Knowledge Graph Linking and Integration

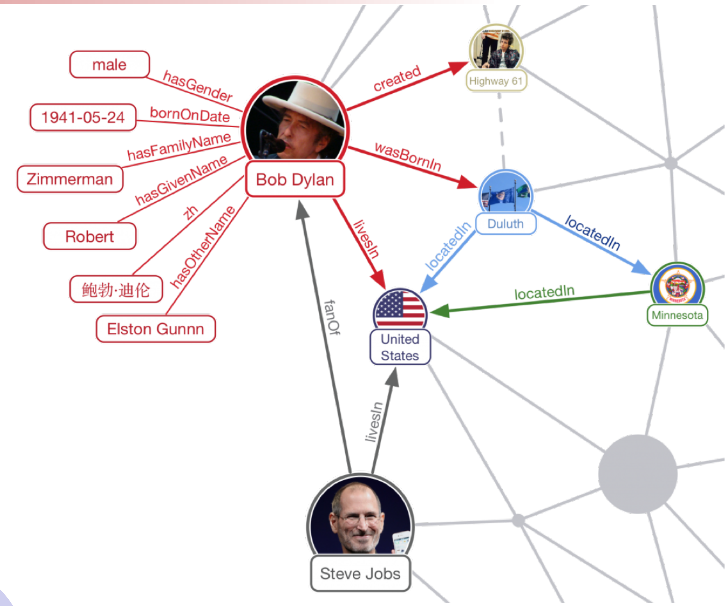
Jing Zhang

School of Information, Renmin University of China

Collaborate with Bo Chen (*RUC*), Xiaobin Tang (*RUC*), Hong Chen (*RUC*), Cuiping Li(*RUC*) and Jie Tang (*THU*)

Knowledge Graph

- A structural form of human knowledge
- Represent by triples:
  - (head entity, relationship, tail entity)
  - (entity, attribute, value)
- Application:
  - Question answering
  - Recommendation system
  - Information retrieval
  - …



DBpedia

Open academic graph (OAG)

Dynamic Knowledge Graph

- Knowledge increases dynamically
  - Link new knowledge to existing entities in knowledge graphs
- Knowledge is distributed in multiple sources
  - Integrate different sources of knowledge graphs

Challenges to be solved

- Ambiguity
  - Ambiguity of author names when linking new papers to exiting authors in OAG
- Heterogeneity
  - Heterogeneity of graphs when integrating multi-lingual knowledge graphs

Ambiguity of author names when linking new papers to exiting authors in OAG

Bo Chen, Jing Zhang, Jie Tang, Lingfan Cai, Zhaoyu Wang, Shu Zhao, Hong Chen, Cuiping Li. CONNA: Addressing Name Disambiguation on the Fly. TKDE'20

Google Scholar

Google 学术搜索

Zhigang Wang

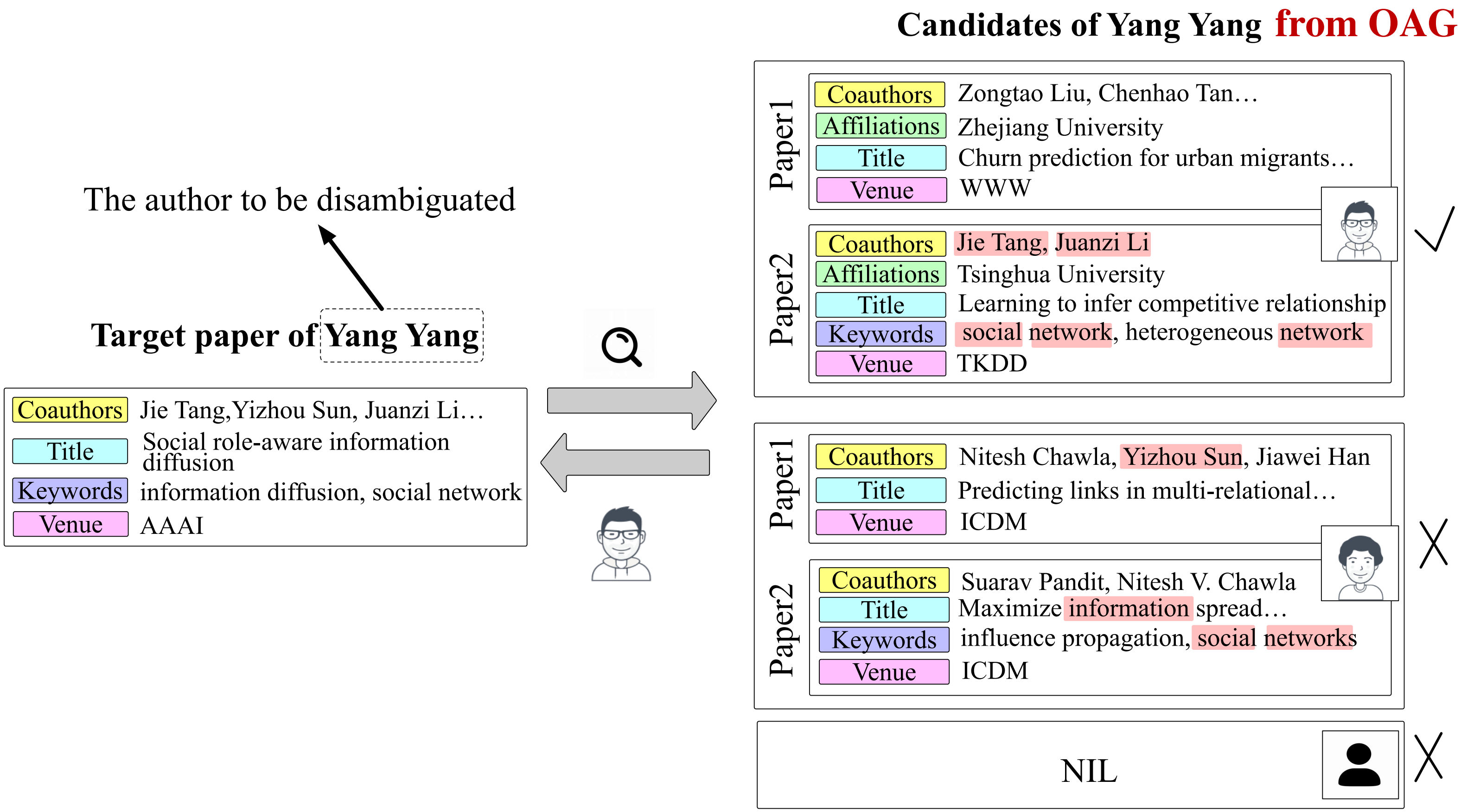
Tsinghua University
在 mails.tsinghua.edu.cn 的电子邮件经过验证
Knowledge Graph

关注

One common coauthor "Jie Tang"

标题	引用次数	年份
Paclitaxel-loaded and A10-3.2 aptamer-targeted poly (lactide-co-glycolic acid) nanobubbles for ultrasound imaging and therapy of prostate cancer M Wu, Y Wang, Y Wang, M Zhang, Y Luo, J Tang, Z Wang, D Wang, ... International journal of nanomedicine 12, 5313	9	2017
Domain specific cross-lingual knowledge linking based on similarity flooding L Pan, Z Wang, J Li, J Tang International Conference on Knowledge Science, Engineering and Management ...	1	2016
Boosting to Build a Large-Scale Cross-Lingual Ontology Z Wang, L Pan, J Li, S Li, M Li, J Tang China Conference on Knowledge Graph and Semantic Computing, 41-53		2016

Linking New Papers to OAG



Matching: how to match a paper and a candidate person?

