南京大学 NANJING UNIVERSITY

Websoft 南京大学万维网软件研究组
The Websoft Research Group, Nanjing University, China

从语义网视角看
知识图谱的近期研究进展

胡 伟

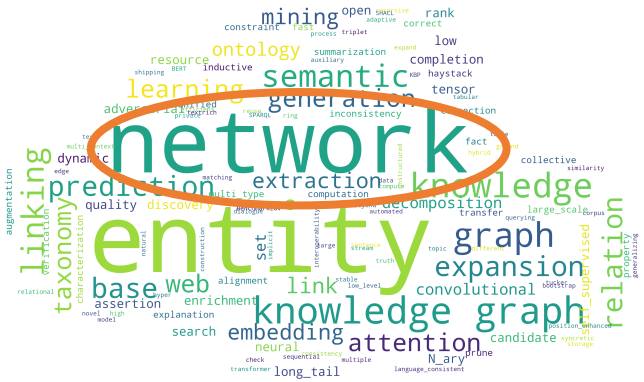
南京大学计算机软件新技术国家重点实验室

南京大学健康医疗大数据国家研究院

调研范围：2020 年语义网领域三大会议

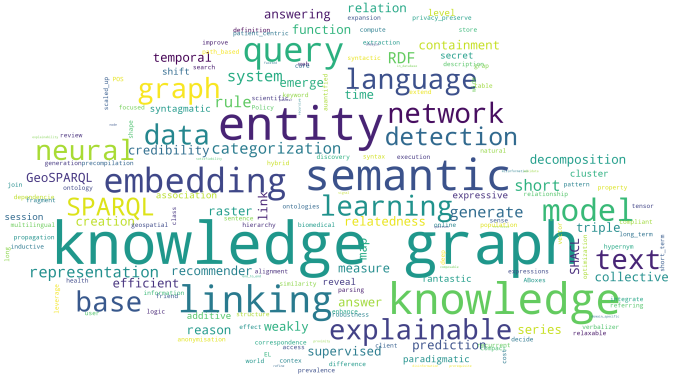
WWW

- The Web Conference
- 研究论文 25 篇
  (Semantics & Knowledge)



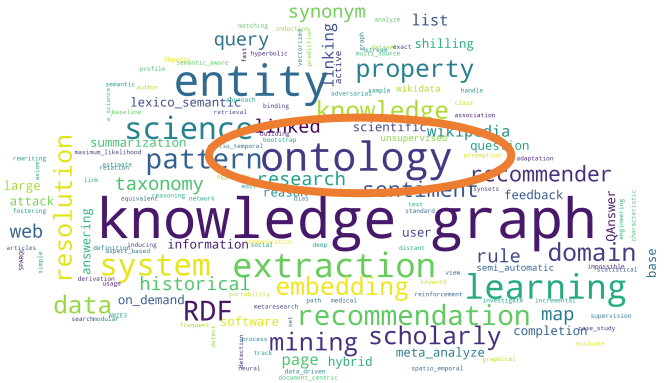
ISWC

- Int’l Semantic Web Conf.
- 研究论文 38 篇



ESWC

- Ext. Semantic Web Conf.
- 研究论文 26 篇



热点分析

知识图谱表示学习

- 多元关系、新兴实体、时序图谱、双曲空间
- 链接预测、实体聚类

知识抽取与图谱构建

- 实体抽取、关系抽取
- 本体构建、分类构建、领域图谱
- 低资源场景、长尾场景

实体对齐

- 多类型实体
- 长路径依赖
- 文本信息
- 主动学习

搜索、摘要与推荐

- 实体排序
- 实体摘要
- SPARQL
- RDF存储
- 推荐增强

知识图谱问答

- 实体链接
- 关系链接
- 知识融合
- 答案补全
- 可移植性

新数据集

1. 知识图谱表示学习

多元关系

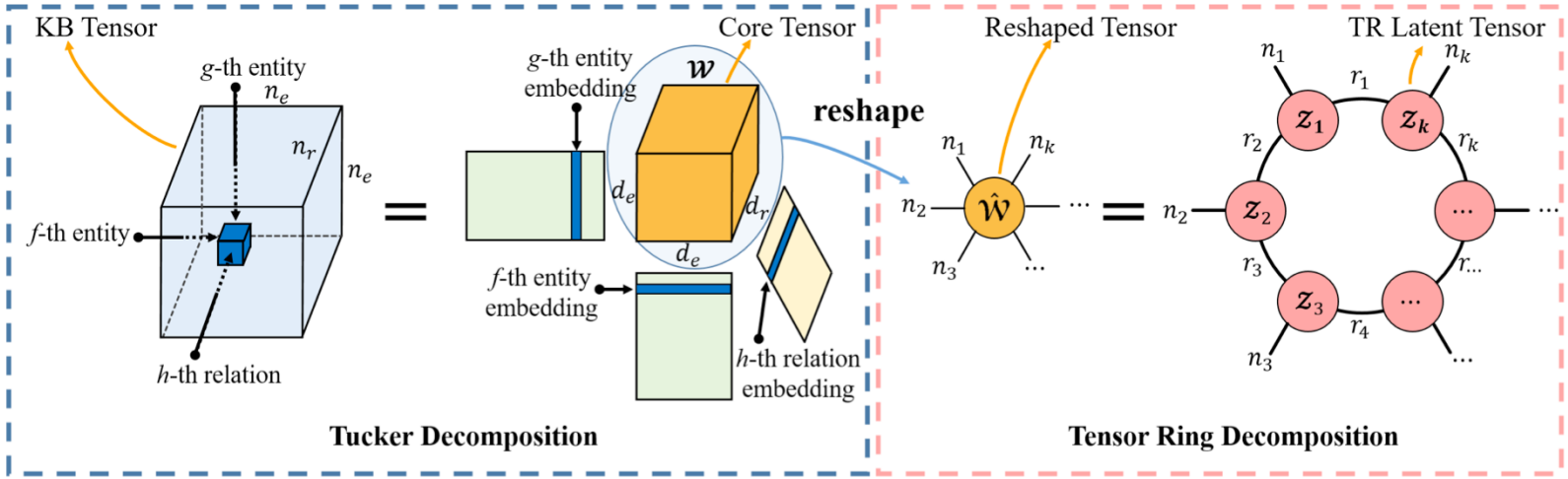
Liu et al. Generalizing Tensor Decomposition for N-ary Relational Knowledge Bases. WWW

- 多元关系
  - 3 元：*Purchase* (Person, Product, Seller)
  - 4 元：*Sports_award* (Player, Team, Award, Season)
- 多元关系知识库的链接预测
  - **问题定义：给定 *n* 元关系和 (*n*–1) 个相关实体，预测缺失的那个实体**
  - 现有工作
    - 基于翻译距离的模型 → 表达能力弱
    - 基于神经网络的模型 → 复杂度高
  - 本文思路：泛化张量分解（2 元 → *n* 元）
    - 一个 *n* 元关系知识库 → (*n*+1) 阶知识库张量

多元关系 (cont'd)

Liu et al. Generalizing Tensor Decomposition for N-ary Relational Knowledge Bases. WWW

- 具体做法
  1. 通过 Tucker 分解来分解原始知识库张量
  2. 将核心张量 W 重塑为 k 阶张量 $\hat{\mathcal{W}}$
  3. 通过 Tucker Ring 分解来分解重塑张量 $\hat{\mathcal{W}}$



- 实验结果
  - 张量分解的方法 GETD, n-CP, n-TuckER 优于基于翻译距离的 RAE 模型和基于神经网络的 NaLP 模型

Table 4: Link prediction results on WikiPeople dataset.

