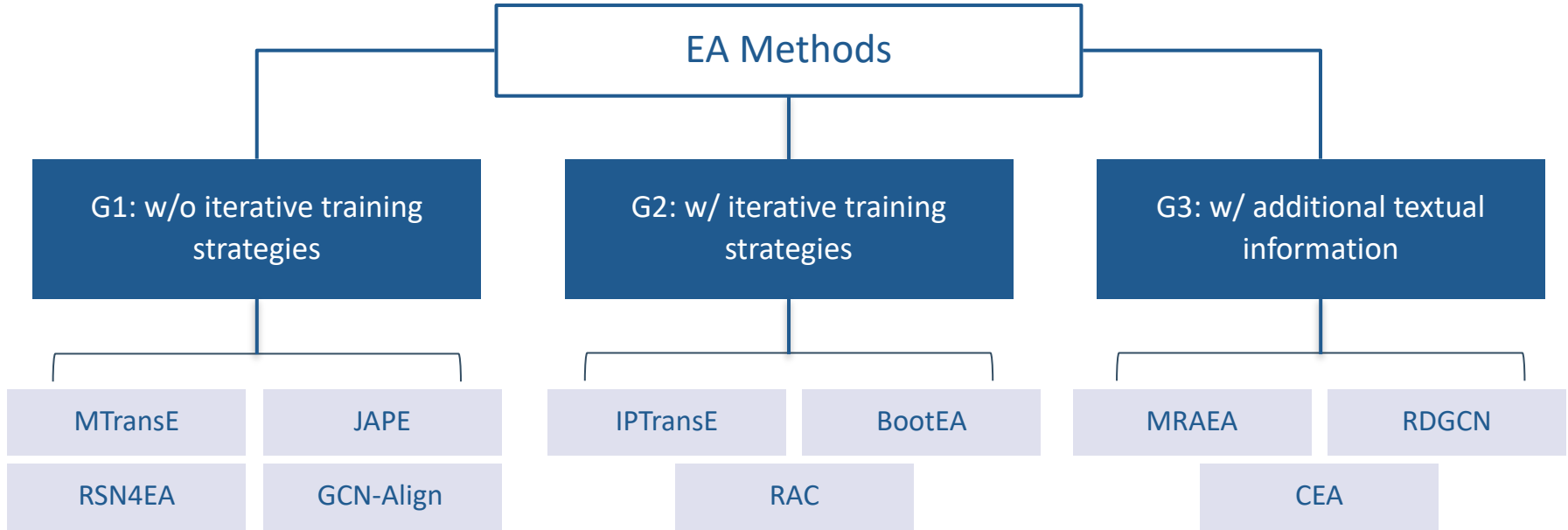
- training/validation/testing: 20%/10%/70%
- Seed alignment pairs: 100 (Selected entities from the leftover training data, guided by the labeling budget B .)

Implementation Details



[1] W. Zeng, X. Zhao, J. Tang, and C. Fan, "Reinforced active entity alignment," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, p. 2477–2486.

Compared Methods



Research Questions (RQs)

- 1 Does SARA outperform **baseline** alignment models under conditions of **limited supervision**?
- 2 What is the impact of different **fine-tuning strategies**, parameter freezing, and early stopping strategies applied to the proposed TitleSimBERT module on entity alignment (EA)?
- 3 To what extent does the inclusion of our proposed **“semantic-assisted” loss (SAL)** impact the model's performance?
- 4 How do the **various modules** in our proposed method influence the overall performance?

Main Results (RQ1)

Table 2: Overall EA results on Hits@k and MRR

Method	DBP-FB			SRPRS _{DBP-YG}			SRPRS _{EN-FR}			VOD-TV			
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	
G1	MTransE [7]	8.5	23.0	0.14	19.6	40.1	0.27	21.3	44.7	0.29	0.1	0.6	0.01
	JAPE [31]	6.5	20.4	0.12	19.3	50.0	0.3	24.1	54.4	0.34	0.5	2.3	0.01
	RSN4EA [11]	25.3	49.7	0.34	39.3	66.5	0.49	35.0	63.6	0.44	-	-	-
	GCN-Align [14]	17.8	42.3	0.26	34.7	64.0	0.45	29.6	59.2	0.40	3.9	14.3	0.08
G2	IPTransE [25]	3.0	10.0	0.06	10.3	26.0	0.16	12.4	30.1	0.18	-	-	-
	BootEA [32]	21.2	42.5	0.29	38.1	65.1	0.47	36.5	64.9	0.46	14.0	33.1	0.21
	RAC [44]	19.9	45.6	0.28	34.3	64.2	0.44	26.5	57.1	0.37	36.7	55.9	0.44
G3	MRAEA [21]	6.2	21.0	0.11	21.8	53.9	0.32	13.9	38.0	0.22	23.0	54.5	0.34
	RDGCN [41]	49.5	68.0	0.56	88.9	95.3	0.91	67.2	78.8	0.71	0.8	4.1	0.02
	CEA [45]	†68.2	†86.1	†0.75	99.5	†99.9	†0.99	†87.0	†93.2	†0.89	†98.6	†99.8	†0.99
	SARA	75.7	97.4	0.82	†99.1	99.9	0.99	91.0	96.1	0.93	99.5	99.8	1.00