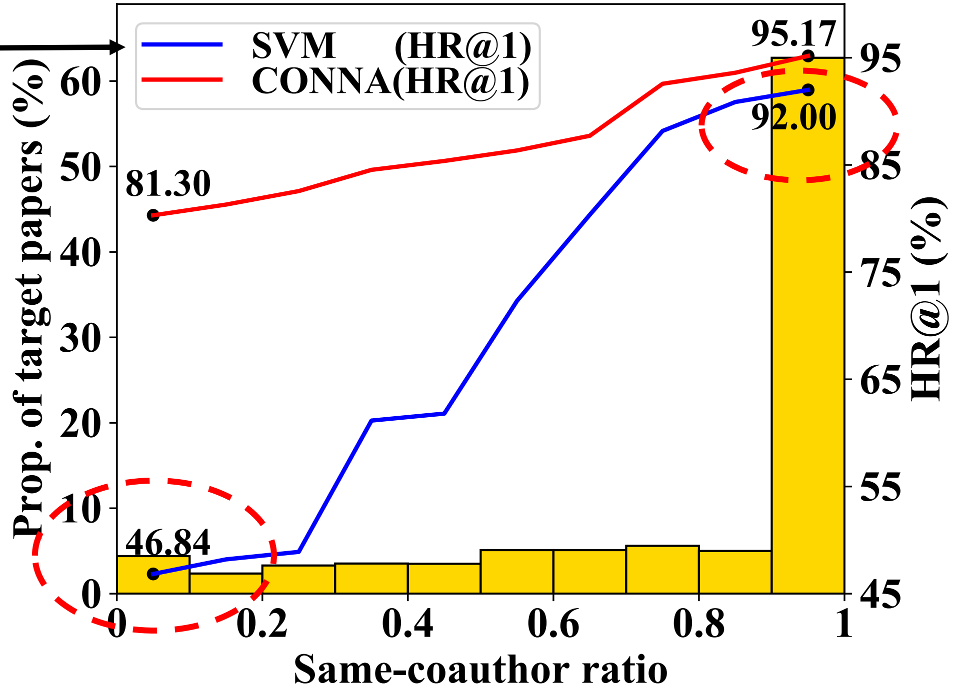
Decision: how to decide to assign the top matched candidate or NIL?

Matching

- How can coauthor names take effect ?

Name is important ! [WSDM, 2013]
56% accounts of same names across the social networks can be correctly linked together

Feature-based

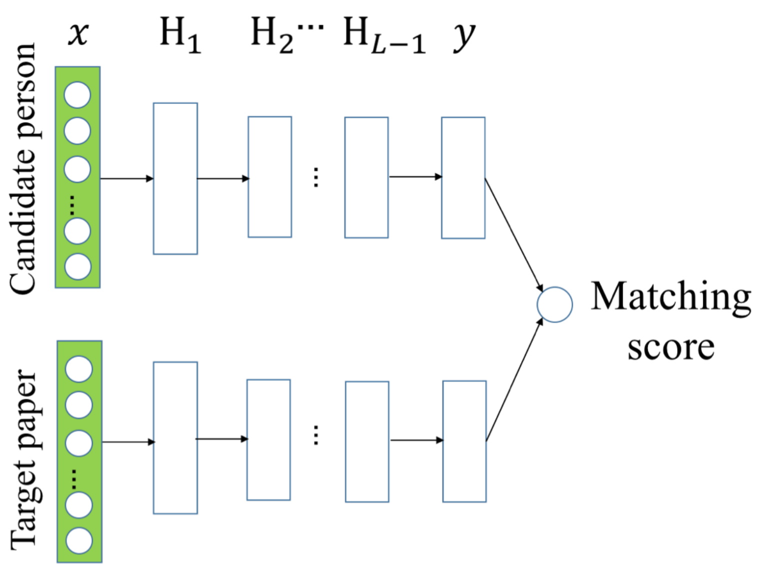


If there are many similar coauthors, it is easy to match

If there are only a few same coauthors, it is hard to match

How to Improve the Matching Performance?

- Feature-based
  - Exact matching the tokens
- Representation-based
  - Semantic matching a paper and a person

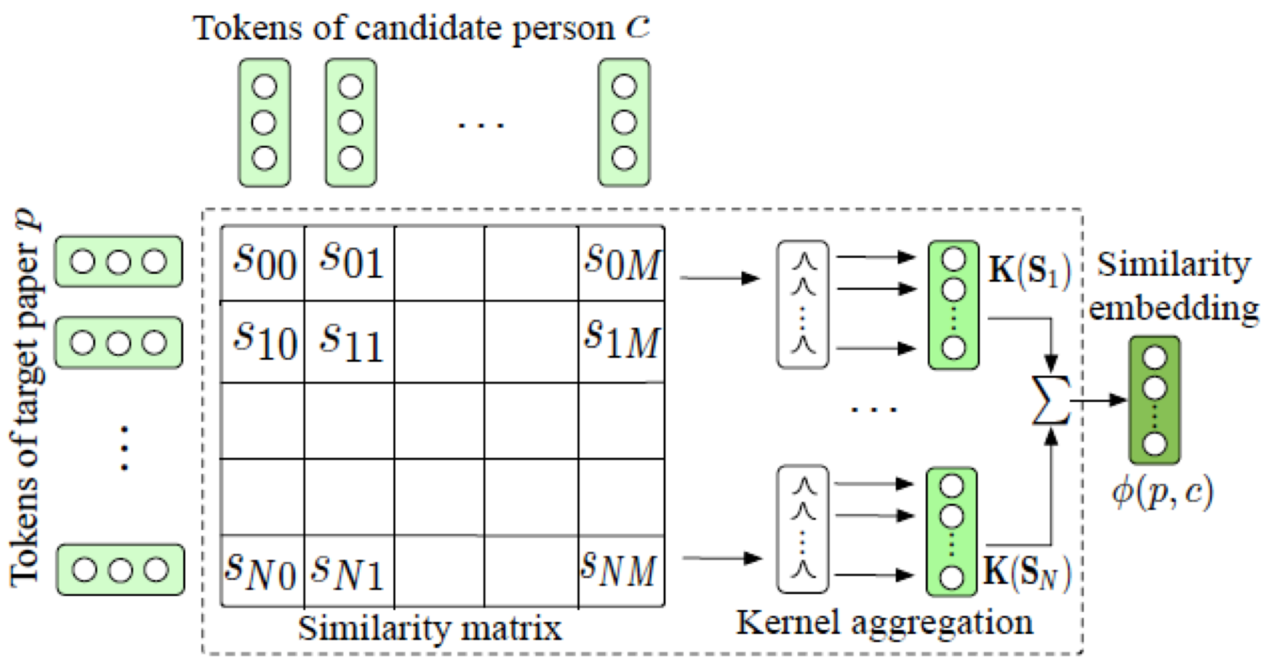


Dilute the effect of the exact matching.

E.g., exact matching is suitable for comparing coauthor names

Basic Interaction Matching Model

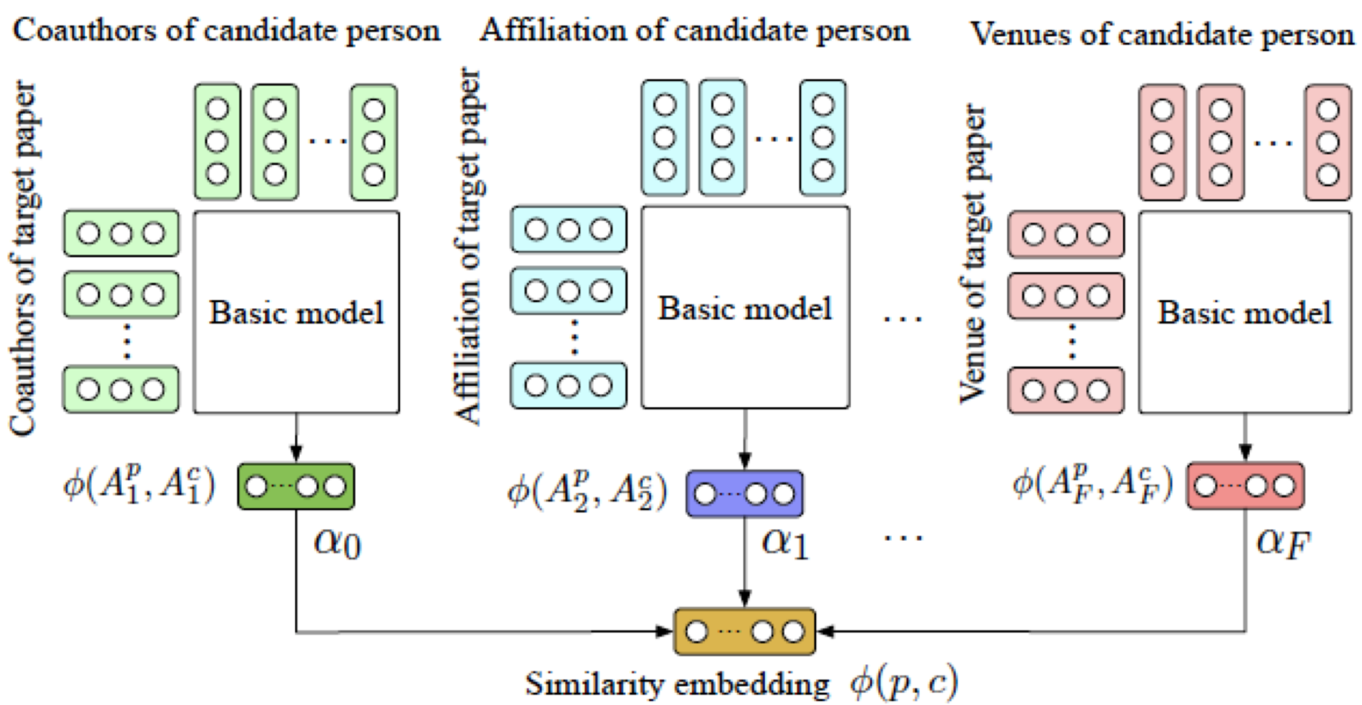
- Capture both the exact and soft matches



