DataFunSummit

知识图谱

在线峰会

2021.03.27（周六） 09: 00~18: 0

2021

2021

小米在知识表示学习方向的探索与实践

吕荣荣 小米-算法工程师

目录

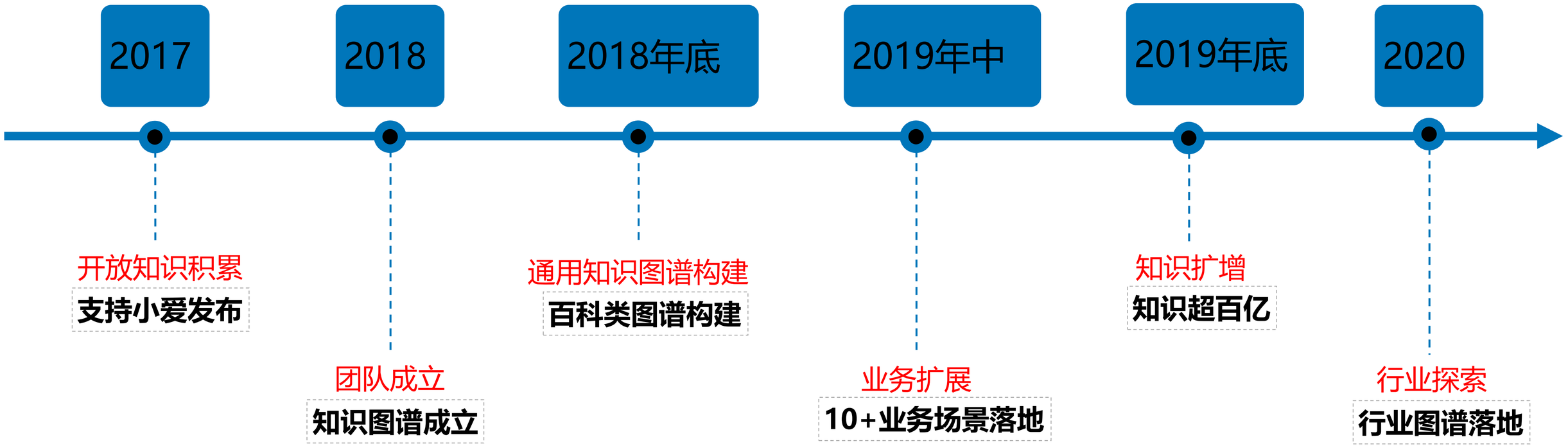
CONTENTS

01

业务介绍

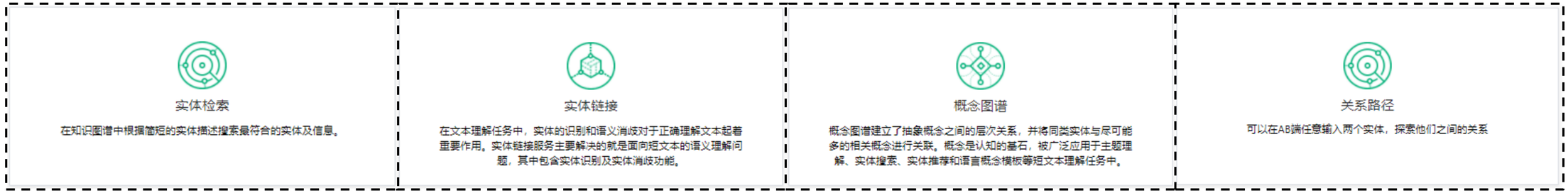
小米知识图谱的架构和业务

小米知识图谱的架构和业务



小米知识图谱的架构和业务

小米知识图谱团队，旨在研究知识图谱在开放领域和行业领域的构建和应用技术，把图谱推广到相关的业务场景上。



我们的用户

小米网　小米有品　小爱同学　智能问答　米答

小米信息流　智能客服　全局搜索　翻译机

小米知识图谱的架构和业务

知识图谱团队为小爱同学赋能

Query:
巩俐的籍贯

下午5:42

人物·视频

巩俐的籍贯

籍贯：山东济南

巩俐

简介:巩俐（GongLi），1965年12月31日出生于中国辽宁省沈阳市，祖籍山东省济南市，华语电影女演员，毕业于中央戏剧学院，联合国促进和平艺术家，联合国全球环境保护大使。1987年因主演电影《红高粱》成名，该片获得第38届柏林国际电影节金熊奖。1992年凭借电影《秋菊打官司》获得第4...

籍　　贯: 山东济南
职　　业: 演员
出生日期: 1965年12月31日
个人信息: 169cm/56kg/摩羯座/A型血

本次回答满意吗?　赞　踩

搜索　视频　图片

人物关系


让·米歇尔·雅尔
丈夫


赵英
母亲


巩固
哥哥


黄和祥
前夫


周星驰
搭档


刘德华
搭档


张国荣
搭档


张艾嘉
搭档


张艺谋
搭档

相关资讯　进入

巩俐,别再整那些烂活儿了

要说华人女星,绕不开一个人,巩俐。江湖地位不用多说,"巩皇"这名头不是白叫的。但近年我们...

离开张艺谋之后,为何章子怡强过巩俐?"天...

说到最强"谋女郎"的比拼,一定是在巩俐和章子怡之间产生,这个因为和张艺谋合作而诞生的女演员称号,是张艺谋整个职业生涯的重要标志。当年...

巩俐领衔顶配阵容 《夺冠》提档9月25日

影片由陈可辛执导,张翼编剧,巩俐、黄渤、吴刚、彭昱畅、白浪、中国女子排球队领衔主演,李现特别出演。 影片《夺冠》原定1月24日上映,1月19...

为什么张艺谋不娶巩俐,而娶了小他31岁的...

阅读全文约需8分钟为什么张艺谋不娶巩俐,而娶了小他31岁的陈婷?文/晏凌羊 来源/晏凌羊每次看到...

上映27天拿下23天日冠,八佰之后,巩俐这部...

