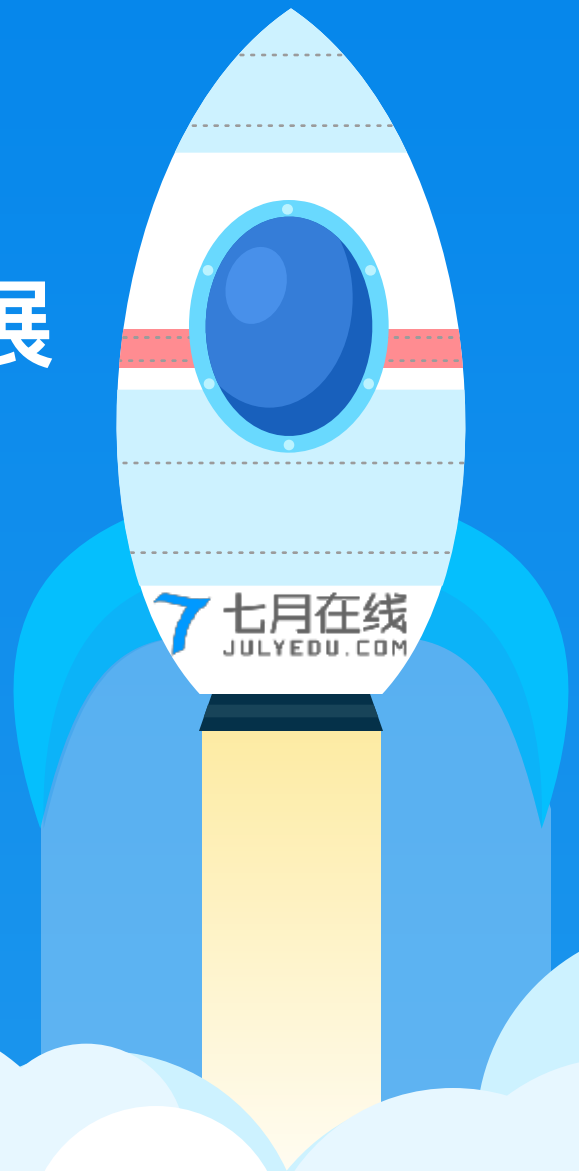
知识图谱中实体链接的方法与进展

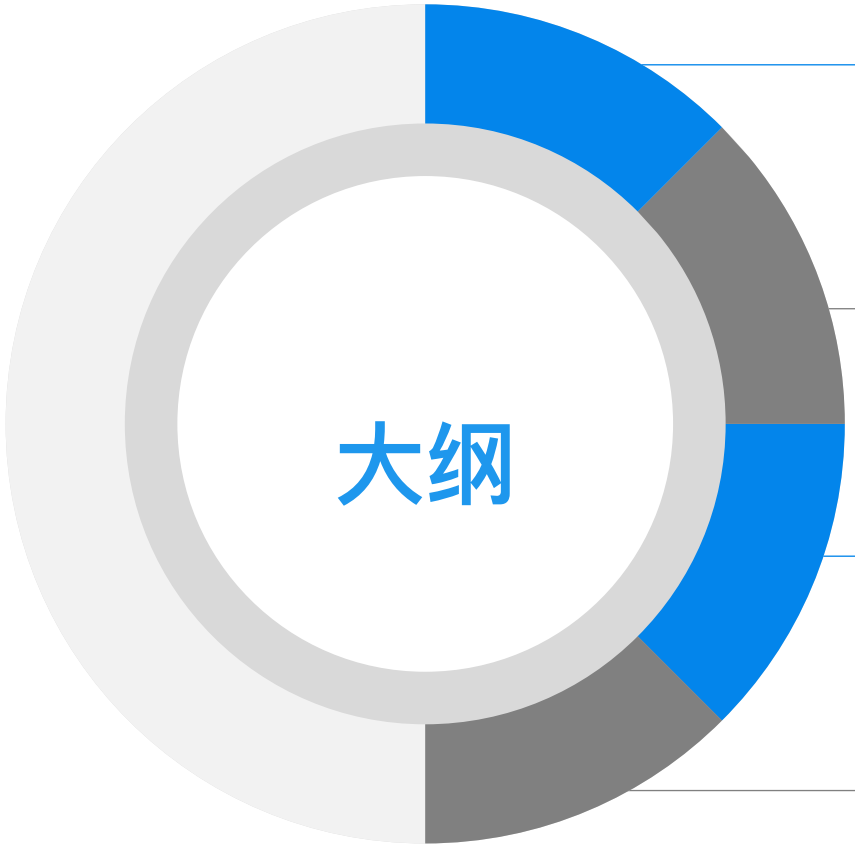
胡老师

https://www.julyedu.com/



七月在线
JULYEDU.COM