Model	WikiPeople-3				WikiePeople-4			
	MRR	Hits@10	Hits@3	Hits@1	MRR	Hits@10	Hits@3	Hits@1
RAE	0.239	0.379	0.252	0.168	0.150	0.273	0.149	0.080
NaLP	0.301	0.445	0.327	0.226	0.342	0.540	0.400	0.237
n-CP	0.330	0.496	0.356	0.250	0.265	0.445	0.315	0.169
n-TuckER	0.365	0.548	0.400	0.274	0.362	0.570	0.432	0.246
GETD	**0.373**	**0.558**	**0.401**	**0.284**	**0.386**	**0.596**	**0.462**	**0.265**

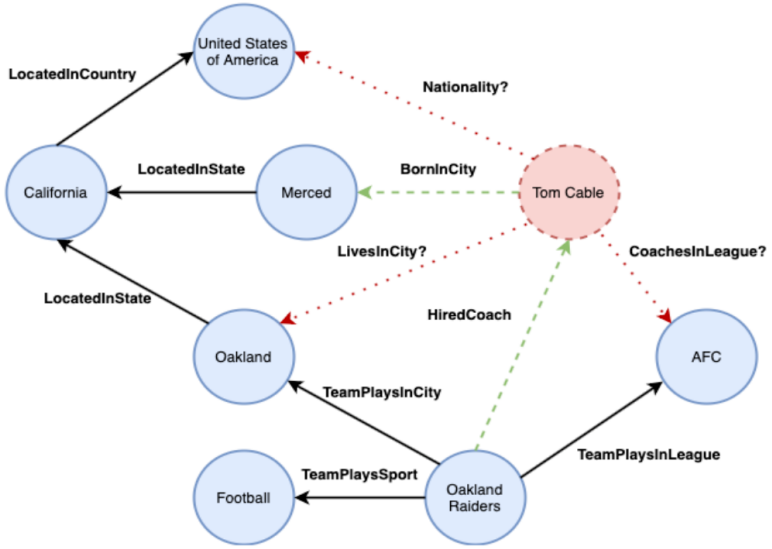
Table 5: Link prediction results on JF17K dataset.

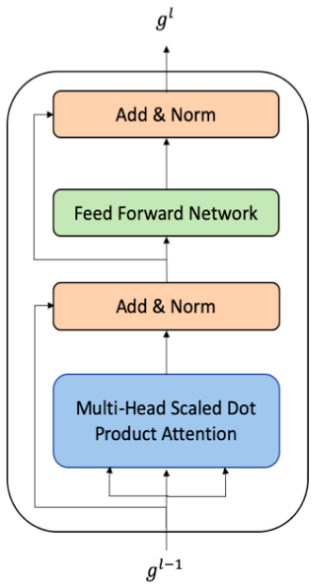
Model	JF17K-3				JF17K-4			
	MRR	Hits@10	Hits@3	Hits@1	MRR	Hits@10	Hits@3	Hits@1
RAE	0.505	0.644	0.532	0.430	0.707	0.835	0.751	0.636
NaLP	0.515	0.679	0.552	0.431	0.719	0.805	0.742	0.673
n-CP	0.700	0.827	0.736	0.635	0.787	0.890	0.821	0.733
n-TuckER	0.727	0.852	0.761	0.664	0.804	0.902	0.841	0.748
GETD	**0.732**	**0.856**	**0.764**	**0.669**	**0.810**	**0.913**	**0.844**	**0.755**

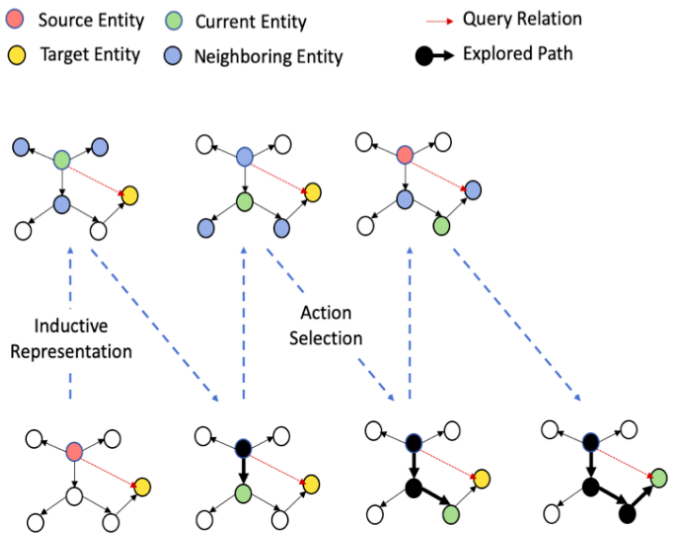
新兴实体

Bhowmik & Melo. Explainable Link Prediction for Emerging Entities in Knowledge Graphs. ISWC

- 新兴实体（emerging entity）
  - 与知识图谱中已有实体的关联非常稀疏
- 一个归纳式表示学习框架
  - 适合于拥有许多新兴实体的动态知识图谱
  - 保留了基于路径的模型的可解释性推理







顶点表示归纳

- Graph Transformer 编码
  - 根据邻域信息与查询关系的相关性来聚合邻域信息

可解释性推理

- 强化学习解码
  - 使用策略梯度来解码得到通往答案实体的推理路径

新兴实体 (cont'd)

Bhowmik & Melo. Explainable Link Prediction for Emerging Entities in Knowledge Graphs. ISWC

- 实验设定
  - FB15K-237、WN18RR 和 NELL-995
  - **对数据集进行重新划分**，使得测试集中主语实体是未见过的新兴实体
- 实验结果

	WN18RR-Inductive				FB15K-237-Inductive				NELL-995-Inductive			
		Hits@N				Hits@N				Hits@N		
Model	MRR	@1	@3	@10	MRR	@1	@3	@10	MRR	@1	@3	@10
TransR [24]	0.8	0.6	0.7	0.9	5.0	4.0	5.2	6.6	5.3	4.9	5.3	6.5
TransH [40]	0.0	0.0	0.0	0.0	6.2	5.4	6.3	8.0	3.6	3.4	3.6	3.6
RotatE [32]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ConvE [7]	1.9	1.1	2.1	3.5	26.3	20.0	28.7	38.8	43.4	32.5	50.3	60.9
R-GCN [29]	14.7	11.4	15.1	20.7	19.1	11.5	20.9	34.3	58.4	50.9	62.9	71.6
SACN [30]	17.5	9.7	20.3	33.5	29.9	20.5	32.8	50.0	42.4	37.0	42.9	53.2
CompGCN [36]	2.2	0.0	2.2	5.2	26.1	19.2	28.5	39.2	42.8	33.1	47.9	61.0
AnyBURL [25]	-	**48.3**	50.9	53.9	-	28.3	43.0	56.5	-	8.7	11.0	12.3
MultiHopKG [23]	45.5	39.4	49.2	56.5	38.6	29.3	43.4	56.7	74.7	69.1	78.3	84.2
Our Model w/ RS	**48.8**	42.1	**52.2**	**60.6**	**39.8**	**30.7**	**44.5**	**57.6**	**75.2**	**69.7**	**79.1**	**84.4**