Xiong, et al. SIGIR'17

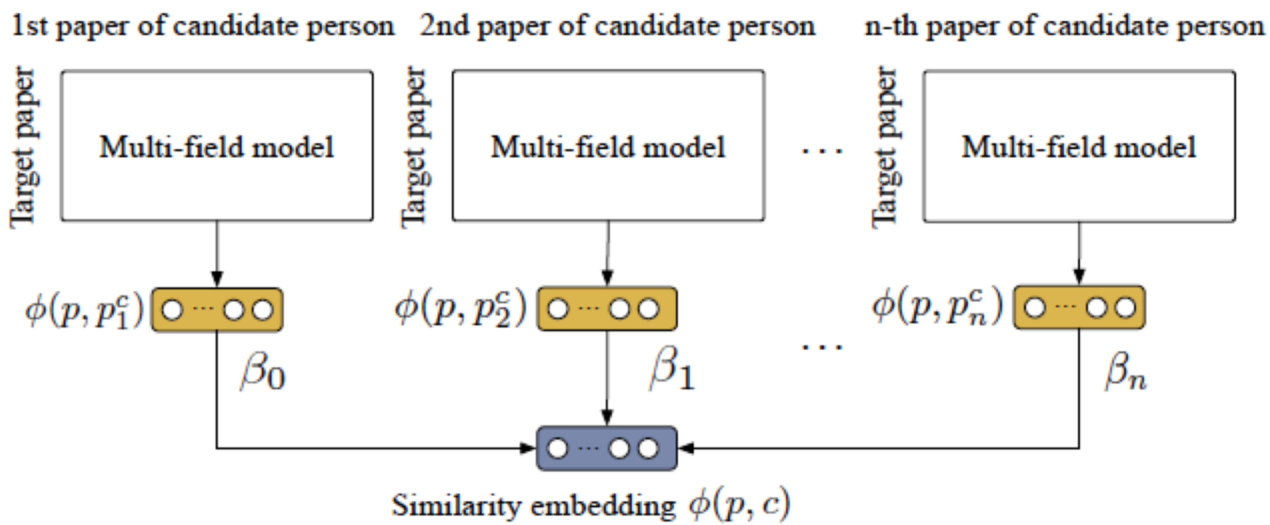
Modeling Multi-field Attributes

- Assumption
  - Different fields of attributes takes different effects.



Modeling Multi-Instances of Papers

- Assumption
  - Different papers of a candidate person takes different effects.



Objective Function: Learning to Rank

- Triplet loss function:

$$\mathcal{L}(\Theta) = \sum_{(p,c^+,c^-)\in\mathcal{D}} \max\{0, g(\phi(p,c^+)) - g(\phi(p,c^-)) + m)\},$$

  - m: margin between positive and negative pairs
  - g: transform feature vector ϕ to a score

Decision

- Decide to assign the top-matched person (y=1) or NIL(y=0).

- Training data: Top-matched person by the matching model
  - Positive instances: $\{(\phi(p, c^{+}), y = 1)\}$
  - Negative instances: $\{(\phi(p, c^{-}), y = 0)\}$

Similarity embedding generated by the matching component

- Objective: train a classification model

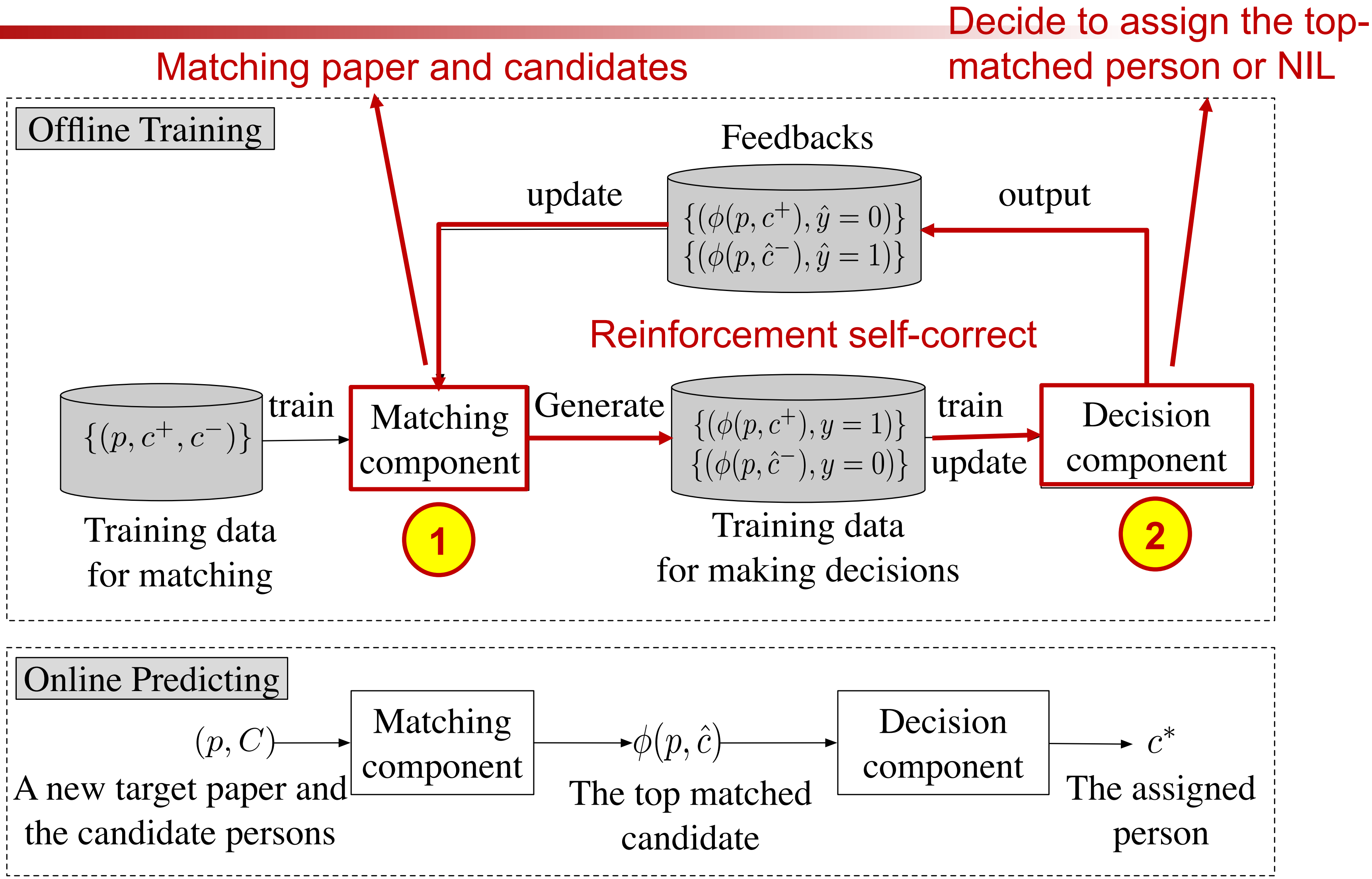
$$h(\psi): \{\phi(p, c)\} \rightarrow \{y\}$$

Reinforcement Self-Correction

- Generator:
  - Treat the matcher as a generator to generate the features $\phi(p, c)$
- Feedback:
  - The Decision classifier $h(\psi)$: $\{\phi(p, c)\} \rightarrow \{y\}$ give feedback to the matcher

$$R(y, \hat{y}) = \begin{cases} 1 & \hat{y} = y; \\ 0 & \text{otherwise.} \end{cases}$$