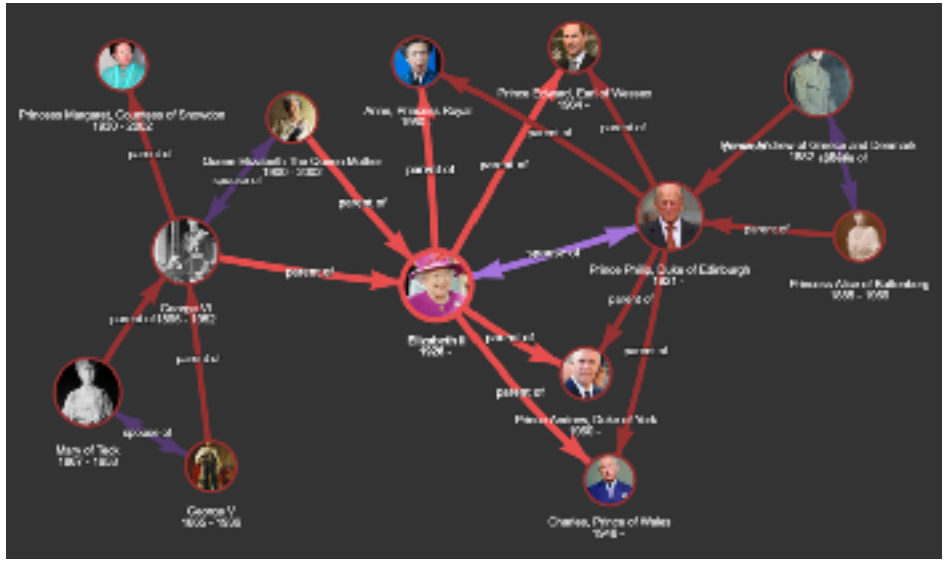
每一小节回顾一下



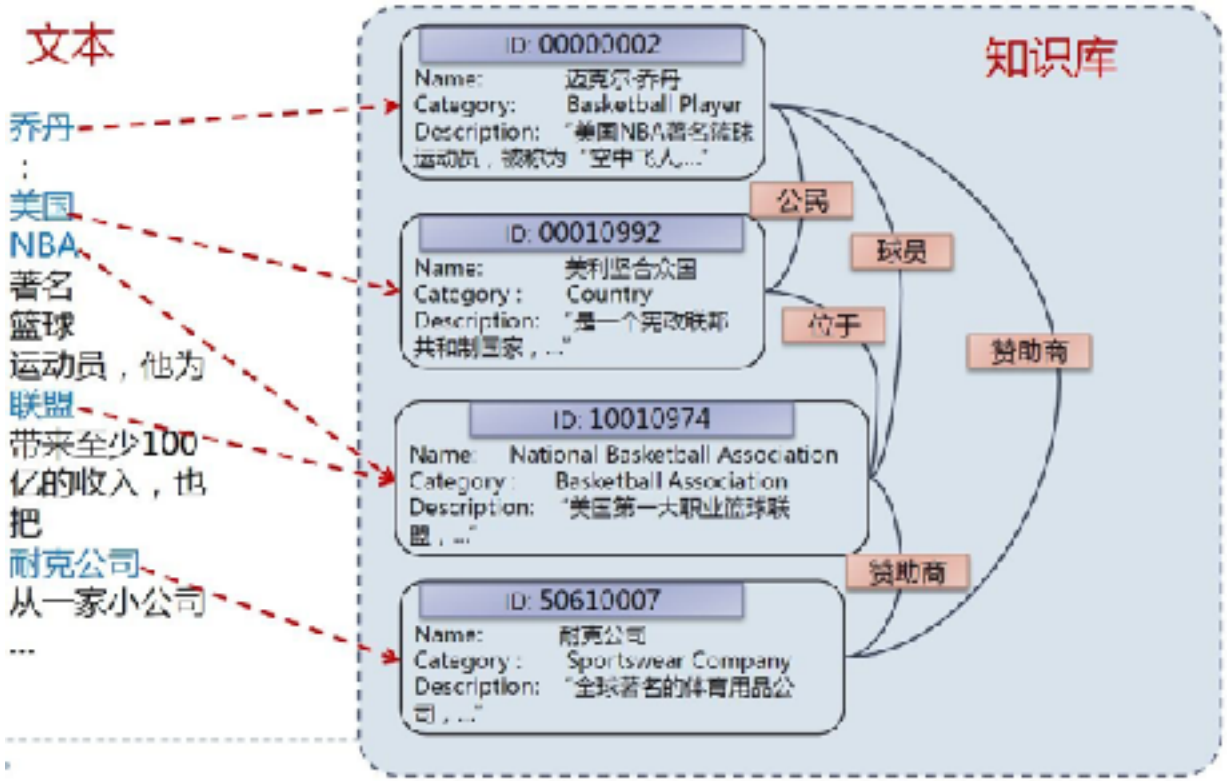
- 知识图谱相关概念的介绍
- 实体链接的挑战及应用
- 实体链接的技术流程
- 实体链接相关技术的前沿进展

