南京大学
NANJING UNIVERSITY

大模型时代的知识融合

孙泽群

南京大学 万维网软件研究组

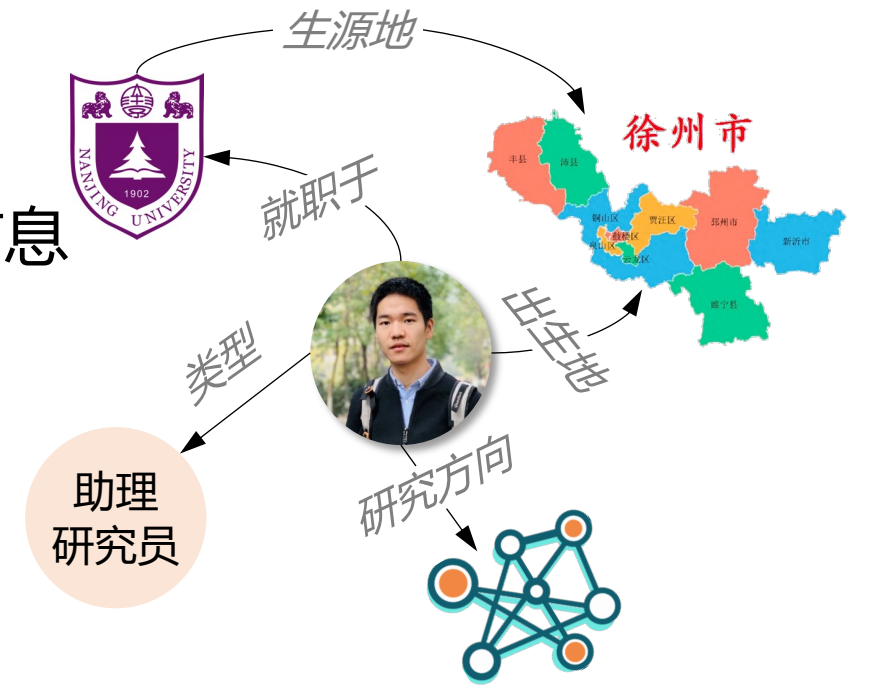
2024年9月20日

提纲

- **研究背景**
- 方法
- 应用
- 总结与展望

知识图谱

- 在知识图谱中
  - 点：**实体**或者**概念**，有名称、描述等属性信息
  - 边：有向，带标签，表示关系
- 事实可以用<主, 谓, 宾>三元组表示
  - <孙泽群，就职于，南京大学>



知识图谱以结构化的方式组织、描述和理解客观世界中的概念、实体及其之间的关系与属性，是结构化的语义知识库。

知识图谱

知识图谱具有天然的**不完备性**，但是不同来源的知识图谱可能存在**互补知识**

Freebase

DBpedia

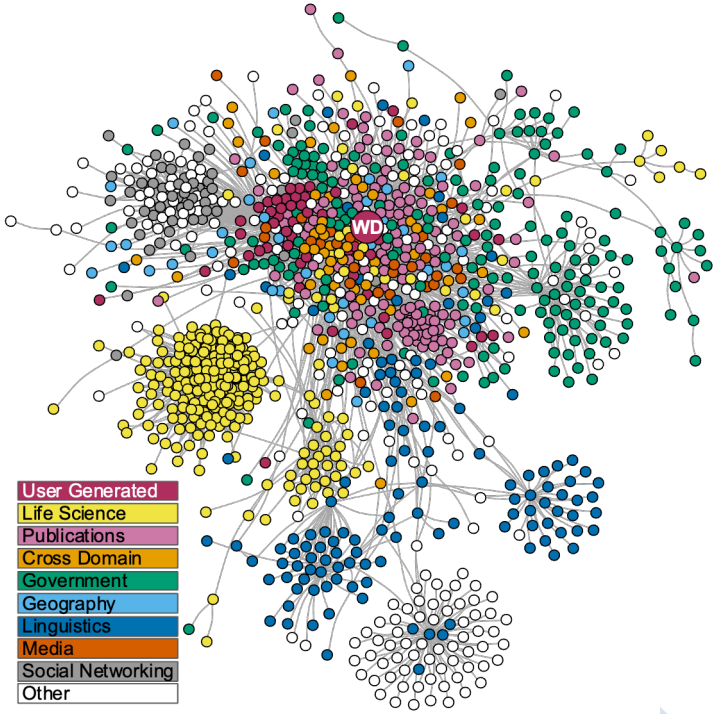
WIKIDATA

yago
select knowledge

ConceptNet
An open, multilingual knowledge graph

WordNet
A Lexical Database for English

Baidu 知识图谱



大数据 → **大知识**

不同来源但共指的实体

https://baike.baidu.com/item/重庆市/

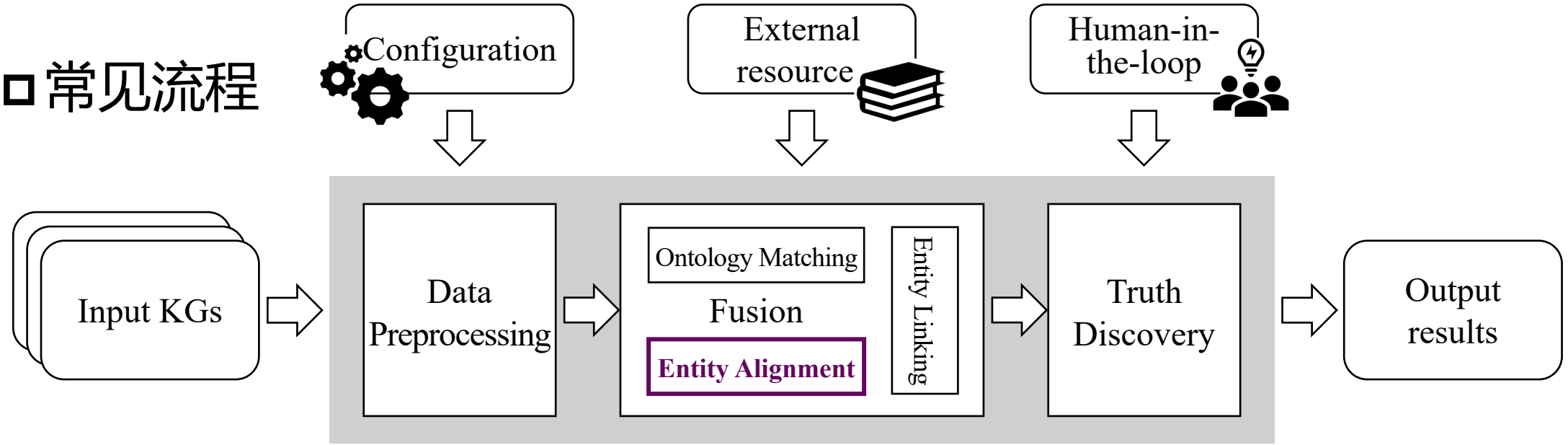
中文名	重庆市
外文名	Chongqing
	Chungking
别　名	山城、渝州、雾都、桥都、江城、巴渝、中国第一网
	红城市、8D魔幻之都
行政区划代码	500000
行政区类别	直辖市
所属地区	中国西南地区
地理位置	长江上游地区
	中国内陆西南部
面　积	82402 km²
下辖地区	26个区、8个县、4个自治县
政府驻地	渝中区人民路232号 [229]
电话区号	023（+86）
邮政编码	400000

https://en.wikipedia.org/wiki/Chongqing

Country	China
Settled	c. 316 BC
Separated from Sichuan	14 March 1997
Municipal seat	Yuzhong District
Divisions	26 districts, 12 counties
– County-level	
– Township-level	
Government	
• Type	Municipality
• Body	Chongqing Municipal People's Congress
• Party Secretary	Yuan Jiajun
• Congress Chairperson	Wang Jiong
• Mayor	Hu Henghua
• Municipal CPPCC Chairperson	Cheng Lihua
• National People's Congress Representation	58 deputies

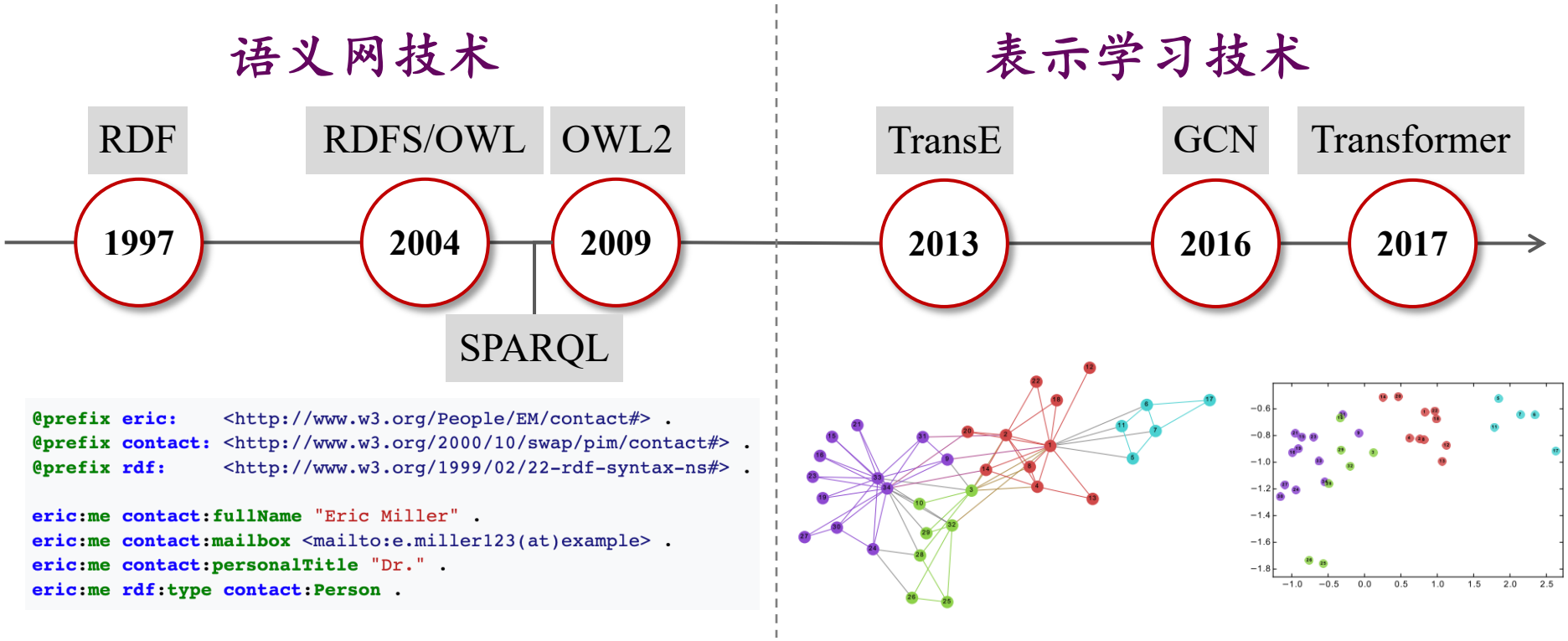
知识图谱融合

- 任务
  - 通过**实体对齐**、**本体匹配**、**真值发现**等操作，将不同知识图谱融合为统一的表达形式
- 常见流程



技术演变

- 知识图谱符号化表示 vs. 向量化表征



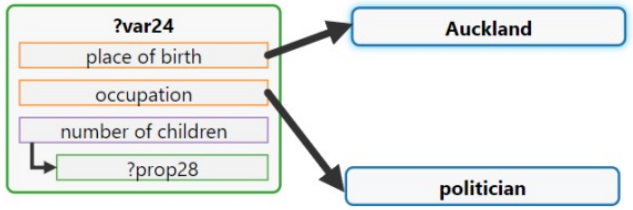
离散的、异质的符号表示与查询 *vs.* **连续的、统一的向量表征与计算**

技术演变

传统知识融合	基于表示学习的知识融合	基于大模型的知识融合
基于字符串相似度	基于图结构表示学习	生成式
基于概率模型	基于文本表示学习	判别式
基于规则	基于多模态表示学习	LLM在回路
人在回路	主动学习	……

目标演变

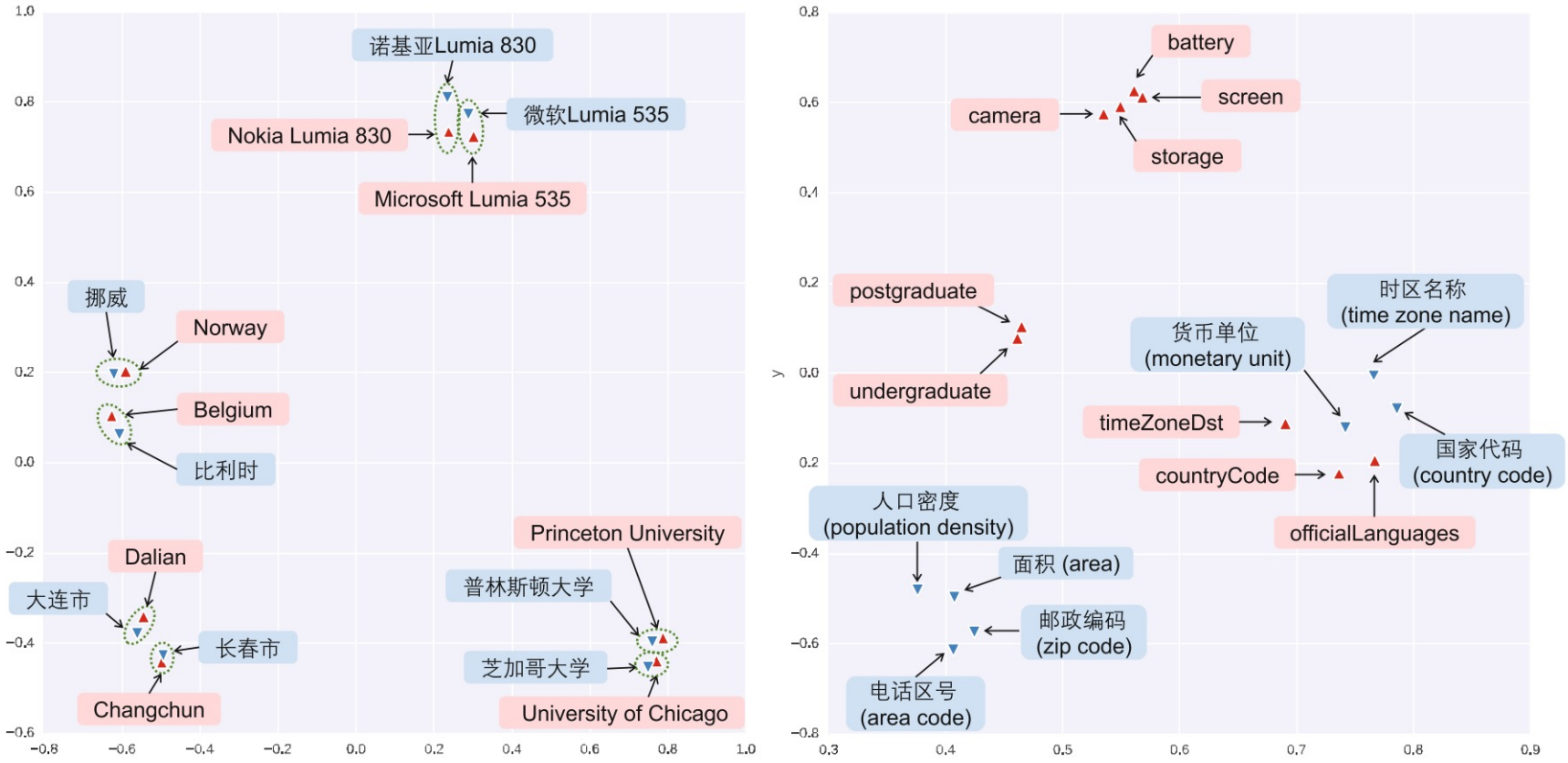
- 语义网时代：知识互操作，帮助机器更好地理解和处理数据



Hernán Vargas, Carlos Buil-Aranda, Aidan Hogan: RDF Explorer: A Visual Query Builder for Semantic Web Knowledge Graphs. ISWC (Satellites) 2019: 229-232

目标演变

- 表示学习时代：在**向量空间**捕捉和计算多源知识图谱的**语义相似度**



Zequn Sun, Wei Hu, Chengkai Li: Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding. ISWC 2017: 628-644

大模型时代的多源知识融合

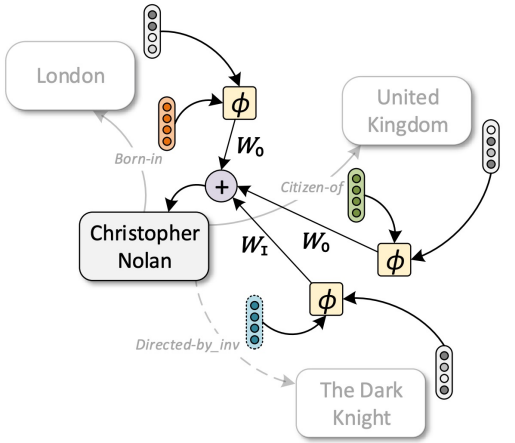
单源知识不完备给知识图谱表征与智能应用带来重大挑战

现有工作：

从“简单浅层”到“复杂深层”

从表示简单的显式知识，到挖掘隐式的、深层关联的复杂知识

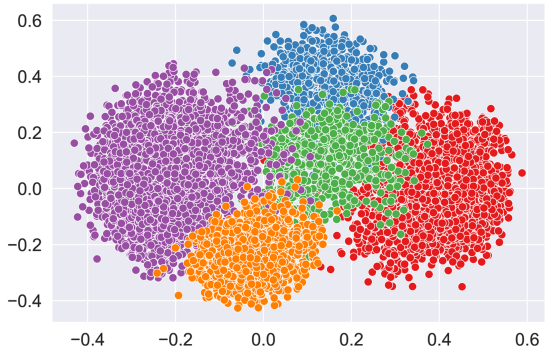
增加模型复杂度，未解决单源知识不完备的本质问题



解决思路：

从“单源同构”到“多源异构”