Framework


Decide to assign the top-matched person or NIL
Matching paper and candidates
Offline Training
Feedbacks
update
output
$\{(\phi(p, c^+), \hat{y} = 0)\}$
$\{(\phi(p, \hat{c}^-), \hat{y} = 1)\}$
Reinforcement self-correct
$\{(p, c^+, c^-)\}$
train
Matching component
Generate
$\{(\phi(p, c^+), y = 1)\}$
$\{(\phi(p, \hat{c}^-), y = 0)\}$
train
update
Decision component
Training data for matching
1
Training data for making decisions
2
Online Predicting
(p, C)
Matching component
$\phi(p, \hat{c})$
Decision component
c^*
A new target paper and the candidate persons
The top matched candidate
The assigned person

Training Process

Algorithm 1: Reinforcement Joint Training

Input: A training set $\{(p, C)\}$.
Output: A matching component and a decision component parametrized by Θ and Φ respectively.

```
Construct 𝒟^r = {(p, c⁺, c⁻)};
Pre-train Θ of the matching component on 𝒟^r;
Construct 𝒟^c = {(φ(p, c), y)} by the matching component;
Pre-train Φ of the decision component on 𝒟^c;
repeat
    for (φ(p, c), y) ∈ 𝒟^c do
        Predict ŷ by the decision component ;
        Calculate R(y, ŷ) by Eq. (9) ;
        Calculate ∇_Θ J(Θ) by Eq. (10);
        Θ → Θ + μ∇_Θ J(Θ), where μ is the lear
    Regenerate 𝒟^c by the matching component;
    Update Φ in the decision component on 𝒟^c;
until Convergence;
```

Pre-train two components

Reward: punish the wrong features, reward the right features

$$R(y, \hat{y}) = \begin{cases} 1 & \hat{y} = y; \\ 0 & \text{otherwise.} \end{cases}$$

Gradient

$$\nabla_\Theta J(\Theta) \quad \sum_{(\phi(p,c),y)\in \mathcal{I}} R(y, \hat{y}) \nabla p_\Theta(\phi(p, c)),$$

Experimental Results

Performance of the Matching Results (%)

Representation matching
or feature engineering

Model	OAG-WhoIsWho			KDD Cup		
	HR@1	HR@3	MRR	HR@1	HR@3	MRR
Camel	41.20	62.00	55.00	44.62	67.19	59.44
HetNetE	46.00	67.00	60.24	51.06	77.44	66.41
GML	70.87	94.53	82.59	72.13	95.34	82.90
GBDT	87.30	98.10	92.71	84.18	92.09	89.59
CONNA^r(BP)	86.20	96.40	92.20	91.12	95.72	93.73
CONNA^r(MFP)	88.00	98.75	93.25	-	-	-
CONNA^r(MFMI)	89.45	98.40	93.82	91.45	95.80	94.03
CONNA	90.45	98.30	94.46	92.10	96.35	94.66
CONNA+Fine-tune	**91.10**	**98.45**	**94.86**	**92.60**	**96.71**	**94.95**

Interaction matching, multi-field, multi-instance components take effect.
Joint training can improving the matching performance.

Experimental Results

Direct classification or heuristic loss

With ground truth

Without ground truth

Performance of the Decision Results (%)

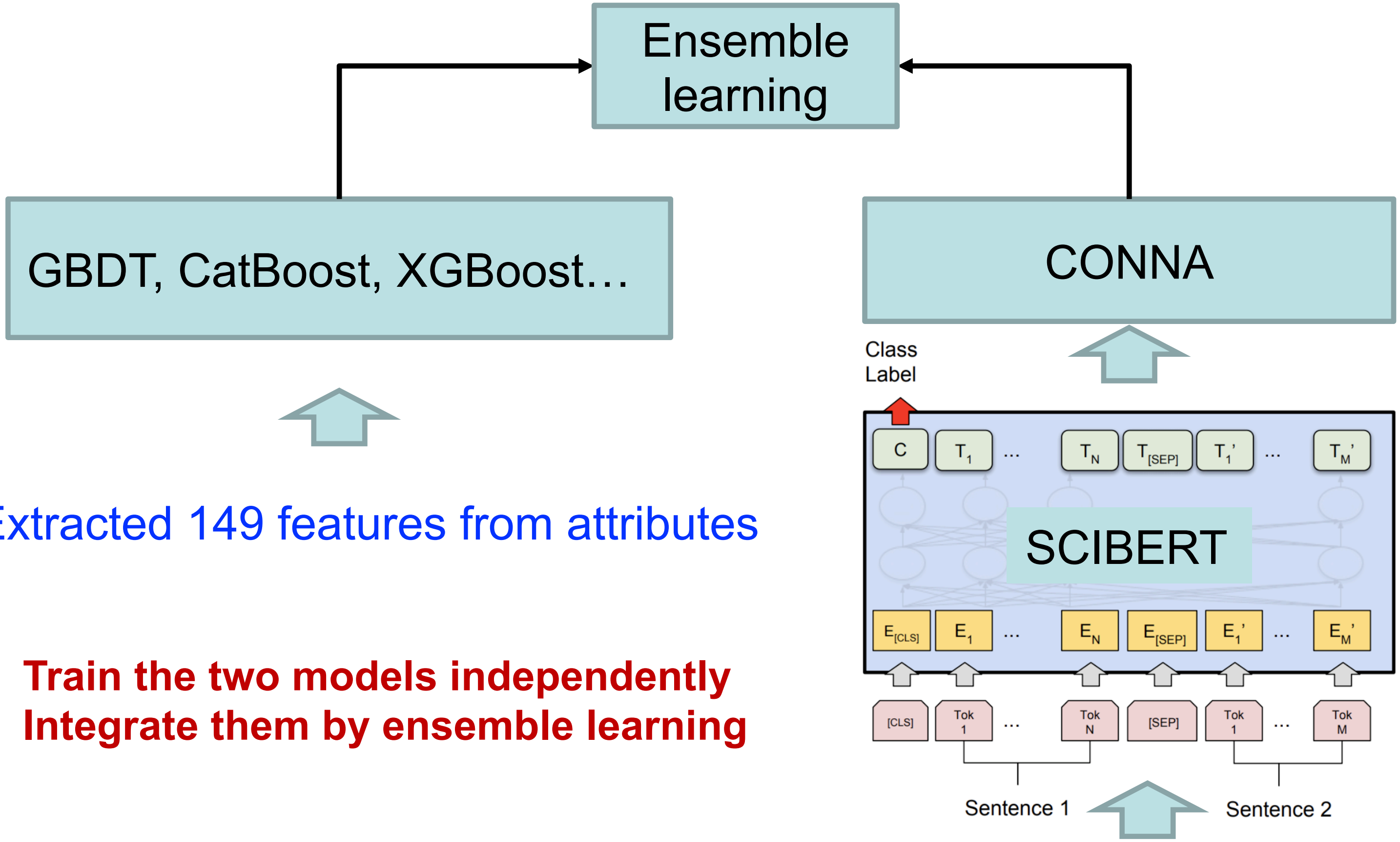
Model	QAG-WhoIsWho						KDD Cup					
	Samples with $c^* = c^+$			Samples with $c^* =$ NIL			Samples with $c^* = c^+$			Samples with $c^* =$ NIL		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
GBDT	82.87	72.40	77.28	75.39	85.04	79.98	83.64	71.64	77.17	75.20	85.98	80.23
Threshold	79.33	57.60	66.38	66.47	84.07	74.24	74.89	71.00	72.90	72.43	76.20	74.27
Heuristic Loss	71.79	78.40	74.95	76.21	69.20	72.54	85.14	69.60	76.59	74.29	87.85	80.50
CrossEntropy	79.42	82.33	80.85	81.66	78.67	80.14	89.60	82.79	86.06	86.15	88.05	87.09
CONNA	79.53	**89.87**	84.38	88.35	76.87	82.21	88.44	**86.20**	87.31	**86.54**	88.73	87.62
CONNA+Fine-tune	**82.47**	**90.33**	**86.22**	**89.31**	80.80	**84.84**	**89.87**	85.73	**87.75**	86.36	**90.33**	**88.30**

The problem is not merely a matching or a classification decision problem. We need to not only keep the relevant order within each candidate list, but also globally distinguish all the positive pairs from all the negative pairs.