基于 embedding 的模型无法处理新兴实体，无法进行归纳式表示学习和推理

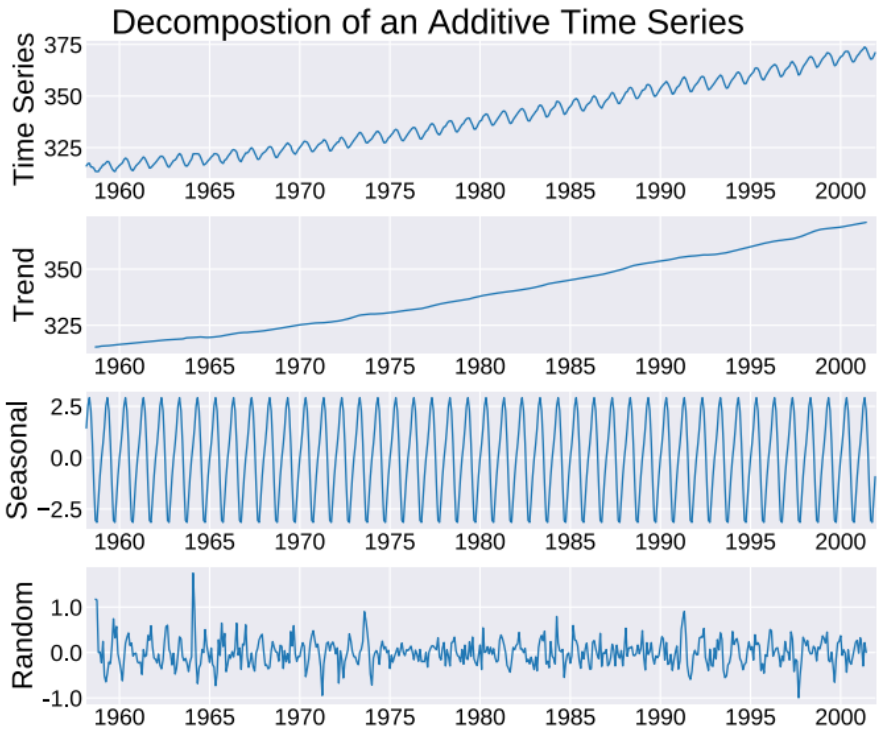
基于图卷积的模型没有显示考虑查询的关系信息

→ 考虑了邻域信息和查询关系的相关性

时序图谱

Xu et al. Temporal Knowledge Graph Completion Based on Time Series Gaussian Embedding. ISWC

- 大多数现有的知识图谱表示学习模型忽略了图谱中有用的时序信息
- 时间感知的事实
  - (*Obama*, *wasBornIn*, *Hawaii*) // August 4, 1961
  - (*Obama*, *presidentOf*, *USA*) // from 2009 to 2017
- ATiSE
  - 使用多维加性时间序列分解来刻画实体/关系的时间演变过程
    - **时间序列 = 趋势 + 周期 + 随机**
  - 考虑实体/关系表示随时间演化的不确定性，将时序知识图谱的表示映射到多维高斯分布的空间中

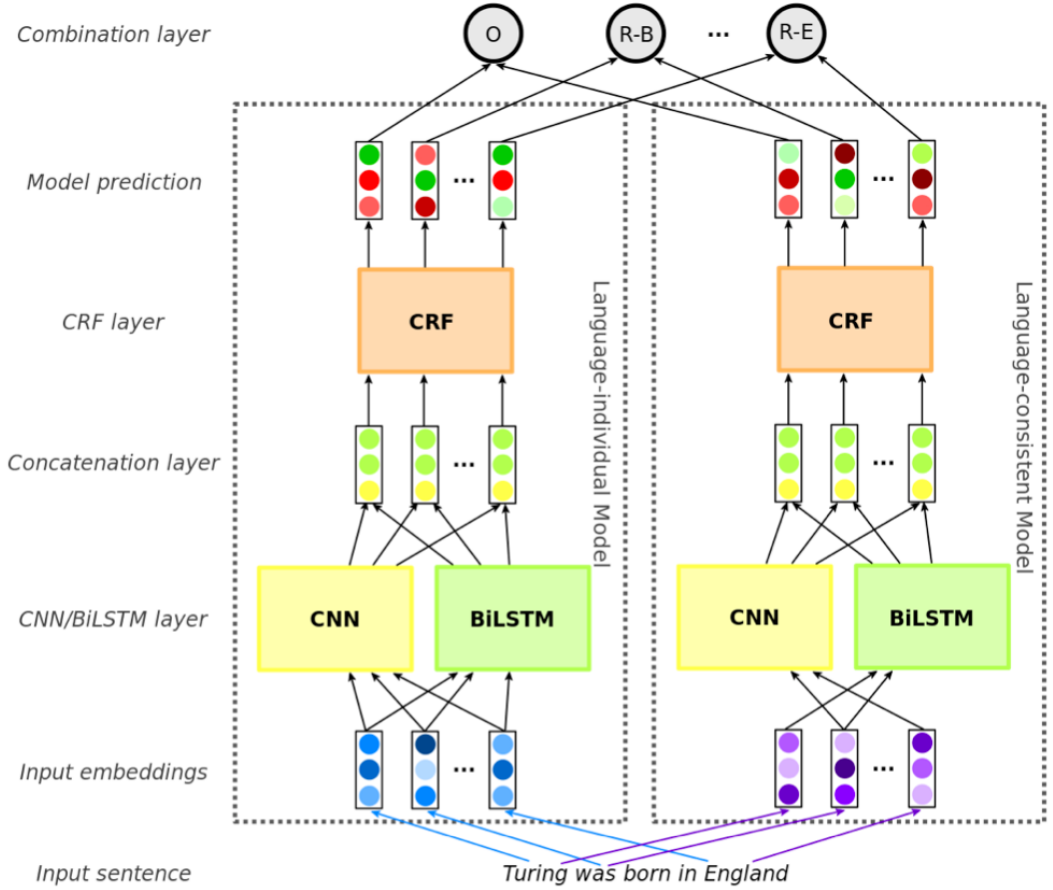


2. 知识抽取与图谱构建

开放关系抽取

Harting et al. LOREM: Language-consistent Open Relation Extraction from Unstructured Text. WWW

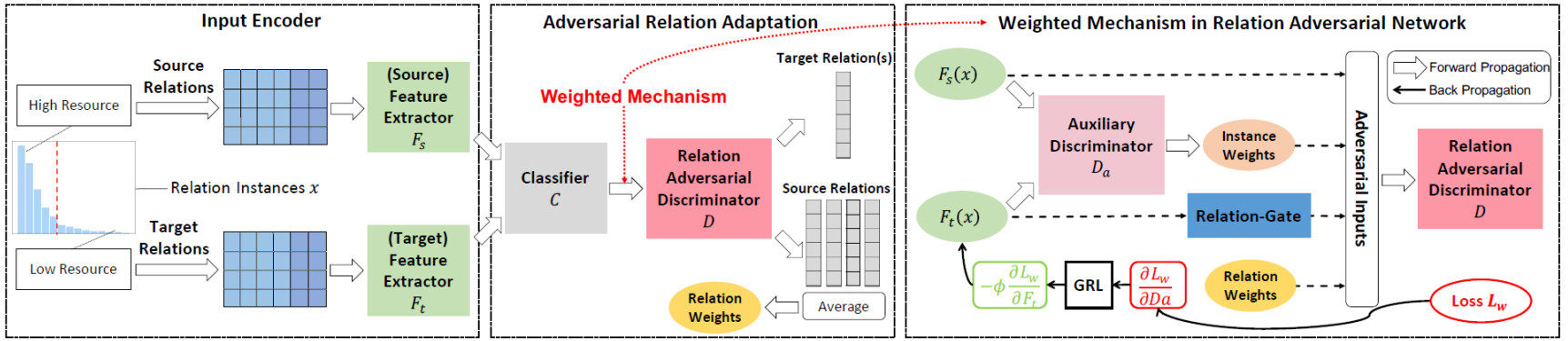
- 开放关系抽取：在无结构文本中发现实体之间的任意语义关联
  - “Turing was born in England in 1912” → <*Turing*, *was born in*, *England*>
- LOREM
  - **特点：不依赖特定语言的知识和外部的 NLP 工具**
    - 利用语言一致的关系结构来提升多语言的性能
    - 使用多语言对齐的词嵌入作为关系抽取器的输入
  - 具体做法
    - 一个面向特定语言的模型
    - 一个面向所有语言的模型
    - 一个组合层