捕捉图谱内部语义，对齐多源表征空间，实现**多源知识融合与迁移**



大模型时代的多源知识融合

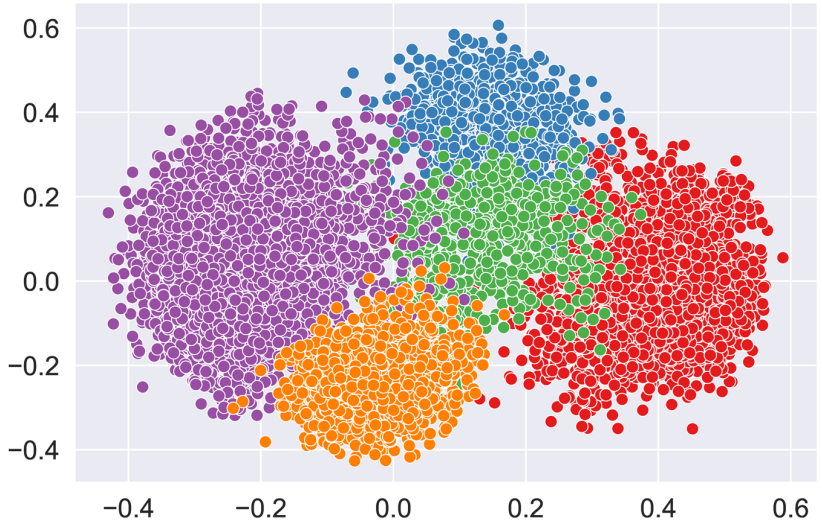
Freebase

DBpedia

WIKIDATA

WordNet

统一的语义计算和知识迁移空间



多模态

知识检索

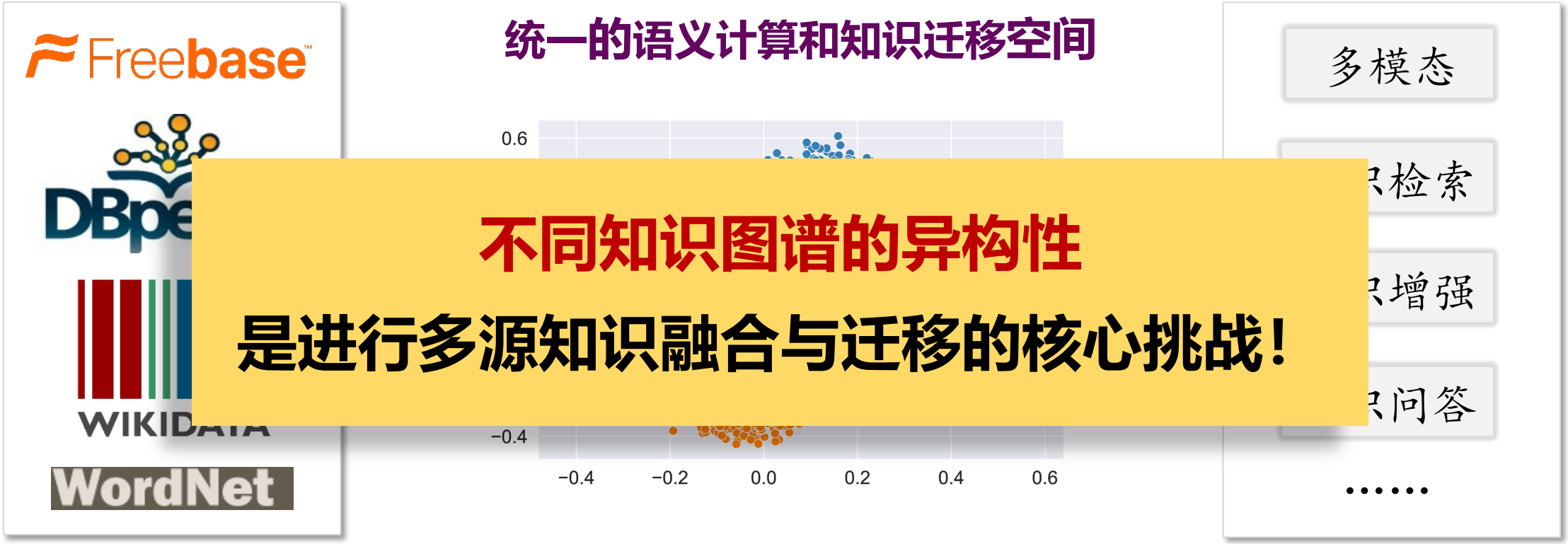
知识增强

知识问答

……

链接开放数据的积累 **人工智能技术的发展** **智能应用的知识需求**

大模型时代的多源知识融合



不同知识图谱的异构性

是进行多源知识融合与迁移的核心挑战！

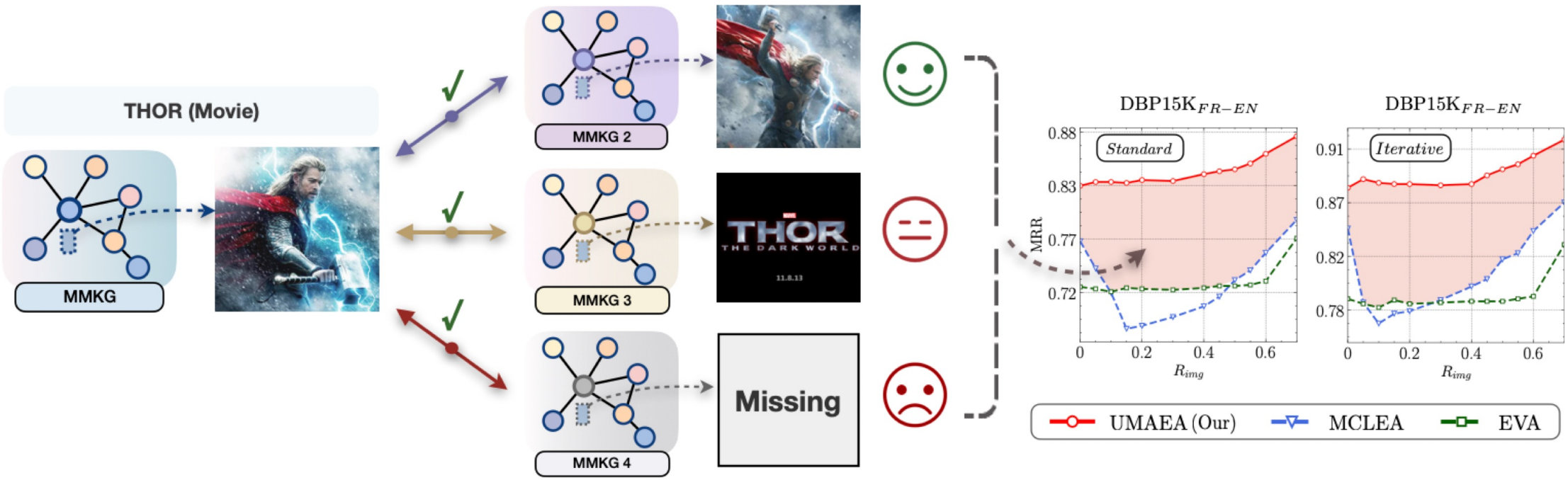
链接开放数据的积累　　人工智能技术的发展　　智能应用的知识需求

提纲

- 研究背景
- **方法**
- 应用
- 总结与展望

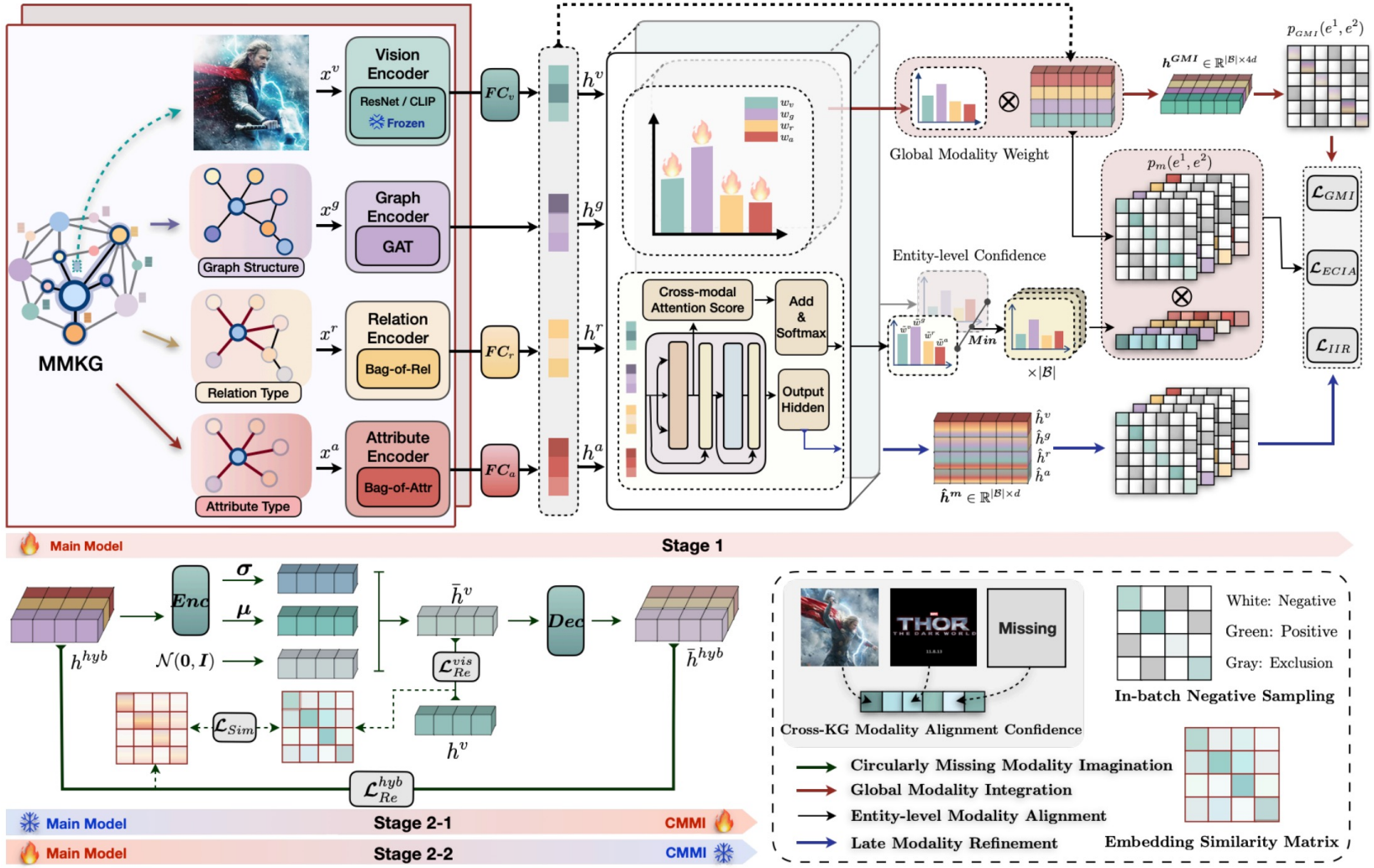
工作1：消解实体模态缺失与歧义

实体的图像信息存在噪音甚至部分实体缺乏图像



Zhuo Chen, Lingbing Guo, Yin Fang, Yichi Zhang, Jiaoyan Chen, Jeff Z. Pan, Yangning Li, Huajun Chen, Wen Zhang: Rethinking Uncertainly Missing and Ambiguous Visual Modality in Multi-Modal Entity Alignment. ISWC 2023: 121-139

工作1：消解实体模态缺失与歧义



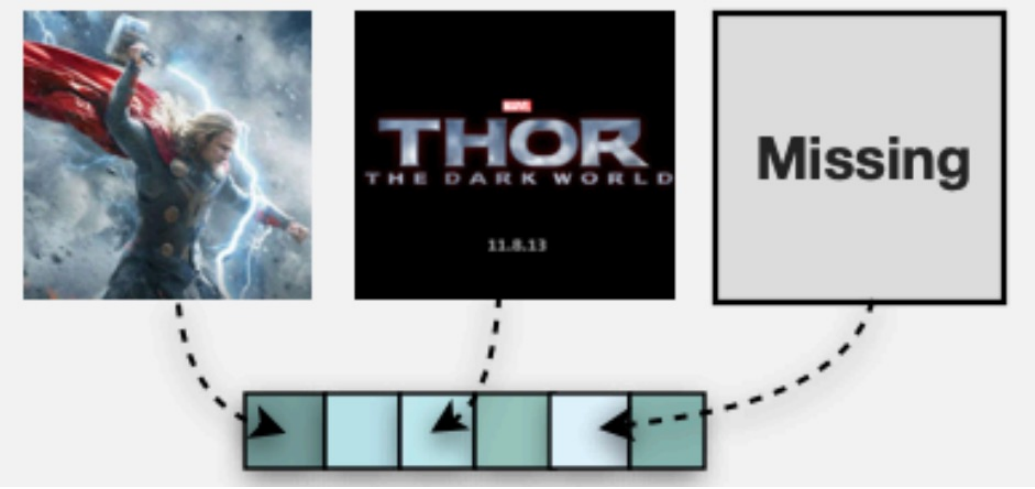
工作1：消解实体模态缺失与歧义

- **多尺度模态混合**
  - **全局模态集成**：通过可学习的全局权重进行自适应对齐，强调每个多模态实体对的全局对齐
  - **实体级模态对齐**：利用对齐种子的最小跨知识图谱置信度来约束模态对齐。通过动态调整模态权重，能够更精确地对齐不同图谱中的实体
  - **后置模态细化**：通过Transformer层输出，进一步增强实体级自适应模态对齐。利用隐式级联细化目标，确保对齐过程的精确性和鲁棒性

工作1：消解实体模态缺失与歧义

循环缺失模态想象

- 从变分自编码器和生成对抗网络中汲取灵感
- 通过生成建模和无监督领域迁移技术，主动生成缺失的模态信息
- 模型在面对模态缺失时仍能有效进行推理和预测



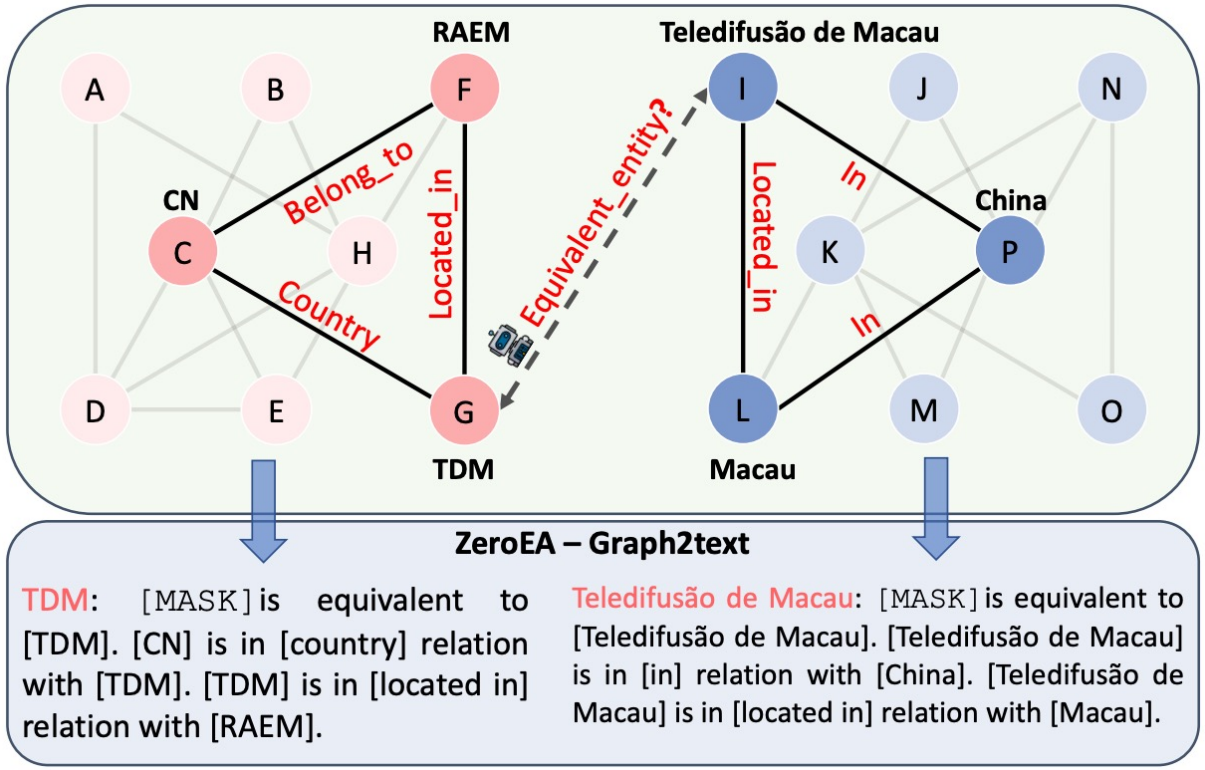
工作1：消解实体模态缺失与歧义

面对模态不完整时，模型容易过拟合噪声，出现性能波动或下降。通过多尺度模态混合和循环缺失模态想象，有效缓解了这些问题

	Models	$R_{img}=0.05$			$R_{img}=0.2$			$R_{img}=0.4$			$R_{img}=0.6$		
		H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
$DBP15K_{ZH-EN}$	MSNEA [7]	.413	.722	.517	.411	.725	.518	.446	.743	.546	.520	.786	.611
	EVA [30]	.623	.878	.715	.624	.878	.716	.623	.875	.714	.625	.876	.717
	MCLEA [29]	.638	.905	.732	.588	.865	.686	.611	.874	.704	.661	.896	.744
	w/o CMMI	.703	.934	.787	.710	.937	.793	.721	.939	.801	.753	.949	.825
	UMAEA	**.720**	**.938**	**.800**	**.727**	**.941**	**.806**	**.727**	**.941**	**.806**	**.758**	**.951**	**.829**
	Improve ↑	8.2%	3.3%	.068	10.3%	6.3%	.090	10.4%	6.6%	.092	9.7%	5.5%	.085
$DBP15K_{JA-EN}$	MSNEA [7]	.313	.643	.425	.311	.644	.422	.369	.678	.472	.480	.744	.569
	EVA [30]	.615	.877	.708	.616	.877	.710	.616	.878	.711	.624	.881	.716
	MCLEA [29]	.599	.897	.706	.579	.846	.675	.613	.867	.703	.686	.898	.761
	w/o CMMI	.708	.943	.794	.712	.947	.798	.730	.950	.810	.772	.962	.843
	UMAEA	**.725**	**.949**	**.807**	**.726**	**.949**	**.808**	**.732**	**.952**	**.813**	**.775**	**.963**	**.845**
	Improve ↑	11.0%	5.2%	.099	11.0%	7.2%	.098	11.6%	7.4%	.102	8.9%	6.5%	.084
$DBP15K_{FR-EN}$	MSNEA [7]	.297	.690	.427	.304	.690	.428	.360	.710	.474	.478	.772	.574
	EVA [30]	.624	.895	.720	.624	.895	.720	.626	.898	.721	.634	.900	.728
	MCLEA [29]	.634	.930	.741	.582	.863	.682	.601	.879	.702	.675	.901	.757
	w/o CMMI	.727	.956	.813	.733	.960	.817	.746	.961	.828	.790	.968	.857
	UMAEA	**.752**	**.970**	**.830**	**.755**	**.960**	**.832**	**.763**	**.962**	**.838**	**.792**	**.970**	**.859**
	Improve ↑	11.8%	4.0%	.089	13.1%	6.7%	.112	13.7%	6.4%	.117	11.7%	6.9%	.102

工作2：大模型编码图结构

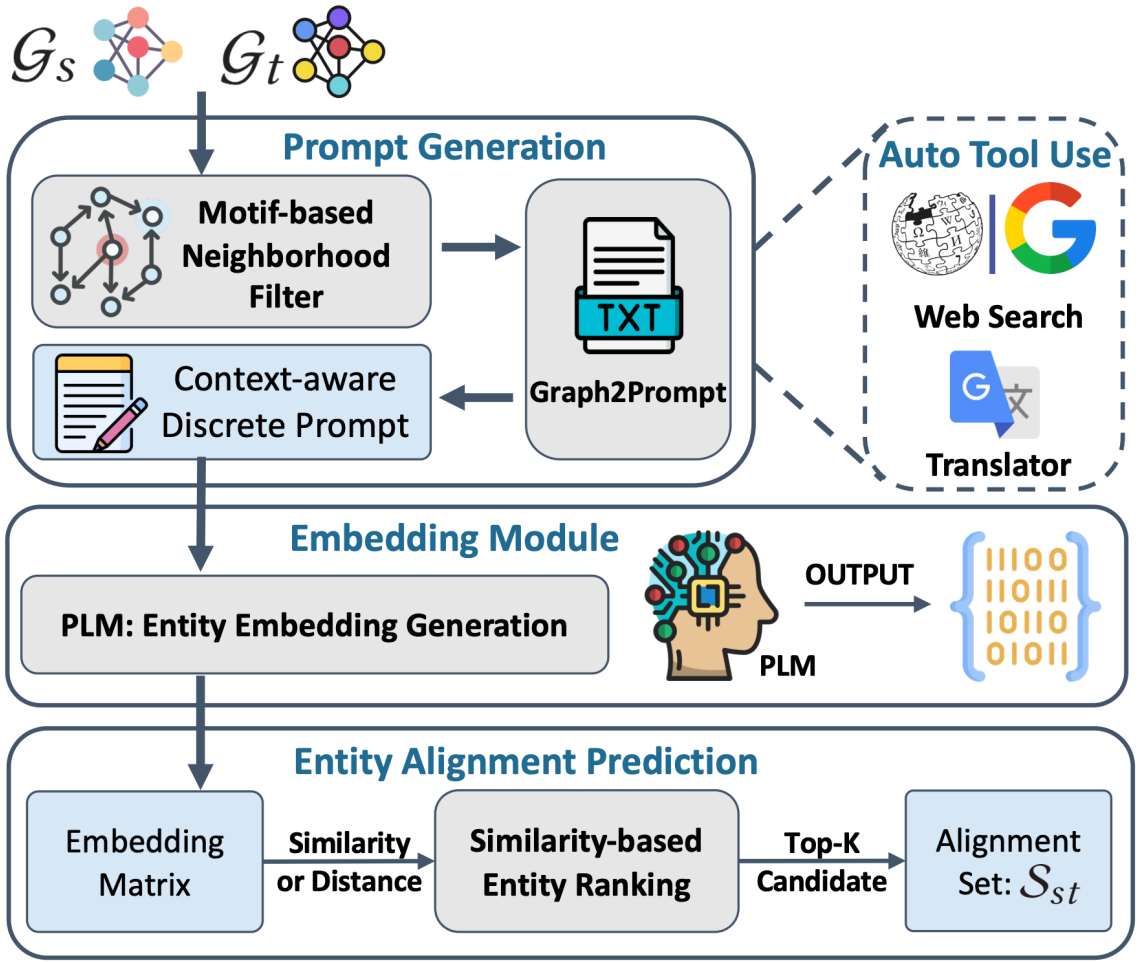
大模型如何统一编码文本和图结构信息



Nan Huo, Reynold Cheng, Ben Kao, Wentao Ning, Nur Al Hasan Haldar, Xiaodong Li, Jinyang Li, Mohammad Matin Najafi, Tian Li, Ge Qu: ZeroEA: A Zero-Training Entity Alignment Framework via Pre-Trained Language Model. Proc. VLDB Endow. 17(7): 1765-1774 (2024)

工作2：大模型编码图结构

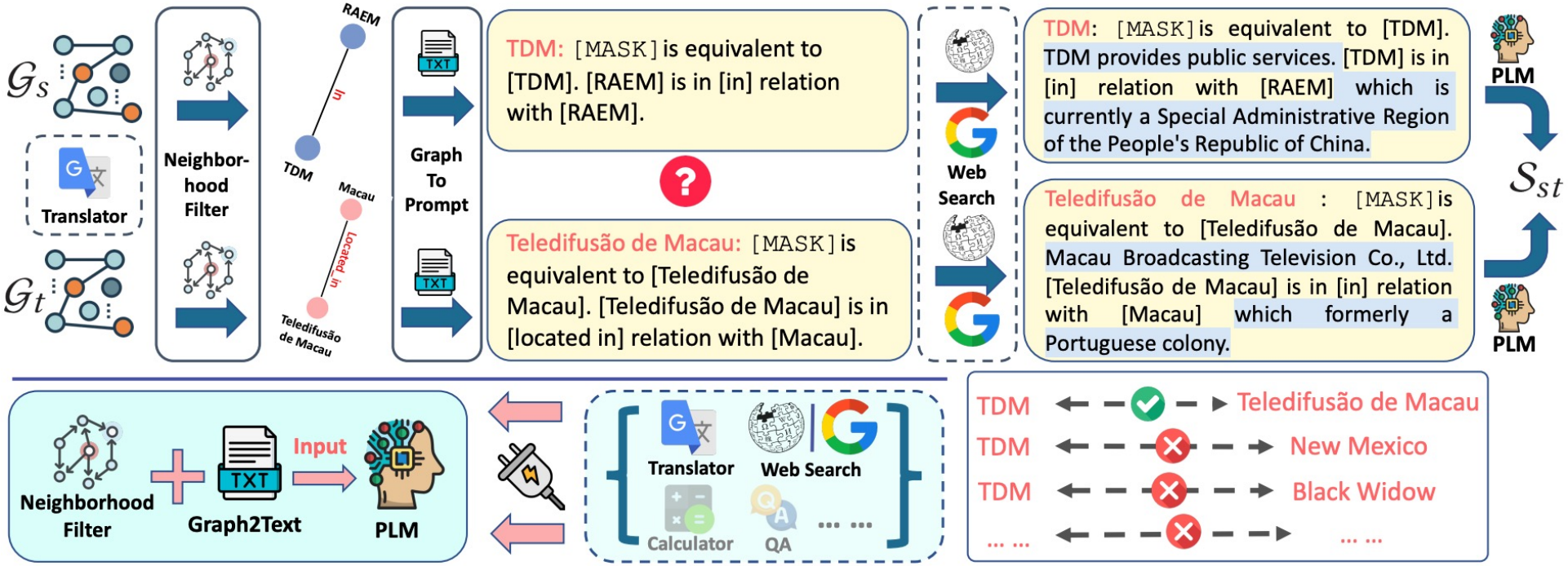
- 思路：
  - 图文转换
  - 通过工具拓展特征，构造带掩码的提示
  - 掩码表示作为实体表征