/01 知识图谱相关概念的介绍

- 什么是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，有时也称为Knowledge Base (知识库)。
- whyKG?
- 现在机器的感知能力已经越来越接近于人类了，但是在认知领域。
- 知识图谱一般由三元组构成：<实体1，关系，实体2> 或者 <实体，属性，属性值>；



- Entity (实体)：实体是知识图谱的基本单元，也是文本中的重要信息单位。
- Mention (提及)：自然文本中表达实体的语言片段。
- Entity Linking(实体链接)：把文本中的mention链接到KG里的entity的任务。



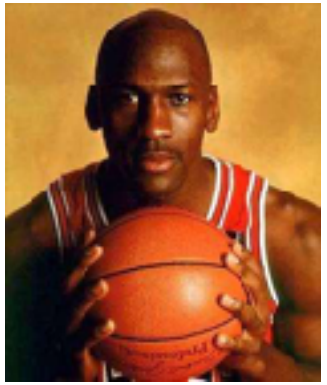
/02 实体链接的挑战及应用

- Mention Variations：同一实体有不同的mention。
- Entity Ambiguity：同一mention对应不同的实体。

唐僧
唐三藏
金蝉子
玄奘
旃檀功德佛
江流儿
长老
唐玄奘

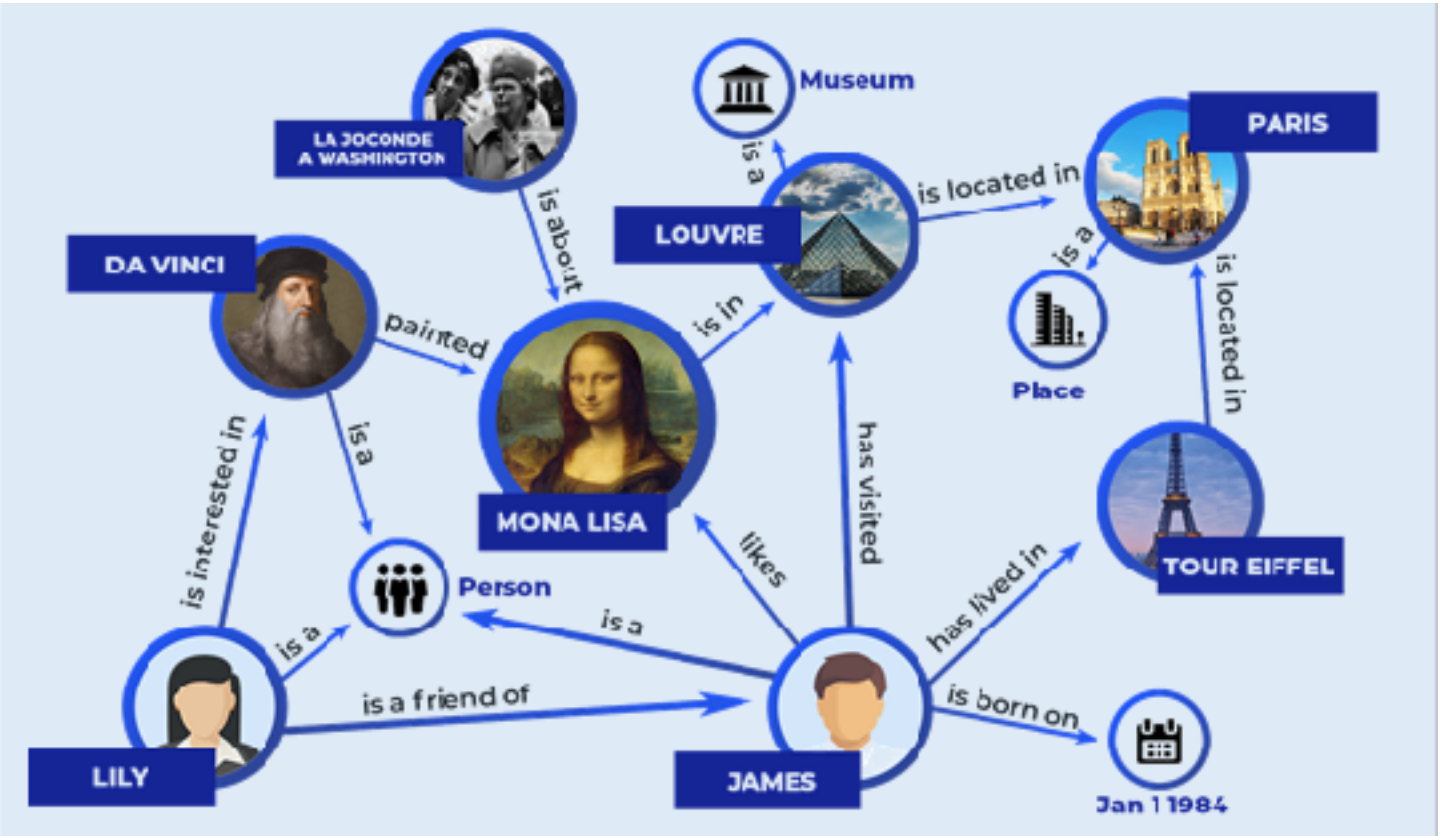
迈克尔·乔丹

机器学习界的开山鼻祖,尊称“乔帮主”，是人工智能领域两位根目录人物之一。



前美国职业篮球运动员，绰号“飞人”（Air Jordan）。

- 问答系统



蒙娜丽莎的作者是谁?
蒙娜丽莎在哪个博物馆?

- 阅读理解

梅拉尼娅·特朗普 - 百度百科

职业：模特
生日：1970年4月26日
毕业院校：斯洛文尼亚卢布尔雅那大学
简介：梅拉尼娅·特朗普，女，1970年4月26日出生
人物经历　个人生活
baike.baidu.com/

川普(美国政治家) - 百度百科

生日：1946年6月14日
代表作品：做生意的艺术，学徒
简介：川普一般指唐纳德·特朗普。唐纳德
6年6月14日-），出生于美国纽约，祖籍
人物经历　为政举措　商业成就　个人作品
baike.baidu.com/