Model effectiveness

1. SARA mostly achieves **the best results** in all evaluation metrics for the four datasets, demonstrating the effectiveness of our model.
2. The only exception to SARA's superior performance is in the case of Hit@1 for SRPRS_{DBP-YG}. However, both **SARA and the most competitive baseline CEA achieve accuracy rates of over 99% in this scenario.**

Main Results (RQ1)

Significant performance improvements for small label budget

Table 3: Performance with varying labeled budgets (Hits@1)

DBP-FB						
Budgets (B)	RAC	MRAEA	RDGCN	CEA	SARA	Improv. (%)
250	4.9	6.2	49.5	†68.2	75.7	11.0%
500	7.2	9.4	49.8	†73.9	78.9	6.8%
750	9.4	12.0	51.2	†76.1	81.1	6.6%
SRPRS _{EN-FR}						
Budgets (B)	RAC	MRAEA	RDGCN	CEA	SARA	Improv.(%)
250	11.8	13.9	67.2	†87.0	91.0	4.6%
500	15.6	18.4	68.1	†88.9	91.6	3.0%
750	18.2	22.3	68.6	†89.5	92.0	2.8%

1. Significant improvement was achieved by **incorporating entity names** as additional information on EA, and we further examined the performance of G3.
2. We additionally include RAC for comparison as it is a framework proposed to address the limited supervision problem.
3. These two datasets were specifically selected because they **exhibit severe naming inconsistencies** compared to the other two datasets.
4. We can observe that our model outperforms the baselines, particularly **with fewer labeled data**.
→ This finding suggests that our proposed method effectively reduces the reliance on manual labeling in EA tasks.

Different Fine-tuning Strategies (RQ2)

Table 4: Impact of fine-tuning methods ($B = 250$)

Dataset	Method	Hits@1	Hits@10	MRR
DBP-FB	SARA	75.7	97.4	0.82
	SARA (direct)	71.4	94.9	0.79
	SARA (fixed)	70.3	93.8	0.77
SRPRS _{DBP-YG}	SARA	99.1	99.9	0.99
	SARA (direct)	96.8	98.6	0.97
	SARA (fixed)	94.4	97.2	0.95
SRPRS _{EN-FR}	SARA	91.0	96.1	0.93
	SARA (direct)	69.3	80.2	0.74
	SARA (fixed)	78.4	86.6	0.81
VOD-TV	SARA	99.5	99.8	1.00
	SARA (direct)	99.1	99.7	0.99
	SARA (fixed)	98.8	99.6	0.99

In the variant called **SARA (direct)**, we adopt a different approach where instead of separately fine-tuning the TitleSimBERT module with (2), we directly fine-tune the Sentence-BERT model using (6).

$$\mathcal{L}_{\text{TitleSimBERT}} = -\frac{1}{M} \sum_{i=1}^M y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i), \quad (2)$$

$$\mathcal{L} = \lambda_a \mathcal{L}_{\text{SAL}} + \lambda_b \mathcal{L}_{\text{structural}}, \quad (6)$$

Updating the entity name embeddings contributes positively to the overall performance

1. Both **SARA** and **SARA (direct)** outperform the variant that keeps the entity name embeddings fixed (i.e., SARA (fixed))
2. The results consistently demonstrate that separately fine-tuning the entity name embeddings using the loss function in (2) achieves the best performance.