Nan Huo, Reynold Cheng, Ben Kao, Wentao Ning, Nur Al Hasan Haldar, Xiaodong Li, Jinyang Li, Mohammad Matin Najafi, Tian Li, Ge Qu: ZeroEA: A Zero-Training Entity Alignment Framework via Pre-Trained Language Model. Proc. VLDB Endow. 17(7): 1765-1774 (2024)

工作2：大模型编码图结构

- 图文转换模块
  - 基于Motif值过滤噪音邻居
  - 基于预定义模板进行图结构的文本化处理



工作2：大模型编码图结构

- 工具模块
  - 机器翻译：适用于跨语言实体对齐
  - Web搜索工具：搜索相关实体的网络信息
- 表征生成模块
  - 将实体的各种信息通过模版整合为大模型的提示
  - 输入提示，将掩码对应的隐藏层表示作为实体的向量表征
- 实体对齐预测模块
  - 近邻检索

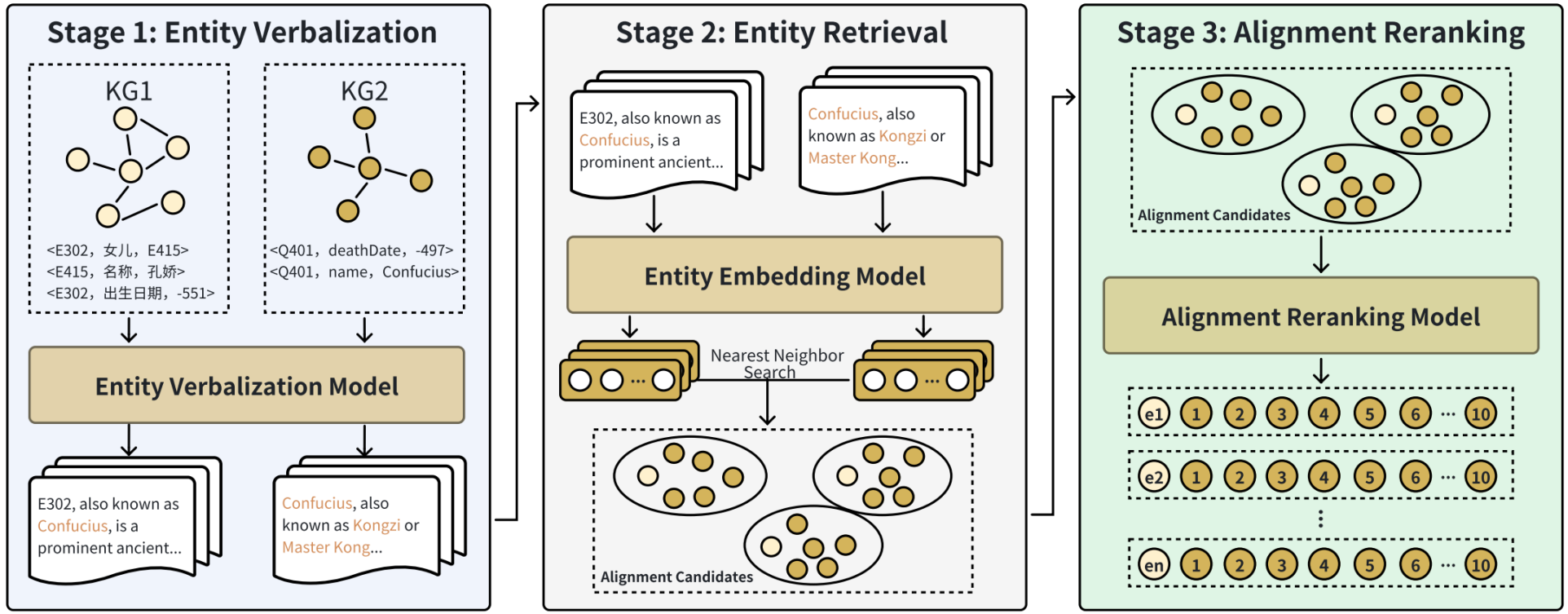
工作2：大模型编码图结构

ZeroEA在多个实体对齐基准数据集上取得了SOTA结果

Model		$DBP15K_{zh\_en}$			$DBP15K_{ja\_en}$			$DBP15K_{fr\_en}$			$DWY100K_{dbp\_wd}$			$DWY100K_{dbp\_yg}$		
		Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR
Supervised																
Trans.	MTransE [7]	0.308	0.614	0.364	0.279	0.575	0.349	0.244	0.556	0.335	0.281	0.520	0.362	0.252	0.493	0.376
	JAPE [44]	0.412	0.745	0.490	0.363	0.685	0.476	0.324	0.667	0.430	0.318	0.589	0.378	0.236	0.484	0.364
	BootEA [45]	0.629	0.848	0.703	0.622	0.854	0.701	0.653	0.874	0.731	0.748	0.791	0.898	0.761	0.894	0.818
	TransEdge [3]	0.735	0.919	0.801	0.719	0.932	0.795	0.710	0.941	0.796	0.788	0.938	0.832	0.792	0.936	0.889
GNN	GCN-Align [53]	0.413	0.744	-	0.399	0.745	-	0.373	0.745	-	0.477	0.562	0.514	0.601	0.642	0.623
	MuGNN [6]	0.494	0.844	0.611	0.501	0.857	0.621	0.495	0.870	0.621	0.616	0.897	0.732	0.741	0.937	0.856
	RDGCN [54]	0.708	0.846	0.746	0.767	0.895	0.812	0.886	0.957	0.911	0.902	0.954	0.923	0.864	0.889	0.973
	CEAFF [58]	0.795	-	-	0.860	-	-	0.964	-	-	1.000	-	-	1.000	-	-
	MEAformer [9]	0.949	0.993	0.965	0.978	0.999	0.986	0.991	1.00	0.995	-	-	-	-	-	-
PLM	BERT-INT [47]	0.968	0.990	0.977	0.964	0.991	0.975	0.995	0.998	0.995	0.992	0.999	0.999	0.999	0.999	0.999
	SDEA [65]	0.870	0.966	0.910	0.848	0.952	0.890	0.969	0.995	0.980	0.980	0.996	0.990	0.999	1.0	1.0
Unsupervised & Self-supervised																
Trans.	MultiKE [61]	0.509	0.576	0.532	0.393	0.489	0.432	0.639	0.712	0.665	0.915	0.974	0.932	0.880	0.962	0.916
GNN	SelfKG [32]	0.829	0.919	-	0.890	0.953	-	0.959	0.992	-	0.983	0.998	-	0.998	1.000	-
PLM	ZeroEA(dir)	0.972	0.990	0.981	0.975	0.992	0.981	0.983	0.992	0.988	0.986	0.991	0.988	0.999	1.000	0.999
	ZeroEA(undir)	**0.985**	**0.993**	**0.991**	**0.982**	**0.995**	**0.989**	**0.998**	**0.999**	**0.998**	**0.998**	**0.999**	**0.996**	**0.999**	**1.000**	**0.999**

工作3：实体检索增强

如何利用生成式大模型进行实体对齐呢?



Zhichun Wang, Xuan Chen: DERA: Dense Entity Retrieval for Entity Alignment in Knowledge Graphs. CoRR abs/2408.01154 (2024)

工作3：实体检索增强

- 实体语言化
  - 将实体的异构三元组转换为统一的自然语言描述
  - 使用生成语言模型（如GPT）训练语言模型
- 实体检索
  - 将实体描述编码为向量，使用实体表征模型（如BGE）进行训练，并基于表征相似性检索最相似的实体
- 对齐重排序
  - 设计一个基于BERT的重排序模型，候选对齐进行重排序，确保对齐的精度

工作3：实体检索增强

实体检索的方法在标准数据集的多个特征设置上都取得了优异的性能

Info.	Model	DBP15K-ZH-EN			DBP15K-JA-EN			DBP15K-FR-EN		
		Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
Attributes	JAPE	0.412	0.745	0.490	0.363	0.685	0.476	0.324	0.667	0.430
	GCN-Align	0.413	0.744	0.549	0.399	0.745	0.546	0.373	0.745	0.532
	JarKA	0.706	0.878	0.766	0.646	0.855	0.708	0.704	0.888	0.768
	DERA(Ours)	**0.946**	**0.982**	**0.961**	**0.921**	**0.959**	**0.937**	**0.949**	**0.985**	**0.964**
Names	GMNN	0.679	0.785	–	0.740	0.872	–	0.894	0.952	–
	SelfKG	0.745	0.866	–	0.816	0.913	–	0.957	0.992	–
	TEA-NSP	0.815	0.953	0.870	**0.890**	**0.967**	**0.920**	0.968	0.995	0.980
	TEA-MLM	0.831	0.957	0.880	0.883	0.966	0.910	0.968	0.994	0.980
	DERA(Ours)	**0.846**	**0.962**	**0.900**	0.866	0.951	0.889	**0.980**	**0.996**	**0.987**
Names & Attributes	HMAN	0.871	0.987	–	0.935	**0.994**	–	0.973	0.998	–
	AttrGNN	0.796	0.929	0.845	0.783	0.921	0.834	0.919	0.978	0.910
	BERT-INT	**0.968**	0.990	0.977	0.964	0.991	0.975	**0.992**	0.998	**0.995**
	ICLEA	0.884	0.972	–	0.924	0.978	–	0.991	**0.999**	–
	TEA-NSP	0.941	0.983	0.960	0.941	0.979	0.960	0.979	0.997	0.990
	TEA-MLM	0.935	0.982	0.950	0.939	0.978	0.950	0.987	0.996	0.990
	DERA(Ours)	**0.968**	**0.994**	**0.979**	**0.967**	0.992	**0.978**	0.989	**0.999**	**0.995**
Translated Names	HGCN-JE	0.720	0.857	–	0.766	0.897	–	0.892	0.961	–
	RDGCN	0.708	0.846	0.746	0.767	0.895	0.812	0.886	0.957	0.911
	NMN	0.733	0.869	–	0.785	0.912	–	0.902	0.967	–
	DERA(Ours)	0.930	0.982	0.950	0.917	0.978	0.941	0.972	0.995	0.982
	DATTI†	0.890	0.958	–	0.921	0.971	–	0.979	0.990	–
	SEU†	0.900	0.965	0.924	0.956	0.991	0.969	0.988	**0.999**	0.992
	EASY†	0.898	0.979	0.930	0.943	0.990	0.960	0.980	0.998	0.990
	CPL-OT†	0.927	0.964	0.940	0.956	0.983	0.970	0.990	0.994	0.990
	UED†	0.915	–	–	0.941	–	–	0.984	–	–
	LightEA†	0.952	0.984	0.964	0.981	0.997	0.987	0.995	0.998	0.996
	DERA†(Ours)	**0.985**	**0.997**	**0.990**	**0.994**	**0.999**	**0.996**	**0.996**	**0.999**	**0.997**

工作4：用大模型进行数据标注

实体对齐标注数据不足的问题严重影响模型效果

Cross-language overlap: Number of instances that are described in multiple languages.

Class	*Instances*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10+*
Person	871.630	676.367	94.339	42.382	21.647	12.936	8.198	5.295	3.437	2.391	4.638
Place	643.260	307.729	150.349	45.836	36.339	20.831	13.523	20.808	31.422	11.262	5.161
Organisation	206.670	160.398	22.661	9.312	5.002	3.221	2.072	1.421	928	594	1.061
Work	360.808	243.706	54.855	23.097	12.605	8.277	5.732	4.007	2.911	1.995	3.623

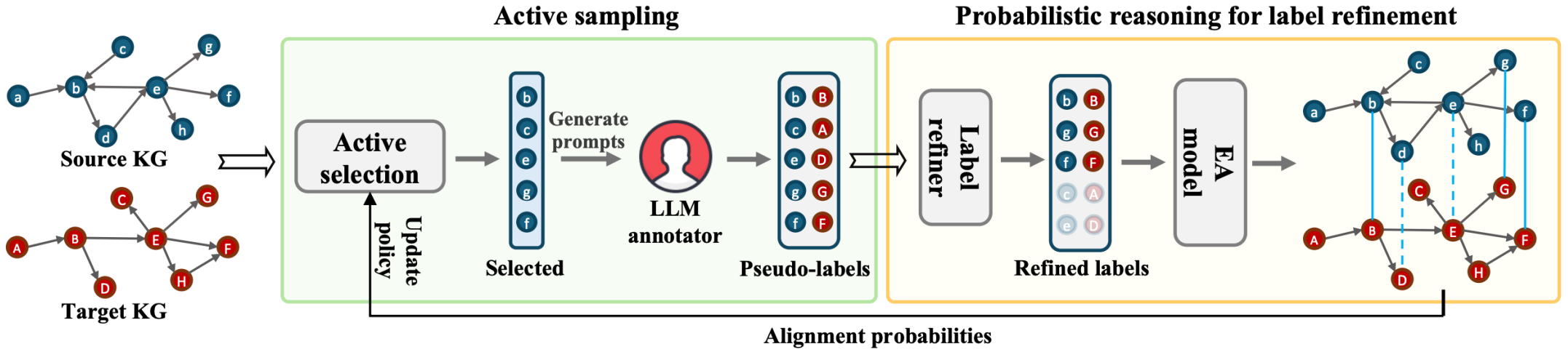
表示只在一种语言版本中出现过的实例数量，即没有实体对齐的数量。

例如，Person实例中，实体对齐约占比(871630-676367)/871630=22.4%。

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, Christian Bizer: DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6(2): 167-195 (2015)

工作4：用大模型进行数据标注

- 思路：LLM-in-the-loop
  - 主动学习选择待标注实体
  - 大模型用于实体对齐标注



Shengyuan Chen, Qinggang Zhang, Junnan Dong, Wen Hua, Qing Li, Xiao Huang: Entity Alignment with Noisy Annotations from Large Language Models. CoRR abs/2405.16806 (2024)

工作4：用大模型进行数据标注

- 实体选择
  - 考虑关系的确定性 $$U_r(e_h) = (1 - P(e_h)) + \sum_{(e_h, r, e_t) \in \mathcal{T}} w_r \, (1 - P(e_t))$$
  - 考虑实体的确定性 $$U_n(e_h) = (1 - P(e_h)) + \sum_{(e_h, r, e_t) \in \mathcal{T}} (1 - P(e_t))$$
- 大模型标注
  - 设计策略过滤掉不可能对齐的实体，构造潜在可能对齐的实体对
  - 构造提示，让大模型判断实体对是否是真的对齐

工作4：用大模型进行数据标注

使用大模型进行数据标注有效提高实体对齐效果

Table 1: Evaluation of entity alignment performance, measured by Hit@K for $K \in \{1, 10\}$, and Mean Reciprocal Rank (MRR), presented in %. Experiment statistics are computed over three trials.

	EN-FR-15K			EN-DE-15K			D-W-15K			D-Y-15K		
	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR
Group1. Entity Alignment with GPT-3.5.												
IMUSE	50.0±0.1	72.6±0.8	57.5±0.4	51.6±4.7	75.9±3.9	60.5±4.5	6.0±0.2	14.6±2.5	9.0±1.0	54.4±2.5	78.9±1.1	63.2±2.0
AlignE	6.6±0.3	24.5±0.5	12.6±0.5	6.2±0.3	18.4±1.0	10.4±0.5	8.0±0.9	24.0±2.7	13.3±1.4	50.1±2.0	76.6±1.4	59.2±1.8
BootEA	44.8±1.1	71.9±1.2	54.2±1.2	68.1±0.2	85.4±0.3	74.3±0.2	60.8±0.2	79.3±0.1	67.4±0.2	87.8±0.1	96.7±0.1	91.2±0.1
GCNAlign	17.4±0.3	43.2±0.4	25.9±0.3	22.2±0.2	46.2±1.1	30.3±0.3	16.9±0.1	39.3±0.3	24.3±0.1	45.3±0.4	68.3±0.6	53.3±0.5
RDGCN	69.3±0.3	82.5±0.3	74.3±0.3	73.3±4.3	84.6±2.6	77.4±3.7	79.2±0.7	89.7±0.5	83.2±0.6	82.6±3.7	91.9±1.3	86.1±2.7
Dual-AMN	51.9±0.3	79.6±0.9	61.6±0.5	70.5±0.7	91.1±0.3	78.9±0.6	62.0±0.1	86.8±0.1	71.9±0.1	85.8±0.3	98.4±0.0	91.4±0.1
LLM4EA	**74.2±0.3**	**92.9±0.4**	**81.0±0.3**	**89.1±0.5**	**97.8±0.1**	**92.6±0.3**	**87.5±0.3**	**96.7±0.1**	**90.9±0.2**	**97.7±0.0**	**99.5±0.0**	**98.3±0.0**
Group2. Entity Alignment with GPT-4.												
IMUSE	52.7±0.9	74.9±1.0	59.8±0.9	59.6±2.6	81.8±1.5	67.9±2.1	21.6±6.1	50.0±10.0	31.1±7.4	86.6±0.5	94.2±0.1	89.2±0.4
AlignE	30.8±2.4	69.1±2.5	43.1±2.5	46.4±5.2	76.5±3.8	56.6±4.8	36.1±3.7	67.8±3.6	46.7±3.7	86.4±0.9	97.0±0.3	90.2±0.6
BootEA	58.2±0.3	83.7±0.3	67.0±0.3	80.5±0.4	92.6±0.2	84.8±0.3	71.6±0.2	88.3±0.2	77.6±0.2	95.0±0.1	98.6±0.0	96.3±0.1
GCNAlign	30.6±0.0	65.3±0.3	42.1±0.2	41.9±0.4	68.6±0.5	51.2±0.4	31.3±0.3	61.6±0.1	41.4±0.2	82.6±0.2	94.9±0.2	87.2±0.1
RDGCN	72.1±0.2	84.5±0.1	76.7±0.2	74.1±1.1	85.1±0.7	78.0±1.0	82.5±1.1	91.4±0.7	85.9±1.0	85.4±0.9	93.2±0.4	88.3±0.8
Dual-AMN	76.7±0.1	94.9±0.3	83.6±0.2	90.7±0.1	97.9±0.2	93.6±0.1	81.5±0.1	94.9±0.2	86.7±0.1	97.5±0.0	99.3±0.1	98.1±0.0
LLM4EA	**80.2±0.3**	**96.0±0.2**	**86.0±0.2**	**93.1±0.5**	**98.7±0.2**	**95.3±0.3**	**89.8±0.3**	**97.9±0.2**	**92.9±0.3**	**97.9±0.1**	**99.6±0.0**	**98.5±0.1**

工作4：用大模型进行数据标注

GPT3.5的标注预算越高，性能提升越大

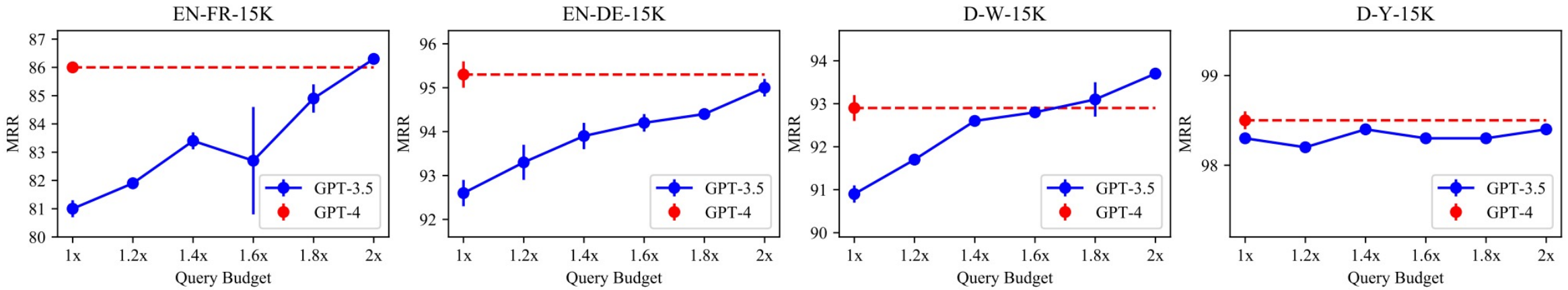
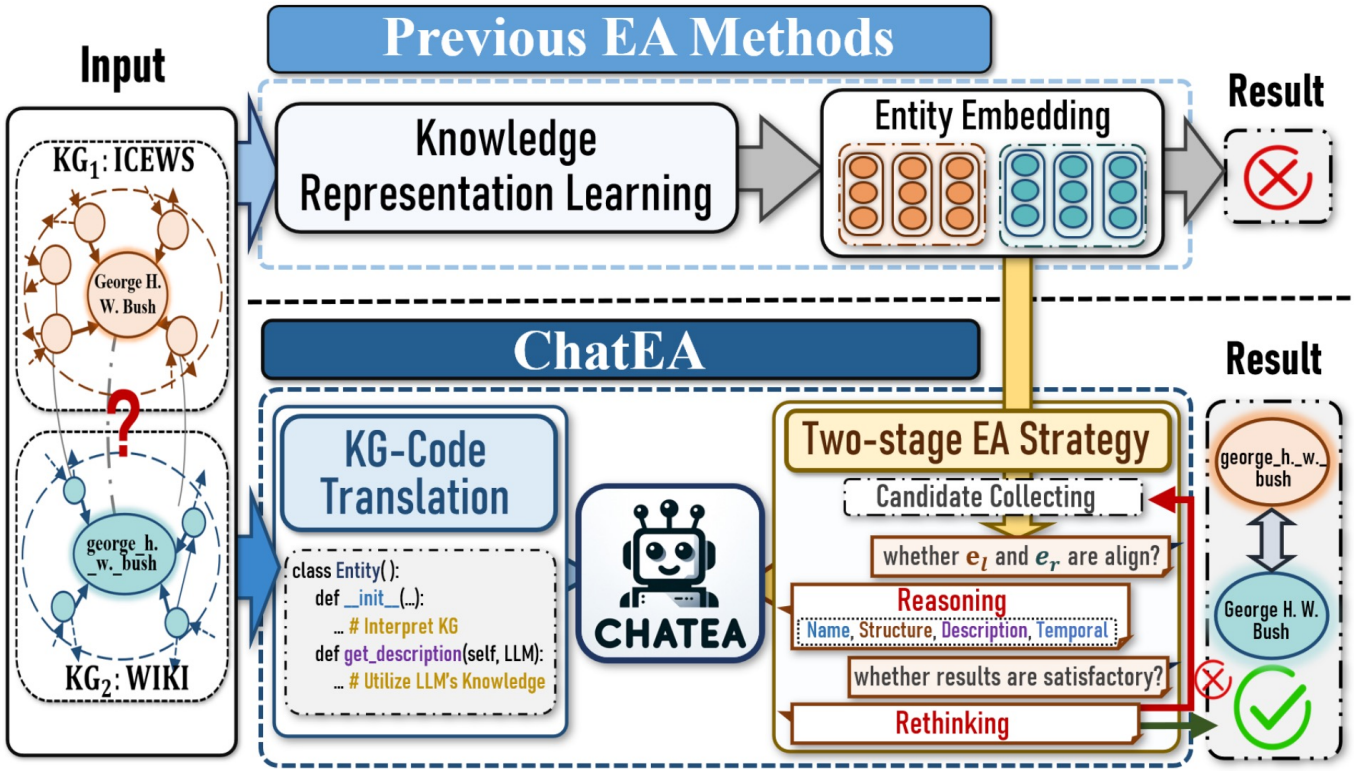


Figure 2: Performance-cost comparison between GPT-3.5 and GPT-4 as the annotator, evaluated by MRR. We increase the budget for GPT-3.5 to evaluate its performance. $[n\times]$ denotes using $n\times$ of the default query budget. Each experiment is repeated three times to show mean and standard deviation.