低资源知识图谱补全

Zhang et al. Relation Adversarial Network for Low Resource Knowledge Graph Completion. WWW

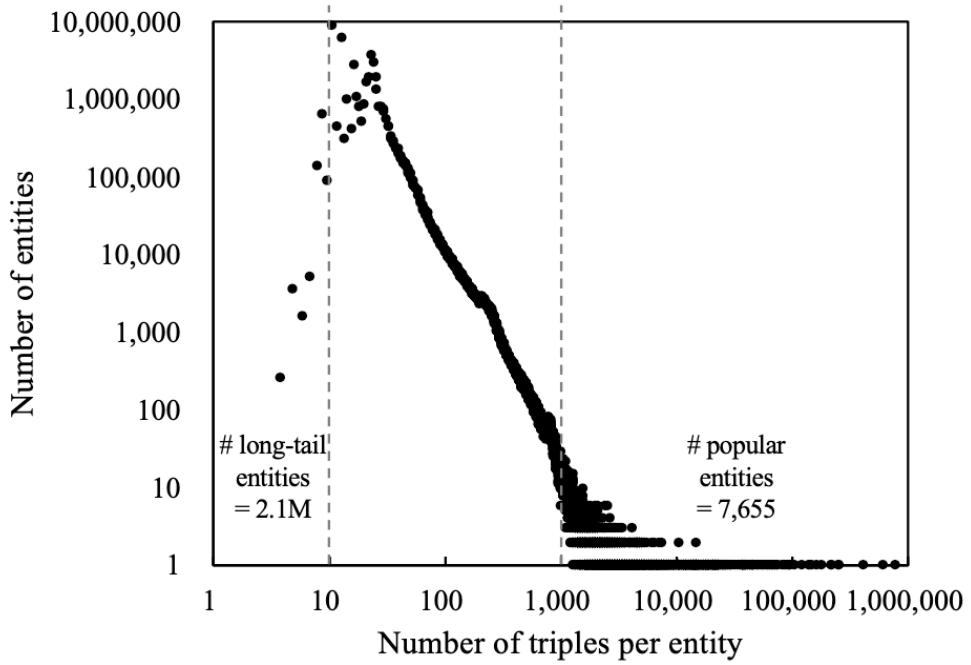
- 知识图谱补全：通过**链接预测或关系抽取**来填补缺失的关联
  - 低资源环境：新添加的关系通常只有很少的训练样例
- 主要思想：利用一个对抗过程来帮助把从高资源关系所学到的知识/特征适配到不同但相关的低资源关系
- 具体做法：加权关系对抗网络
  - 对抗关系适配：寻找可以区分具有不同关系分布的样本的关系判别器
  - 加权关系适配：识别无关的源关系/样本并自动降低其重要性，以解决负迁移问题并鼓励正迁移



长尾实体丰富

Cao et al. Open Knowledge Enrichment for Long-tail Enrichment. WWW

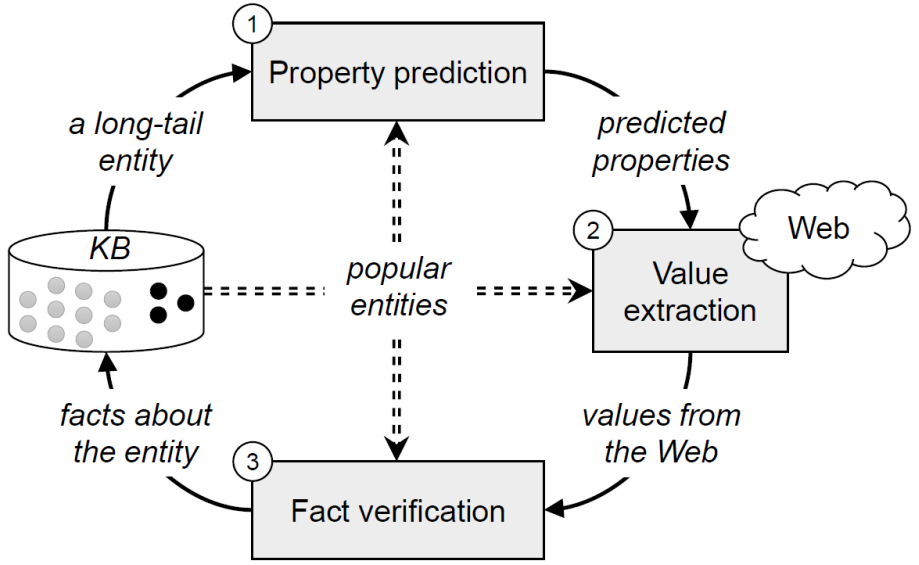
- 长尾实体
  - Freebase 中约有 210 万个实体少于10个的事实，而有 7,655 个实体超过一千个的事实
    - 幂律分布 (power-law)
- 现有工作
  - 仅针对实体丰富的部分环节
  - 缺乏对长尾实体的特别处理
- 本文思路
  - 借助于相似的流行实体和广泛可用的 Web 数据
    - 要找出一个人缺少什么，就看看其他人拥有什么
    - 一些长尾实体只是在知识库中缺乏事实，而不是在真实世界中也缺乏事实



长尾实体丰富 (cont'd)

Cao et al. Open Knowledge Enrichment for Long-tail Enrichment. WWW

- OKELE：**全流程**
  - 基于图神经网络的属性预测
  - 面向结构化/半结构化/文本数据的取值抽取
  - 基于概率图模型的事实验证
- 实验结果
  - Freebase 选取 10 个类，每个类 50 个长尾实体
  - 为每个长尾实体预测 10 个属性及可能的取值



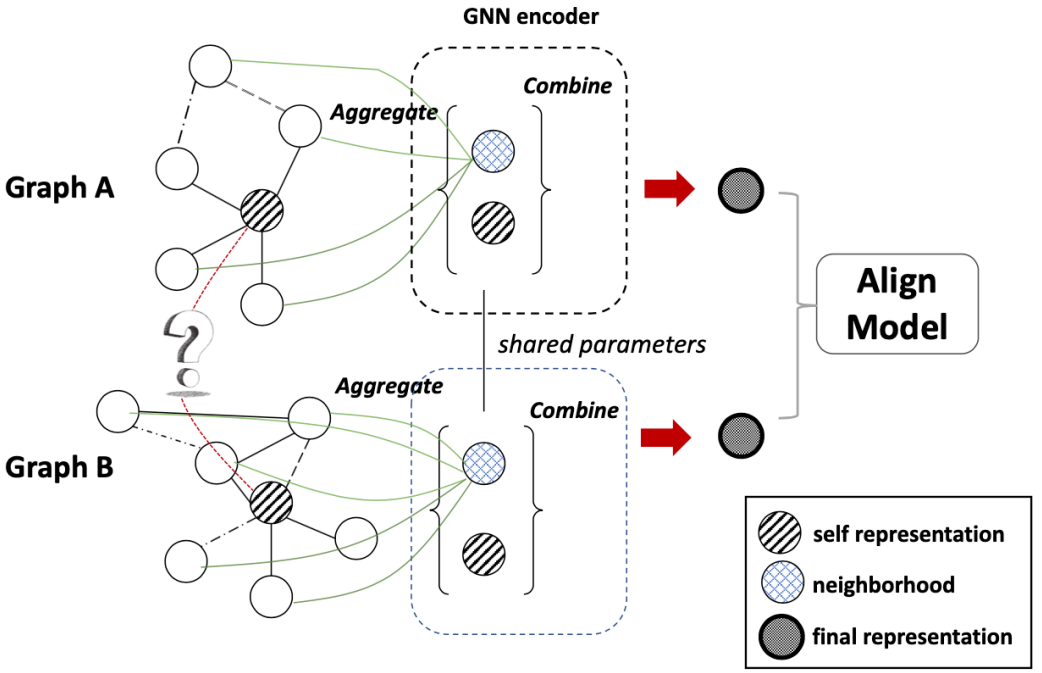
	Models	actor	album	book	building	drug	film	food	mountain	ship	software	Avg.
#Verified props.	GMF+CATD	280	134	205	218	170	417	65	183	254	207	4.27
	OKELE	264	167	266	209	170	432	70	182	260	199	4.44
#Verified facts	GMF+CATD	485	153	228	328	375	722	402	275	303	248	7.04
	OKELE	508	198	418	320	547	1,027	615	272	301	247	8.91
Precision	GMF+CATD	**0.845**	0.204	0.312	0.527	0.710	0.846	0.501	0.440	0.837	0.444	0.567
	OKELE	0.805	**0.290**	**0.464**	**0.531**	**0.831**	**0.890**	**0.665**	**0.446**	**0.870**	**0.446**	**0.624**