Different Fine-tuning Strategies (RQ2)

Table 5: Impact of parameter freezing and early stopping
($B = 250$)

Dataset	Method	Hits@1	Hits@10	MRR
DBP-FB	SARA	75.7	97.4	0.82
	SARA (w/o PZ & ES)	42.6	56.2	0.55
	SARA (w/o ES)	72.9	97.1	0.81
	SARA (w/o PZ)	49.5	64.2	0.53
SRPRS _{DBP-YG}	SARA	99.1	99.9	0.99
	SARA (w/o PZ & ES)	86.5	89.1	0.88
	SARA (w/o ES)	98.7	99.4	0.99
	SARA (w/o PZ)	94.9	96.5	0.95
SRPRS _{EN-FR}	SARA	91.0	96.1	0.93
	SARA (w/o PZ & ES)	11.3	17.6	0.14
	SARA (w/o ES)	89.4	91.6	0.91
	SARA (w/o PZ)	42.7	48.8	0.45
VOD-TV	SARA	99.5	99.8	1.00
	SARA (w/o PZ & ES)	82.2	87.5	0.83
	SARA (w/o ES)	93.1	95.8	0.93
	SARA (w/o PZ)	95.3	99.6	0.95

SARA consistently achieves the highest performance across all datasets.

The removal of either parameter freezing (PF) or early stopping (EZ) substantially impacts the model's performance.

When both PF and ES are removed, SARA experiences a considerable drop in performance, with reduction of 43.7%, 12.7%, 87.6%, and 17.4% on the DBP-FB, SRPRS_{DBP-YG}, SRPRS_{EN-FR}, VOD-TV datasets, respectively.

→ This emphasizes the effectiveness of utilizing these two fine-tuning techniques.

Different Fine-tuning Strategies (RQ2)

Table 5: Impact of parameter freezing and early stopping
($B = 250$)

Dataset	Method	Hits@1	Hits@10	MRR
DBP-FB	SARA	75.7	97.4	0.82
	SARA (w/o PZ & ES)	42.6	56.2	0.55
	SARA (w/o ES)	72.9	97.1	0.81
	SARA (w/o PZ)	49.5	64.2	0.53
SRPRS _{DBP-YG}	SARA	99.1	99.9	0.99
	SARA (w/o PZ & ES)	86.5	89.1	0.88
	SARA (w/o ES)	98.7	99.4	0.99
	SARA (w/o PZ)	94.9	96.5	0.95
SRPRS _{EN-FR}	SARA	91.0	96.1	0.93
	SARA (w/o PZ & ES)	11.3	17.6	0.14
	SARA (w/o ES)	89.4	91.6	0.91
	SARA (w/o PZ)	42.7	48.8	0.45
VOD-TV	SARA	99.5	99.8	1.00
	SARA (w/o PZ & ES)	82.2	87.5	0.83
	SARA (w/o ES)	93.1	95.8	0.93
	SARA (w/o PZ)	95.3	99.6	0.95

The impact of frozen parameters on the model's performance is relatively more substantial than that of early stopping.

For instance, on the DBP-FB and SRPRS_{EN-FR} datasets, the removal of ES results in a performance drop of 4% and 2%, respectively.

However, when PF is removed, the performance of SARA experiences a more substantial decline of 35% and 53%, respectively.

Effects of the Semantic-assisted Loss (RQ3)

Table 6: Sensitivity analysis on SAL ($B = 250$)

λ_a	DBP-FB		VOD-TV		SRPRS _{EN-FR}		SRPRS _{DBP-YG}	
	Hits@1	MRR	Hits@1	MRR	Hits@1	MRR	Hits@1	MRR
0.2	71.9	0.78	98.8	0.99	90.1	0.92	99.5	0.99
0.4	71.9	0.78	99.1	0.99	89.9	0.92	99.0	0.99
0.6	72.8	0.8	98.6	0.99	90.7	0.93	99.1	0.99
0.8	75.4	0.81	98.8	0.99	90.9	0.93	99.4	0.99
1.0	75.7	0.82	99.5	1.00	91.0	0.93	99.1	0.99
1.2	75.4	0.81	98.6	0.99	91.0	0.93	99.2	0.99

$$\mathcal{L} = \lambda_a \mathcal{L}_{\text{SAL}} + \lambda_b \mathcal{L}_{\text{structural}}, \quad (6)$$

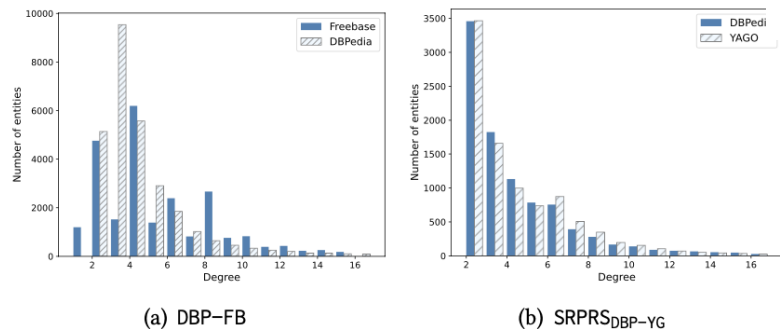


Figure 2: Degree distributions of DBP-FB and SRPRS_{DBP-YG}

Varying the weight of SAL has a minimal impact on the performance of SARA for datasets where it already achieves high accuracy (i.e., Hits@1 above 90%)

However, for the DBP-FB dataset, increasing the weight of SAL does lead to some improvement in the learning process, with $\lambda_a = 1$ yielding the best performance.