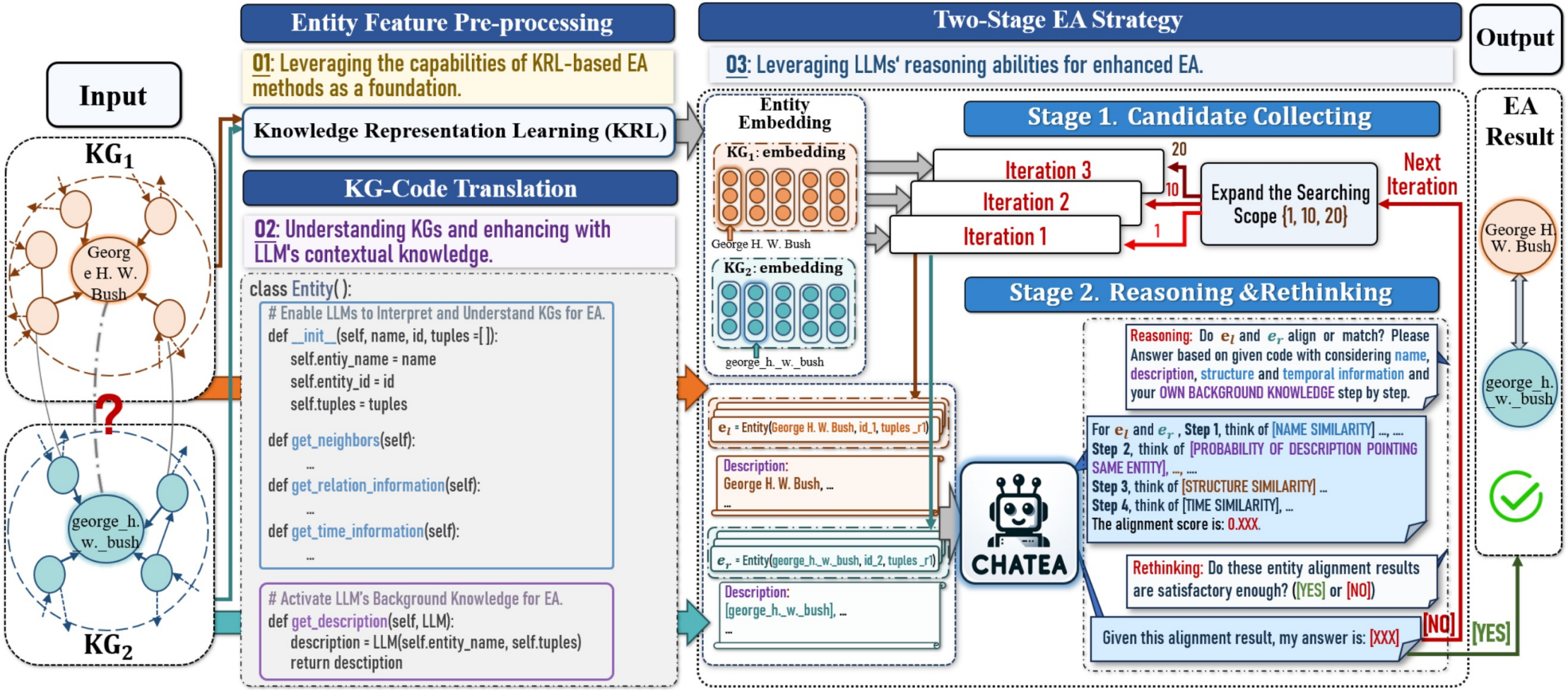
工作5：实体对齐的多轮推理

端到端的实体对齐方法没有反思和纠错机会



Xuhui Jiang, Yinghan Shen, Zhichao Shi, Chengjin Xu, Wei Li, Zixuan Li, Jian Guo, Huawei Shen, Yuanzhuo Wang: Unlocking the Power of Large Language Models for Entity Alignment. ACL 2024: 7566-7583

工作5：实体对齐的多轮推理



工作5：实体对齐的多轮推理

- **多视图实体表示学习**
- **KG-Code翻译**
  - 将知识图谱转换为代码格式，使大语言模型能够理解图结构数据
- **两阶段实体对齐**
  - 候选实体收集：通过实体表征的近邻搜索构造潜在的候选实体对
  - 推理和反思：利用大模型进行多步推理，评估实体对齐的可能性，并决定是否扩大搜索范围进行进一步迭代

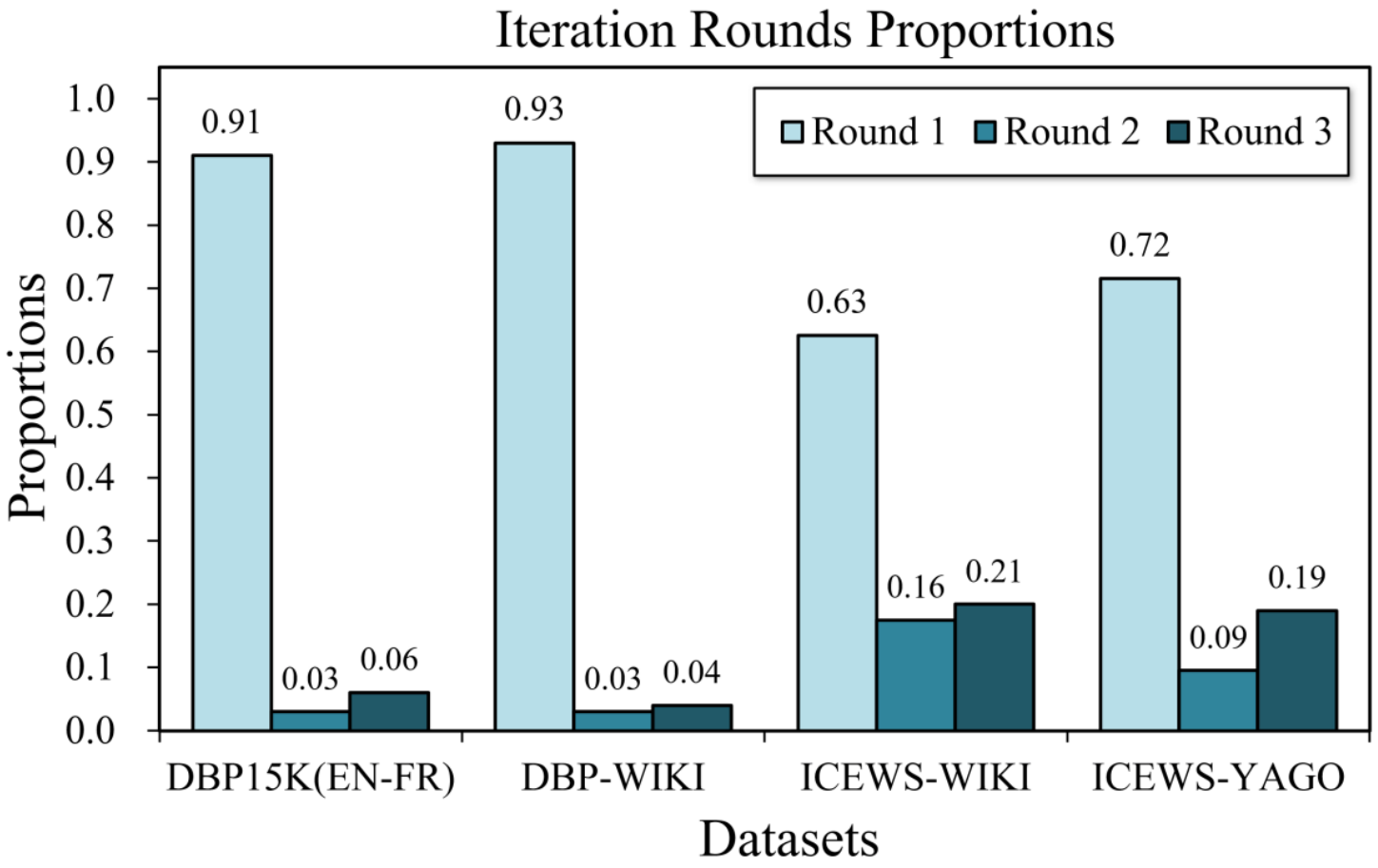
工作5：实体对齐的多轮推理

ChatEA在普通实体对齐和时序实体对齐上都取得了优异的性能

Models	DBP15K(EN-FR)			DBP-WIKI			ICEWS-WIKI			ICEWS-YAGO		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
MTransE	0.247	0.577	0.360	0.281	0.520	0.363	0.021	0.158	0.068	0.012	0.084	0.040
AlignE	0.481	0.824	0.599	0.566	0.827	0.655	0.057	0.261	0.122	0.019	0.118	0.055
BootEA	0.653	0.874	0.731	0.748	0.898	0.801	0.072	0.275	0.139	0.020	0.120	0.056
GCN-Align	0.411	0.772	0.530	0.494	0.756	0.590	0.046	0.184	0.093	0.017	0.085	0.038
RDGCN	0.873	0.950	0.901	0.974	0.994	0.980	0.064	0.202	0.096	0.029	0.097	0.042
Dual-AMN	0.954	0.994	0.970	0.983	0.996	0.991	0.083	0.281	0.145	0.031	0.144	0.068
TEA-GNN	-	-	-	-	-	-	0.063	0.253	0.126	0.025	0.135	0.064
TREA	-	-	-	-	-	-	0.081	0.302	0.155	0.033	0.150	0.072
STEA	-	-	-	-	-	-	0.079	0.292	0.152	0.033	0.147	0.073
BERT	0.937	0.985	0.956	0.941	0.980	0.963	0.546	0.687	0.596	0.749	0.845	0.784
FuAlign	0.936	0.988	0.955	0.980	0.991	0.986	0.257	0.570	0.361	0.326	0.604	0.423
BERT-INT	0.990	0.997	0.993	**0.996**	0.997	0.996	0.561	0.700	0.607	0.756	0.859	0.793
Simple-HHEA	0.959	0.995	0.972	0.975	0.991	0.988	0.720	0.872	0.754	0.847	0.915	0.870
ChatEA	**0.990**	**1.000**	**0.995**	0.995	**1.000**	**0.998**	**0.880**	**0.945**	**0.912**	**0.935**	**0.955**	**0.944**

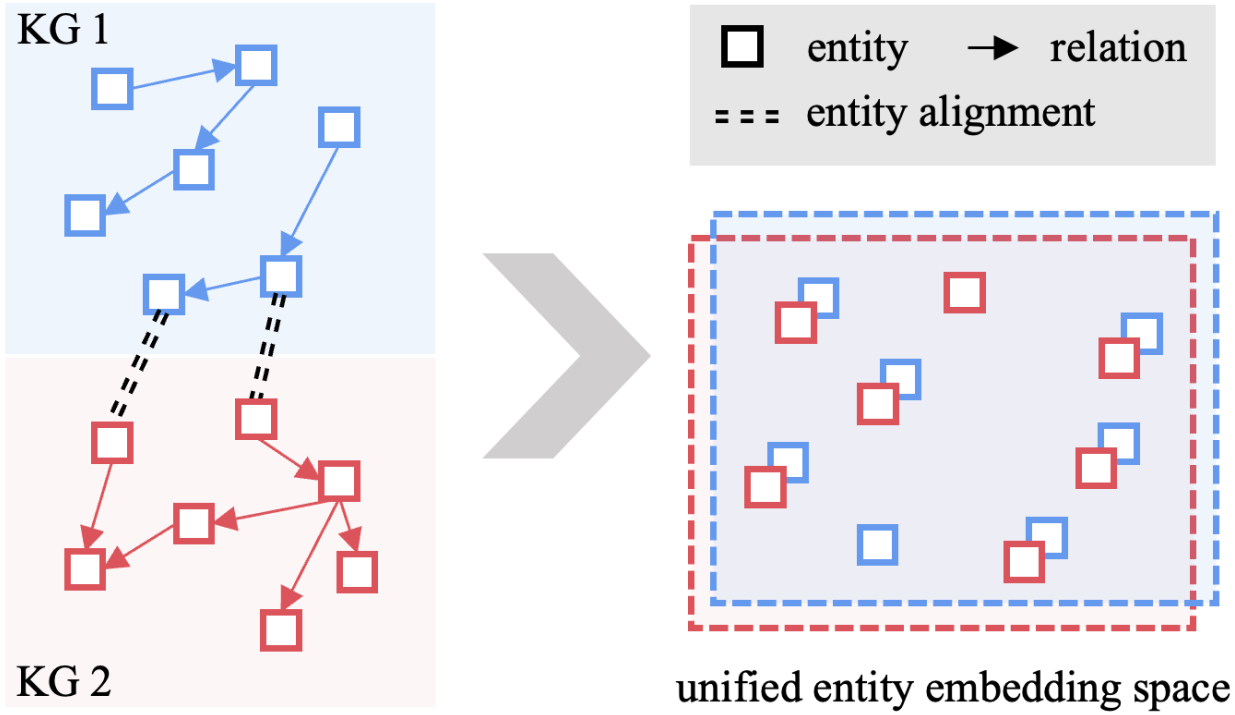
工作5：实体对齐的多轮推理

- 简单数据集，大部分实体的对齐推理只需一轮迭代
- 复杂时序数据集上，大量实体需要两轮或三轮迭代



工作6：解释实体对齐（补充）

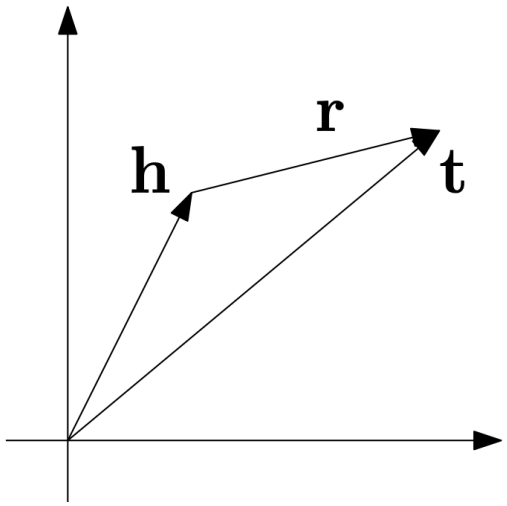
表示学习为什么可以捕捉异构实体的相似性?



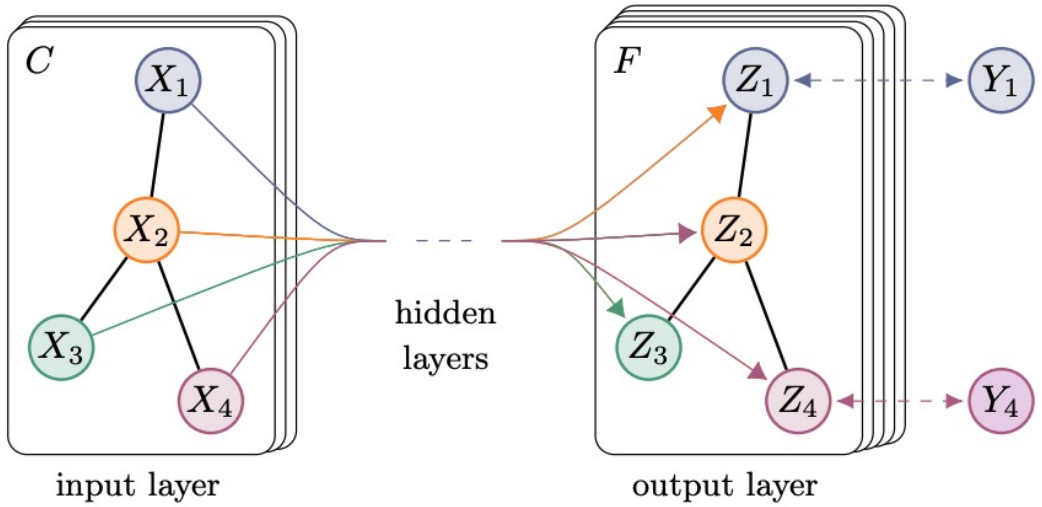
Zequn Sun, Jiacheng Huang, Xiaozhou Xu, Qijin Chen, Weijun Ren, Wei Hu: What Makes Entities Similar? A Similarity Flooding Perspective for Multi-sourced Knowledge Graph Embeddings. ICML 2023: 32875-32885

工作6：解释实体对齐（补充）

- 主要挑战：不同表示学习方法差异很大（TransE vs GCN）



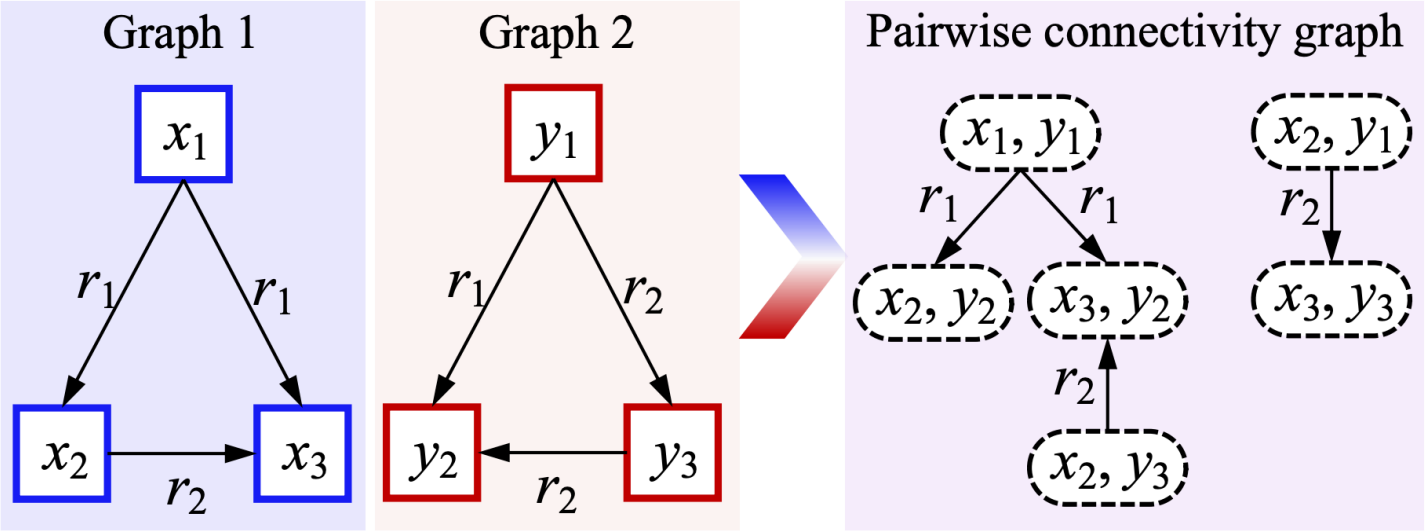
TransE



GCN

工作6：解释实体对齐（补充）

- 思路：**消解关系表征**，以实体表示实体，计算相似度传播过程
  - 一种基于不动点计算的迭代式图匹配技术



$$\sigma^{i+1}(x_1, y_1) = \sigma^i(x_1, y_1) + \varphi(\sigma^i(x_2, y_2), \sigma^i(x_3, y_2))$$

相似度传播方法

$$\Omega = \mathrm{normalize}(\Omega_0 + \Omega + \varphi(\Omega_0 + \Omega))$$

相似度矩阵　　初始相似度矩阵

工作6：解释实体对齐（补充）

- 方法
  - 从表示学习模型的优化目标导出实体的数学表示，然后计算实体相似度

- 实体的数学表示
  - 基于TransE的方法: $\mathbf{e} = \frac{1}{|\mathcal{T}_e|} \sum_{(e,r,o)\in\mathcal{T}_e} \left(\mathbf{o} - \frac{1}{|\mathcal{T}_r|} \sum_{(s',r,o')\in\mathcal{T}_r} (\mathbf{o}' - \mathbf{s}') \right)$
  - 基于GCN的方法: $\mathbf{e} = \frac{1}{|N(e)|} \sum_{e'\in N(e)} \mathbf{e}'$

基于TransE或GCN，一个实体可以表示为其他相关实体表示的组合

工作6：解释实体对齐（补充）

解释1：基于TransE或GCN的实体对齐方法以表示学习计算实体相似度不动点。

证明：令实体表示为 $\mathbf{x}_i = \lambda_{i,1}\mathbf{x}_1 + \lambda_{i,2}\mathbf{x}_2 + \cdots + \lambda_{i,n}\mathbf{x}_n = \sum_{k=1}^{n} \lambda_{i,k}\mathbf{x}_k$，其中$\lambda$表示组合系数，可由TransE或GCN导出。则实体x_i和y_j的相似度为：

$$\omega_{i,j} = \mathbf{x}_i \cdot \mathbf{y}_j = \sum_{k=1}^{n}\sum_{l=1}^{m} \lambda_{i,k}\lambda'_{j,l}\,\mathbf{x}_k \cdot \mathbf{y}_l = \sum_{k=1}^{n}\sum_{l=1}^{m} \lambda_{i,k}\lambda'_{j,l}\omega_{k,l}.$$