2020年10月2日凌晨，特朗普证实他与夫人梅拉尼娅感染新冠病毒，并于当天傍晚被送往沃尔特·里德国家军事医疗中心接受治疗。

沃尔特·里德国家军事医学中心 - 百度百科

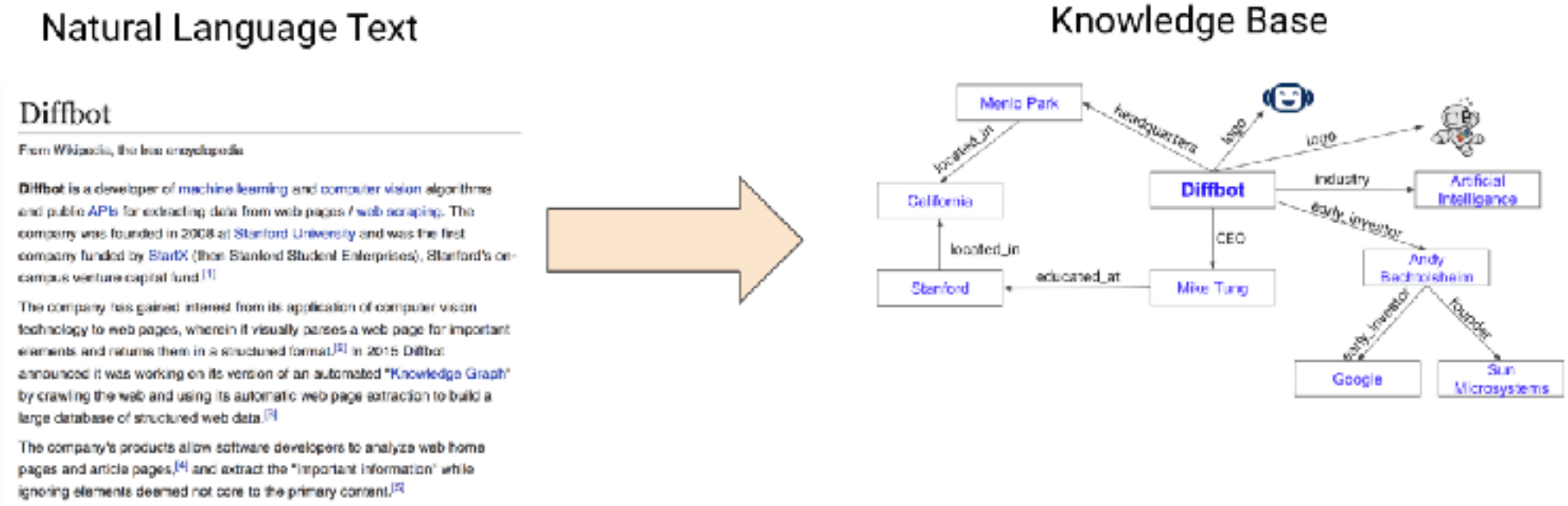
沃尔特·里德国家军事医学中心，2011年由沃尔特·里德
塞斯达的海军医疗中心合并成立，被称为"总统的医院"。
发展历史　诊治总统
baike.baidu.com/

- 舆情监督


Bai du 百度
金三胖
网页 资讯 视频 图片 知道 文库
百度为您找到相关结果约0个
很抱歉，没有找到与"金三胖"相关的网页。
温馨提示：
请检查您的输入是否正确
如网页未收录或者新站未收录，请提交网址给我们
如有任何意见或建议，请及时反馈给我们


Google
金三胖
www.youtube.com › watch
今天，金三胖有最新消息！俄羅斯聯邦
27 avr. 2020 - 今天，金三胖有最新消息！俄羅斯聯邦共
6,835 views6.8K views. • Apr 27, 2020. 41 5. Share Sa
www.youtube.com › watch
金三胖病危朝鲜政局多變，貓神斷定太
7:18
金三胖病危朝鮮政局多變，貓神斷定
率. 51,964 views51K views. • Apr 2
23 avr. 2020 - Ajouté par 貓神【猫
www.youtube.com › watch
金正恩金三胖心臟手術死亡真的假的?
24 avr. 2020 - 金正恩金三胖心臟手術死亡真的假的?

- 知识图谱构建与扩展


Natural Language Text
Diffbot
From Wikipedia, the free encyclopedia
Diffbot is a developer of machine learning and computer vision algorithms and public APIs for extracting data from web pages (web scraping). The company was founded in 2008 at Stanford University and was the first company funded by StartX (then Stanford Student Enterprises), Stanford's on-campus venture capital fund.[1]
The company has gained interest from its application of computer vision technology to web pages, wherein it visually parses a web page for important elements and returns them in a structured format.[2] In 2015 Diffbot announced it was working on its version of an automated "Knowledge Graph" by crawling the web and using its automatic web page extraction to build a large database of structured web data.[3]
The company's products allow software developers to analyze web home pages and article pages,[4] and extract the "important information" while ignoring elements deemed not core to the primary content.[5]
Knowledge Base
Menlo Park
headquarters
ceo
founded_in
California
Diffbot
industry
Artificial Intelligence
CEO
located_in
Stanford
educated_at
Mike Tung
Andy Bechtolsheim
founder
Google
Sun Microsystems