→ This dataset presents additional challenges due to its high structural heterogeneity. (Fig. 2)

Ablation Study (RQ4)

Table 7: Ablation study ($B = 250$)

Methods	DBP-FB			VOD-TV		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
SARA	75.7	97.4	0.82	99.5	99.8	1.00
SARA (w/o PF)	72.1	96.1	0.78	97.8	98.7	0.98
SARA (w/o SE)	5.4	16.8	0.10	51.0	68.5	0.56
SARA (w/o SE & PF)	4.6	13.9	0.08	38.6	57.8	0.44

Methods	SRPRS _{EN-FR}			SRPRS _{DBP-YG}		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
SARA	91.0	96.1	0.93	99.1	99.9	0.99
SARA (w/o PF)	89.9	94.3	0.91	99.1	99.5	0.99
SARA (w/o SE)	13.8	31.8	0.20	19.8	42.9	0.28
SARA (w/o SE & PF)	11.7	28.6	0.18	19.0	41.4	0.27

SE

$$\mathcal{L}_{\text{SAL}} = \sum_{(s,t) \in L} \sum_{(s',t') \in L'_{(s,t)}} \max(d(\mathbf{g}_s, \mathbf{g}_t) + \gamma \quad (4)$$

$$\mathbf{g}_s = \mathbf{e}_s \parallel \mathbf{o}_s \quad \mathbf{g}_t = \mathbf{e}_t \parallel \mathbf{o}_t$$

$$-d(\mathbf{g}'_s, \mathbf{g}'_t)(1 + P(\mathbf{e}'_s, \mathbf{e}'_t)), 0)$$

$$P(\mathbf{e}'_s, \mathbf{e}'_t) = \cos(\mathbf{e}'_s, \mathbf{e}'_t).$$

PF

Ablation Study (RQ4)

Table 7: Ablation study ($B = 250$)

Methods	DBP-FB			VOD-TV		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
SARA	75.7	97.4	0.82	99.5	99.8	1.00
SARA (w/o PF)	72.1	96.1	0.78	97.8	98.7	0.98
SARA (w/o SE)	5.4	16.8	0.10	51.0	68.5	0.56
SARA (w/o SE & PF)	4.6	13.9	0.08	38.6	57.8	0.44

Methods	SRPRS _{EN-FR}			SRPRS _{DBP-YG}		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
SARA	91.0	96.1	0.93	99.1	99.9	0.99
SARA (w/o PF)	89.9	94.3	0.91	99.1	99.5	0.99
SARA (w/o SE)	13.8	31.8	0.20	19.8	42.9	0.28
SARA (w/o SE & PF)	11.7	28.6	0.18	19.0	41.4	0.27

Both modules positively impact the performance of the SARA model.

Removing either of these modules leads to a decline in performance.

Specifically, when removing the name embeddings (i.e., SARA (w/o SE)), there is a significant drop in performance.

→ This highlights the crucial role that entity name embeddings play in the EA task.

While the impact of removing the penalty function (i.e., SARA (w/o PF)) is relatively smaller compared to the name embeddings, there is still a noticeable decrease in performance.

→ Emphasize the importance of both modules and demonstrate their contributions to the overall performance of the SARA model.

Conclusions

Conclusions

Addresses the two main challenges

- **The heavy reliance on labeled data**
→ Reinforced active learning to select valuable entity pairs for training
- **The underutilization of semantic information**
→ Effectively utilizes entity semantic information with an advanced deep contextualized language model to enhance model performance.

Experiment results

- 3 Benchmark datasets and 1 real-world dataset
- Effectiveness of SARA in addressing limited supervision problem
- Outperforming other state-of-the-art semantic-enhanced models

Strengths

- **Handling scenarios with inconsistent entity names**
- Dealing with heterogeneous knowledge graphs

→ Our framework can **alleviate the model's dependence on label data** by making **full use of semantic information**, effectively improving the model performance of EA in real life.

Thanks for your attention.
Q&A