可以看到其相似度受到相关实体相似度的影响。令$\Lambda = (\lambda_{i,j})_{i=1,j=1}^{n,n}$表示源图谱的$\lambda$值，$\Lambda' = (\lambda'_{i,j})_{i=1,j=1}^{m,m}$是目标图谱的$\lambda$值，二者的实体相似度矩阵为$\Omega = (\omega_{i,j})_{i=1,j=1}^{n,m}$，则表示学习方法的实体相似度传播为：$\Lambda\Omega(\Lambda')^{\top} = \Omega$，这表明，表示学习模型学到的实体表示相似度是$\Omega$的不动点。

工作6：解释实体对齐（补充）

定理 3 上述基于表示学习的方法找到的实体对齐构成函数 $f:\{1,2,\cdots,n\}\to\{0,1,2,\cdots,m\}$，使得 $\forall i,j,f(i)>0\wedge f(j)>0\to\lambda'_{f(i),f(j)}\approx\lambda_{i,j}$。

证明 3 考虑将 $\mathcal{E}_2$ 与自身对齐，则有

$$\Lambda'\mathbf{I}_m(\Lambda')^\top\approx\mathbf{I}_m, \tag{6-16}$$

其中 $\mathbf{I}_m$ 是单位矩阵。假设上述基于表示学习的模型找到的对齐是 $\hat{S}$，可以用一个*0-1*矩阵 $\hat{\Omega}$ 来表示它，使得 $\hat{\Omega}_{i,j}=1$ 当且仅当 $(x_i,x_j)\in\hat{S}$。与大多数实体对齐设置类似，本章假设在 $\hat{S}$ 中，每个实体最多与另一个知识图谱中的一个实体对齐。请注意，$\hat{\Omega}$ 近似等于公式 *(6-9)* 的一个不动点。因此，有

$$\hat{\Omega}^\top\Lambda\hat{\Omega}(\Lambda')^\top\approx\hat{\Omega}^\top\hat{\Omega}=\hat{\mathbf{I}}_m, \tag{6-17}$$

其中 $\hat{\mathbf{I}}_m$ 是对角矩阵，$\hat{\mathbf{I}}_{j,j}=1$ 当且仅当 y_j 出现在 $\hat{S}$ 中的一组实体对齐。考虑到 $\Lambda'\mathbf{I}_m(\Lambda')^\top\approx\mathbf{I}_m$，有 $\hat{\Omega}^\top\Lambda\hat{\Omega}\hat{\mathbf{I}}_m\Lambda'$。设 f 是定义如下的函数：

$$f(i)=\begin{cases}j, & (x_i,y_j)\in\hat{S}\\ 0, & \forall y_j\in\mathcal{E}_2,(x_i,y_j)\notin S\end{cases}. \tag{6-18}$$

当 $f(i)>0$ 且 $f(j)>0$ 时，本章有 $(\hat{\Omega}^\top\Lambda\hat{\Omega})_{f(i),f(j)}=\lambda_{i,j}$，即 $\lambda'_{f(i),f(j)}\approx\lambda_{i,j}$。

表示学习可以缓解多源图谱结构上的异构性

实体对齐方法为每个知识图谱构造一个实体表示相关性矩阵Λ，并找到一个映射函数f使得需要对齐的图谱的Λ矩阵相同。该函数确定最后的对齐结果。如果将这些矩阵视为知识图谱中节点之间的边权重，则这些基于表示学习的实体对齐方法在数学角度等同于进行图匹配

工作6：解释实体对齐（补充）

表示学习的方法更新实体表征，以求解相似度不动点。那么，能不能不用表征，直接更新相似度呢?

Input: K_1, K_2, seed entity alignment S, the maximum number of iterations T, a threshold ϵ for algorithm termination, an embedding model $\mathcal{M}$

/* Represent an entity as the composition of other entities as defined in Eq. (6-11). */

Let the gradients of $\mathcal{M}$ be zero to get entity representations;

Compute lambda values Λ and Λ' to represent entities; /* Initialize the similarity matrix. Set the similarity of seed entity alignment to 1. */

$\Omega_0 \leftarrow (0)_{i=1,j=1}^{n,m}$;

for $i, j \in S$ **do**

  $\Omega_{0\,i,j} \leftarrow 1$;

/* Compute fixpoint using Eq. (6-13). */

for $t = 1, 2, \cdots, T$ **do**

  $\Omega_t \leftarrow \mathtt{normalize}\left(\Lambda \Omega_{t-1} (\Lambda')^{\top}\right)$;

  if $\triangle(\Omega_t, \Omega_{t-1}) < \epsilon$ **then**

    break;

return Ω_t;

工作6：解释实体对齐（补充）

- 基于分析结果，构造了两个不需要学习表征的方法TransFlood和GCNFlood
- 部分指标上取得了比他们各自基础表征模型更好的性能
- **表示学习模型未达到各自的理论性能上限，还有提高空间**

Table 1. EA results on DBP15K as well as OpenEA D-W and D-Y. The best scores in each group are marked in bold. The results of MTransE are taken from (Sun et al., 2017). The results of GCN-Align are taken from its paper. "-" denotes their unreported metrics.

Models	DBP15K ZH-EN			DBP15K JA-EN			DBP15K FR-EN			OpenEA D-W 15K			OpenEA D-Y 15K		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
MTransE	0.308	0.614	-	0.279	0.575	-	0.244	0.556	-	0.259	-	0.354	0.463	-	0.559
TransFlood (ours)	**0.315**	**0.707**	0.451	**0.372**	**0.757**	0.505	**0.347**	**0.752**	0.484	**0.294**	0.699	**0.427**	**0.503**	0.880	**0.641**
GCN-Align	**0.413**	0.744	-	**0.399**	0.745	-	**0.373**	0.745	-	**0.364**	-	0.461	0.465	-	0.536
GCNFlood (ours)	0.349	**0.761**	0.490	0.376	**0.770**	0.512	0.349	**0.761**	0.490	0.358	0.739	**0.486**	**0.478**	0.754	**0.583**

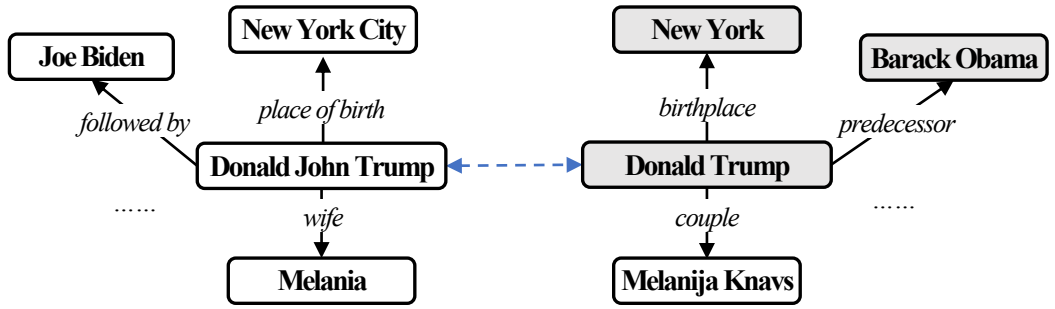
工作6：解释实体对齐（补充）

是否可以利用大模型解释实体对齐的结果呢?

Xiaobin Tian, Zequn Sun, Wei Hu: Generating Explanations to Understand and Repair Embedding-Based Entity Alignment. ICDE 2024: 2205-2217

工作6：解释实体对齐（补充）

Entity Alignment



Prompt Construction

Query

Please find important matching triples below as explanations for entity alignment
(Donald John Trump, *sameAs*, Donald Trump)

Triples

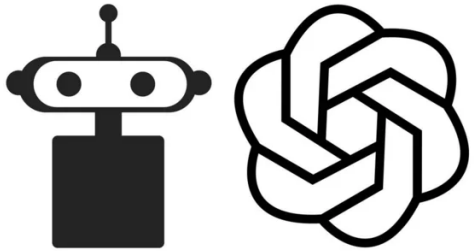
The triples of Donald John Trump are (Donald John Trump, *wife*, Melania),…..
The triples of Donald Trump are (Donald Trump, *couple*, Melanija Knavs),…..

Examples

……

Output

The explanations are
(Donald John Trump, *wife*, Melania) matches (Donald Trump, *couple*, Melanija Knavs)
(Donald John Trump, *place of birth*, New York City) matches (Donald Trump, *birthplace*, New York)
……



ChatGPT

Xiaobin Tian, Zequn Sun, Wei Hu: Generating Explanations to Understand and Repair Embedding-Based Entity Alignment. ICDE 2024: 2205-2217

工作6：解释实体对齐（补充）

大模型可以有效解释实体对齐结果，且进一步增强传统解释方法

EA models	Ver. methods	ZH-EN			DBP-WD		
		Prec.	Recall	F1	Prec.	Recall	F1
MTransE	ChatGPT	0.823	0.862	0.842	0.841	0.970	0.901
	ExEA	0.918	0.938	0.928	0.846	0.966	0.902
	ChatGPT + ExEA	**0.982**	**0.986**	**0.984**	**0.960**	**0.996**	**0.977**
Dual-AMN	ChatGPT	0.816	0.826	0.821	0.815	0.944	0.875
	ExEA	0.879	0.940	0.905	0.911	0.978	0.943
	ChatGPT + ExEA	**0.970**	**0.984**	**0.977**	**0.967**	**0.996**	**0.981**

Xiaobin Tian, Zequn Sun, Wei Hu: Generating Explanations to Understand and Repair Embedding-Based Entity Alignment. ICDE 2024: 2205-2217

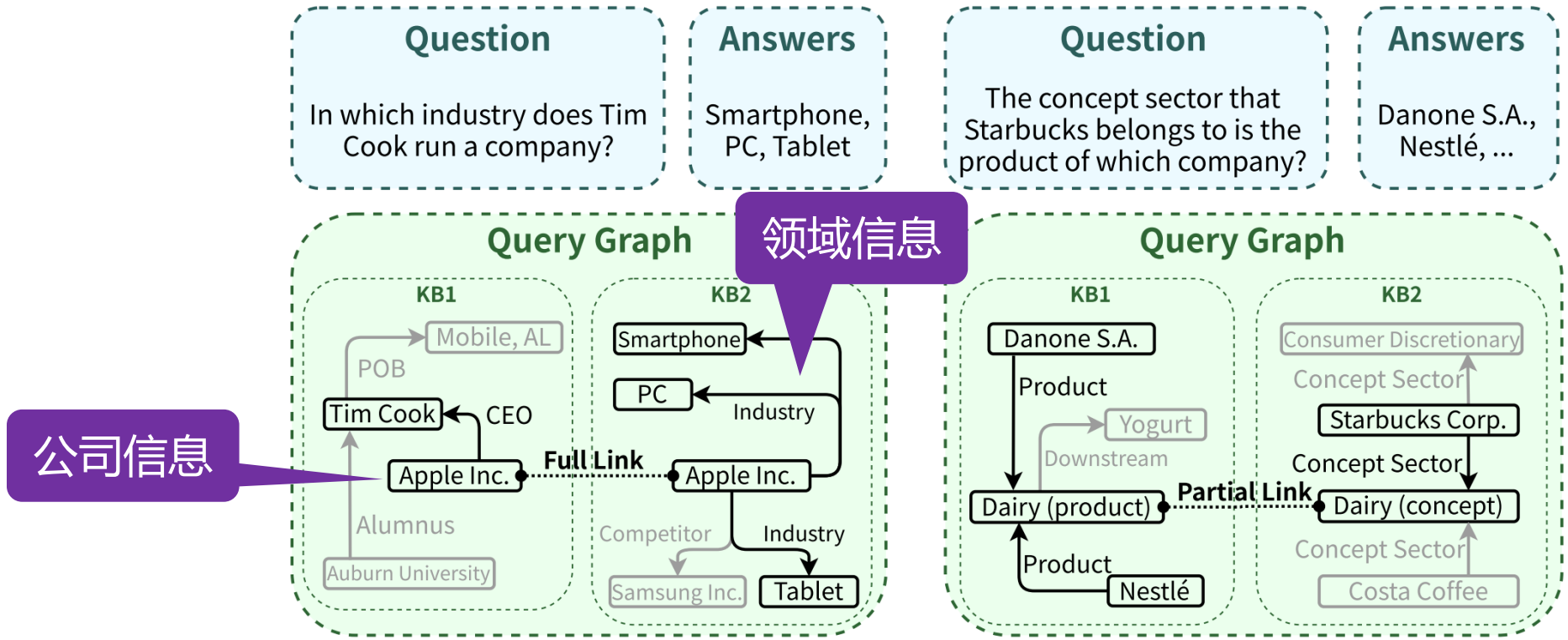
提纲

- 研究背景
- 方法
- **应用**
- 总结与展望

核心思想

应用1：多源知识图谱问答

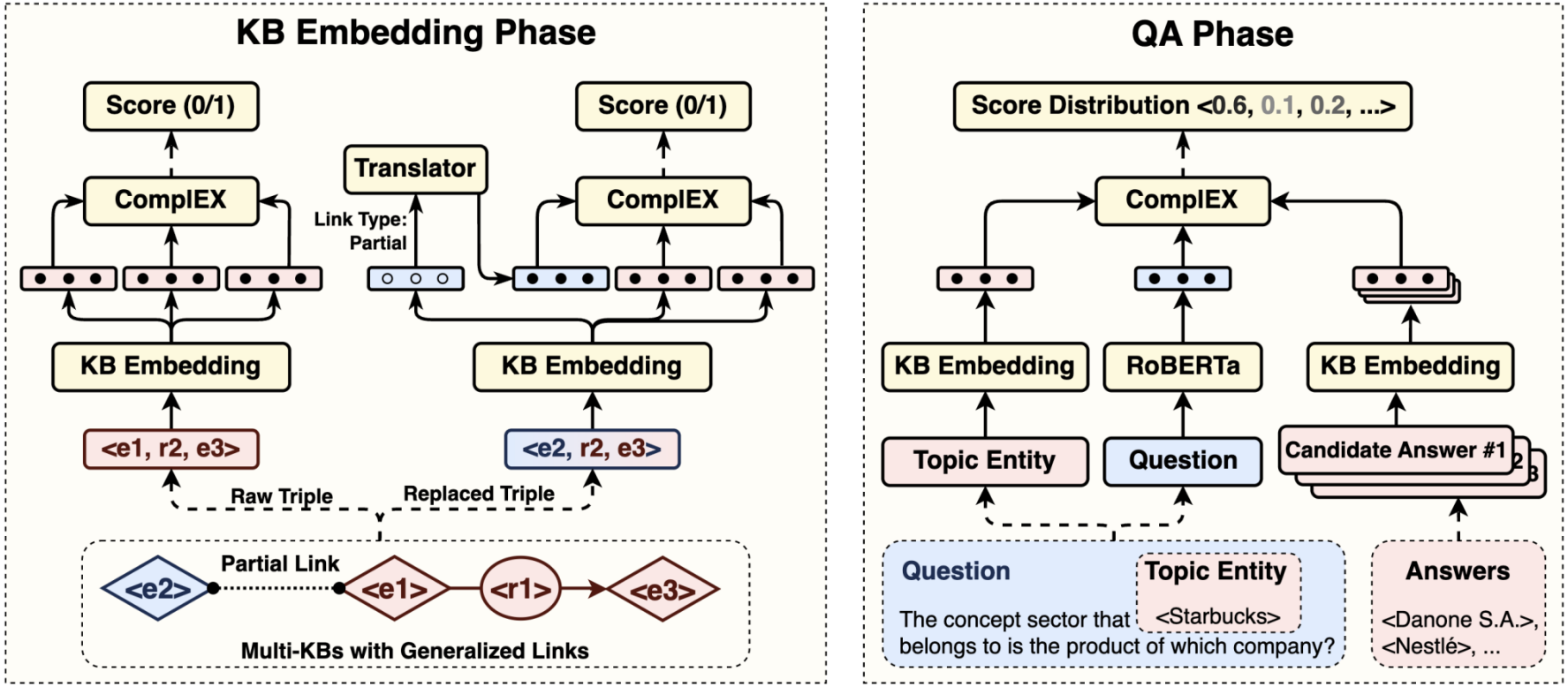
单源知识图谱的不完备严重影响复杂问题的理解和回答



Minhao Zhang, Yongliang Ma, Yanzeng Li, Ruoyu Zhang, Lei Zou, Ming Zhou: Two is Better Than One: Answering Complex Questions by Multiple Knowledge Sources with Generalized Links. CoRR abs/2309.05201 (2023)

应用1：多源知识图谱问答

- 思路：对多源图谱和自然语言问句进行**联合表示学习与对齐**

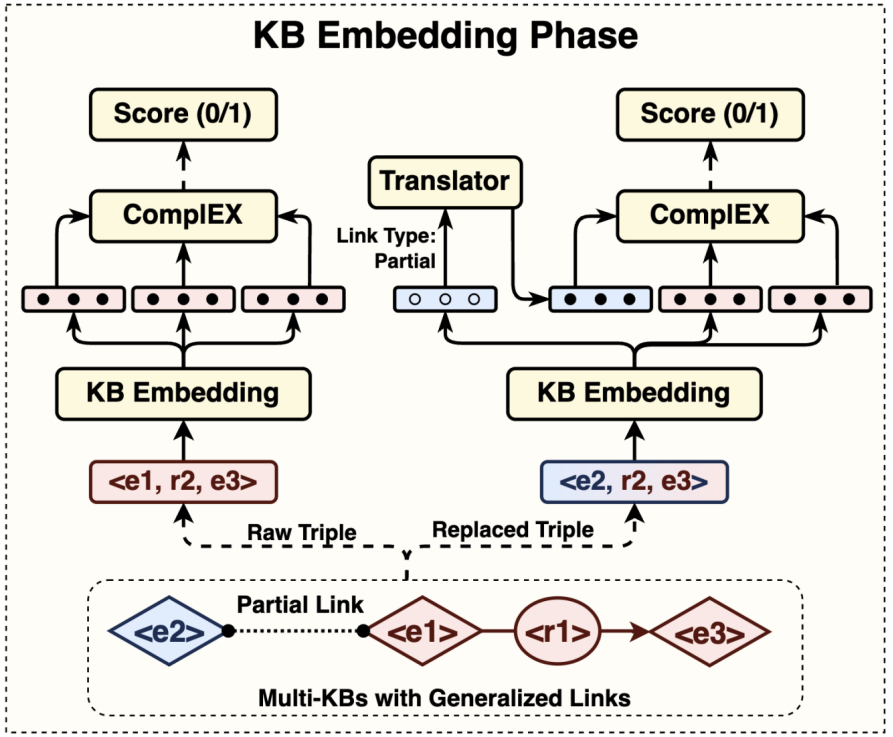


应用1：多源知识图谱问答

- 多源知识图谱表示学习
  - 交换对齐实体扩充三元组
  - 用ComplEx (Trouillon et al., 2016) 进行学习

$$\lambda_{s,r,o} = \text{Sigmoid}(\text{ComplEx}(\mathbf{h_s}, \mathbf{h_r}, \mathbf{h_o})) \in \mathbb{R}$$

$$\ell_{raw} = \sum_{<s,r,o>\in \text{KB}} -log(\lambda_{s,r,o}) - \sum_{i=1}^{k} log(1-\lambda_{s,r,\tilde{o}_i})$$



应用1：多源知识图谱问答

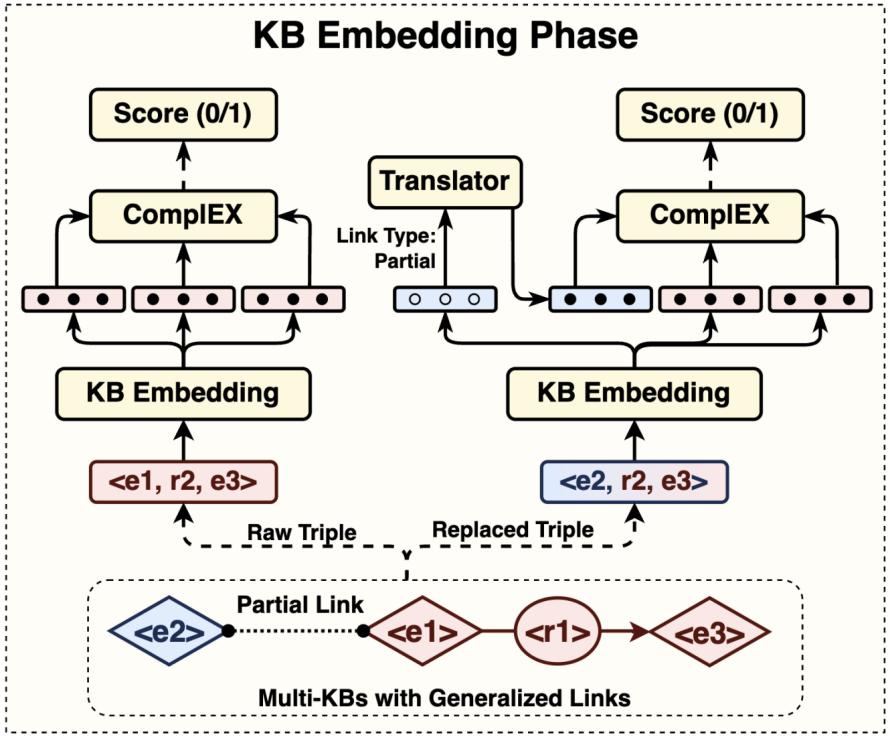
- 多源知识图谱表示学习
  - 对于局部对齐的实体（partial links）
    - 不是严格语义等价，引入转换器进行软对齐

$$\tilde{\mathbf{h}}_{\hat{s}} = Trans(\mathbf{h}_{\hat{s}} \oplus E_t(t)) \in \mathbb{C}^h$$