/03 实体链接的技术流程

粗排

基于一些粗粒度的方法获取潜在的实体候选项。

- 基于词典
- 基于模糊匹配
- 基于词向量

引用表构建

细筛

基于一些细粒度的方法获取最可能的实体项。

- 提取特征
- 建模

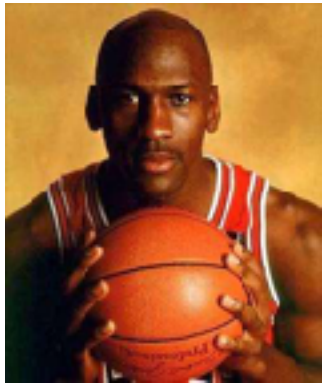
如何获取最佳实体项?

实体相关的特征有哪些?

实体相关的特征有哪些?

迈克尔·乔丹

机器学习界的开山鼻祖,尊称“乔帮主”，是人工智能领域两位根目录人物之一。



前美国职业篮球运动员，绰号“飞人”（Air Jordan）。

李娜

李娜，中国女子网球运动员，亚洲第一位大满贯女子单打冠军。

李娜，出生于河南省郑州市，毕业于河南省戏曲学校，曾是中国大陆女歌手，出家后法名释昌圣。

实体知名度

实体相关的特征有哪些?

实体上下文

实体相关的特征有哪些?

李娜 武汉市 华中科技大学 北京奥运会 法国网球公开赛 澳大利亚网球公开赛 中国网球巡回赛

李娜 郑州市 河南省戏曲学校 豫剧《好人一生平安》《青藏高原》《嫂子颂》《赵四小姐和张学良》

实体语义关联度

实体相关的特征有哪些?

苹果

Topic(Computer)	Topic(Video)	Topic(Software)	Topic(Wine)	Topic(Food)	Topic(Plant)
Computer	*Video*	*Computer software*	*Wine*	*Food*	*Plant*
CPU	*Mobile phone*	*Microsoft Windows*	*Grape*	*Restaurant*	*Flower*
Hardware	*Mass media*	*Linux*	*Vineyard*	*Meat*	*Leaf*
Personal computer	*Music*	*Web browser*	*Winery*	*Cheese*	*Tree*
Computer memory	*Television*	*Operating system*	*Apple*	*Vegetable*	*Fruit*

实体相关主题

实体相关的特征有哪些?

思考：还有哪些有用特征呢?

- 文本出处（报纸、公众号（表达更加活泼，生活化一些）、书籍（提及更加正式一些））
- 文本作者背景（记者（公众人物），粉丝（爱豆））
- 文本产生的时期（（不同的人在不同的时期火的程度不一样））
- 文本的类型（网文、小说、微博、新闻）
- ...

成本VS提升性能

如何获取最佳实体项?

- 局部实体链接分数：通过提及及其相关信息与某一候选实体间的匹配程度进行匹配。
- 全局实体链接分数：通过提及及其相关信息与所有候选实体间的匹配程度进行匹配。

如何获取最佳实体项?

局部实体链接

p(候选实体，提及）= p（候选实体）* p（提及知识|候选实体）* p（提及上下文|候选实体）

p(李娜【运动员】，“李娜”)