Online Deployment

https://www.aminer.cn/

- Combine traditional feature engineering model with embedding model



Extracted 149 features from attributes

Raw attributes

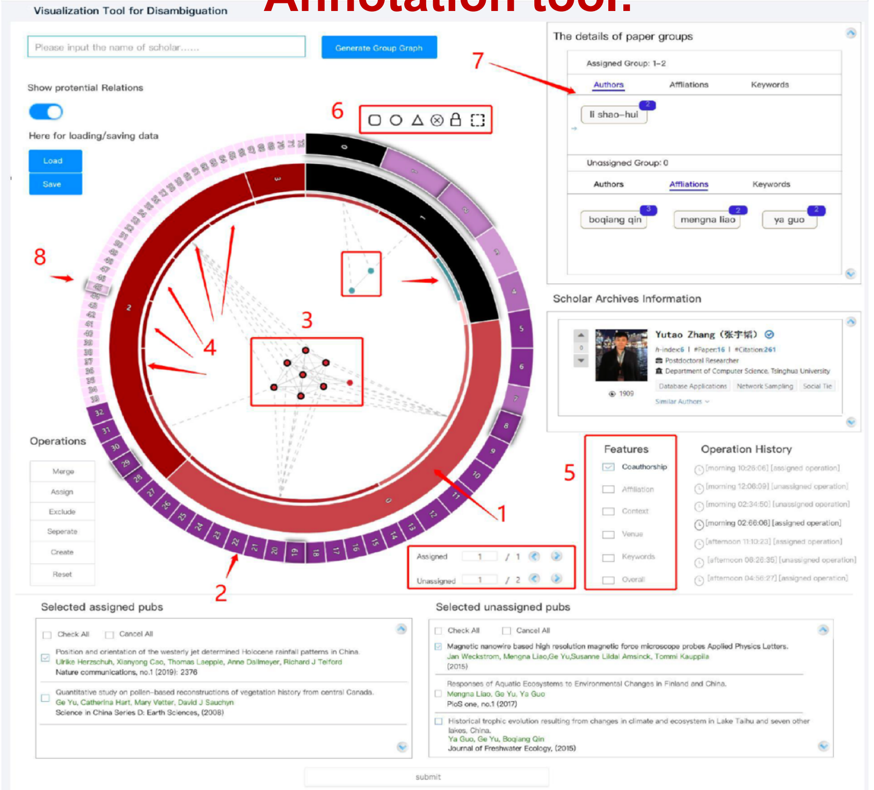
1. **Train the two models independently**
2. **Integrate them by ensemble learning**

Dataset: WhoIsWho

- https://www.aminer.cn/whoiswho

Dataset Version	#Name	#Author	#Paper	Date
na-v1	421	45,187	399,255	2019-10-01
na-v2	231	13,662	221,802	2020-05-20

Annotation tool:

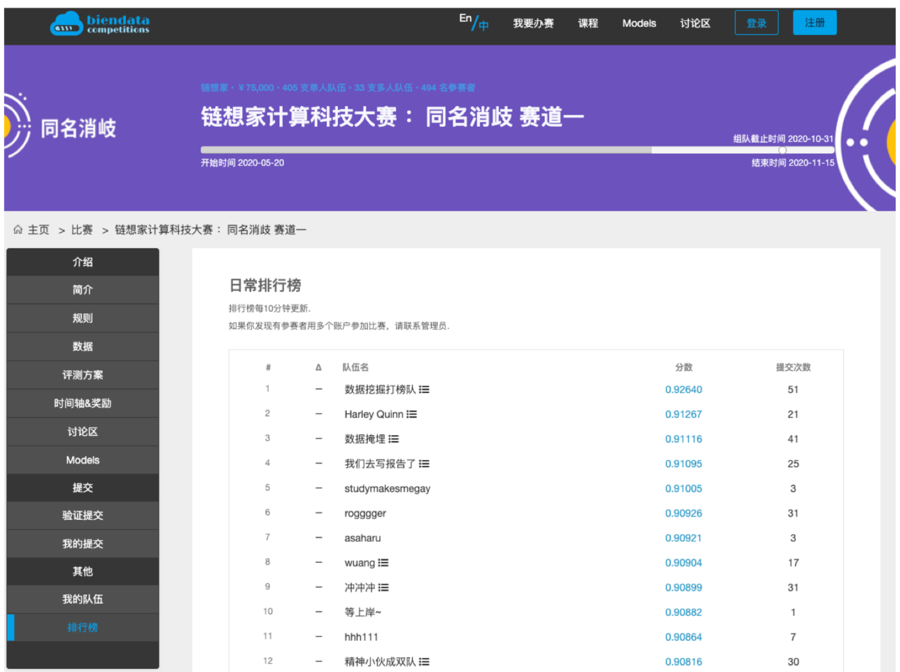


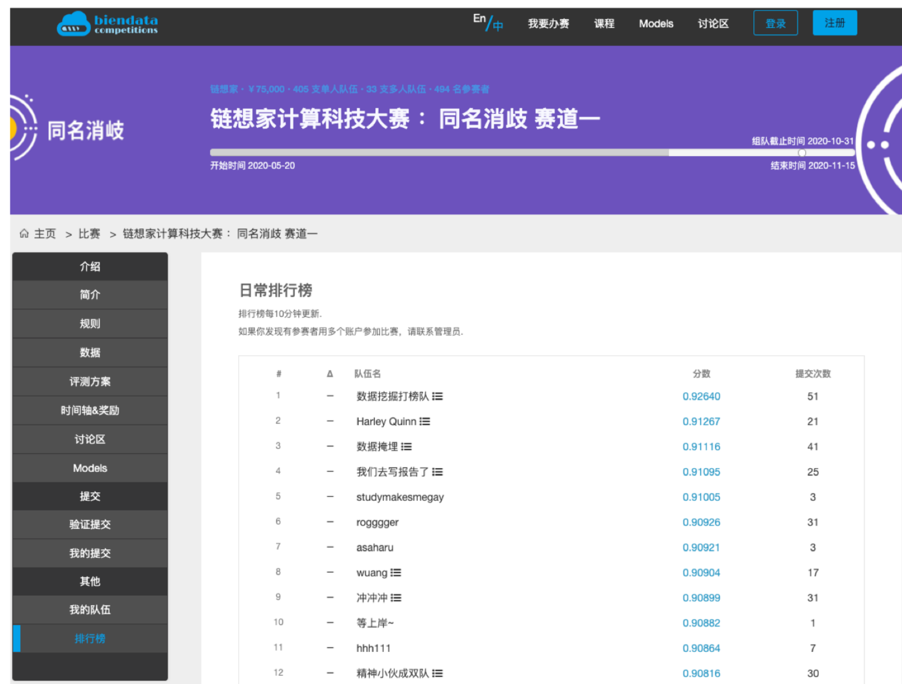
1. **Collect** all the assigned papers.
2. **Remove** the wrongly assigned papers or **split** the papers of an author.
3. **Annotate** the unassigned papers or **merge** the papers of two authors.

Challenges

Deadline:2020-11-15

- Name disambiguation from scratch
  https://www.biendata.xyz/competition/chaindream_nd_task1/
- Continuous name disambiguation
  https://www.biendata.xyz/competition/chaindream_nd_task2/





Heterogeneity of graphs when integrating multi-lingual knowledge graphs

Xiaobin Tang, Jing Zhang, Bo Chen, Yang Yang, Hong Chen, Cuiping Li. BERT-INT: A BERT-based Interaction Model For Knowledge Graph Alignment. IJCAI’20