    - 然后同样复用ComplEx进行学习

$$\ell_{link} = \sum_{<\hat{s},r,o>\in R} -log(\lambda_{\hat{s},r,o}) - \sum_{i=1} log(1-\lambda_{\hat{s},r,\tilde{o}_i})$$

  - 最后优化目标：最小化$\ell_{raw} + \ell_{link}$



应用1：多源知识图谱问答

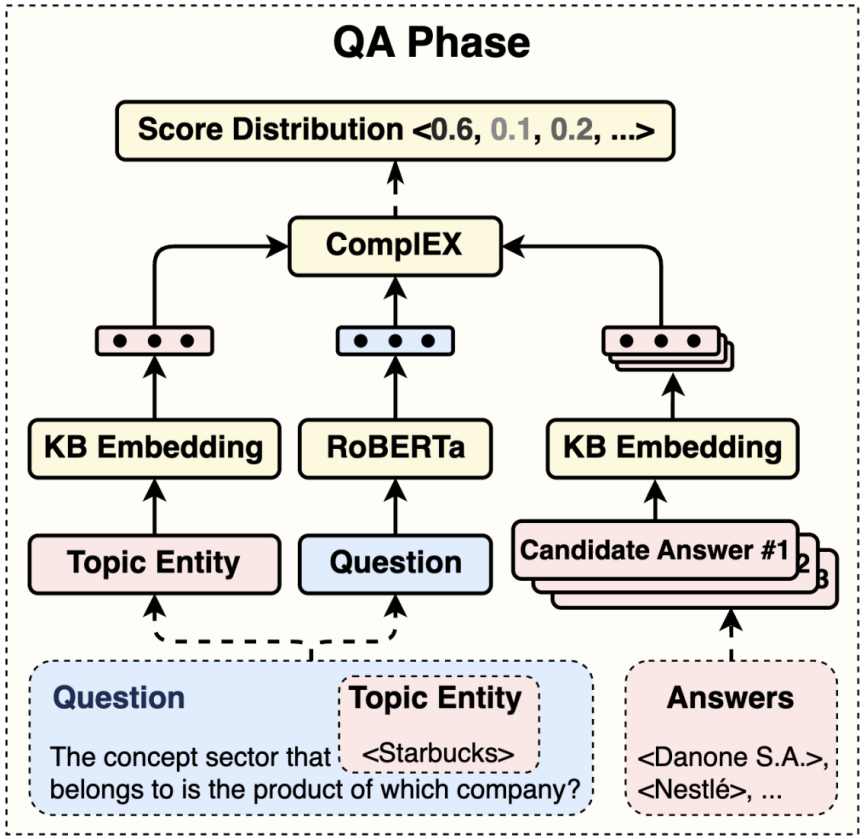
- 多源知识图谱问答
  - 使用预训练语言模型编码问题
  - 通过全连接层进行维度映射

$$\mathbf{h_q} = R2C(E_{rb}(\mathbf{q})) \in \mathbb{C}^h$$

  - 通过实体链接获得问题涉及的主题实体
  - 问题当作“关系”，用ComplEx进行学习

$$\lambda_{e_i,a_j} = \text{Sigmoid}(\text{ComplEx}(\mathbf{h_{e_i}}, \mathbf{h_q}, \mathbf{h_{a_j}})) \in \mathbb{R}$$

$$\ell_q = -\sum_{i,j} log(\lambda_{e_i,a_j}) - \sum_{i,k} log(1 - \lambda_{e_i,\tilde{a}_k})$$



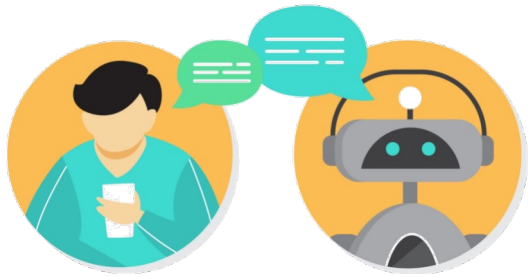
应用1：多源知识图谱问答

背景知识图谱的完备性和知识问答的精度呈正相关

Method	Dev Set		Test Set	
	MRR	Hits@1	MRR	Hits@1
No-Link	.192 ± .000	.113 ± .011	.166 ± .010	.088 ± .011
Merge-KB	.351 ± .004	.265 ± .011	.350 ± .011	.242 ± .011
Full-Link	.371 ± .011	.261 ± .009	.374 ± .021	.254 ± .022
Multi-KB	**.494 ± .003**	**.373 ± .007**	**.488 ± .017**	**.363 ± .021**

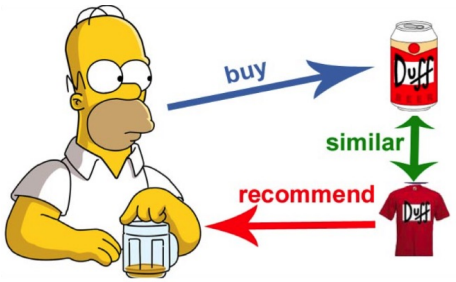
应用2：多源知识图谱预训练

不同任务/不同图谱之间难以进行模型复用和知识迁移



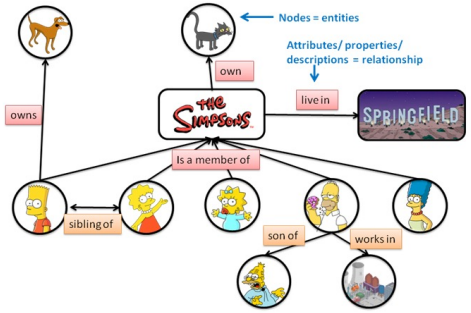
QA systems

Freebase



Recommendation systems

WIKIDATA



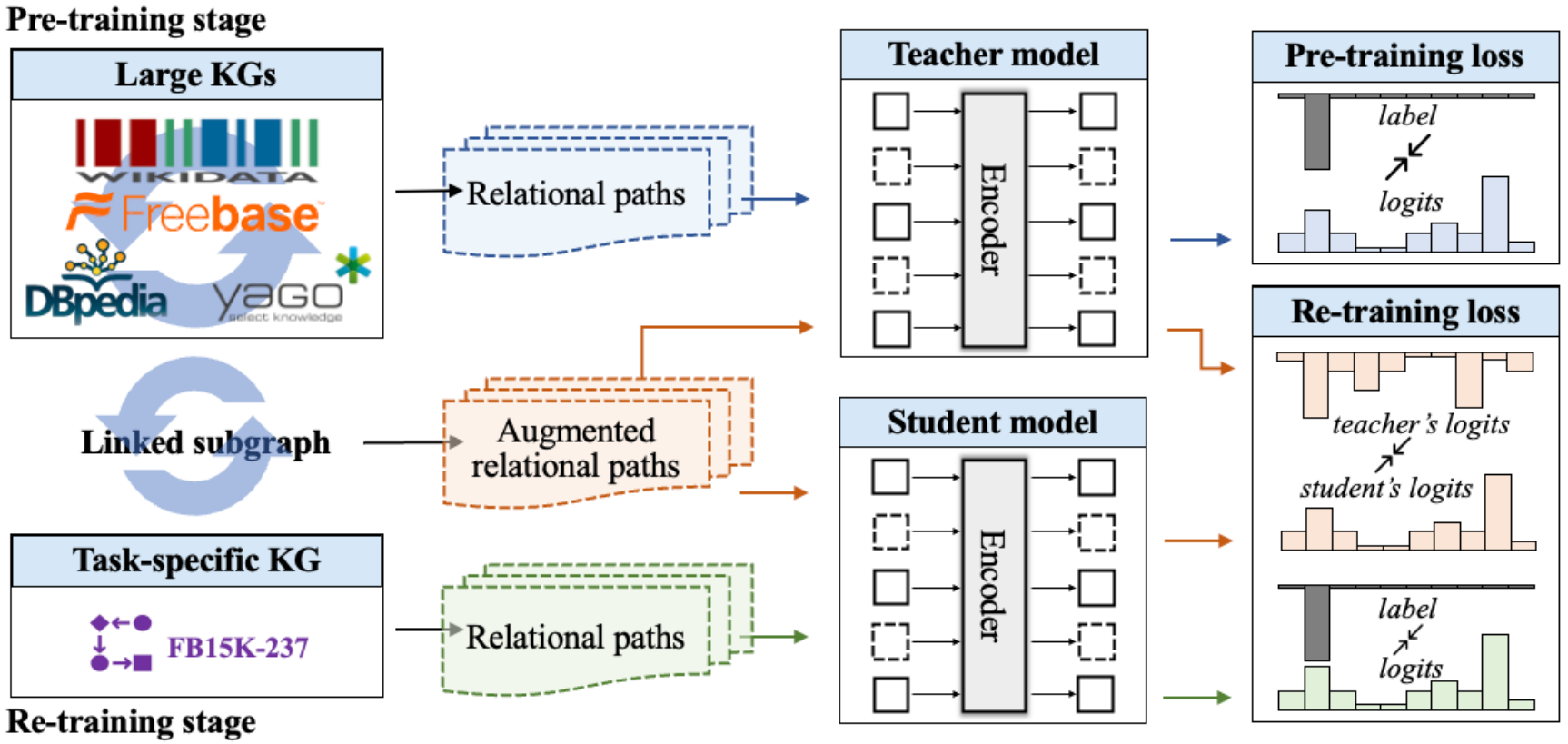
Semantic search

DBpedia

Zequn Sun, Jiacheng Huang, Jinghao Lin, Xiaozhou Xu, Qijin Chen, Wei Hu: Joint Pre-training and Local Re-training: Transferable Representation Learning on Multi-source Knowledge Graphs. KDD 2023: 2132-2144

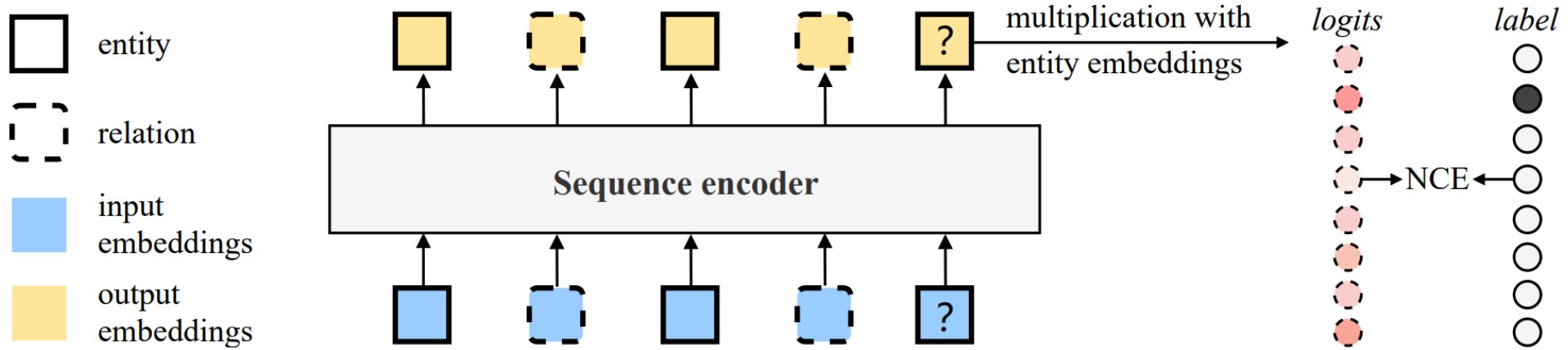
应用2：多源知识图谱预训练

- 思路：在多源图谱进行**联合预训练**，在下游图谱进行**局部重训练**



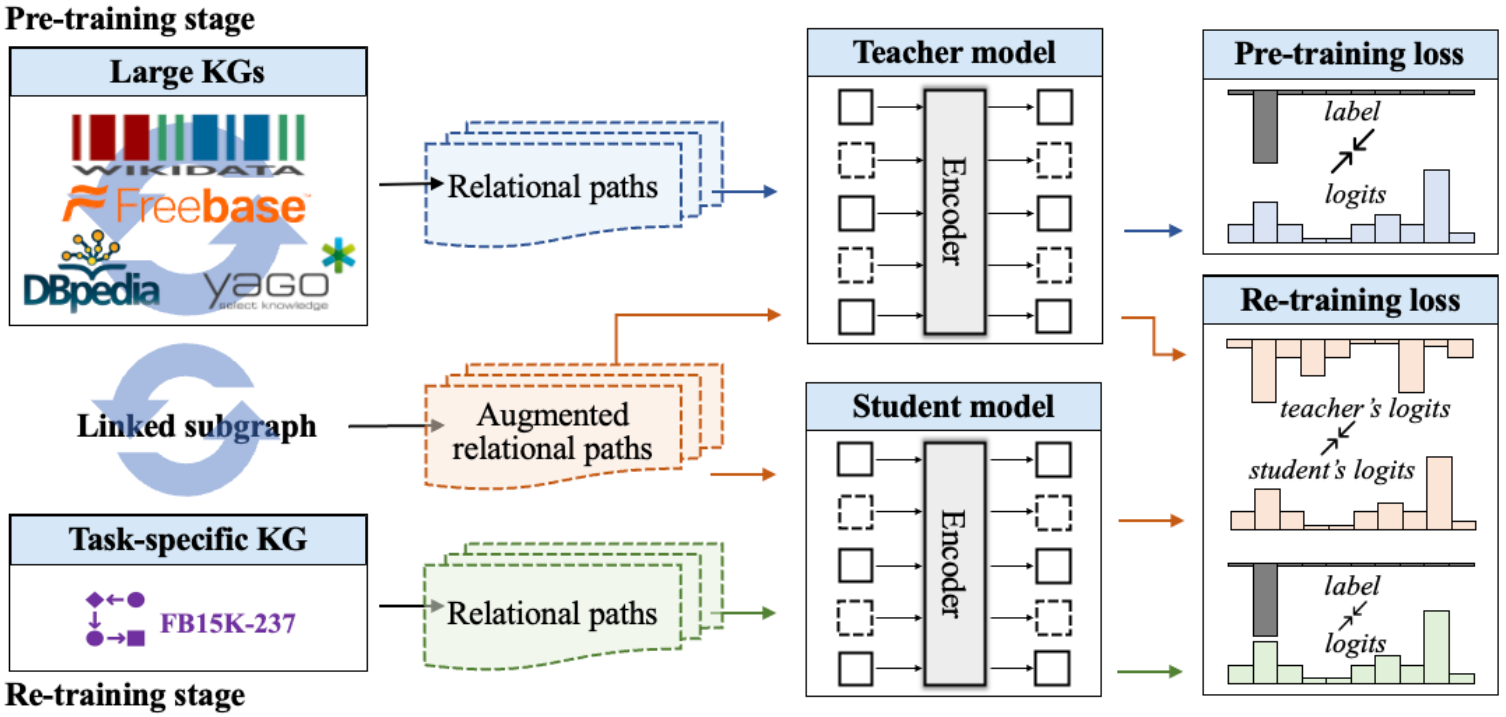
应用2：多源知识图谱预训练

- 联合预训练
  - 基于实体对齐混合多源图谱，构建**实体关系路径**，以实体预测和关系预测为自监督任务，训练一个大的**教师模型**



应用2：多源知识图谱预训练

- 局部重训练：
  - 复用教师图谱的**部分子图参与训练**，且构建该子图的统一计算空间
  - 通过**多层级知识蒸馏**指导小的学生模型的训练



应用2：多源知识图谱预训练

- 所提出的方法在两个数据集上取得了正向迁移效果，提高了链接预测性能
- 在WN18RR上未观测到明显效果，因为没有直接可迁移的知识
- 效率上远高于基于联合训练的方法

Table 4: Results on FB15K-237, WN18RR and YAGO3-10 with Wikidata5M and DBpedia5M (WikiDBP10M) as background KGs.

Setting	Model	FB15K-237			WN18RR			YAGO3-10		
		MRR	*H*@10	*H*@1	*MRR*	*H*@10	*H*@1	*MRR*	*H*@10	*H*@1
JointLP	TransE	0.362 (25.7%)	0.574 (20.8%)	0.248 (–)	0.235 (4.9%)	0.526 (3.1%)	0.060 (–)	0.501 (35.4%)	0.703 (14.9%)	0.314 (29.8%)
	ConvE	0.375 (15.4%)	0.572 (14.2%)	0.270 (13.9%)	0.451 (4.9%)	0.513 (-1.3%)	0.419 (4.8%)	0.479 (8.9%)	0.681 (9.8%)	0.403 (14.8%)
	RotatE	0.384 (13.6%)	0.585 (9.8%)	0.276 (14.5%)	**0.484** (1.7%)	**0.558** (2.3%)	0.443 (3.5%)	0.516 (4.2%)	0.725 (8.2%)	0.439 (9.2%)
	TuckER	0.422 (17.9%)	0.605 (13.1%)	0.327 (22.9%)	0.471 (0.2%)	0.520 (-1.1%)	**0.445** (0.5%)	0.553 (9.5%)	0.701 (6.1%)	0.473 (12.1%)
	MuKGE (RNN)	0.417 (46.3%)	0.620 (**43.2%**)	0.310 (47.6%)	0.400 (-1.5%)	0.476 (**6.5%**)	0.362 (-5.2%)	0.685 (53.2%)	0.851 (**37.9%**)	0.586 (65.5%)
	MuKGE (RSN)	0.432 (46.4%)	0.625 (33.5%)	0.329 (55.9%)	0.412 (-3.7%)	0.487 (-0.6%)	0.372 (-6.5%)	0.693 (37.5%)	0.870 (33.0%)	0.584 (38.4%)
	MuKGE (TF)	0.437 (42.3%)	0.650 (35.1%)	0.335 (52.2%)	0.446 (0.5%)	0.504 (2.0%)	0.414 (-0.7%)	0.717 (37.1%)	0.878 (29.5%)	0.620 (41.2%)
PR4LP	MuKGE (RNN)	0.397 (39.3%)	0.600 (38.6%)	0.297 (41.4%)	0.421 (**5.2%**)	0.478 (0.4%)	0.388 (**7.2%**)	0.687 (**53.7%**)	0.841 (36.3%)	0.591 (**66.9%**)
	MuKGE (RSN)	0.446 (**51.2%**)	0.653 (39.5%)	0.340 (**61.1%**)	0.407 (-4.9%)	0.488 (-0.4%)	0.365 (-6.5%)	0.693 (37.5%)	0.872 (33.3%)	0.594 (40.8%)
	MuKGE (TF)	**0.454** (47.9%)	**0.663** (37.8%)	**0.348** (58.2%)	0.446 (0.5%)	0.509 (3.0%)	0.415 (-0.5%)	**0.722** (38.0%)	**0.880** (29.8%)	**0.628** (43.1%)

应用2：多源知识图谱预训练

Wikidata到YAGO3-10数据集可以挖掘到105 1-hop和58 2-hop的跨图谱规则，而到WN18RR，只有1个1-hop和86个2-hop的跨图谱规则

Table 8: Examples of the 1-hop and 2-hop rules mined from the joint graph of YAGO3 (YG) and Wikidata (WD).

Rule head		Rule body	Conf.
YG : founder (X, Y)	⟸	*WD : foundedBy* (X, Y)	0.84
YG : parentOrganization (X, Y)	⟸	*WD : subsidiary* (Y, X)	0.85
YG : musicBy (X, Y)	⟸	*WD : composer* (X, Y)	0.97
YG : byArtist (X, Y)	⟸	*WD : performer* (X, Y)	0.99
YG : spouse (X, Y)	⟸	*WD : father* (Z, Y) ∧ *WD : mother* (Z, X)	0.82
YG : author (X, Y)	⟸	*WD : followedBy* (Z, X) ∧ *WD : notableWork* (Y, Z)	0.84
YG : byArtist (X, Y)	⟸	*WD : follows* (X, Z) ∧ *WD : performer* (Z, Y)	0.97
YG : partOfSeries (X, Y)	⟸	*WD : followedBy* (X, Z) ∧ *WD : partOfTheSeries* (Z, Y)	0.98

Table 9: Rule examples of WN18RR (WN) and WD.

Rule head		Rule body	Conf.
WN : memberOfDomainRegion (X, Y)	⟸	*WD : countryOf Origin* (X, Y)	1.00
WN : hasPart (X, Y)	⟸	*WD : headquartersLocation* (Y, Z) ∧ *WD : ownedBy* (Z, X)	0.80
WN : hasPart (X, Y)	⟸	*WD : executiveBody* (Y, Z) ∧ *WD : country* (Z, X)	0.63
WN : memberMeronym (X, Y)	⟸	*WD : constellation* (Z, X) ∧ *WD : partOf* (Z, Y)	1.00
WN : synsetDomainTopicOf (X, Y)	⟸	*WD : narrativeLocation* (Z, X) ∧ *WD : presentInWork* (Z, Y)	1.00

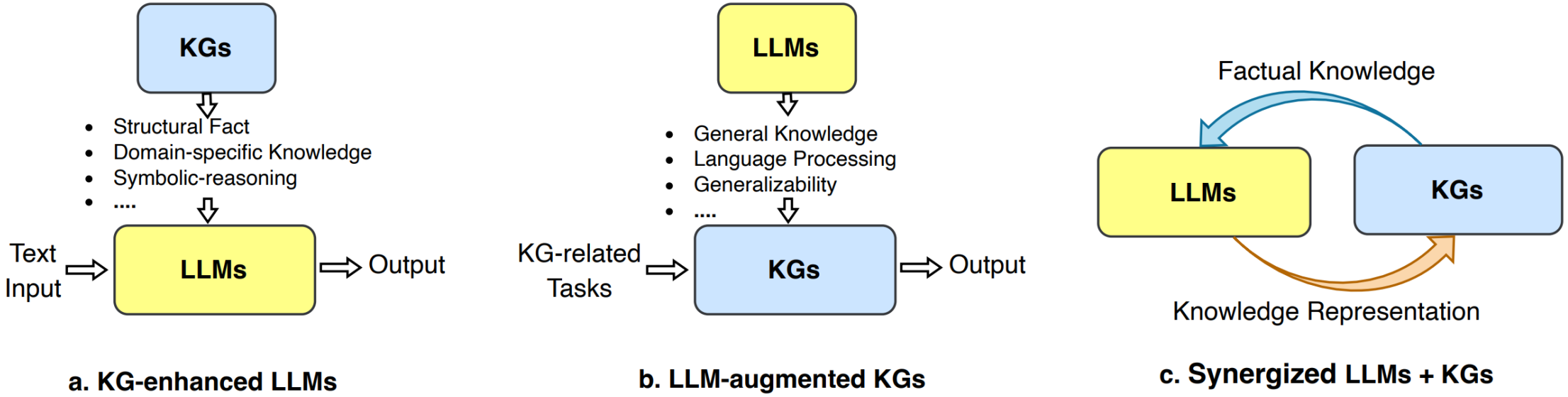
把MuKGE应用到基于表示学习的KBQA任务的简单方法上，观察到明显的性能提升

Table 7: QA accuracy on WebQuestionsSP. Our model is pretrained with Wikidata5M as the background KG.