p(李娜【运动员】) = 0.8

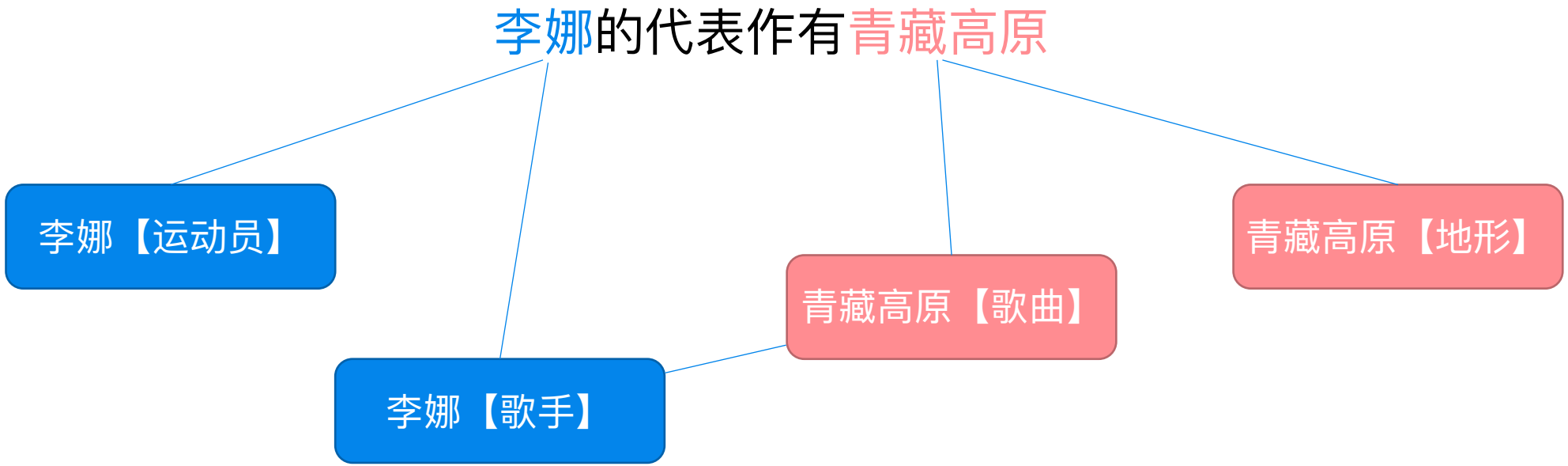
p(体育|李娜【运动员】) = 0.7

p(网球|李娜【运动员】) = 0.9

如何获取最佳实体项?