3. 实体对齐

多类型实体

Zhu et al. Collective Multi-type Entity Alignment Between Knowledge Graphs. WWW

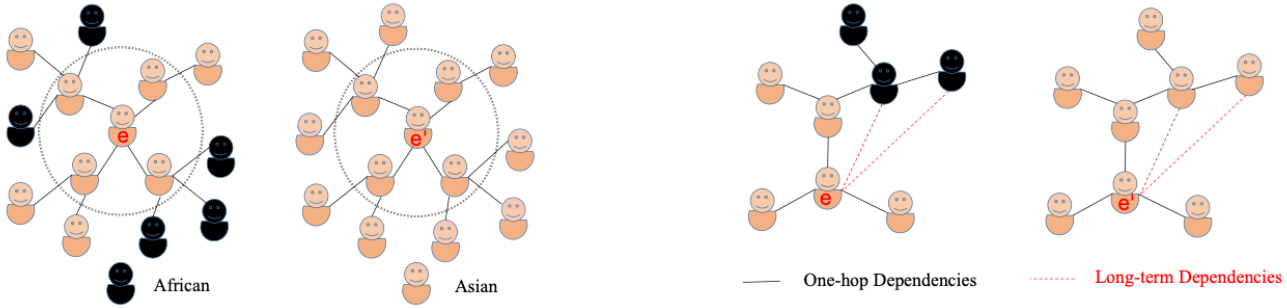
- 实体对齐：识别不同知识图谱中指称真实世界相同对象的实体
- 本文动机
  - 针对不同实体类型的对齐策略可能不同
  - 当前的实体对齐模型不能通过单一模型来对齐多类型实体
- 本文思想：集体决策
  - **精心设计的注意力机制**：有效利用共享邻居信息作为正面证据，也不会忽略重要的负面证据
- CG-MuAlign
  - Node-level cross-graph attention: 相似邻居
  - Edge-level relation-aware self-attention: 负面证据



短期区别和长期依赖

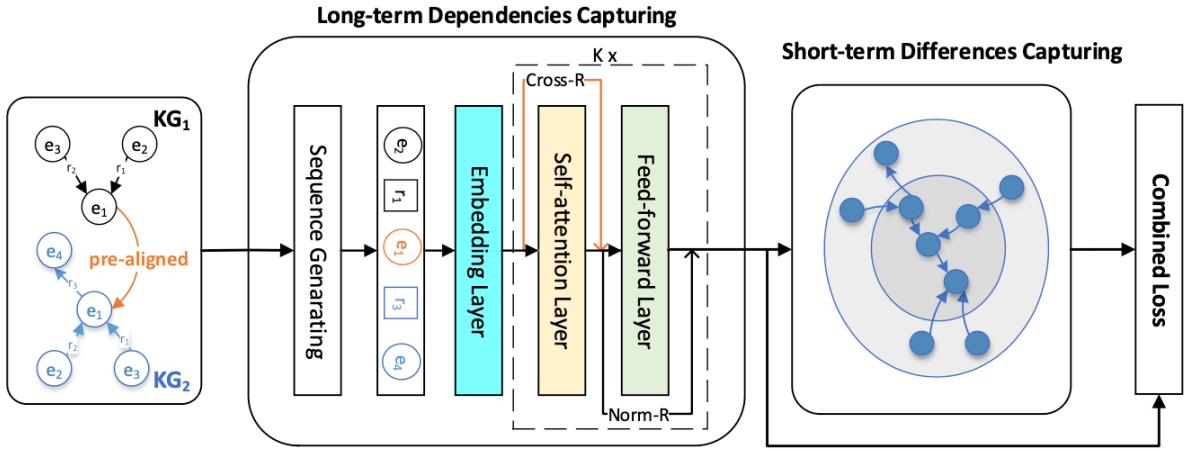
Chen et al. Learning Short-term Differences and Long-term Dependencies for Entity Alignment. ISWC

- 主要思想：捕获周围实体和多跳实体中隐含的实体之间的一些高级交互
  - 短期区别和长期依赖



(a) Short-term differences　　(b) Long-term dependencies

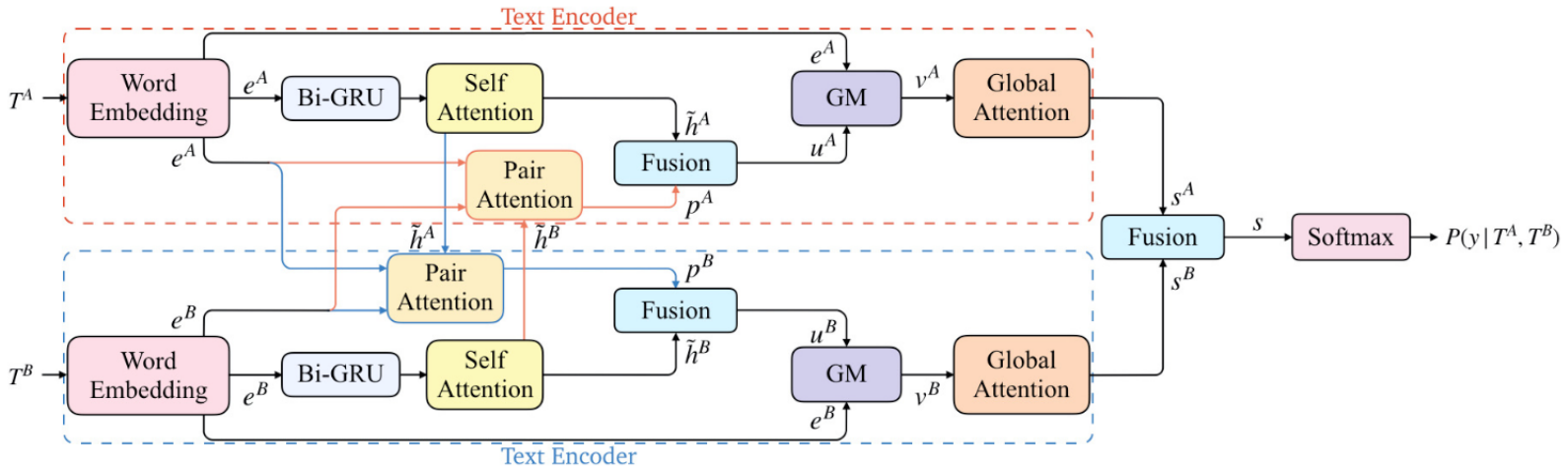
- 具体做法
  - 长期依赖：随机游走 + 自注意力
  - 短期区别：GNN