Motivation

Multiple KGs exist in real world

- A single KG is far from complete
- Different KGs are supplementary in contents

Freebase DBpedia yago select knowledge

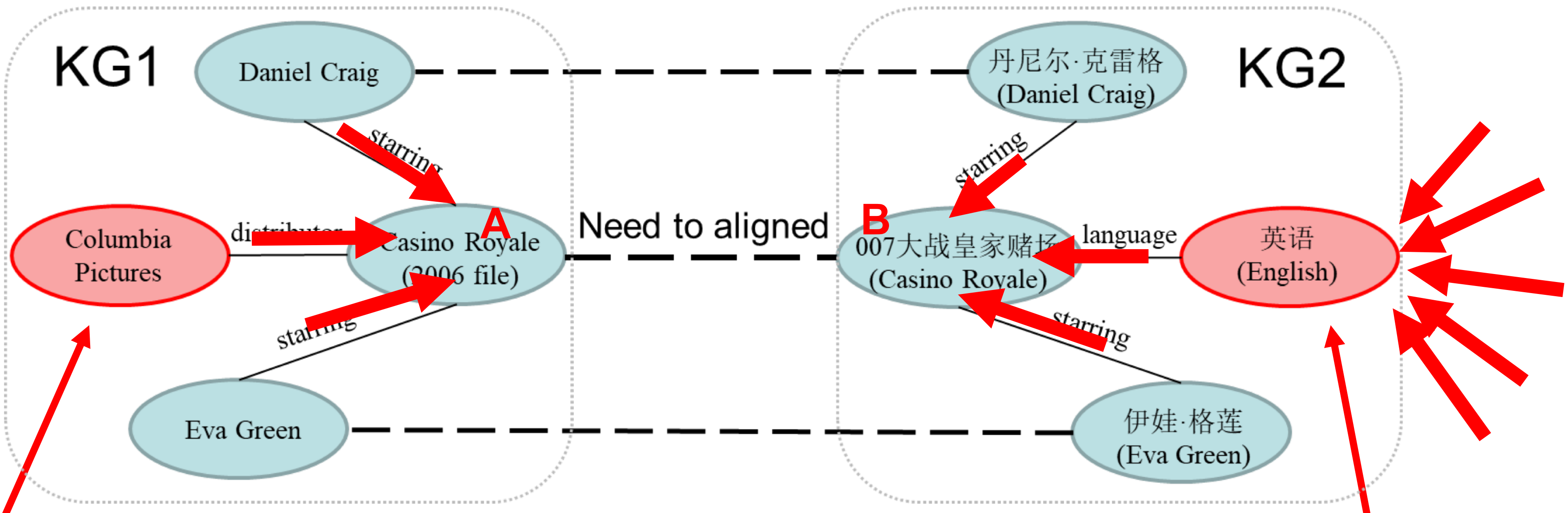
CYC KB WIKIDATA BabelNet

We demand a way to align different KGs!

Challenges

Different KGs are highly heterogeneous.

Existing works apply variant GCN to update node embedding **by aggregating all neighbors**’ embedding. **It may introduce noise and harm performance**



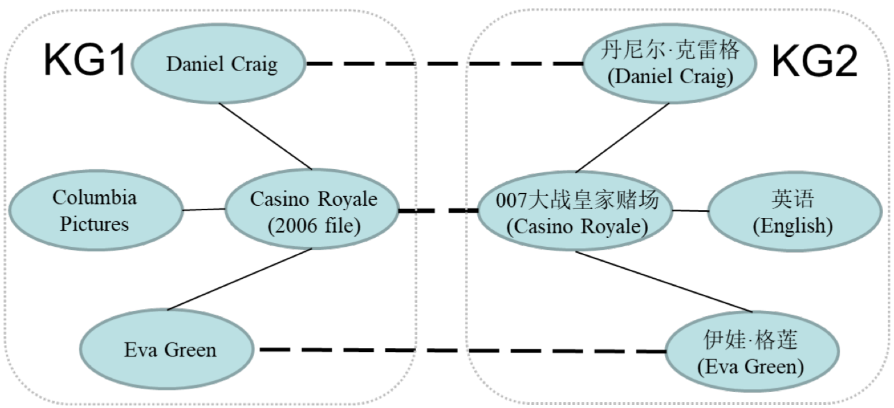
No counterpart can be found in neighbor of “007大战皇家赌场”

Hub Entity “英语” has 800+ neighbors

Solution

Compute interactions between neighbors

- Capture the fine-grained exact matches between neighbors
- Get rid of negative influence from dissimilar neighbors



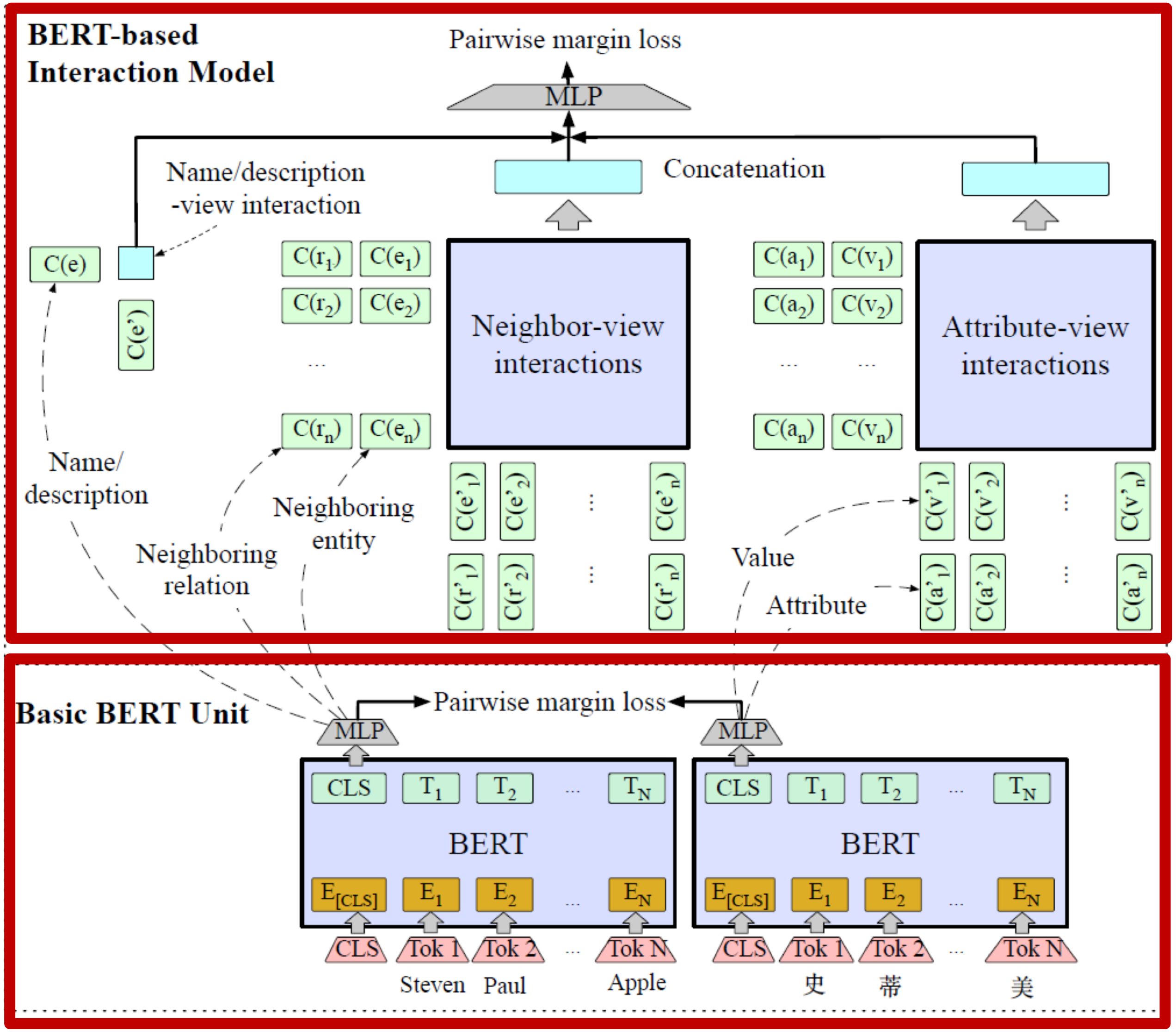
	丹尼尔·克雷格 (Daniel Craig)	伊娃·格莲 (Eva Green)	英语 (English)
Daniel Craig	1.0	0.0	0.0
Columbia Pictures	0.0	0.0	0.0
Eva Green	0.0	**1.0**	0.0

Capture the fine-grained exact match between “Eva Green” and “伊娃·格莲”