而《夺冠》这部由巩俐主演的电影,虽然经历了改

主要作品


片名: 霸王别姬
首播: 1993-01-01(中国香港) /
评分: 9.6
主演: 张国荣、张丰毅、巩俐、葛优、英达

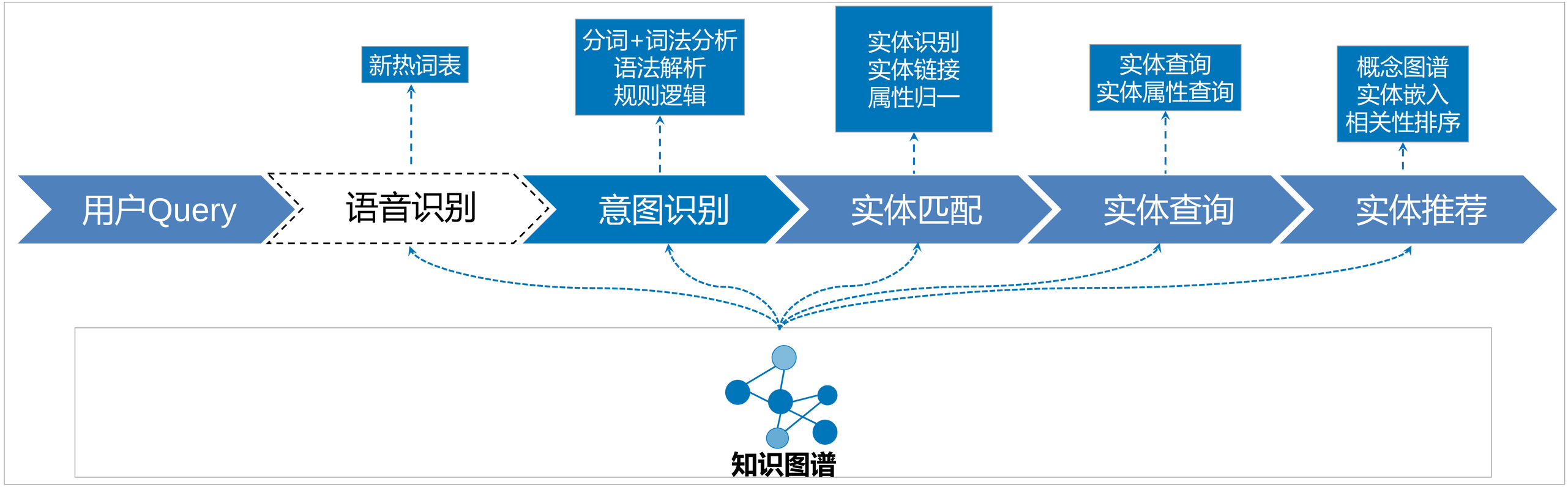

片名: 活着
首播: 1994-05-17(戛纳电影节) /
评分: 9.2
主演: 葛优、巩俐、姜武、牛犇、郭涛


片名: 唐伯虎点秋香
首播: 1993-07-01(香港)
评分: 8.5
主演: 周星驰、巩俐、陈百祥、郑佩佩、朱咪咪

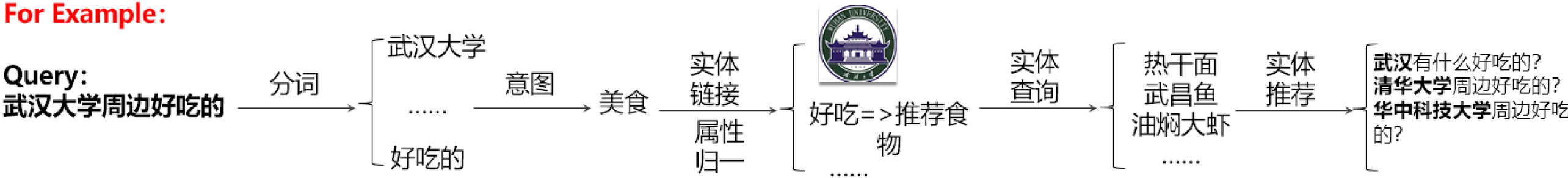

片名: 艺伎回忆录
首播: 2005-12-10(日本)

小米知识图谱的架构和业务

知识图谱团队为小爱同学赋能



For Example:

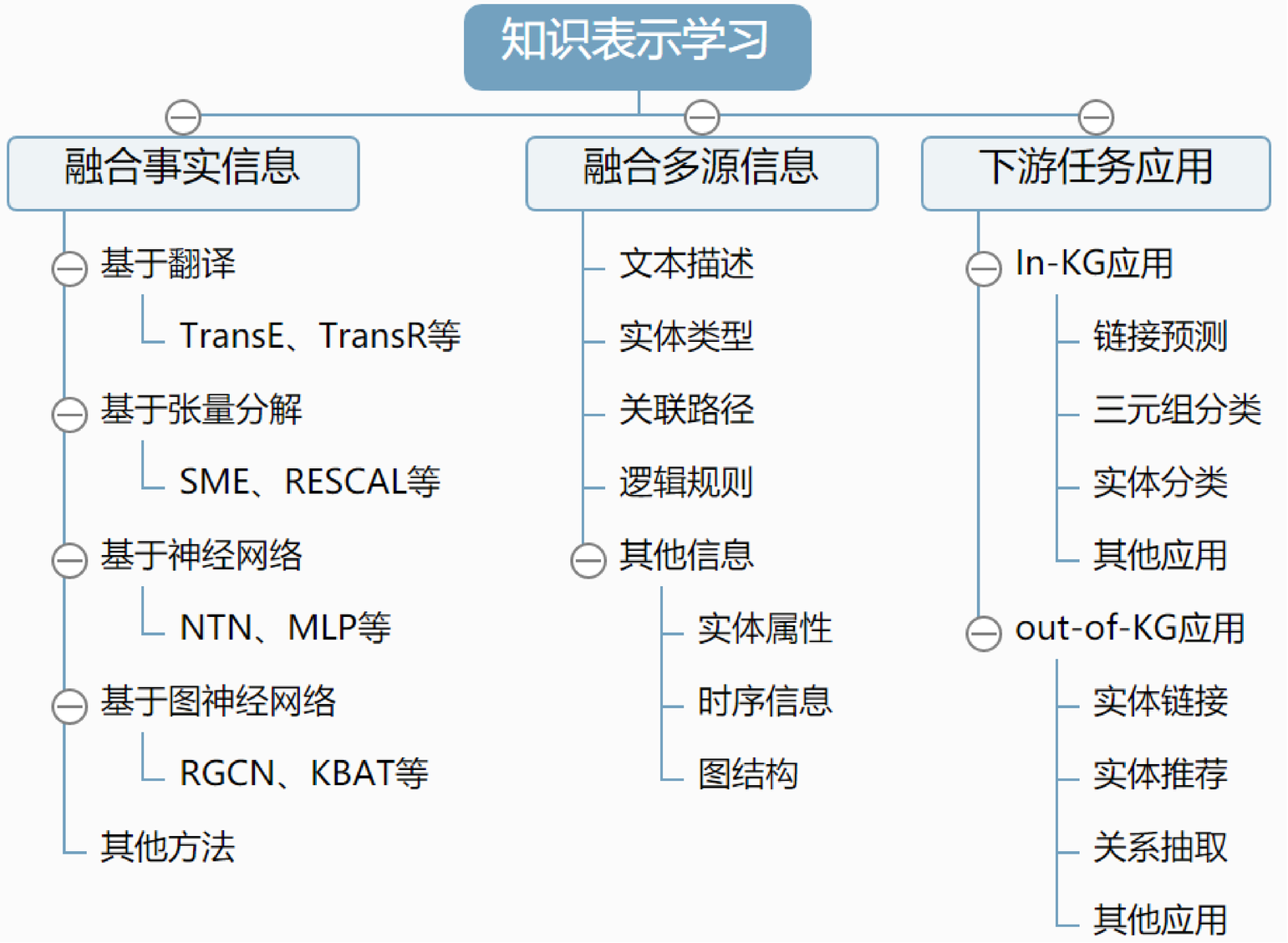


02

算法介绍

融合多源信息的知识表示学习方法

融合文本和知识图谱的知识表示学习方法



问题：
仅利用知识图谱的三元结构信息进行表示学习，数据稀疏问题严重，尚有大量与知识有关的其他信息没有得到有效利用。

解决方法：

1. 利用知识库中的其他信息，如实体和关系的描述信息等。
2. 利用知识库外的海量信息，如互联网文本等包含大量与知识库实体和关系有关的信息。

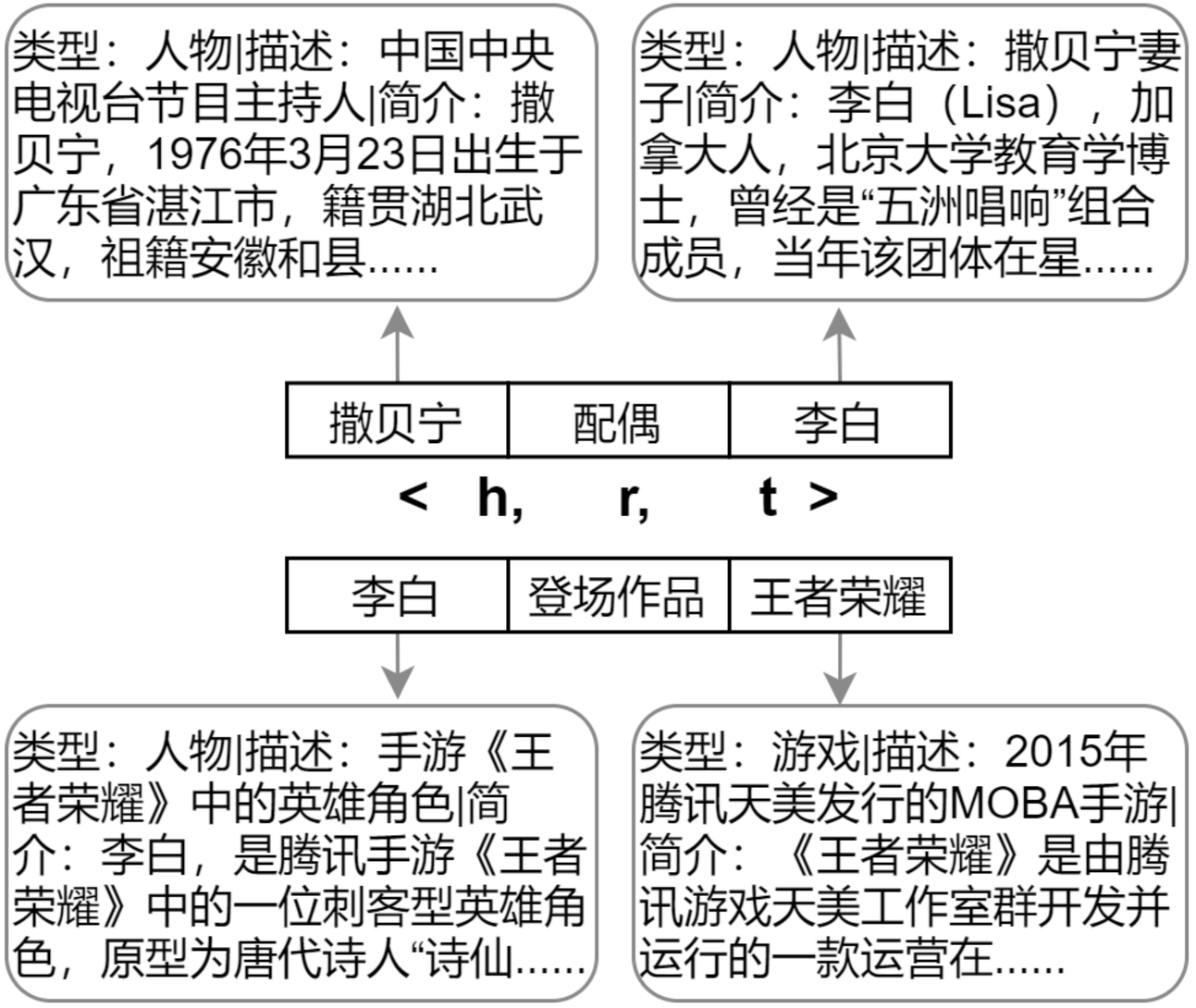
融合文本和知识图谱的知识表示学习方法

融入文本描述的优势：

1. 可以发掘实体间的语义相关性，精确的语义表述能够提升三元组的可区分性。
2. 可以解决zero-shot问题

实体的描述文本：

将所有三元组的“属性-属性值”或“关系-实体提及”都拼成一个字符串，当作该实体的文本描述。由于 type 字段，义项描述和摘要字段的信息更重要，描述文本中都按照 type、义项描述、摘要和其他三元组的顺序进行拼接。



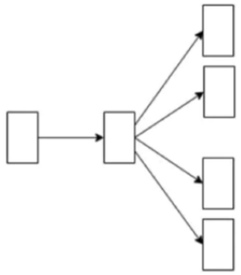
融合文本和知识图谱的知识表示学习方法

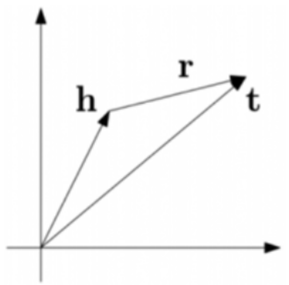
Text Model: Skip-Gram

$$z(w,v) = b - 0.5 \cdot \|\mathbf{w} - \mathbf{v}\|_2^2$$

$$\Pr(w|v) = \frac{\exp\{z(w,v)\}}{\sum_{\tilde{w}\in\mathcal{V}} \exp\{z(\tilde{w},v)\}}$$

$$\mathcal{L}_T = -\sum_{(w,v)} \log \Pr(w|v)$$





Knowledge Model: TransE

$$z(h,r,t) = b - 0.5 \cdot \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2$$

$$\Pr(h|r,t) = \frac{\exp\{z(h,r,t)\}}{\sum_{\tilde{h}\in\mathcal{I}} \exp\{z(\tilde{h},r,t)\}}$$

$$\mathcal{L}_K = -\sum_{(h,r,t)} \big[\log \Pr(h|r,t) + \log \Pr(t|h,r) + \log \Pr(r|h,t)\big]$$

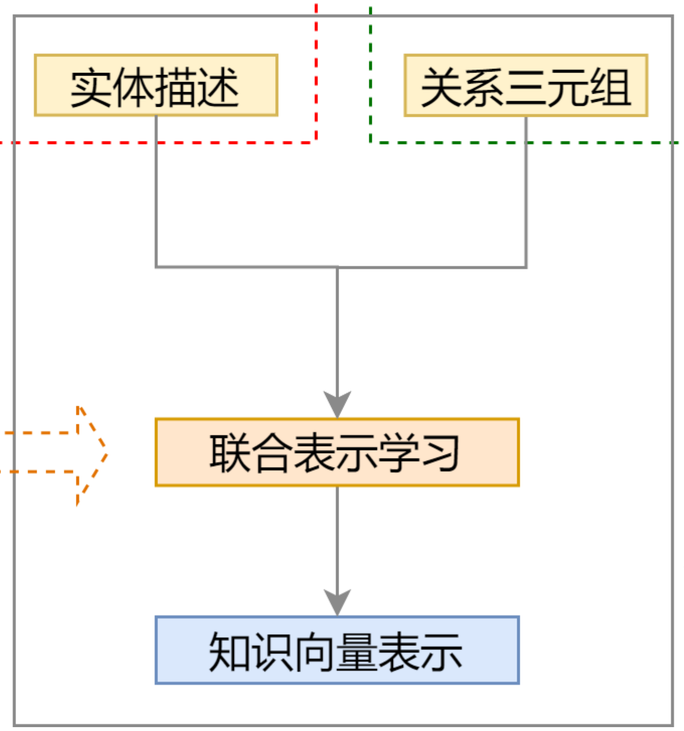
$$z(e,w) = b - 0.5 \cdot \|\mathbf{e} - \mathbf{w}\|_2^2$$

$$\Pr(w|e) = \frac{\exp\{z(e,w)\}}{\sum_{\tilde{w}\in\mathcal{V}} \exp\{z(e,\tilde{w})\}}$$

$$\Pr(e|w) = \frac{\exp\{z(e,w)\}}{\sum_{\tilde{e}\in\mathcal{E}} \exp\{z(\tilde{e},w)\}}$$

$$\mathcal{L}_A = -\sum_{e\in\mathcal{E}} \sum_{w\in D_e} [\log \Pr(w|e) + \log \Pr(e|w)]$$

$$\mathcal{L}(\{\mathbf{e}_i\},\{\mathbf{r}_j\},\{\mathbf{w}_l\}) = \mathcal{L}_K + \mathcal{L}_T + \mathcal{L}_A$$

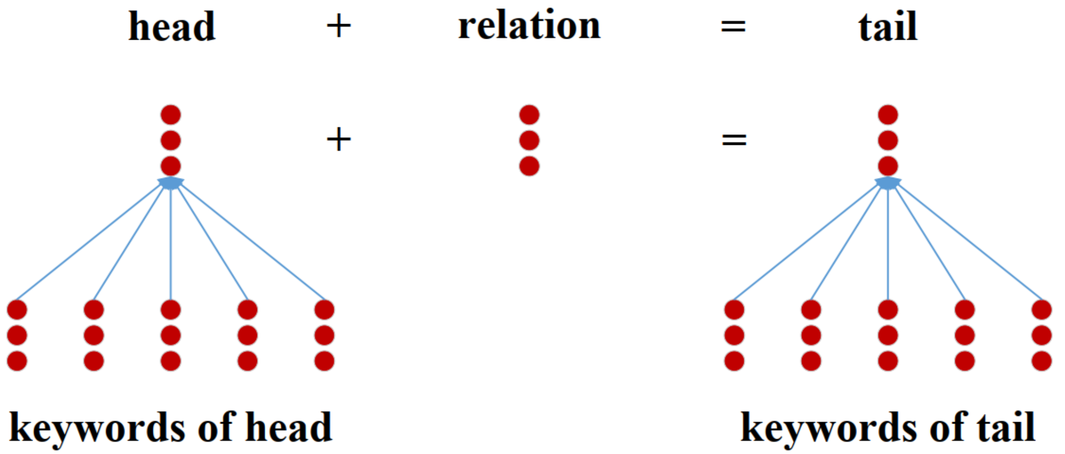


1. 第一部分是文本嵌入模型，采用Skip-Gram模型，距离公式采用的是两个单词之间的欧氏距离。
2. 第二部分是知识嵌入模型，采用transE模型。
3. 第三部分是对齐模型，利用文本描述对齐，确保实体，关系能够和文本中的单词在同一个语义空间中。