多上下文注意力

Zhang et al. Multi-Context Attention for Entity Matching. WWW

- Entity matching with textual instances (**数据库领域**)
  - DeepER [VLDB 2018]：GloVe + LSTM
  - DeepMatcher [SIGMOD 2018]：通过注意力机制扩展 RNN
- MCA (multi-context attention)
  - 主要思想：对于实体文本描述对，充分挖掘嵌入向量的语义上下文
  - 具体做法
    - Self-attention
    - Pair-attention
    - Global-attention
    - Attribute-attention



4. 搜索、摘要与推荐

实体摘要

Liu et al. Entity Summarization with User Feedback. ESWC

- 实体摘要
  - 面临挑战：当实体摘要的质量无法满足用户的信息需求时，缺乏改进实体摘要的机制
- DRESSED
  - 主要思想：将用户引入摘要过程，获得用户反馈
  - 交互过程：“删除-替代”场景
    - 用户删掉“删除项”，系统替换上“替代项”

实体摘要 (cont'd)

Liu et al. Entity Summarization with User Feedback. ESWC

- 具体做法
  - 将“删除-替代”场景建模为 Markov 决策过程
  - 强化学习
    - 策略网络

state: $Z_i = \langle S_i, F_i, C_i, f_i \rangle$,

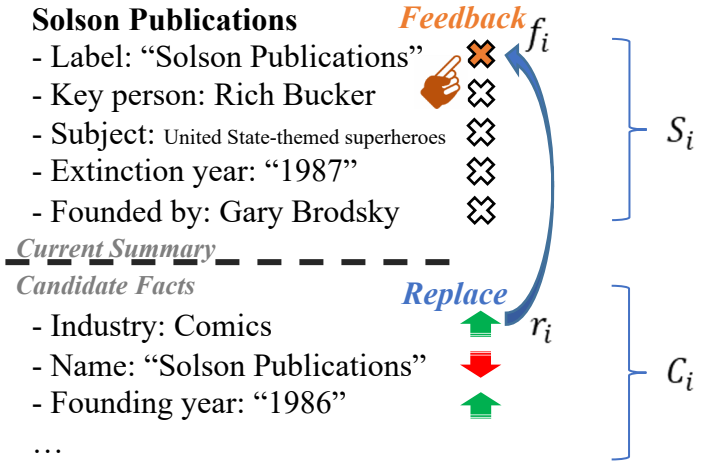
action: $A_i = r_i$,

policy: $\pi_{\boldsymbol{\theta}}(t|Z_i) = \frac{\exp(\texttt{score}(t|Z_i, \boldsymbol{\theta}))}{\sum_{t' \in C_i} \exp(\texttt{score}(t'|Z_i, \boldsymbol{\theta}))}$,

reward: $R_{i+1} = \rho(Z_i, A_i) = \frac{\texttt{rel}(r_i)}{\log(i+2)}$,

transition: $Z_{i+1} = \tau(Z_i, A_i) = \langle S_{i+1}, F_{i+1}, C_{i+1}, f_{i+1} \rangle$,

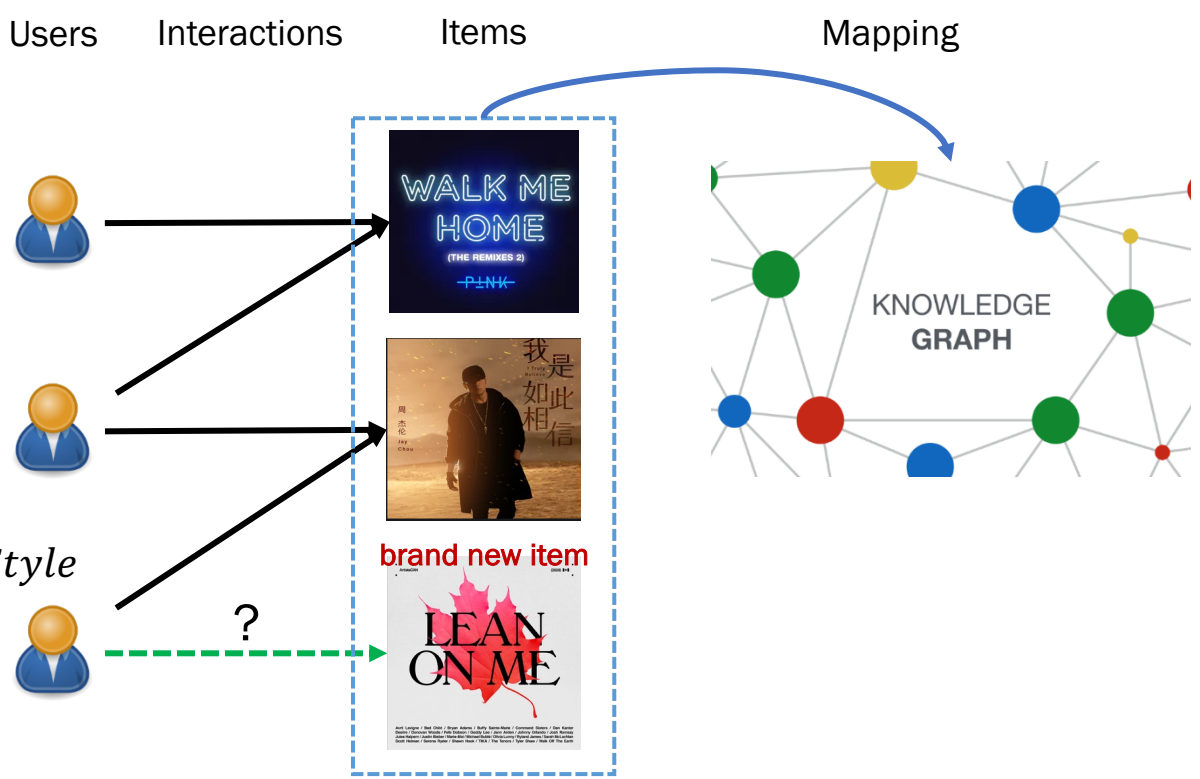
initialization: $Z_0 = \langle S_0, \emptyset, (\texttt{Desc}(e) \setminus S_0), f_0 \rangle$.



知识图谱增强推荐

Lyu et al. Rule-Guided Graph Neural Networks for Recommender Systems. ISWC

- 推荐中的冷启动问题
  - 知识图谱提供了物品间的大量关系
    - 假设存在物品到知识图谱的映射
  - 更好地推荐很少与用户互动的全新物品
- 本文思路：规则学习 + 图神经网络
  - 规则
    - $user \xrightarrow{interacts} Red \xrightarrow{singer} Taylor\ Swift \xleftarrow{singer} Style$
    - 捕获用户和物品间的显式长程语义
  - 图神经网络
    - 保留各种关联，提供更丰富的信息