	Half-KG	Full-KG	Full-KG w/ rel. pruning
EmbedKGQA [40]	0.485	0.587	0.666
NSM [22]	–	–	0.743
μKG [32]	0.547	0.646	0.723
EmbedKGQA + MuKGE	0.518	0.632	0.746

应用3：多源知识注入的指令微调

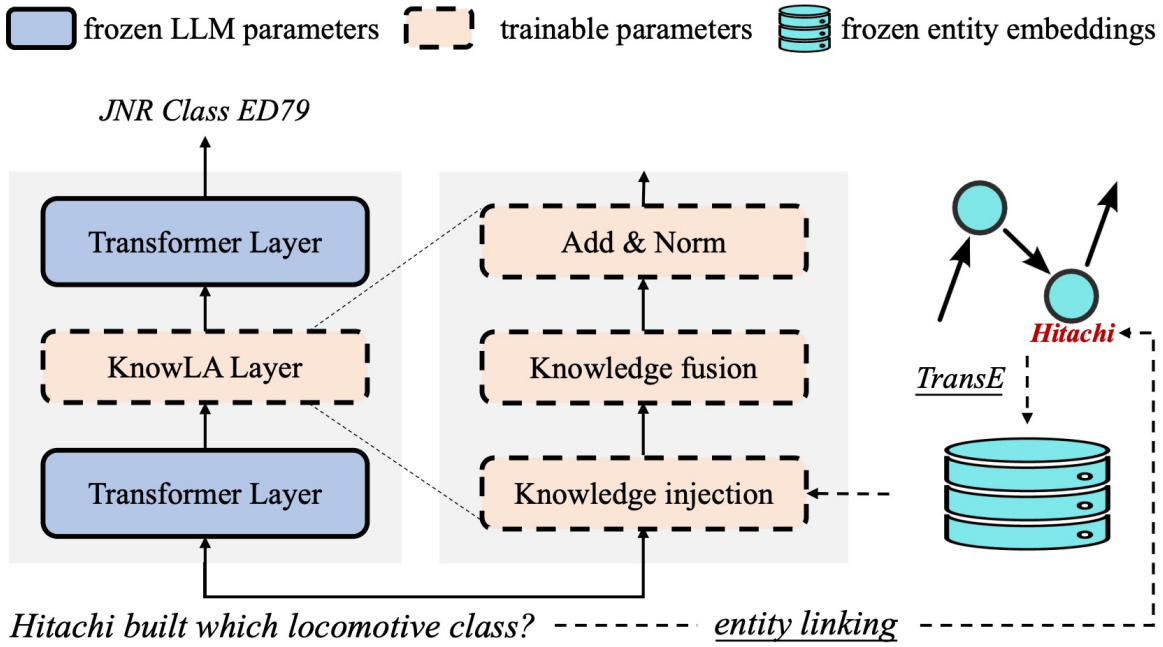
知识图谱表征如何与生成式大语言模型集成是一个挑战



Xindi Luo, Zequn Sun, Jing Zhao, Zhe Zhao, Wei Hu: KnowLA: Enhancing Parameter-efficient Finetuning with Knowledgeable Adaptation. NAACL-HLT 2024: 7153-7166

应用3：多源知识注入的指令微调

- 思路：在生成式大语言模型中间插入**知识适配层**，进行**增量微调**



Xindi Luo, Zequn Sun, Jing Zhao, Zhe Zhao, Wei Hu: KnowLA: Enhancing Parameter-efficient Finetuning with Knowledgeable Adaptation. NAACL-HLT 2024: 7153-7166

应用3：多源知识注入的指令微调

- 实体链接
  - 进行实体链接，获取输入文本中相关实体及其预训练的向量表征
- 知识映射与注入
  - 将实体表征映射到大模型的表示空间，进行知识注入
- 知识融合
  - 将注入的实体表征与大模型的文本表征进行融合，增强知识表达能力

应用3：多源知识注入的指令微调

在同水平参数量增加的情况下，集成知识图谱表征微调大模型可以提高问答精度。

Methods	#Parameters	CommonsenseQA		SIQA		BIG-Bench Hard	
		Accuracy	Score	Accuracy	Score	Accuracy	Score
Llama 2 (7B)	7B	45.37	36.40	46.42	40.58	26.95	24.87
Alpaca2 ($r = 16$)	+0.24%	56.92	46.55	52.61	46.18	28.93	**25.42**
Alpaca2 ($r = 32$)	+0.50%	57.90	46.81	53.17	46.21	28.79	25.36
Contriever (WordNet)	+0.50%	57.15	46.09	52.58	46.13	-	-
Contriever (ConceptNet)		57.06	45.30	52.51	45.51	-	-
KAPING (WordNet)	+0.50%	57.21	45.91	52.51	45.89	-	-
KAPING (ConceptNet)		57.58	45.64	52.66	46.15	-	-
KnowLA (Random)	+0.55%	57.49	47.82	52.61	46.56	29.26	25.34
KnowLA (WordNet)		58.07	**48.35**	**53.22**	46.76	30.00	25.39
KnowLA (ConceptNet)		**58.39**	48.19	**53.22**	**46.81**	**30.19**	25.29
KnowLA (Wikidata)		57.90	47.39	53.21	46.64	29.39	**25.42**

应用3：多源知识注入的指令微调

KnowLA能够激活大模型存储的潜在知识，在集成实体表征后，可以使大模型在FFN层中捕获更多知识，尤其是在较高层次上。

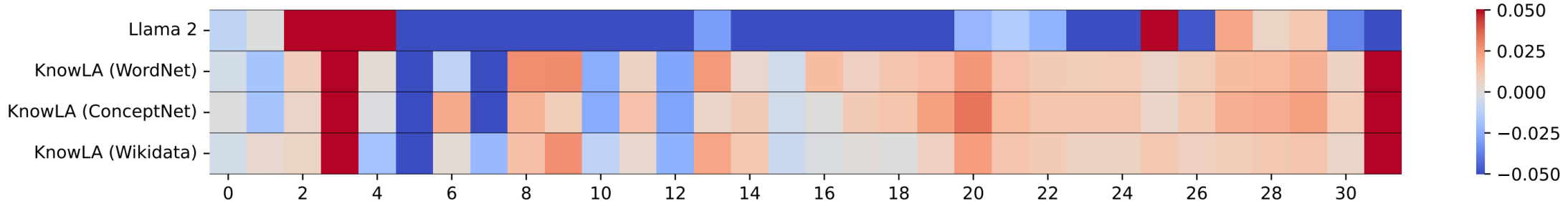


Figure 4: The heatmap indicates the capabilities of KnowLA and Llama 2 in capturing knowledge compared to Alpaca2, which is measured by averaging the changes in cosine similarities of the last token representations from 100 queries across all FFN layers. The x-axis denotes the 32 layers of Llama 2.

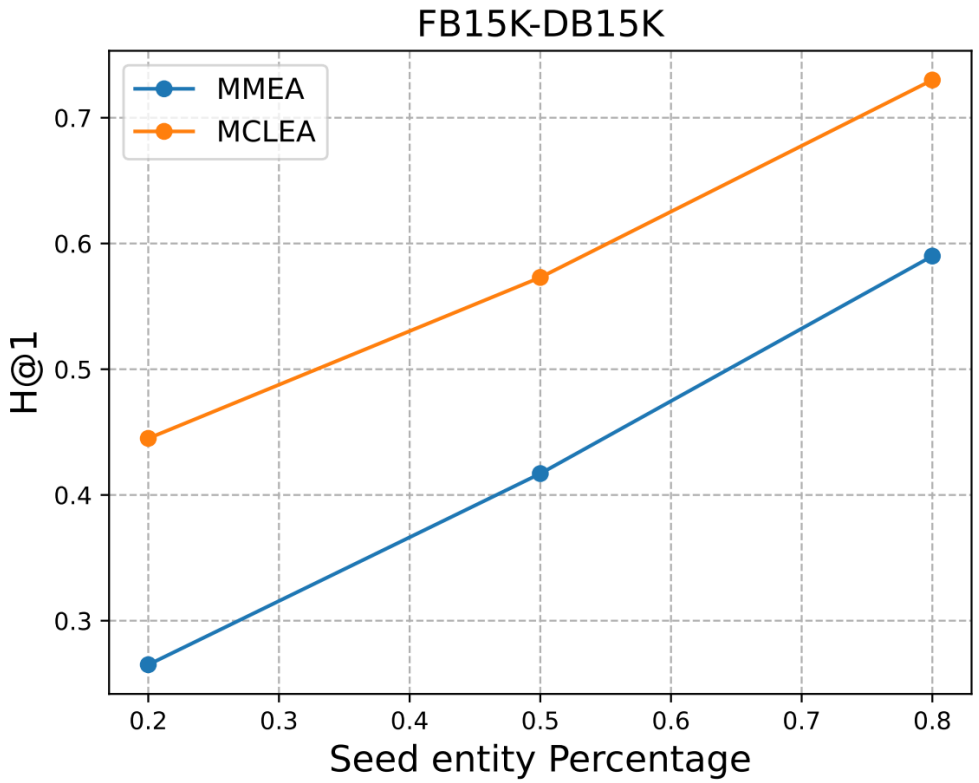
应用3：多源知识注入的指令微调

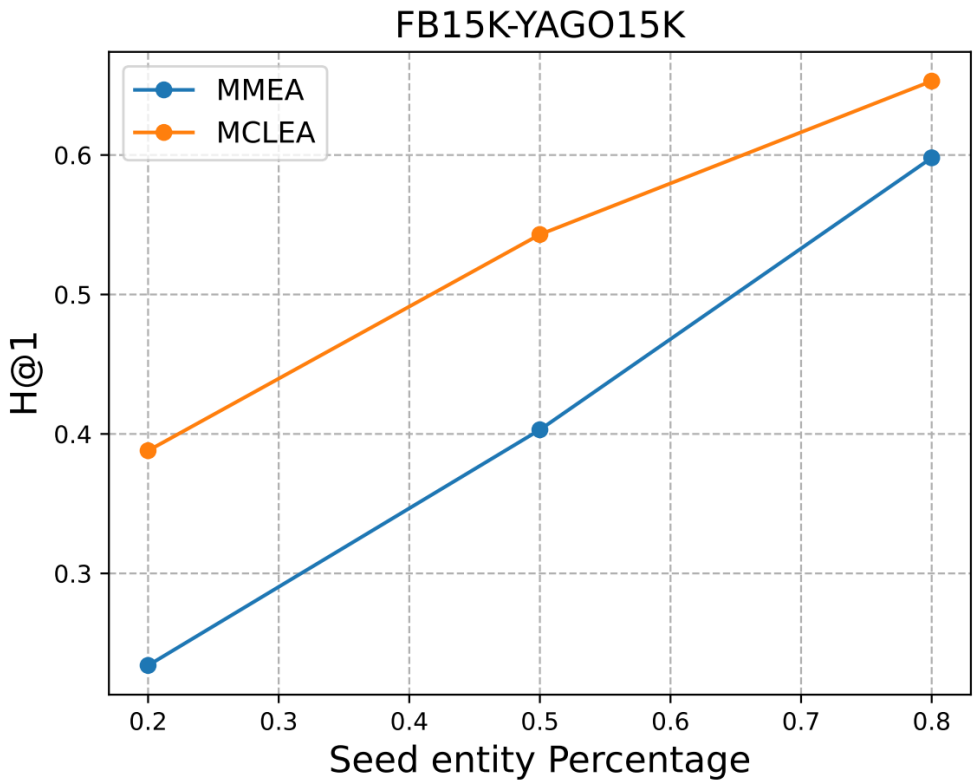
- 拓展到多源知识图谱版本
  - 利用实体对齐合并多个图谱进行表示学习
- 实验结果
  - 相较于单源知识增强的KnowLA相比，多源知识图谱可以提高效果

数据集	CSQA		SIQA	
指标	Acc.	Score	Acc.	Score
KnowLA (单图谱)	57.90	47.39	53.21	46.64
KnowLA (多源图谱)	**59.21**	**48.41**	**53.48**	**46.79**

应用4：多模态知识图谱数据增强

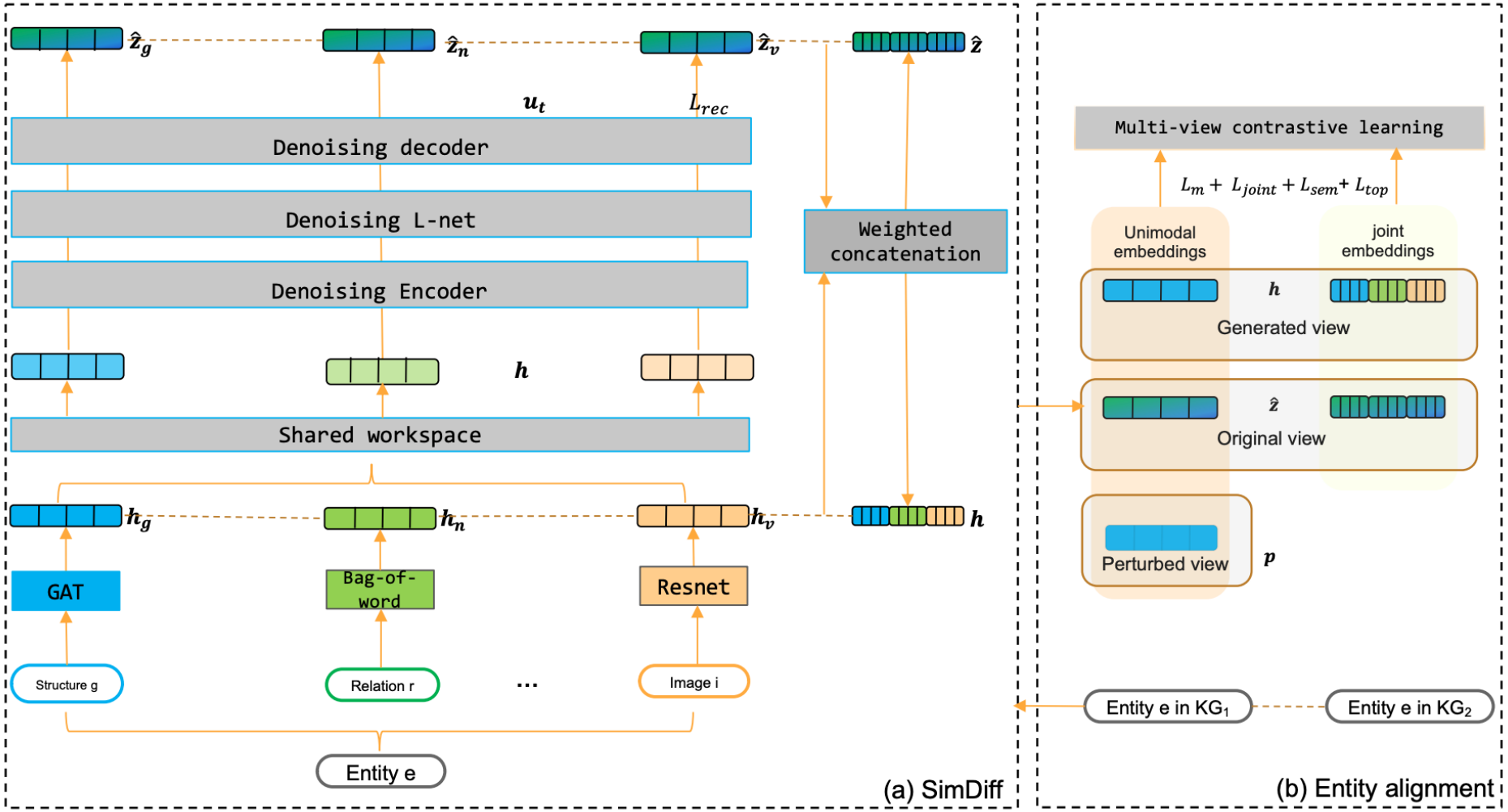
训练数据不足、模态数据不足影响图谱任务的效果





Ran Li, Shimin Di, Lei Chen, Xiaofang Zhou: SimDiff: Simple Denoising Probabilistic Latent Diffusion Model for Data Augmentation on Multi-modal Knowledge Graph. KDD 2024: 1631-1642

应用4：多模态知识图谱数据增强



应用4：多模态知识图谱数据增强

- 解耦和单模态嵌入
  - 将图结构、图像和其它属性解耦，使用不同的编码器获取单模态表征
- 去噪概率扩散模型
  - 通过共享工作空间处理所有模态表征，促进信息交换
  - 使用线性层作为去噪网络，确保在有限数据情况下的稳定训练
  - 通过扩散过程生成新的表征，使用重构损失进行训练
- 生成和应用
  - 将表征输入扩散模型以生成新的表征进行应用

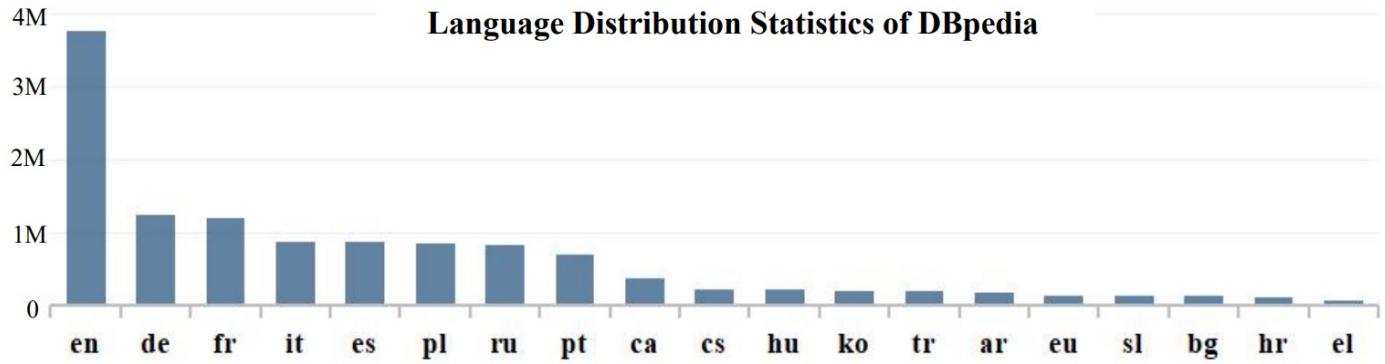
应用4：多模态知识图谱数据增强

通过生成新的多模态数据，有效减少了对训练数据的依赖，提高了泛化能力

	Methods	20% Hits@1	20% Hits@10	20% MRR	50% Hits@1	50% Hits@10	50% MRR	80% Hits@1	80% Hits@10	80% MRR
FB15K-DB15K	MMEA	0.265	0.541	0.357	0.417	0.703	0.512	0.590	0.869	0.685
	EVA*	0.134	0.338	0.201	0.223	0.471	0.307	0.370	0.585	0.444
	HMEA	0.127	0.369	–	0.262	0.581	–	0.417	0.786	–
	PoE	0.126	0.251	0.170	0.464	0.658	0.533	0.666	0.820	0.721
	MSNEA*	0.158	0.403	0.242	0.365	0.662	0.462	0.572	0.821	0.659
	MCLEA	0.445	0.705	0.534	0.573	0.800	0.652	0.730	0.883	0.784
	SimDiff(Ours)	**0.615**	**0.820**	**0.678**	**0.731**	**0.880**	**0.786**	**0.829**	**0.929**	**0.865**
	Improv. best %	38.20	16.31	26.97	27.57	10.00	20.55	13.56	5.21	10.33
FB15K-YAGO15K	MMEA	0.234	0.480	0.317	0.403	0.645	0.486	0.598	0.839	0.682
	EVA*	0.098	0.276	0.158	0.240	0.477	0.321	0.394	0.613	0.471
	HMEA	0.105	0.313	–	0.265	0.581	–	0.433	0.801	–
	PoE	0.113	0.229	0.154	0.347	0.536	0.414	0.573	0.746	0.635
	MSNEA*	0.145	0.357	0.221	0.389	0.660	0.479	0.605	0.821	0.677
	MCLEA	0.388	0.641	0.474	0.543	0.759	0.616	0.653	0.835	0.715
	SimDiff(Ours)	**0.530**	**0.736**	**0.595**	**0.659**	**0.820**	**0.716**	**0.743**	**0.886**	**0.791**
	Improv. best %	36.6	14.82	25.53	21.36	8.037	16.23	13.78	6.108	10.63