论文：《Aligning Knowledge and Text Embeddings by Entity Descriptions》

融合文本和知识图谱的知识表示学习方法

DKRL(CBOW): The CBOW Encoder



$$E_S = || h_s + r + h_s ||$$

$$E_{DD} = || h_d + r - t_d ||$$

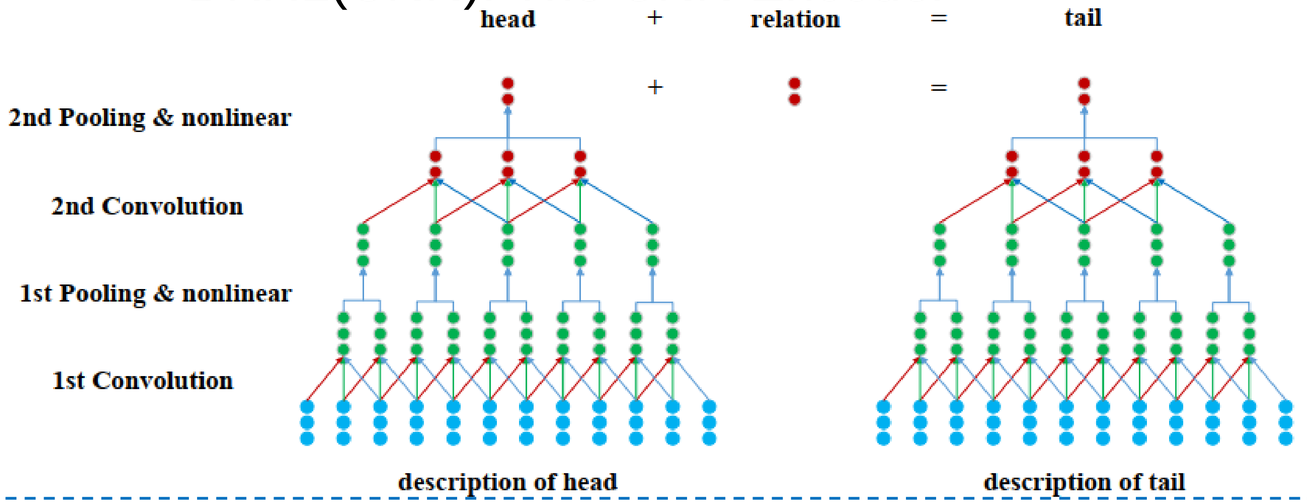
$$E_{DS} = || h_d + r - t_s ||$$

$$E_{SD} = || h_s + r - t_d ||$$

$$E = E_S + E_D$$

$$= E_S + E_{DD} + E_{DS} + E_{SD}$$

DKRL(CNN): The CNN Encoder



$$L = \sum_{(h,r,t)\in T} \sum_{(h',r',t')\in T'} \max(\gamma + d(h+r,t) - d(h'+r',t'), 0),$$

$$T' = \{(h',r,t)|h' \in E\} \cup \{(h,r,t')|t' \in E\} \cup \{(h,r',t)|r' \in R\}$$

论文：《Representation Learning of Knowledge Graphs with Entity Descriptions》

融合文本和知识图谱的知识表示学习方法

需求目标：

1. 得到实体向量
2. 得到词向量
3. 词向量和实体向量之间可以计算相似度
4. 实体向量和实体向量之前可以计算相似度

Table 1: Mean Rank and HITS@10 of Knowledge Graph Completion (For Predicting Entity) on FB15K

FB15K	Mean Rank		HITS@10	
TransE	210	119	48.5	66.1
TransH	212	87	45.7	64.4
Jointly	167 [1]	39 [1]	51.7 [1]	77.3 [1]
DKRL(BOW)	200	113	44.3	57.6
DKRL(ALL)	181	91	49.6	67.4

词向量与实体向量的相似度

```
input word: 王者荣耀
entity and score:
李白 手游《王者荣耀》中的英雄角色 0.6871914871010902
李白 漫画《尸兄》中的角色 0.4868442647918333
李白 QQ游戏《英雄杀》卡牌 0.47989461648584353
李白 南宋徐钧所写诗歌 0.46257176254889987
李白 李荣浩演唱歌曲 0.4526289976719955
李白 撒贝宁妻子 0.4438043542986664
李白 唐代诗人 0.35362615078817616
李白 北京人艺经典话剧 0.34886872301725935
李白 中共党员，上海地下党联络员 0.2934343603235514
```

```
input word: 将进酒
entity and score:
李白 唐代诗人 0.7954981104666751
李白 南宋徐钧所写诗歌 0.6594177266802413
李白 李荣浩演唱歌曲 0.5860849811810142
李白 手游《王者荣耀》中的英雄角色 0.5421541413927163
李白 北京人艺经典话剧 0.5303117604279298
李白 QQ游戏《英雄杀》卡牌 0.4655070254651763
李白 漫画《尸兄》中的角色 0.43050382437348667
李白 撒贝宁妻子 0.4236790438669058
李白 中共党员，上海地下党联络员 0.2748187908448527
```

实体向量与实体向量的相似度

```
input entity: 小米 禾本科狗尾草属一年生草本
entity and score:
苹果 蔷薇科苹果属植物 0.8354706297404012
苹果 伊朗1998年莎米拉·玛克玛尔巴夫执导电影 0.6325797571
苹果 韩国2008年康理贯执导电影 0.5702162436015609
苹果 安与骑兵演唱歌曲 0.5607941649198853
苹果公司 0.5381222371491942
苹果 网游《天堂梦》中人物 0.5211885607135783
苹果 2007年李玉执导电影 0.47746202503045454
苹果 动漫《男子高中生的日常》中角色 0.47035311617841247
```

```
input entity: 北京小米科技有限责任公司
entity and score:
苹果公司 0.79185107258104
苹果 安与骑兵演唱歌曲 0.6309651542813589
苹果 伊朗1998年莎米拉·玛克玛尔巴夫执导电影 0.594317807
苹果 韩国2008年康理贯执导电影 0.5558840673715467
苹果 网游《天堂梦》中人物 0.533240821748632
苹果 蔷薇科苹果属植物 0.4861939169898948
苹果 2007年李玉执导电影 0.45654442486800906
苹果 动漫《男子高中生的日常》中角色 0.3878173304782322
```