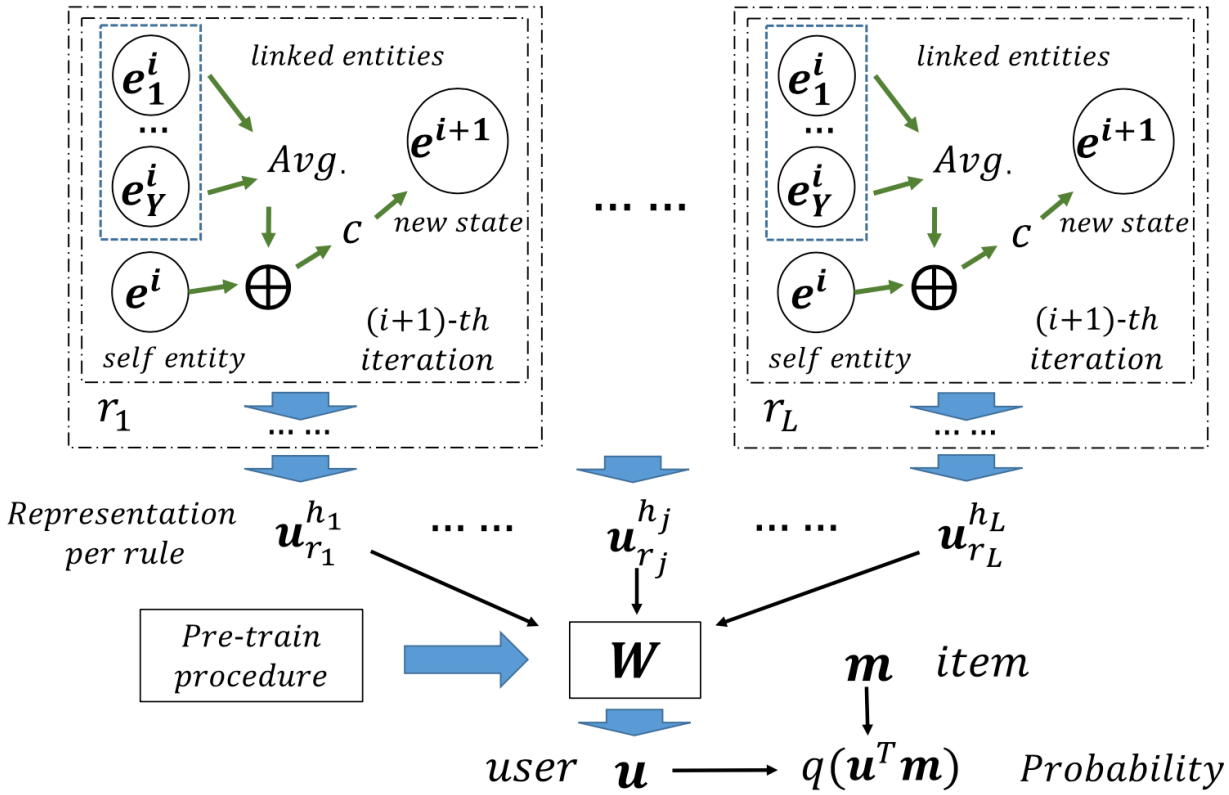


知识图谱增强推荐 (cont'd)

Lyu et al. Rule-Guided Graph Neural Networks for Recommender Systems. ISWC

- RGRec
  - 规则学习
  - 基于单条规则的用户表示
  - 多维表示的聚合
  - 规则权重的预训练
- 实验结果

Dianping-Food	AUC			F1		
	20%	40%	60%	20%	40%	60%
SVD	0.709	0.762	0.787	0.648	0.704	0.729
LibFM	0.812	0.814	0.809	0.761	0.766	0.766
LibFM+TransE	0.798	0.819	0.820	0.747	0.760	0.761
CKE	0.710	0.743	0.773	0.614	0.671	0.703
RKGE	0.703	0.811	0.847	0.628	0.719	0.766
KGCN	0.774	0.807	0.842	0.719	0.742	0.774
RGRec	0.882	0.884	0.884	0.808	0.809	0.809

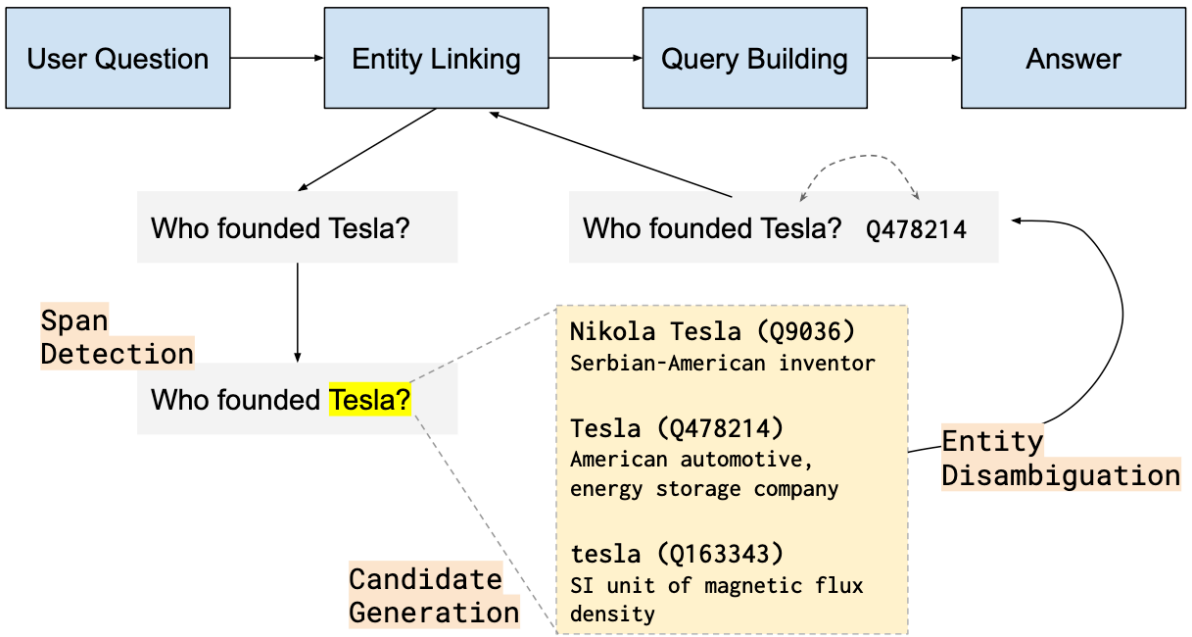


5. 知识图谱问答

实体链接

Banerjee et al. PNEL: Pointer Network based End-To-End Entity Linking over Knowledge Graphs. ISWC

- 基于知识图谱的问答



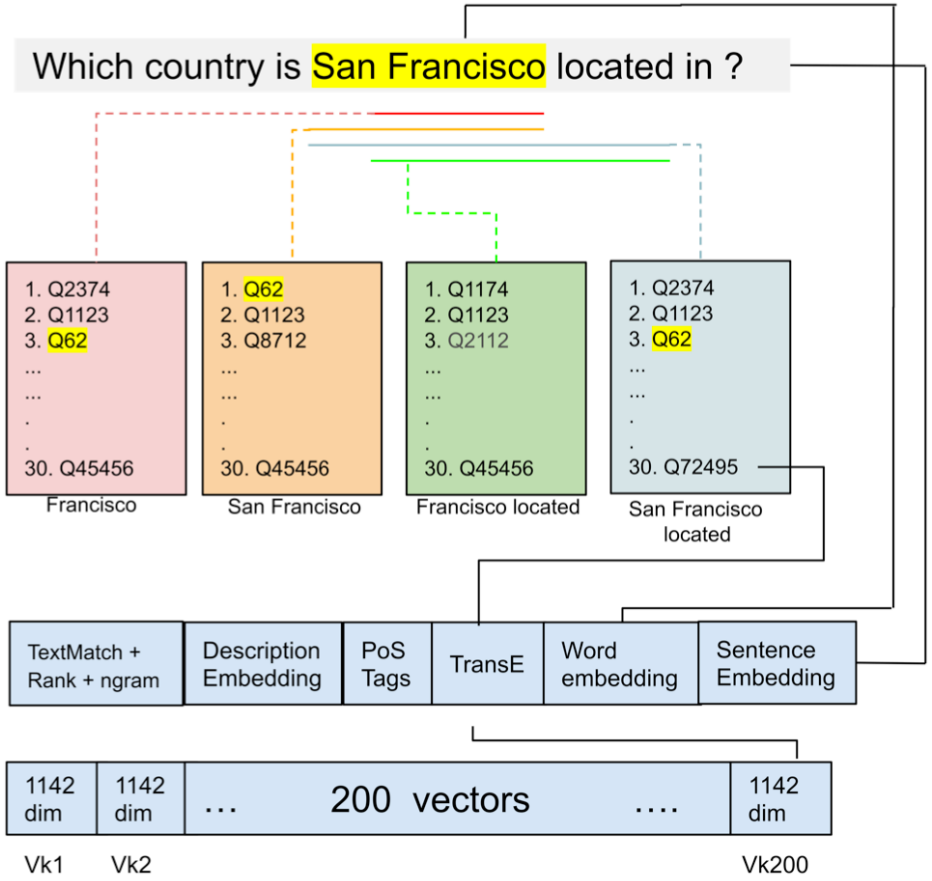
- Entity Linking = Span Detection + Entity Disambiguation
  - 管道模型：误差传递
  - 端到端模型：缺少 Span Detection，因而产生大量候选实体