应用5：多语言知识图谱补全

不同语言的知识完备性差异很大，低资源语言尤其不完备

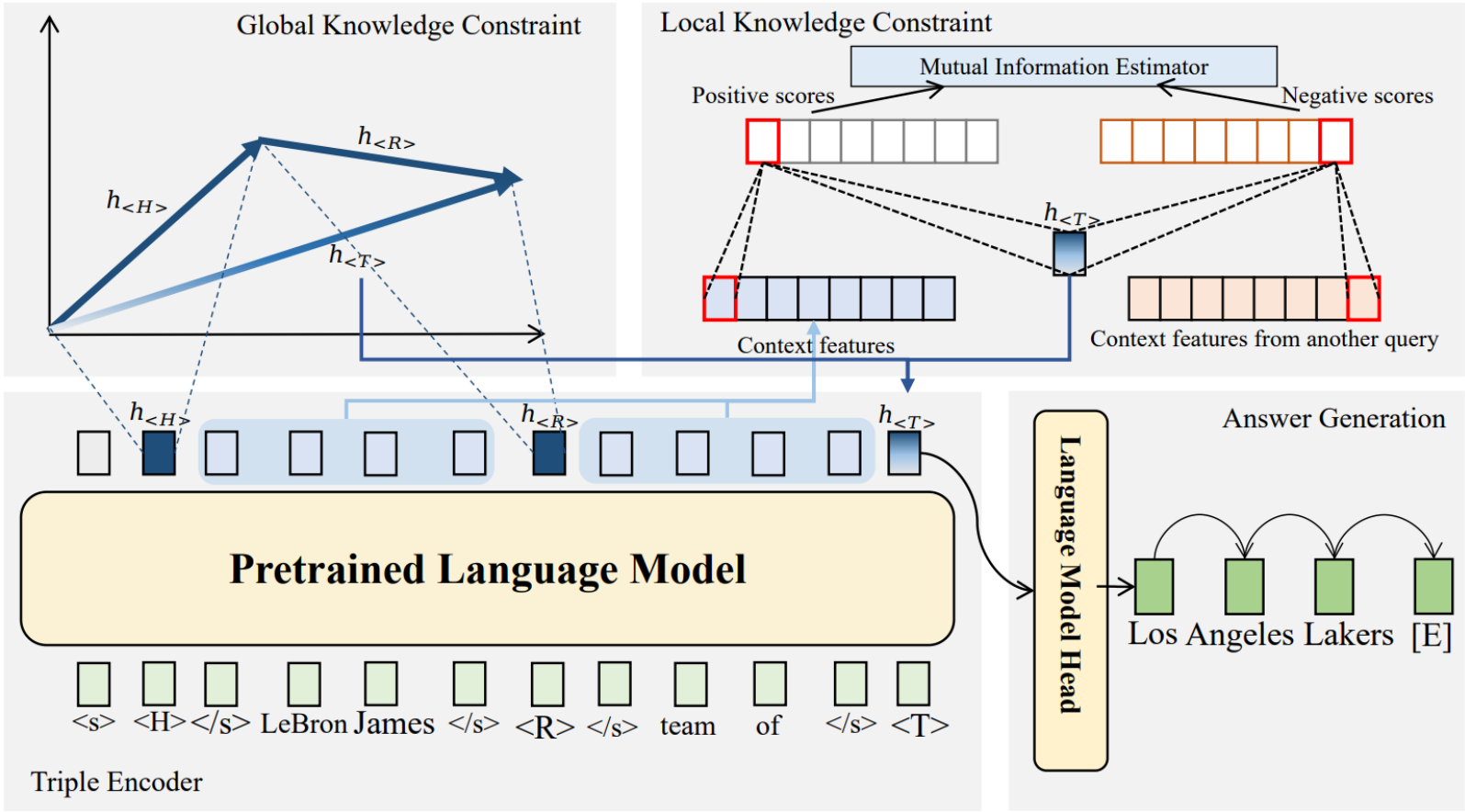


	Results of Prix-LM	Results of Ours
JA	**Query**: (第86回全日本サッカー選手権大会, スタジアム, ?) 86th All Japan Football Championship Stadium	**Golden Anwser:** 鳥取市営サッカー Axis Bird Stadium
	筑波大学蹴球部 (A University Club) University of Tsukuba Football Club	鳥取市営サッカー場 (A Japanese Stadium) Axis Bird Stadium
	貴陽市 (A Chinese City) Guiyang City	栃木市総合運動公園陸上競技場 (A Japanese Stadium) Tochigi City Stadium
	鳥取市営サッカー場 (A Japanese Stadium) Axis Bird Stadium	長居球技場 (A Japanese Stadium) Yodoko Sakura Stadium

Ran Song, Shizhu He, Shengxiang Gao, Li Cai, Kang Liu, Zhengtao Yu, Jun Zhao: Multilingual Knowledge Graph Completion from Pretrained Language Models with Knowledge Constraints. ACL (Findings) 2023: 7709-7721

应用5：多语言知识图谱补全

思路：利用**多语言大模型**进行知识迁移



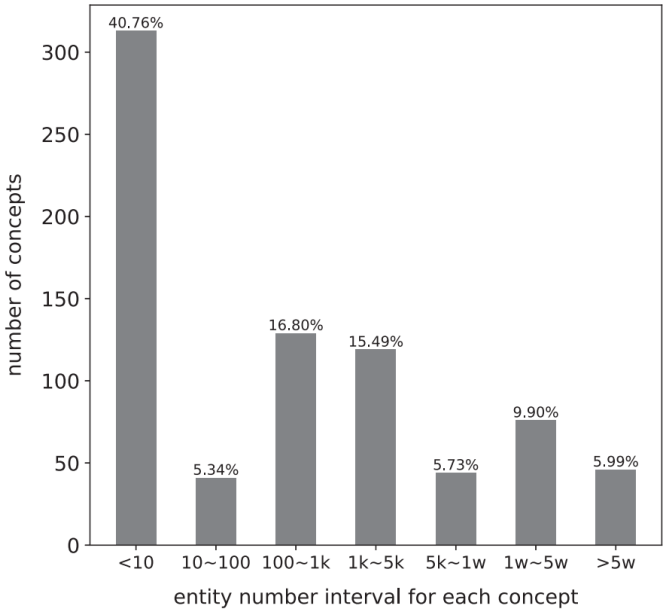
应用5：多语言知识图谱补全

多语言知识迁移可以显著提高各个语言下的知识补全效果

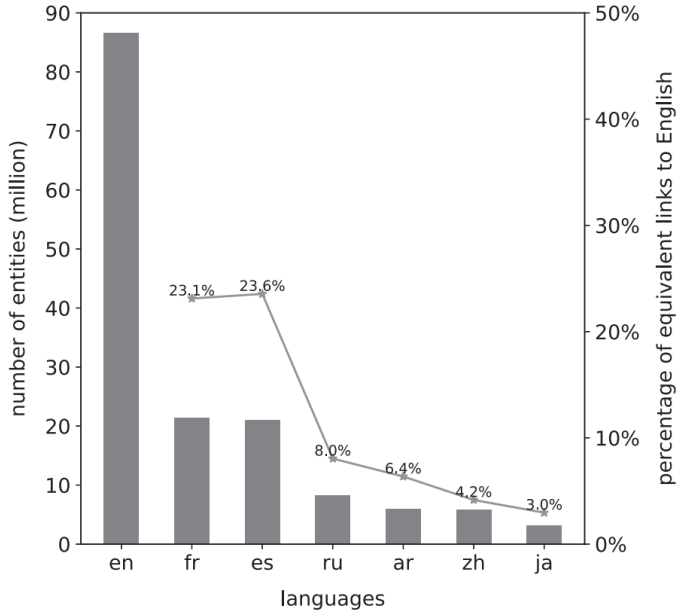
	MODEL	DE	FI	FR	HU	IT	JA	TR	AVG
Hits@1	TransE	0.00	0.01	0.02	0.03	0.04	0.02	0.06	0.02
	ComplEx	4.09	2.45	2.50	3.28	2.87	2.41	1.00	2.65
	RotatE	6.72	5.87	8.40	16.27	6.91	6.21	6.85	8.17
	Prix-LM (Single)	12.86	19.81	18.01	28.72	16.21	19.81	23.79	19.88
	Prix-LM	14.32	18.78	16.47	29.68	14.32	18.19	21.57	19.04
	Ours	**17.54**	**20.74**	**18.34**	**30.91**	**14.98**	**22.05**	**25.20**	**21.39**
Hits@3	TransE	6.14	6.54	6.60	14.91	5.95	7.22	8.20	7.93
	ComplEx	8.47	5.28	5.19	6.70	4.31	4.68	2.11	5.24
	RotatE	10.52	7.42	14.62	21.75	12.11	9.75	11.29	12.49
	Prix-LM (Single)	23.09	28.75	24.75	38.44	25.32	29.02	33.05	28.91
	Prix-LM	23.68	29.54	23.15	39.80	25.46	27.01	31.45	28.58
	Ours	**30.40**	**29.74**	**26.36**	**44.18**	**27.03**	**30.79**	**35.48**	**31.99**
Hits@10	TransE	17.54	17.80	15.26	29.00	14.16	20.65	19.35	19.10
	ComplEx	9.35	8.21	8.91	16.96	8.76	8.23	5.24	9.38
	RotatE	14.61	8.61	19.49	28.31	18.48	14.44	17.13	17.29
	Prix-LM (Single)	33.82	38.91	34.04	47.31	36.61	38.81	38.50	38.28
	Prix-LM	33.91	41.29	32.25	46.23	35.18	36.12	37.50	37.49
	Ours	**41.81**	**43.44**	**35.15**	**58.00**	**39.15**	**42.45**	**44.55**	**43.50**

应用6：知识图谱构建

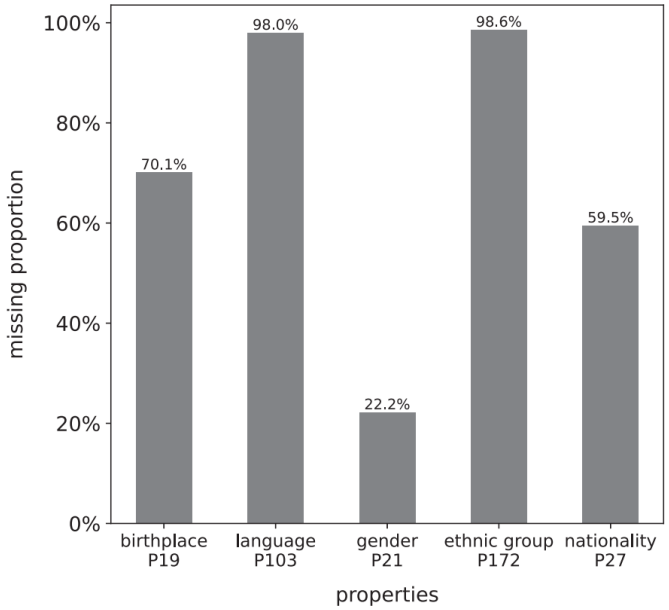
单源知识不完备的问题严重影响知识图谱构建的质量



(a) Concept statistics in DBpedia



(b) Equivalent links in Wikidata

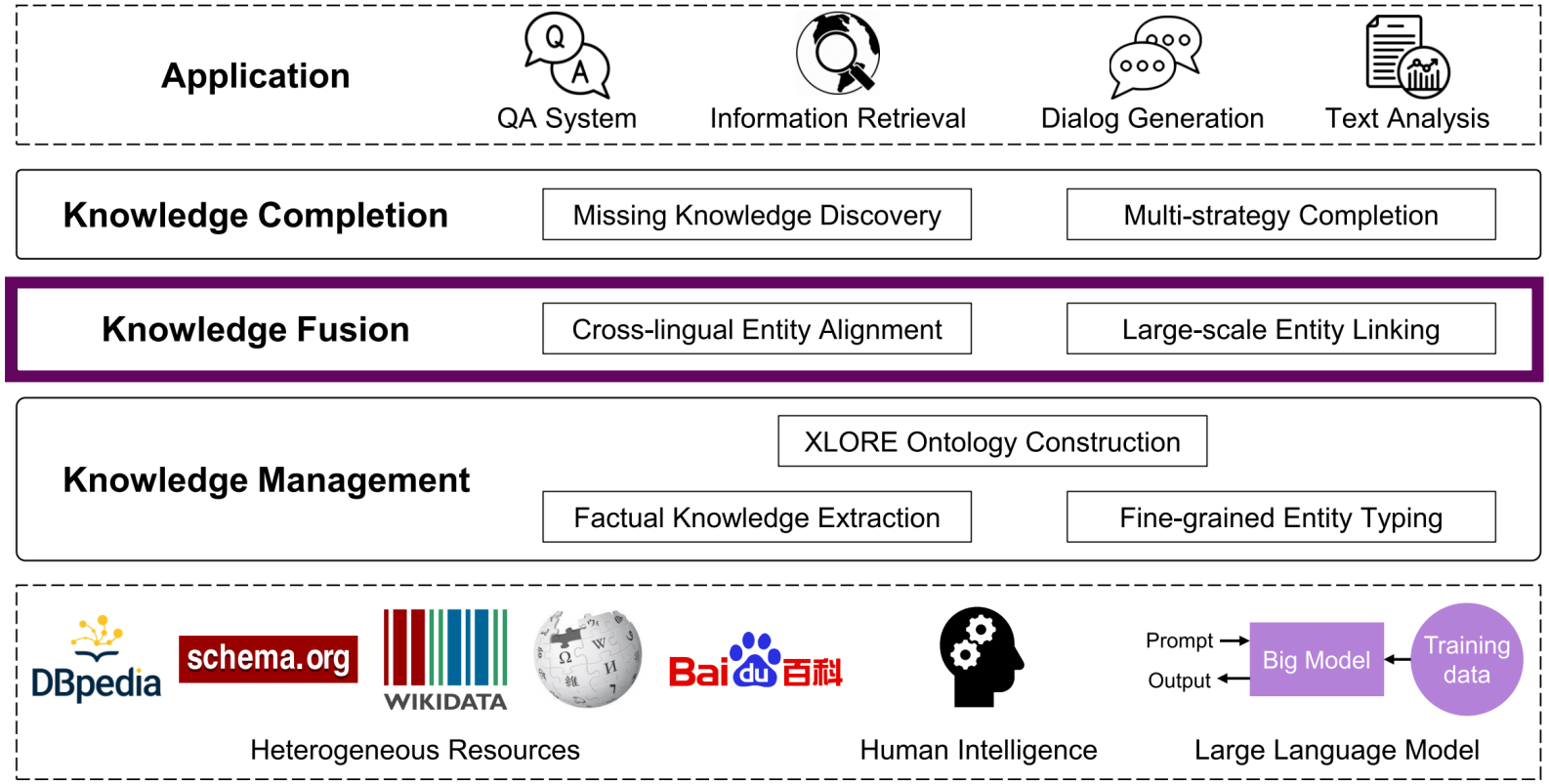


(c) Incompleteness in Wikidata

Kaisheng Zeng, Hailong Jin, Xin Lv, Fangwei Zhu, Lei Hou, Yi Zhang, Fan Pang, Yu Qi, Dingxiao Liu, Juanzi Li, and Ling Feng. 2024. XLORE 3: A Large-Scale Multilingual Knowledge Graph from Heterogeneous Wiki Knowledge Resources. ACM Trans. Inf. Syst. 42, 6, Article 145

应用6：知识图谱构建

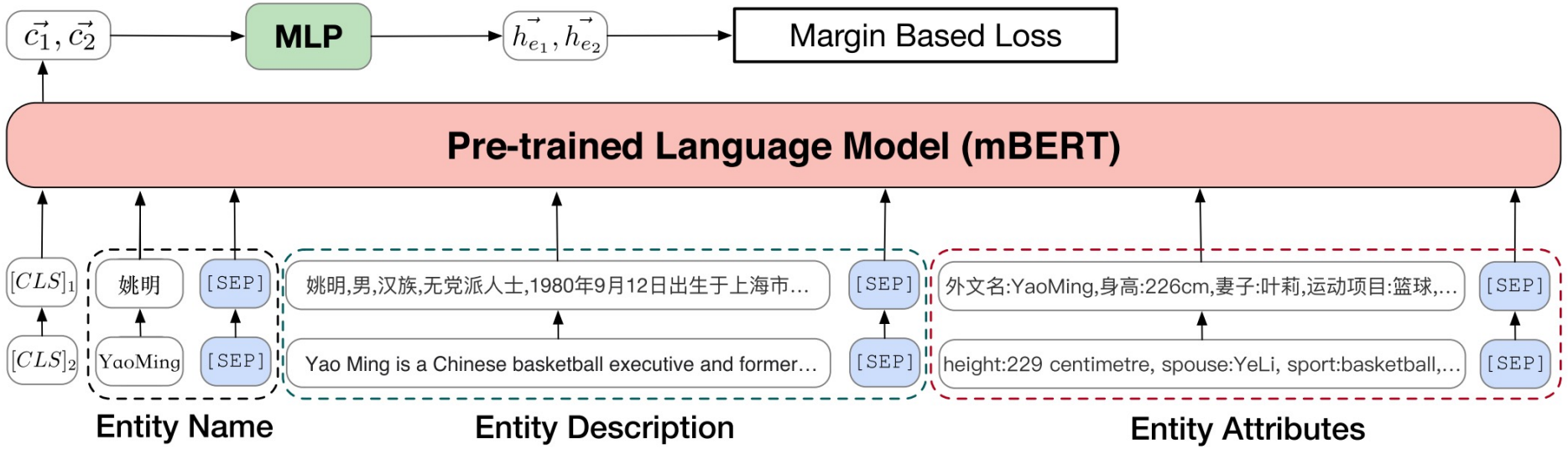
- 思路：**集成多源数据**，进行实体对齐等知识融合操作



应用6：知识图谱构建

- 集成多粒度的文本特征，包括名称、描述和属性
- 微调多语言大模型学习实体表征：

$$\mathcal{L} = \sum_{(e_w, e_b^+, e_b^-) \in D_{train}} \text{MAX}\left\{0, f(\vec{h}_{e_w}, \vec{h}_{e_b^+}) + \gamma - f(\vec{h}_{e_w}, \vec{h}_{e_b^-})\right\}$$



应用6：知识图谱构建

- 基于大模型的实体对齐方法性能优越

Methods	$DBP^{1}_{ZH\_EN}$		$DBP^{2}_{ZH\_EN}$		$DBP^{3}_{ZH\_EN}$	
	Hits@1	Hits@10	Hits@1	Hits@10	Hits@1	Hits@10
MMEA	18.6	38.8	25.1	45.8	26.8	46.1
MTransE	27.8	59.6	17.5	50.6	19.9	51.8
BootEA	29.2	59.5	16.1	43.4	18.4	44.5
GCN_Align	24.7	56.7	14.3	46.0	17.3	50.6
HAKE	26.9	58.3	19.4	52.1	22.0	52.8
Dual_Amn	41.7	68.4	33.1	60.7	36.6	62.8
JAPE	52.6	72.8	52.5	70.6	53.1	71.6
RREA	54.4	78.5	39.8	67.4	45.0	70.5
EVA	53.6	77.4	54.3	71.9	55.3	72.2
SelfKG	62.1	80.1	63.9	81.1	64.0	81.3
ICLEA	83.3	95.0	84.9	94.8	84.1	94.8
Ours	**90.3**	**96.1**	**91.9**	**97.8**	**91.2**	**98.1**

应用6：知识图谱构建

- 多源知识融合极大充实了所构建的知识图谱

Knowledge element	# Entity	# Property	# Concept	# Fact	# Qualifier	Equivalent link
XLORE 2	14,951,135	512,883	1,371,272	163.43M	0	423,974
XLORE 3	66,237,822	2,608	446	1998.6M	342	750,281

- 在线网站：https://xlore.cn/index

提纲

- 研究背景
- 方法
- 应用
- **总结与展望**

总结与展望

- 近期工作尝试解决如下问题：

模态歧义与缺失	图表征	图文异质性	数据标注噪音	多步推理错误积累	缺乏解释
ISWC'23 (best paper candidate)	VLDB'24	arXiv'24	arXiv'24	ACL'24	ICML'23, ICDE'24

- 近期工作更多关注知识融合的进一步应用，如：

多源知识问答	多源图谱预训练	知识补全	指令微调	数据增强	图谱构建
arXiv'23	KDD'23	ACL'23	NAACL'24	KDD'24	TOIS'24

- 个人认为，未来工作可以考虑

多源知识图谱联合对齐 (n>2)　知识计算基座模型　大模型与知识图谱的融合计算

其它参考文献

关于多源知识图谱对齐

- Yaming Yang, Zhe Wang, Ziyu Guan, Wei Zhao, Weigang Lu, Xinyan Huang: Aligning Multiple Knowledge Graphs in a Single Pass. CoRR abs/2408.00662 (2024)

关于知识图谱基座模型

- Mikhail Galkin, Xinyu Yuan, Hesham Mostafa, Jian Tang, Zhaocheng Zhu: Towards Foundation Models for Knowledge Graph Reasoning. ICLR 2024
- 崔员宁, 孙泽群, 胡伟. 基于规则提示的知识图谱通用推理预训练模型[J]. 计算机研究与发展, 2024, 61(8): 2030-2044. DOI: 10.7544/issn1000-1239.202440133

关于大语言模型与知识图谱融合

- Shiyu Tian, Yangyang Luo, Tianze Xu, Caixia Yuan, Huixing Jiang, Chen Wei, Xiaojie Wang: KG-Adapter: Enabling Knowledge Graph Integration in Large Language Models through Parameter-Efficient Fine-Tuning. ACL (Findings) 2024: 3813-3828
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Jonathan Larson: From Local to Global: A Graph RAG Approach to Query-Focused Summarization. CoRR abs/2404.16130 (2024)

相关综述文献

知识图谱融合

- Pavel Shvaiko, Jérôme Euzenat: Ontology Matching: State of the Art and Future Challenges. IEEE Trans. Knowl. Data Eng. 25(1): 158-176 (2013)
- 庄严, 李国良, 冯建华. 知识库实体对齐技术综述[J]. 计算机研究与发展, 2016, 53(1): 165-192. DOI: 10.7544/issn1000-1239.2016.20150661
- Kaisheng Zeng, Chengjiang Li, Lei Hou, Juanzi Li, Ling Feng: A comprehensive survey of entity alignment for knowledge graphs. AI Open 2: 1-13 (2021)

知识图谱表示学习

- Quan Wang, Zhendong Mao, Bin Wang, Li Guo: Knowledge Graph Embedding: A Survey of Approaches and Applications. IEEE Trans. Knowl. Data Eng. 29(12): 2724-2743 (2017)
- Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, Philip S. Yu: A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. IEEE Trans. Neural Networks Learn. Syst. 33(2): 494-514 (2022)

知识图谱与大模型协同

- Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, Xindong Wu: Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Trans. Knowl. Data Eng. 36(7): 3580-3599 (2024)

相关工具资源

- 实体对齐工具库OpenEA
  - 支持20+实体对齐方法
  - https://github.com/nju-websoft/OpenEA
- 多源知识图谱表示学习与应用库μKG
  - 支持10+表示学习方法、10+知识图谱融合方法
  - 支持4个主要下游应用及其对应的数据集
  - https://github.com/nju-websoft/muKG

谢谢！ 请批评指正！

孙泽群
sunzq@nju.edu.cn
http://ws.nju.edu.cn