03

算法应用

知识表示学习在小米的应用

a) 实体链接

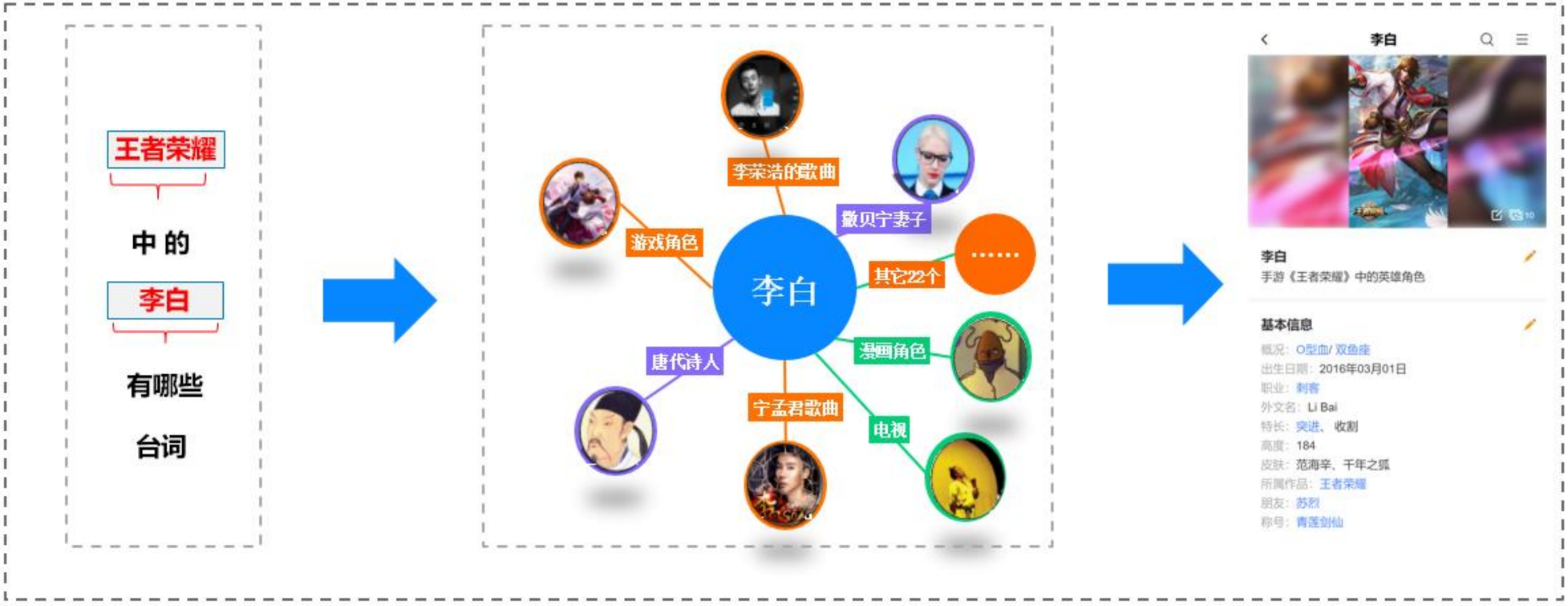
b) 实体推荐

c) 知识补全

实体链接中的应用

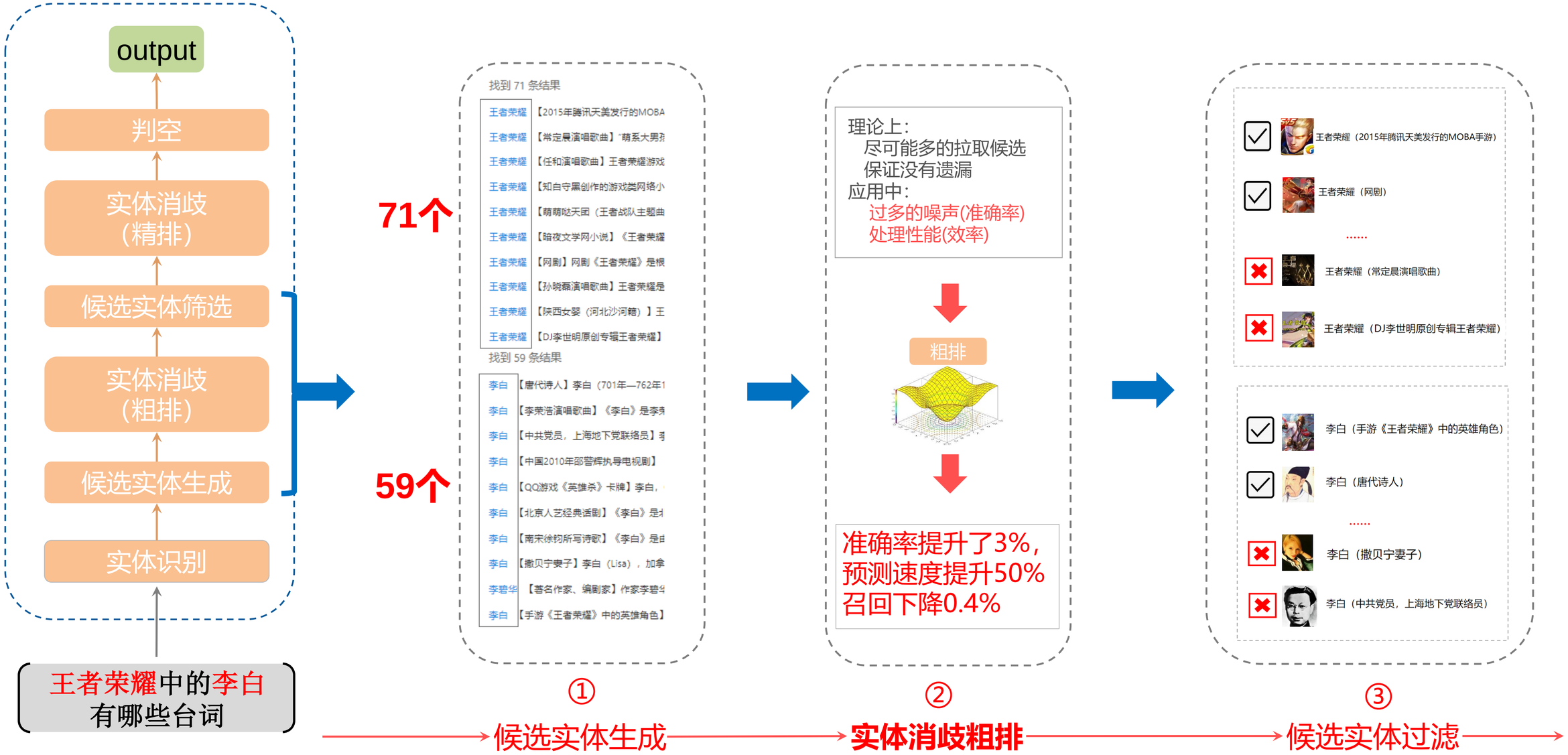
实体链接，就是把文本中的实体指称链接到知识图谱对应的实体上的任务。

Knowledge Graph（知识图谱）： 一种语义网络，旨在描述客观世界的概念实体及其之间的关系。
Mention（提及、实体指称）： 自然文本中表达实体的语言片段。
Entity（实体）： 实体是知识库的基本单元，也是文本中承载信息的重要语言单位，由多个三元组组成