Framework

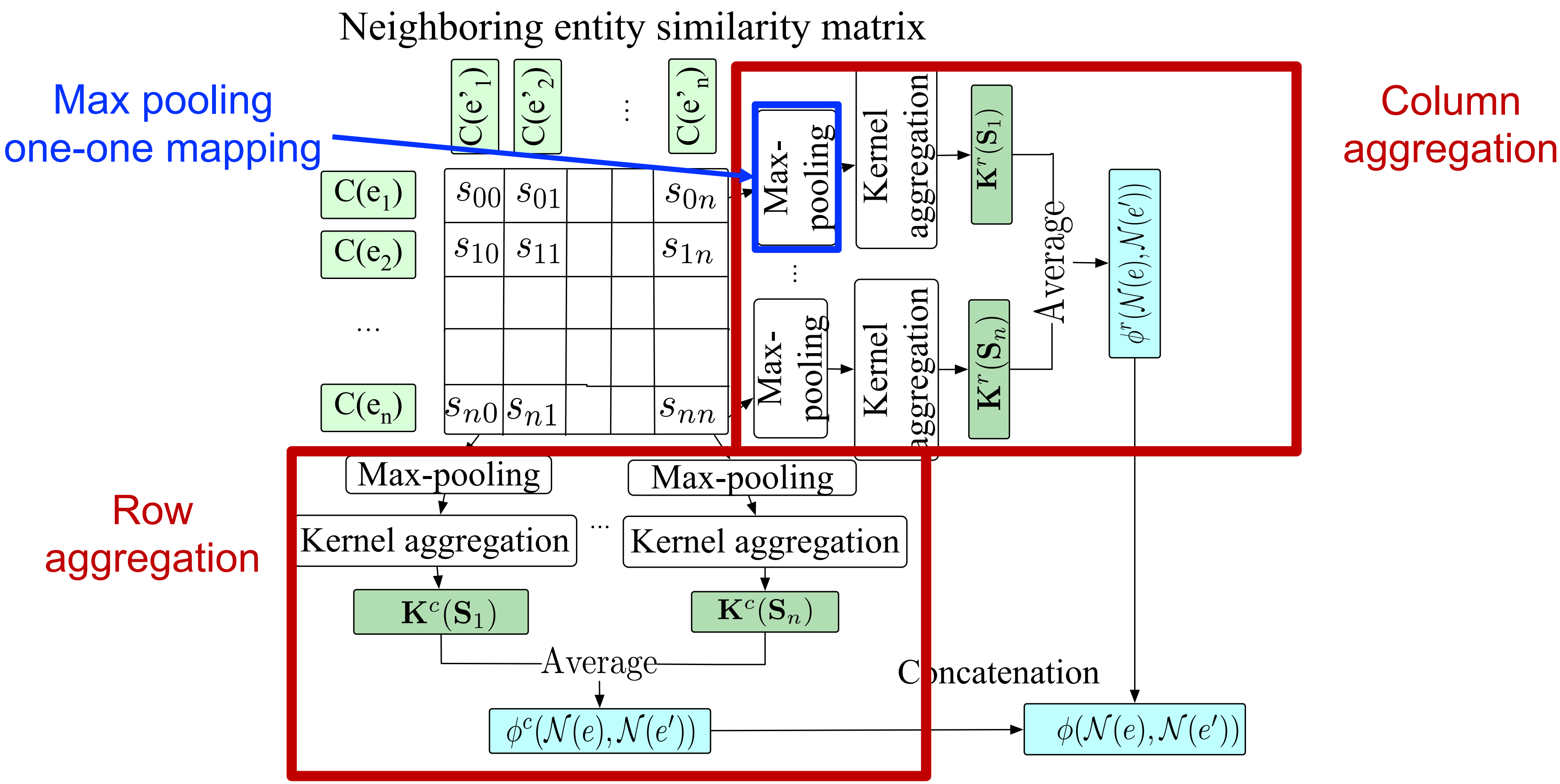
Compute the interactions of different views

Embed name/description/attribute/value of an entity



Neighbor-view Interactions

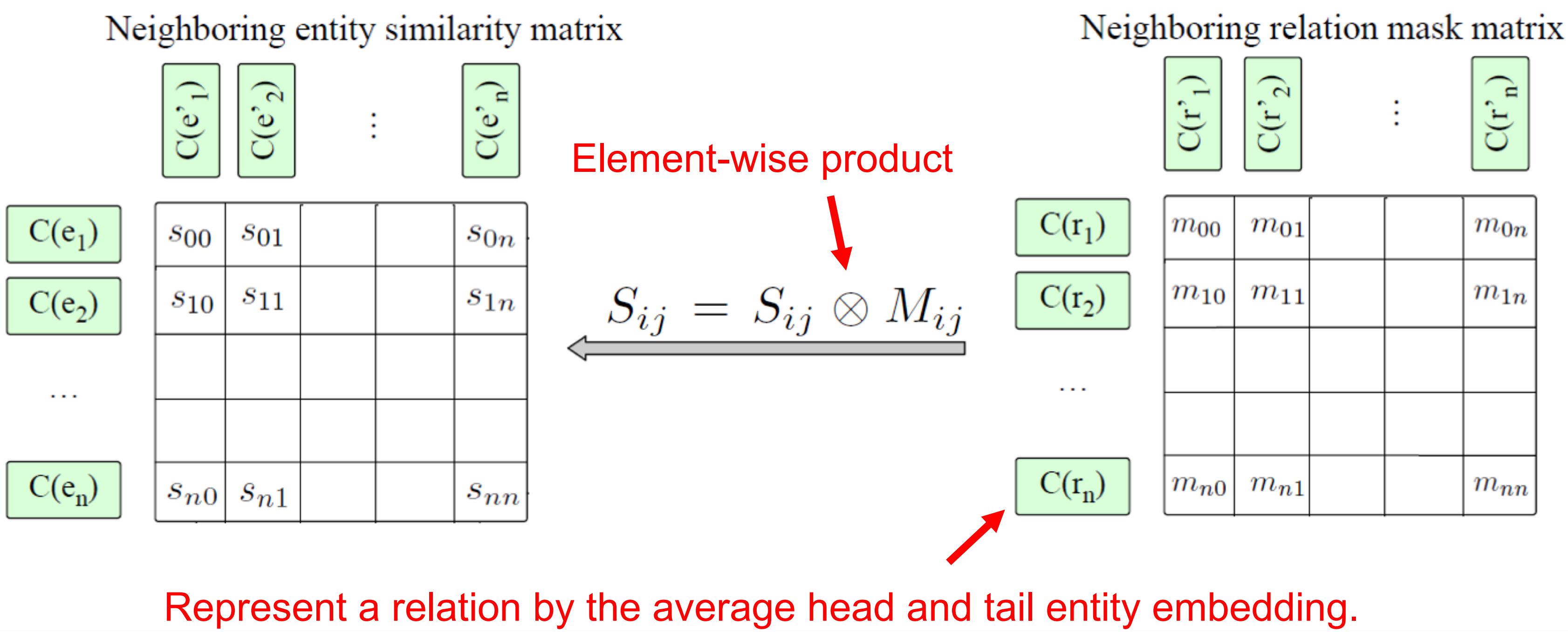
Propose a dual aggregation function to capture exact and soft matches between each pair of neighbors



Neighbor-view Interactions

A pair of neighbors are more convincing for supporting the alignment of two entities if the corresponding relations are also similar.

For triplet (e, r_i, e_i) and (e', r_j', e_j'), if r_i is similar to r_j' and e_i is similar to e_j', e and e' will be more likely to be aligned.



Dataset

- DBP15K
- https://github.com/nju-websoft/JAPE
- Each pair of cross-lingual KGs has 15,000 inter-lingual links (use 30% seed alignment to predict the rest).

Datasets		Entities	Relationships	Attributes	Rel. triples	Attr. triples
$\text{DBP15K}_{\text{ZH-EN}}$	Chinese	66,469	2,830	8,113	153,929	379,684
	English	98,125	2,317	7,173	237,674	567,755
$\text{DBP15K}_{\text{JA-EN}}$	Japanese	65,744	2,043	5,882	164,373	354,619
	English	95,680	2,096	6,066	233,319	497,230
$\text{DBP15K}_{\text{FR-EN}}$	French	66,858	1,379	4,547	192,191	528,665
	English	105,889	2,209	6,422	278,590	576,543

DBP15K: https://github.com/nju-websoft/JAPE

Experiment Result

Model	DBP15K$_{ZH-EN}$ HR1	DBP15K$_{ZH-EN}$ HR10	DBP15K$_{ZH-EN}$ MRR	DBP15K$_{JA-EN}$ HR1	DBP15K$_{JA-EN}$ HR10	DBP15K$_{JA-EN}$ MRR	DBP15K$_{FR-EN}$ HR1	DBP15K$_{FR-EN}$ HR10	DBP15K$_{FR-EN}$ MRR
Only use graph structures by variant TransE									
MTransE [Chen *et al.*, 2017]	0.308	0.614	0.364	0.279	0.575	0.349	0.244	0.556	0.335
IPTransE [Zhu *et al.*, 2017]	0.406	0.735	0.516	0.367	0.693	0.474	0.333	0.685	0.451
BootEA [Sun *et al.*, 2018]	0.629	0.848	0.703	0.622	0.854	0.701	0.653	0.874	0.731
RSNs [Guo *et al.*, 2019]	0.508	0.745	0.591	0.507	0.737	0.590	0.516	0.768	0.605
TransEdge [Sun *et al.*, 2019]	0.735	0.919	0.801	0.719	0.932	0.795	0.710	0.941	0.796
MRPEA [Shi and Xiao, 2019]	0.681	0.867	0.748	0.655	0.859	0.727	0.677	0.890	0.755
Only use graph structures by variant TransE plus GCN									
MuGNN [Cao *et al.*, 2019]	0.494	0.844	0.611	0.501	0.857	0.621	0.495	0.870	0.621
NAEA [Zhu *et al.*, 2019]	0.650	0.867	0.720	0.641	0.873	0.718	0.673	0.894	0.752
KECG [Li *et al.*, 2019]	0.478	0.835	0.598	0.490	0.844	0.610	0.486	0.851	0.610
AliNet [Sun *et al.*, 2020]	0.539	0.826	0.628	0.549	0.831	0.645	0.552	0.852	0.657
BERT-INT	**0.968**	**0.990**	**0.977**	**0.964**	**0.991**	**0.975**	**0.992**	**0.998**	**0.995**