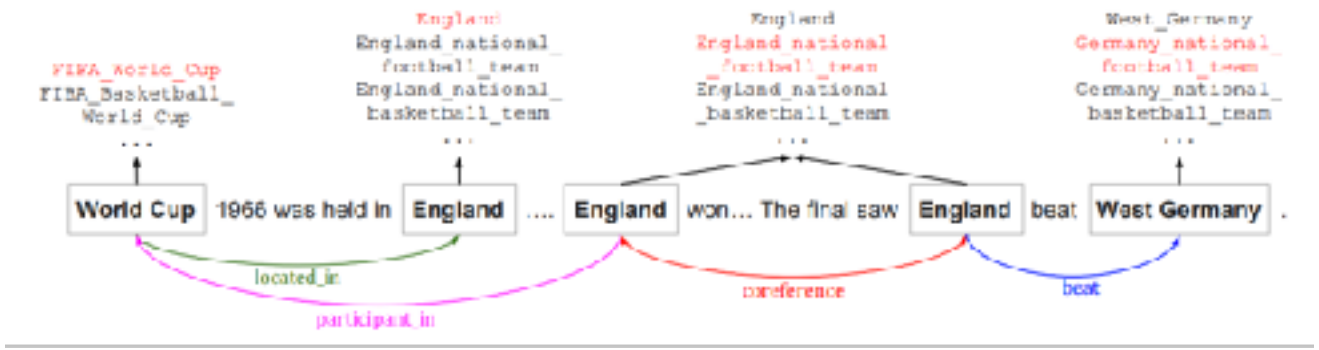
全局实体链接

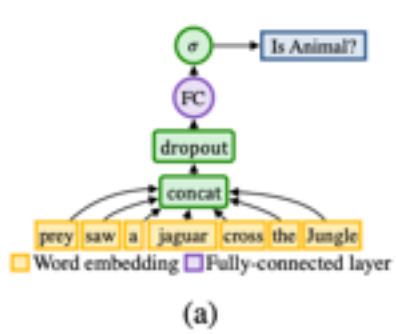
查找边权和最大的子图（提及与实体一一对应）



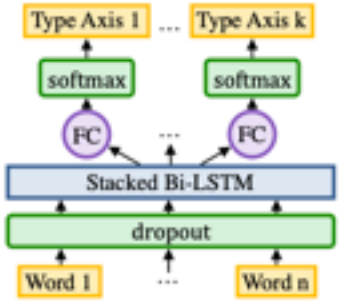
/04 实体链接相关技术的前沿进展

实体链接相关技术的前沿进展

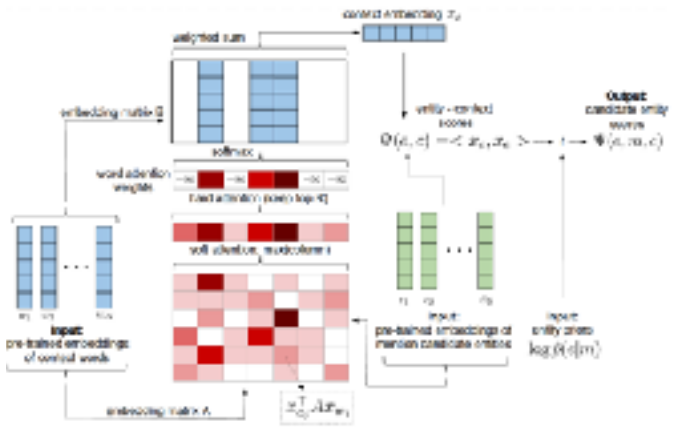




(a)



(b)



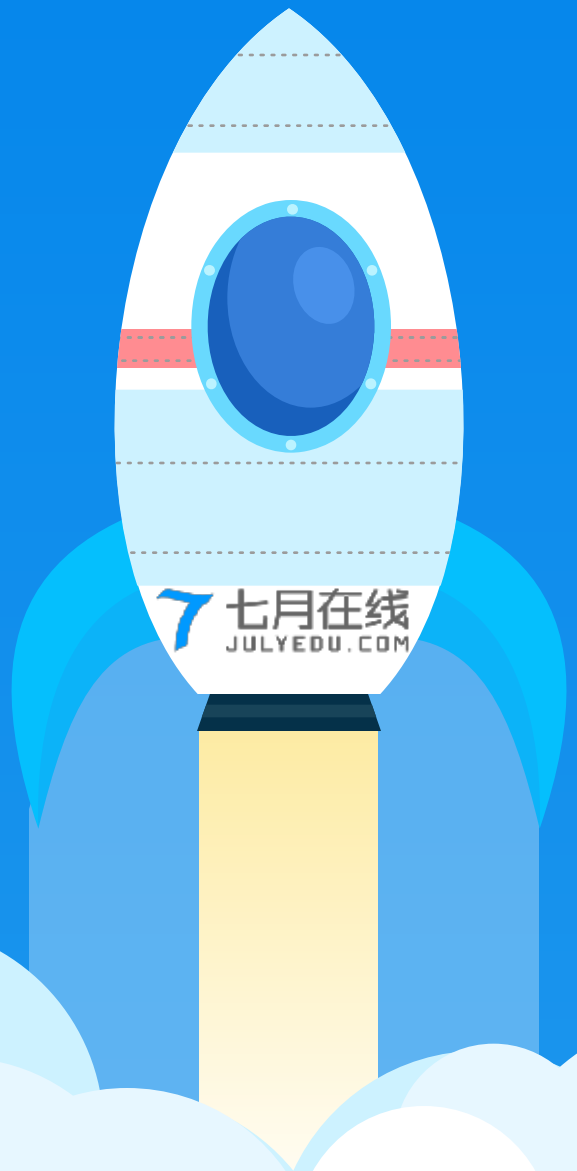
- 引入词汇/短语的分布式表征
- 引入更多的外部知识
- 转化为序列标注问题（端到端）
- ...

NLP就业5期

- 内容系统：五大阶段，分别从NLP基础、PyTorch实战、词向量与预训练模型、机器翻译、结构化预测、到问答系统和聊天机器人；
- 项目实战：提供文本分类、机器翻译、问答系统、FAQ问答机器人、知识图谱、聊天机器人等项目实战、聊天机器人中的语义理解、文本推荐系统，以及一个开放式项目；
- 就业指导：就业部辅助BAT大咖讲师做简历指导、面试辅导、就业内推。

http://m.julyedu.com/detail?id=317

微信扫一扫关注我们



THANKS

Speaker name and title

https://www.julyedu.com

七月在线 JULYEDU.COM