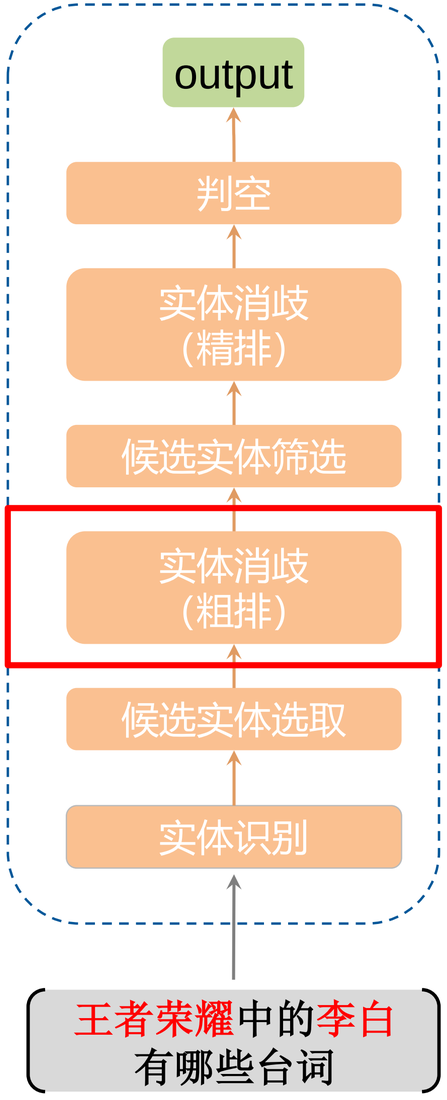


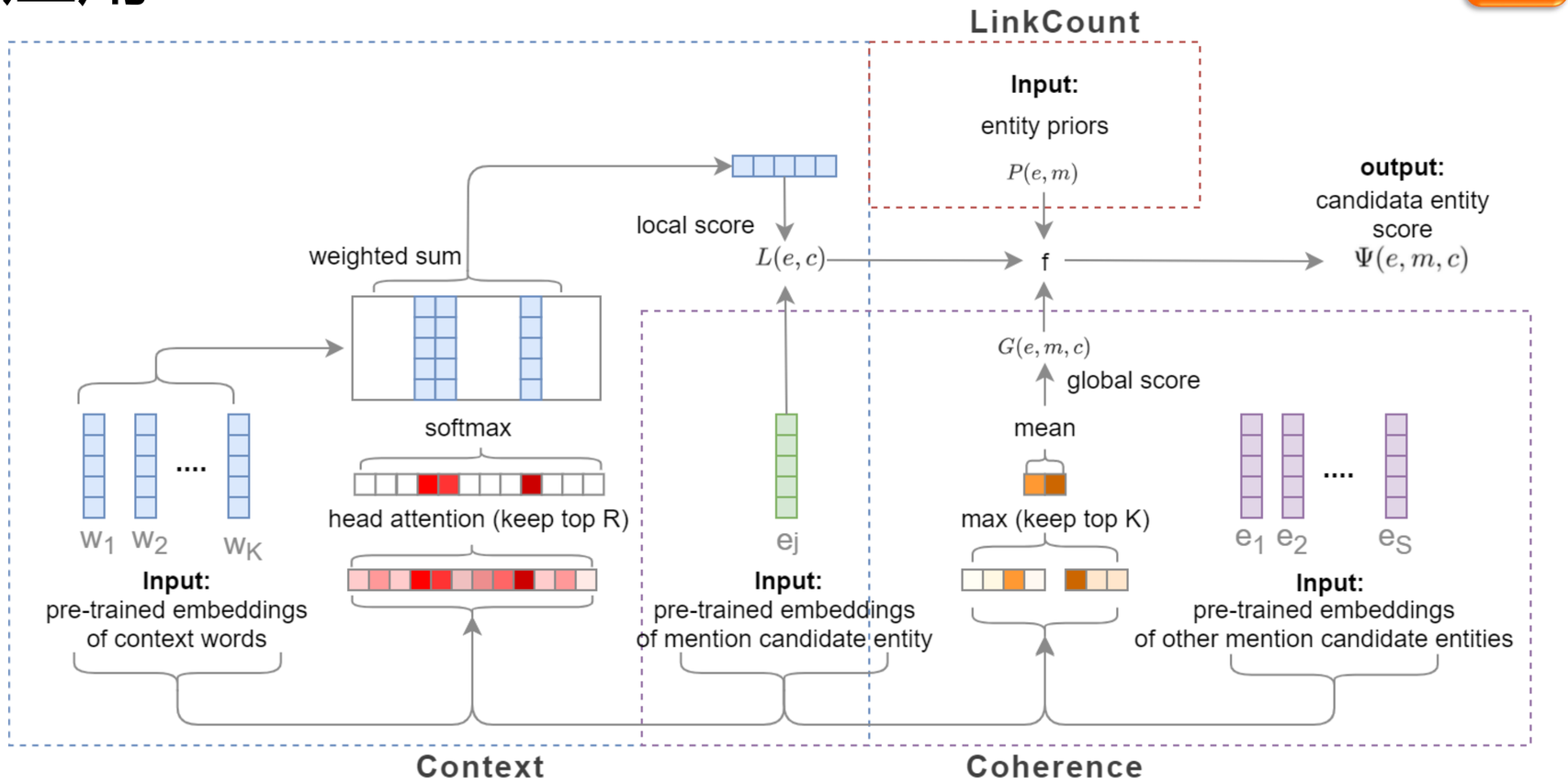
难点：**Mention Variations**：同一entity有不同的mention。（<李白>：青莲居士、李太白）
Entity Ambiguity：同一mention对应不同的entity。（“李白”：王者荣耀中李白的技能；李白和杜甫并称为什么？）

实体链接中的应用


output
判空
实体消歧
（精排）
候选实体筛选
实体消歧
（粗排）
候选实体生成
实体识别
王者荣耀中的李白
有哪些台词
71个
59个
找到 71 条结果
王者荣耀 【2015年腾讯天美发行的MOBA
王者荣耀 【常定晨演唱歌曲】"萌系大男孩
王者荣耀 【任和演唱歌曲】王者荣耀游戏
王者荣耀 【知白守黑创作的游戏类网络小
王者荣耀 【萌萌哒天团（王者战队主题曲
王者荣耀 【暗夜文学网小说】《王者荣耀
王者荣耀 【网剧】网剧《王者荣耀》是根
王者荣耀 【孙晓磊演唱歌曲】王者荣耀是
王者荣耀 【陕西女婿（河北沙河籍）】王
王者荣耀 【DJ李世明原创专辑王者荣耀】
找到 59 条结果
李白 【唐代诗人】李白（701年—762年1
李白 【李荣浩演唱歌曲】《李白》是李荣
李白 【中共党员，上海地下党联络员】李
李白 【中国2010年部警辉执导电视剧】
李白 【QQ游戏《英雄杀》卡牌】李白，
李白 【北京人艺经典话剧】《李白》是
李白 【南宋徐钧所写诗歌】《李白》是
李白 【撒贝宁妻子】李白（Lisa），加拿
李碧华 【著名作家、编剧家】作家李碧华
李白 【手游《王者荣耀》中的英雄角色】
理论上：
尽可能多的拉取候选
保证没有遗漏
应用中：
过多的噪声(准确率)
处理性能(效率)
粗排
准确率提升了3%，
预测速度提升50%
召回下降0.4%
王者荣耀（2015年腾讯天美发行的MOBA手游）
王者荣耀（网剧）
.....
王者荣耀（常定晨演唱歌曲）
王者荣耀（DJ李世明原创专辑王者荣耀）
李白（手游《王者荣耀》中的英雄角色）
李白（唐代诗人）
.....
李白（撒贝宁妻子）
李白（中共党员，上海地下党联络员）
①
候选实体生成
②
实体消歧粗排
③
候选实体过滤

实体链接中的应用





$$u(w) = e^T X_w$$

$$\bar{c} = \{w \in c | u(w) \in \text{topR}(\mathbf{u})\}$$

$$\beta(w) = \begin{cases} \frac{\exp[u(w)]}{\sum_{v \in \bar{c}} \exp[u(v)]}, & \text{if } w \in \bar{c} \\ 0 & \text{otherwise} \end{cases}$$

$$L(e,c) = \sum_{w \in \bar{c}} \beta(w) e^T X_w$$

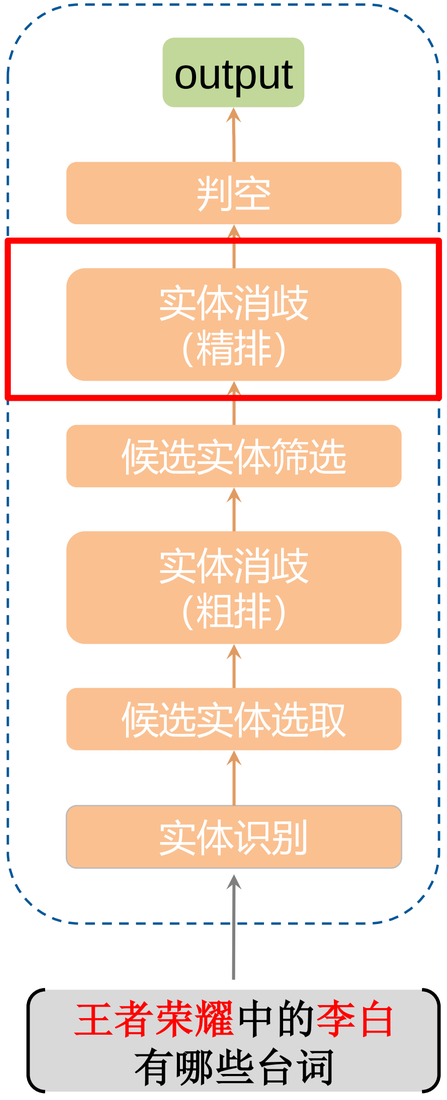
$$M_K = \{\overline{m} \mid \overline{m} \in c \wedge \overline{m} \neq m\}$$

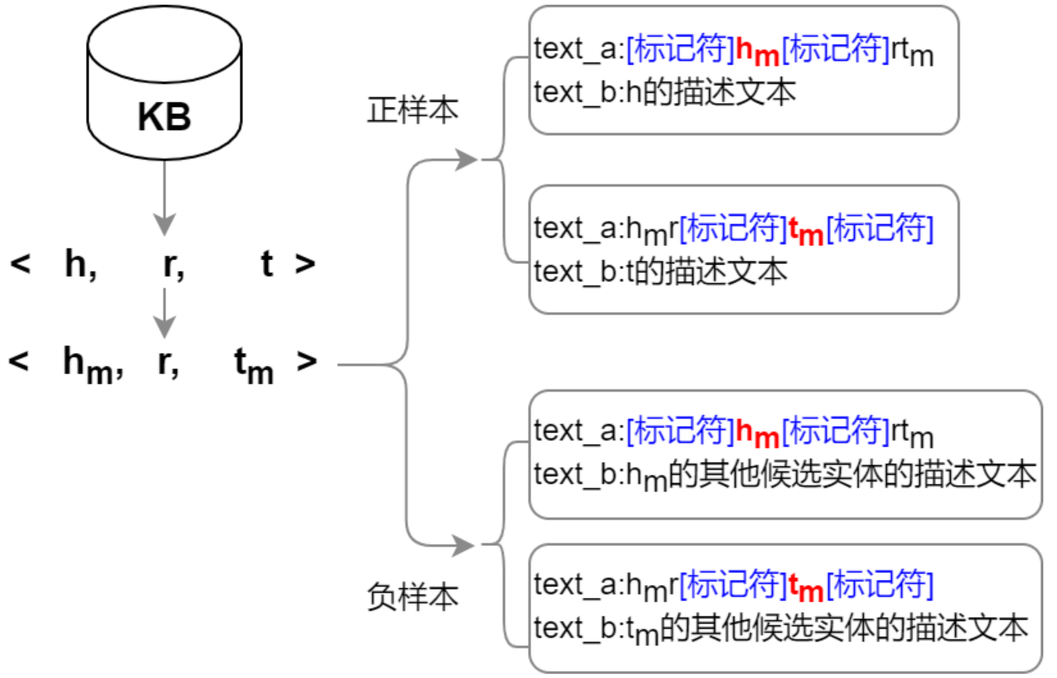
$$G(e,m,c) = \frac{\sum_{\overline{m} \in M_K} max(e^T X_e^{\overline{m}})}{K}$$