实体链接 (cont'd)

Banerjee et al. PNEL: Pointer Network based End-To-End Entity Linking over Knowledge Graphs. ISWC

- PNEL
  - 主要思想：使用指针网络解决端到端实体链接
  - 具体做法：特征工程
    1. 4 *n*-grams + top-*L* matches
    2. 对每个候选，计算 9 个特征，拼接成 1142 维向量
- 实验结果
  - 第一个报告了 LC-QuAD 2.0 数据集上的结果

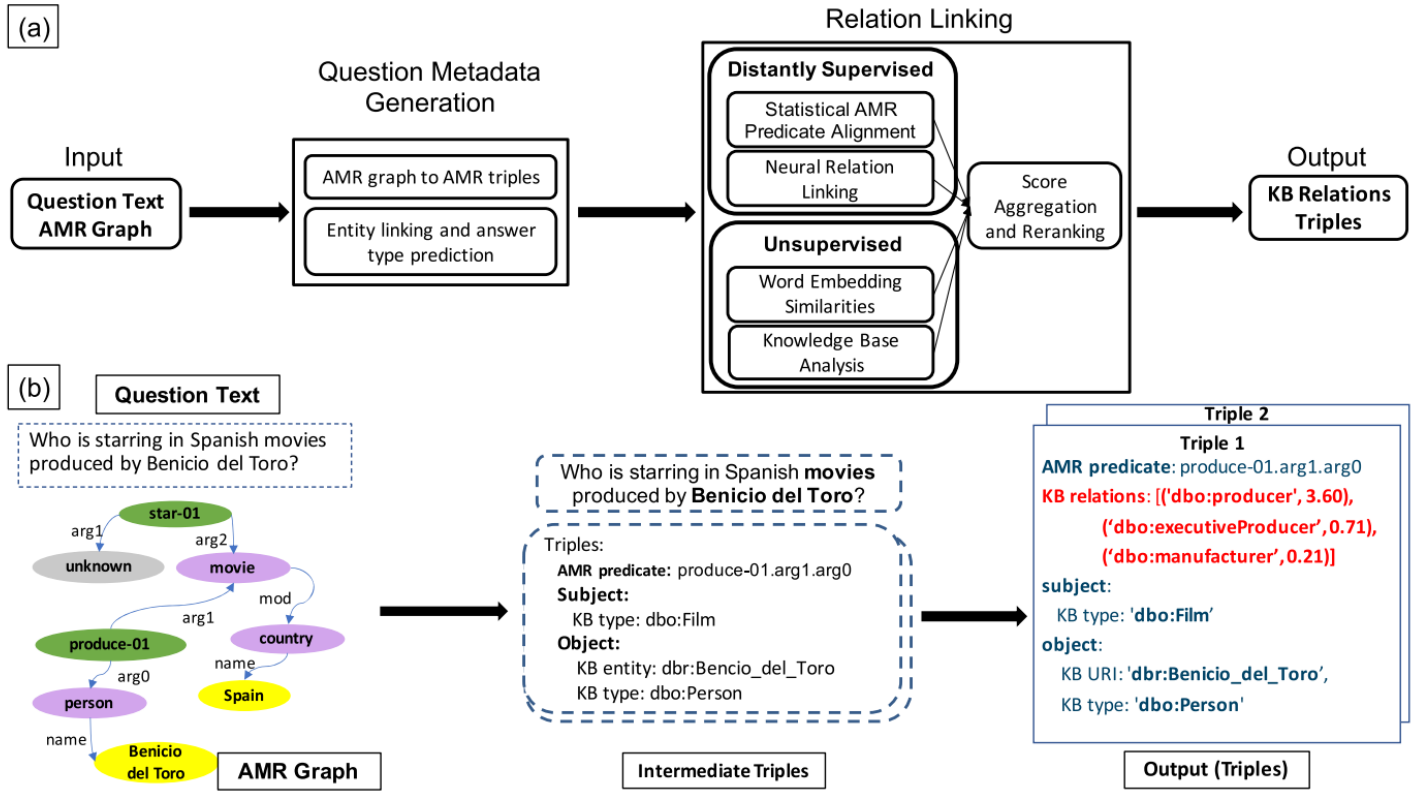
LC-QuAD 2.0	Precision	Recall	F1
VCG	0.516	0.432	0.470
OpenTapioca	0.237	0.411	0.301
Falcon 2.0	0.418	0.476	0.445
PNEL-L	0.688	0.516	0.589



关系链接

Mihindukulasooriya et al. Leveraging Semantic Parsing For Relation Linking Over Knowledge Bases. ISWC

- 关系链接
  - 挑战：自然语言的歧义性、训练数据的缺乏
- SLING
  - 主要思想：AMR、远程监督
  - 具体做法
    1. 输入问题文本和 AMR 图
    2. QMG 模块产生三种元数据
       - AMR 三元组、知识库实体和类型、答案类型预测
    3. RL 模块包括两种远程监督模型和两种无监督模型
       - 整合每种模型的得分并排序



6. 新数据集

新数据集

研究方向	名字	特点
实体对齐	SemTab [ESWC]	针对表格数据到知识图谱匹配任务的挑战，包含 Column-Type Annotation、Cell-Entity Annotation、Columns-Property Annotation 三个子任务以及四个基准数据集，分别有 64、11924、2161、817 张表格
本体推理	OWL2Bench [ISWC]	用于对推理机的三个方面进行基准测试：支持的 OWL 2 语言构造、Abox 的可伸缩性以及查询性能；其支持 All OWL 2 Profiles (EL/QL/RL/DL)；同时支持推理机和 SPARQL 查询引擎的测试，是当前支持**类型最全**的本体推理数据集
搜索与摘要	LOVBench [WWW]	针对本体术语排序任务构建的基准测试集，其依靠 LOV 平台上真实用户的反馈信息构建真值，包含超过 7000 条查询，数据规模**目前最大**
	ESBM [ESWC]	**目前最大**的用于评估实体摘要系统的基准数据集，其包含 175 个异构实体，利用 30 位专家构建了 2100 个通用的实体摘要真值
知识图谱问答	VQuAnDa [ESWC]	**第一个**为问题–查询对提供自然语言化答案的 KBQA 数据集

总结

- 调研了 2020 年语义网领域三大会议 (WWW、ISWC 和 ESWC) 的研究论文和数据集
- 从 (1) 知识图谱表示学习、(2) 知识抽取与图谱构建、(3) 实体对齐、(4) 搜索、摘要与推荐、(5) 知识图谱问答 这五个方面梳理了研究进展

- 趋势总结
  1. 长尾/新兴实体
     - 数据稀疏、冷启动问题
  2. 复杂场景
     - 多语言、多类型、多元关系
  3. 知识增强的机器学习

谢谢！

下载：https://github.com/nju-websoft/KGProgress2020fromSemWeb

致谢：王狄烽、刘扬