Model	DBP15K$_{ZH-EN}$ HR1	DBP15K$_{ZH-EN}$ HR10	DBP15K$_{ZH-EN}$ MRR	DBP15K$_{JA-EN}$ HR1	DBP15K$_{JA-EN}$ HR10	DBP15K$_{JA-EN}$ MRR	DBP15K$_{FR-EN}$ HR1	DBP15K$_{FR-EN}$ HR10	DBP15K$_{FR-EN}$ MRR
Only use graph structures by variant TransE plus adversarial learning									
AKE [Lin *et al.*, 2019]	0.325	0.703	0.449	0.259	0.663	0.390	0.287	0.681	0.416
SEA [Pei *et al.*, 2019]	0.424	0.796	0.548	0.385	0.783	0.518	0.400	0.797	0.533
Combine graph structures and side information by variant GCN									
GCN-Align [Wang *et al.*, 2018]	0.413	0.744	0.549	0.399	0.745	0.546	0.373	0.745	0.532
GM-Align [Xu *et al.*, 2019]	0.679	0.785	-	0.740	0.872	-	0.894	0.952	-
RDGCN [Wu *et al.*, 2019a]	0.708	0.846	0.746	0.767	0.895	0.812	0.886	0.957	0.911
HGCN [Wu *et al.*, 2019b]	0.720	0.857	0.768	0.766	0.897	0.813	0.892	0.961	0.917
DGMC [Fey *et al.*, 2020]	0.772	0.897	-	0.774	0.907	-	0.891	0.967	-
Combine graph structures and side information by multi-view learning									
JAPE [Sun *et al.*, 2017]	0.412	0.745	0.490	0.363	0.685	0.476	0.324	0.667	0.430
MultiKE [Zhang *et al.*, 2019]	0.509	0.576	0.532	0.393	0.489	0.426	0.639	0.712	0.665
JarKA [Chen *et al.*, 2020]	0.706	0.878	0.766	0.646	0.855	0.708	0.704	0.888	0.768
HMAN [Yang *et al.*, 2019]	0.871	0.987	-	0.935	0.994	-	0.973	0.998	-
CEAFF [Zeng *et al.*, 2020]	0.795	-	-	0.860	-	-	0.964	-	-
BERT-INT	**0.968**	**0.990**	**0.977**	**0.964**	**0.991**	**0.975**	**0.992**	**0.998**	**0.995**

BERT-INT outperforms the best baselines by 9.7%-1.9% in HR1 on three datasets respectively

Ablation study

Model	DBP15K$_{ZH\text{-}EN}$			DBP15K$_{JA\text{-}EN}$			DBP15K$_{FR\text{-}EN}$		
	HR1	HR10	MRR	HR1	HR10	MRR	HR1	HR10	MRR
BERT-INT	**0.968**	**0.990**	**0.977**	**0.964**	**0.991**	**0.975**	**0.992**	**0.998**	**0.995**
Remove components									
-max pooling	0.962	0.989	0.973	0.959	0.991	0.973	0.992	0.998	0.995
-column aggregation	0.960	0.989	0.971	0.959	0.990	0.971	0.991	0.998	0.994
-neighbors	0.947	0.987	0.963	0.937	0.986	0.956	0.988	0.998	0.992
-attributes	0.919	0.984	0.945	0.938	0.987	0.957	0.983	0.998	0.990
-neighbors & attributes	0.830	0.970	0.883	0.848	0.974	0.897	0.965	0.995	0.978
Change the interaction component to variant GCN									
BERT-GCN	0.736	0.950	0.799	0.767	0.960	0.824	0.914	0.992	0.936
BERT-RDGCN	0.847	0.974	0.896	0.857	0.969	0.900	0.952	0.990	0.967
BERT-HMAN	0.911	0.993	0.943	0.937	0.994	0.960	0.982	0.999	0.989
Add components									
+relation mask	0.966	0.989	0.975	0.962	0.990	0.973	0.992	0.998	0.995
+attribute mask	0.942	0.986	0.959	0.950	0.990	0.966	0.989	0.998	0.993
+2-hop neighbors	0.965	0.990	0.975	0.964	0.991	0.975	0.992	0.998	0.995

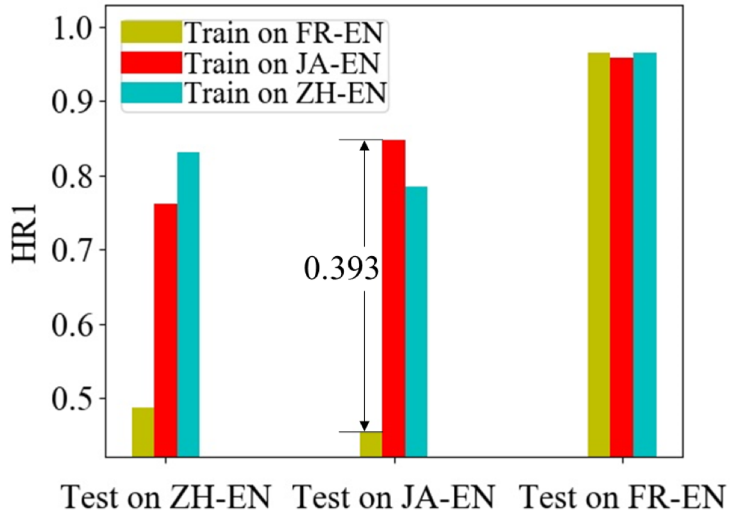
Max pooling and column aggregation take effect.

Interaction model outperforms GNNs

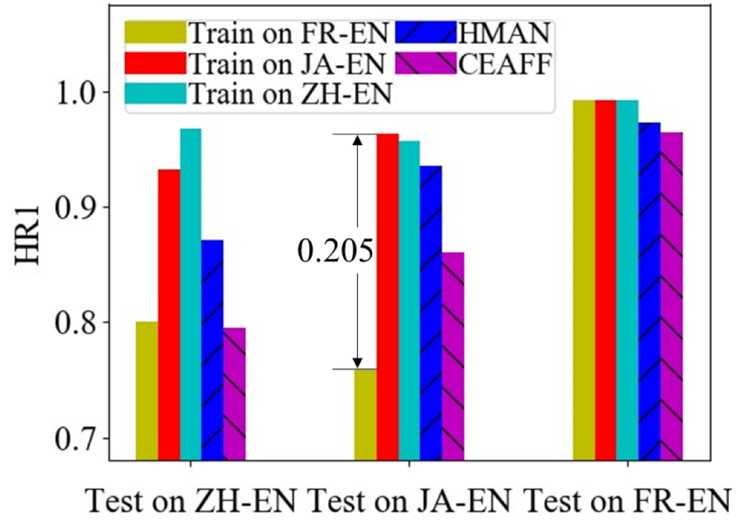
The effect of relation/attribute mask and high-order neighbors is not significant.

Inductive learning

We train the basic BERT unit and BERT-INT on one dataset of DBP15K, directly transfer and evaluate them on the other two datasets.



The basic BERT unit



BERT-INT

- Basic BERT unit and BERT-INT have inductive capacity
- The inductive capacity is not only caused by multi-lingual BERT, but also caused by the proposed interaction model

Conclusion

- Ambiguity and Heterogeneity
  - A multi-field multi-instance interaction model to match a paper and a person.
  - A BERT-based interaction model to match neighbors of entities.
  - Can capture fine-grained matches, and avoid the heterogeneity of graph structures
  - Jointly train the matching and the decision component to boost the both performance.

Thank you!

Knowledge Graph Linking and Integration

Jing Zhang, School of Information, Renmin University of China

Dataset:

https://www.aminer.cn/whoiswho

Challenges:

https://www.biendata.xyz/competition/chaindream_nd_task1/
https://www.biendata.xyz/competition/chaindream_nd_task2/

Code:

https://github.com/BoChen-Daniel/TKDE-2019-CONNA

https://github.com/kosugi11037/bert-int