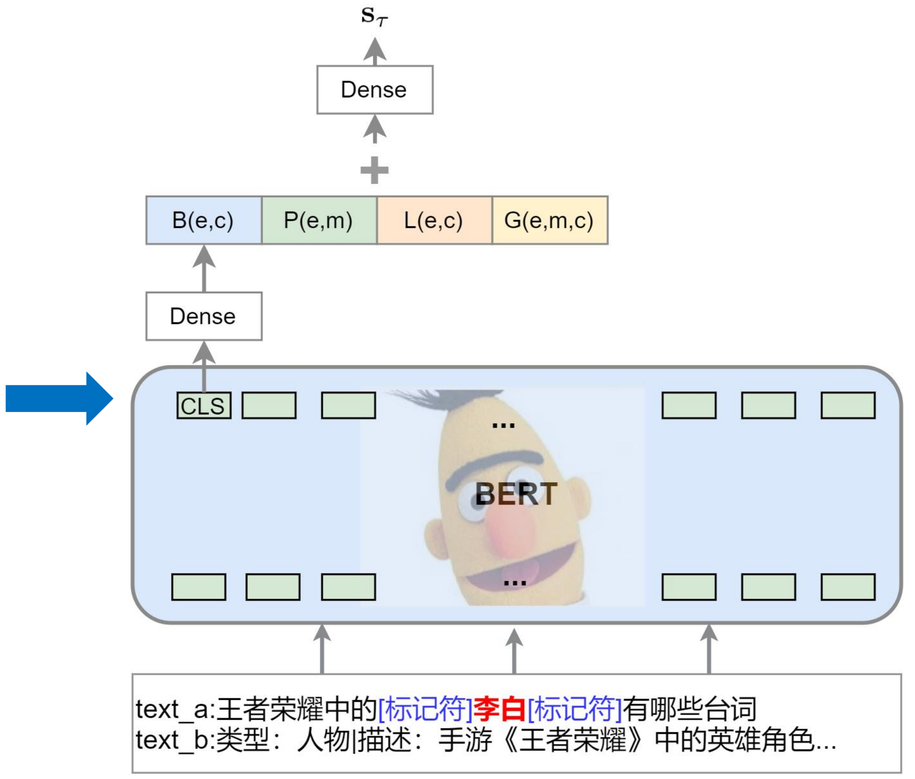
$$P(e,m) = \frac{count_m(e)}{\sum_{e \in X_e} count_m(e)}$$

$$\Psi(e,m,c) = f(P(e,m), L(e,c), G(e,m,c))$$

实体链接中的应用







二次训练：

第一次基于KB中的关系三元组训练

第二次基于标注的线上query训练

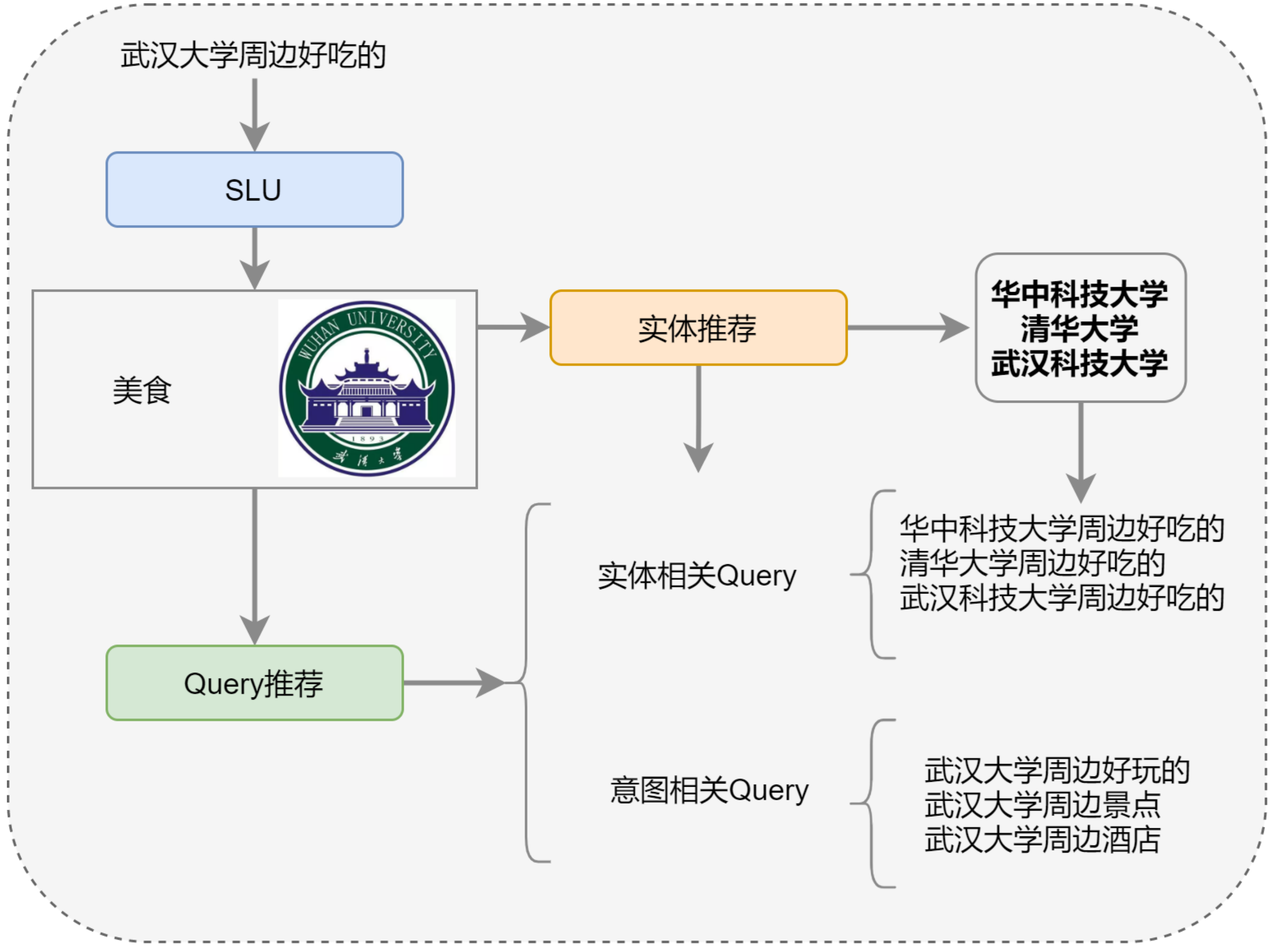
$$\mathbf{s}_\tau = f(h, r, t) = \text{sigmoid}(CW^T)$$

$$\mathcal{L} = -\sum_{\tau \in \mathbb{D}^+ \cup \mathbb{D}^-} (y_\tau \log(s_{\tau 0}) + (1 - y_\tau) \log(s_{\tau 1}))$$

实体的描述文本：将所有三元组的“属性-属性值”或“关系-实体提及”都拼成一个字符串，当作该实体的文本描述。由于 type 字段，义项描述和摘要字段的信息更重要，描述文本中都按照 type、义项描述、摘要和其他三元组的顺序进行拼接。

实体推荐中的应用

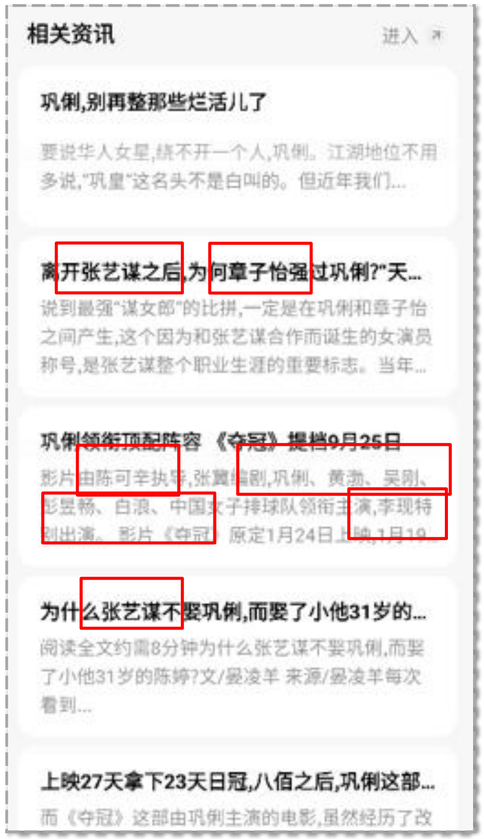
实体推荐，就是根据给定的实体，推荐一系列相关实体。

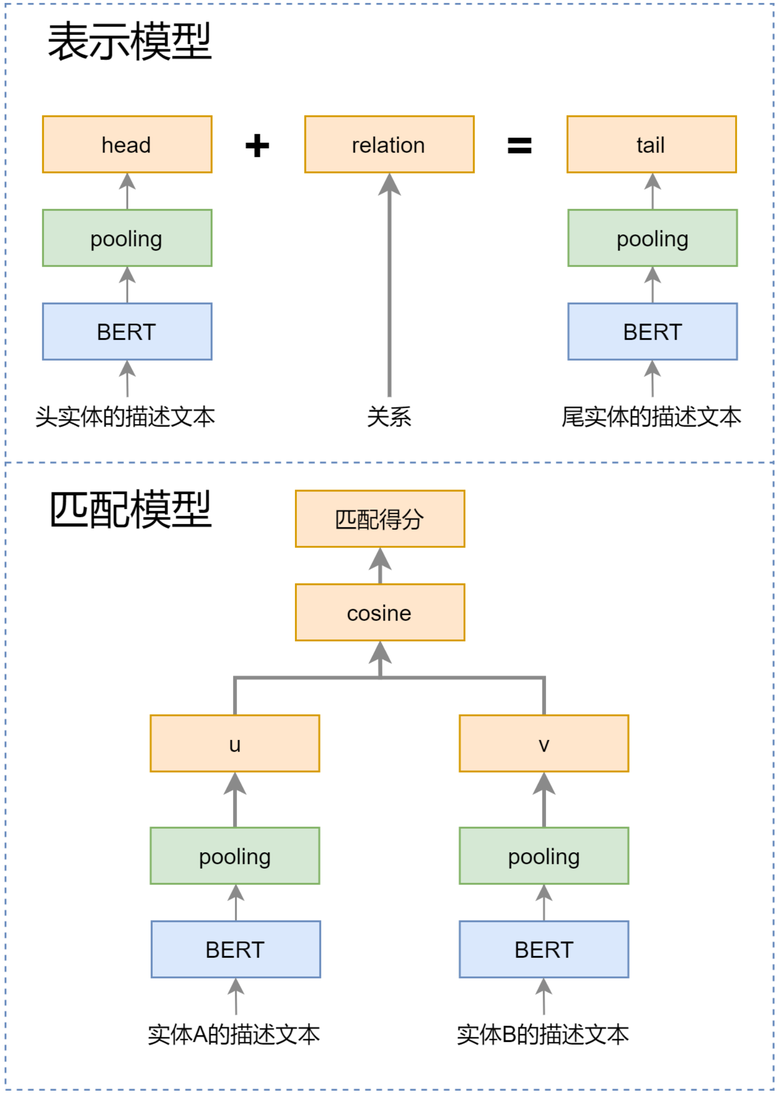


实体推荐中的应用

实体推荐冷启动问题：如何寻找实体的相关实体？

方案：通过百科页面、关系三元组、经过实体链接处理的新闻中的共现实体对，经过类别过滤，作为正样本进行训练。





知识补全中的应用

知识图谱补全（Knowledge Graph Completion，KGC）目前主要被抽象成一个预测问题，即预测出三元组中缺失的部分。

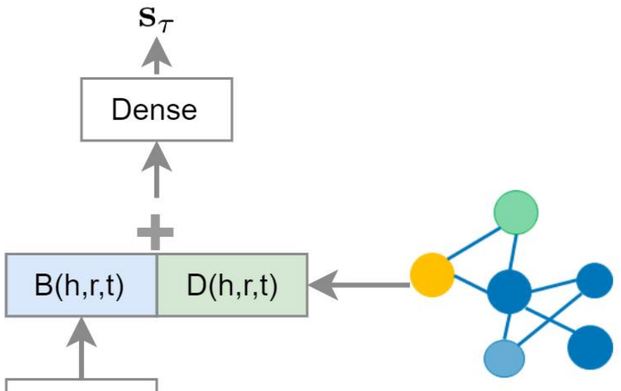
数据清洗

属性归一

目标任务：
补充缺失的信息
输入：
头实体、关系以及尾实体的mention
输出：
知识库中对应的尾实体
方案：

1. 利用scheme限定尾实体的类别
2. 通过t_m筛选尾实体的候选实体数量
3. 基于三元组分类模型判断是否存在正确三元组
4. 排序后选择top1且预测正确的三元组

知识补全中的应用



text_a:类型：人物|描述：中国中央电视台节目主持人|简介：撒贝宁，1976年3月23日出生于广东省湛江市，籍贯湖北武汉，祖籍安徽和县....
text_b:配偶
text_c:类型：人物|描述：撒贝宁妻子|简介：李白（Lisa），加拿大人，北京大学教育学博士，曾经是“五洲唱响”组合成员，当年该团体在星.....

类型：人物|描述：中国中央电视台节目主持人|简介：撒贝宁，1976年3月23日出生于广东省湛江市，籍贯湖北武汉，祖籍安徽和县.....

撒贝宁 | 配偶 | **李白**

< h, r, t_m >

类型：人物|描述：撒贝宁妻子|简介：李白（Lisa），加拿大人，北京大学教育学博士，曾经是“五洲唱响”组合成员，当年该团体在星.....

类型：人物|描述：唐代诗人|简介：李白（701年 - 762年），字太白，号青莲居士，又号“谪仙人”，唐代伟大的浪漫主义诗人.....

- - - -

类型：人物|描述：手游《王者荣耀》中的英雄角色|简介：李白，是腾讯手游《王者荣耀》中的一位刺客型英雄角色，原型为唐代.....

1	0.9	✓
0	0.2	
0	-	
0	0.1	

$$z(h,r,t) = b - 0.5 \cdot \|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_2^2, b = 7$$

$$D(h,r,t) = \frac{exp\{z(h,r,t)\}}{\sum_{\bar{t}\in E_m} exp\{z(h,r,\bar{t})\}}$$

$$\mathcal{L} = -\sum_{\tau\in\mathbb{D}^+\cup\mathbb{D}^-} (y_\tau \log(s_{\tau 0}) + (1-y_\tau)\log(s_{\tau 1}))$$

论文：《KG-BERT: BERT for Knowledge Graph Completion》

04

总结

总结与展望

总结与展望

1. 简单介绍了知识表示学习在实体链接、实体推荐和知识补全中的应用
2. 和word2vec一样，通过知识表示学习得到的实体向量可以应用到很多场景中
3. 工业界实用最重要，很多场景下，对模型的复杂性很敏感，选择模型需要综合考量
4. 知识表示学习的探索之路还有很长，同志们加油！！！

参考文献

1. Wang Z, Zhang J, Feng J, et al. Knowledge graph and text jointly embedding[C] //Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1591-1601.
2. Zhong H, Zhang J, Wang Z, et al. Aligning knowledge and text embeddings by entity descriptions[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 267-272.
3. Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2016, 30(1).
4. Xiao H, Huang M, Meng L, et al. SSP: semantic space projection for knowledge graph embedding with text descriptions[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2017, 31(1).
5. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks[J]. arXiv preprint arXiv:1908.10084, 2019.
6. Yao L, Mao C, Luo Y. KG-BERT: BERT for knowledge graph completion[J]. arXiv preprint arXiv:1909.03193, 2019.
7. 刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 247.

THANKS!

今天的分享就到这里...

Ending