第八章 知识图谱

戴洪良

计算机科学与技术学院/人工智能学院

hongldai@nuaa.edu.cn

实体链接

- 将文本中的实体提及（mention）链接到知识图谱中的对应实体（entity）

2023年8月27日是南航新生报到日。

南京航空航天大学(中国江苏省南京市境内... - 百度百科

南京航空航天大学（Nanjing University of Aeronautics and Astronautics），简称南航（NUAA），位于江苏省南京市，是中华人民共和国工业和信息化部直属的一所具有航空航天民航特色、以理工类为主的综合性全国重点大学，由工业和信息化部、教育部、江苏省共建... 详情 >

历史沿革 办学规模 学术研究 文化传统

中国南方航空股份有限公司 - 百度百科

中国南方航空股份有限公司（China Southern Airlines，简称南航），总部设在广州，成立于1995年3月25日，以蓝色垂直尾翼镶红色木棉花为公司标志，是中国运输飞机最多、航线网络最发达、年客运量最大的航空公司。南航年客运量居亚洲第一、世界第三；机队规模居... 详情 >

现任领导 历史沿革 运营基地 基本规模

- 可应用于信息检索、QA系统、知识图谱构建等

实体链接的流程

- 一般分两步实现实体链接

2023年8月27日是南航新生报到日。

1. 候选生成：

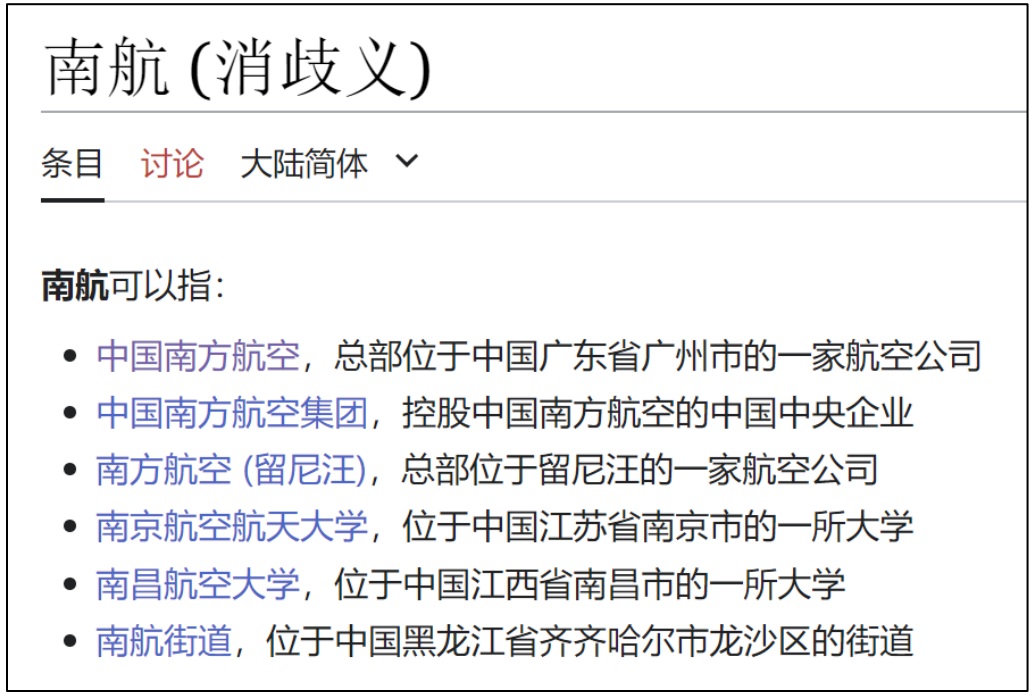
	2. 实体消歧：
中国南方航空 总部位于中国广东省广州市的一家航空公司	0.3
南方航空 (留尼汪) 总部位于留尼汪的一家航空公司	0.1
南京航空航天大学 位于中国江苏省南京市的一所大学	0.8 ✓
南昌航空大学 位于中国江西省南昌市的一所大学	0.6
南航街道 位于中国黑龙江省齐齐哈尔市龙沙区的街道	0.2

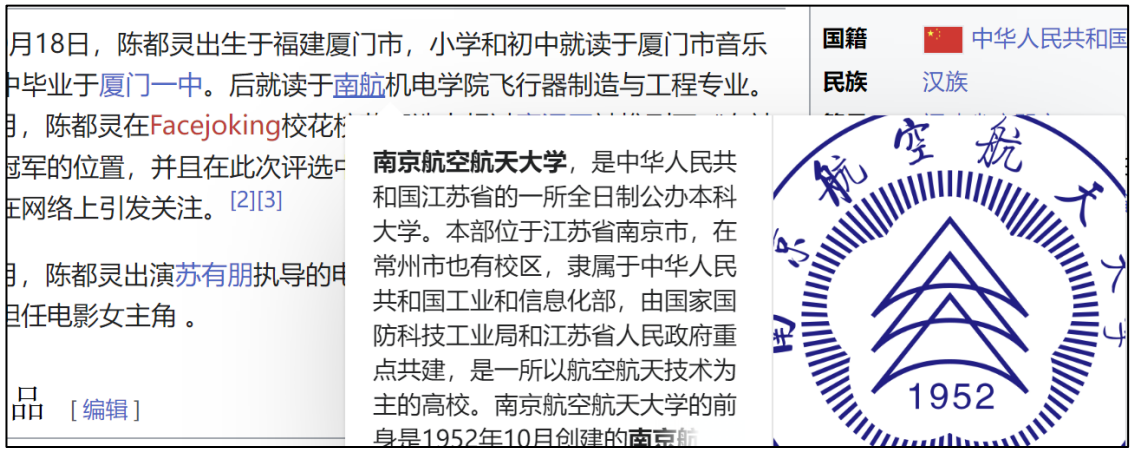
候选生成

- 基于别名表的方法
- 基于向量表示相似度的方法

候选生成

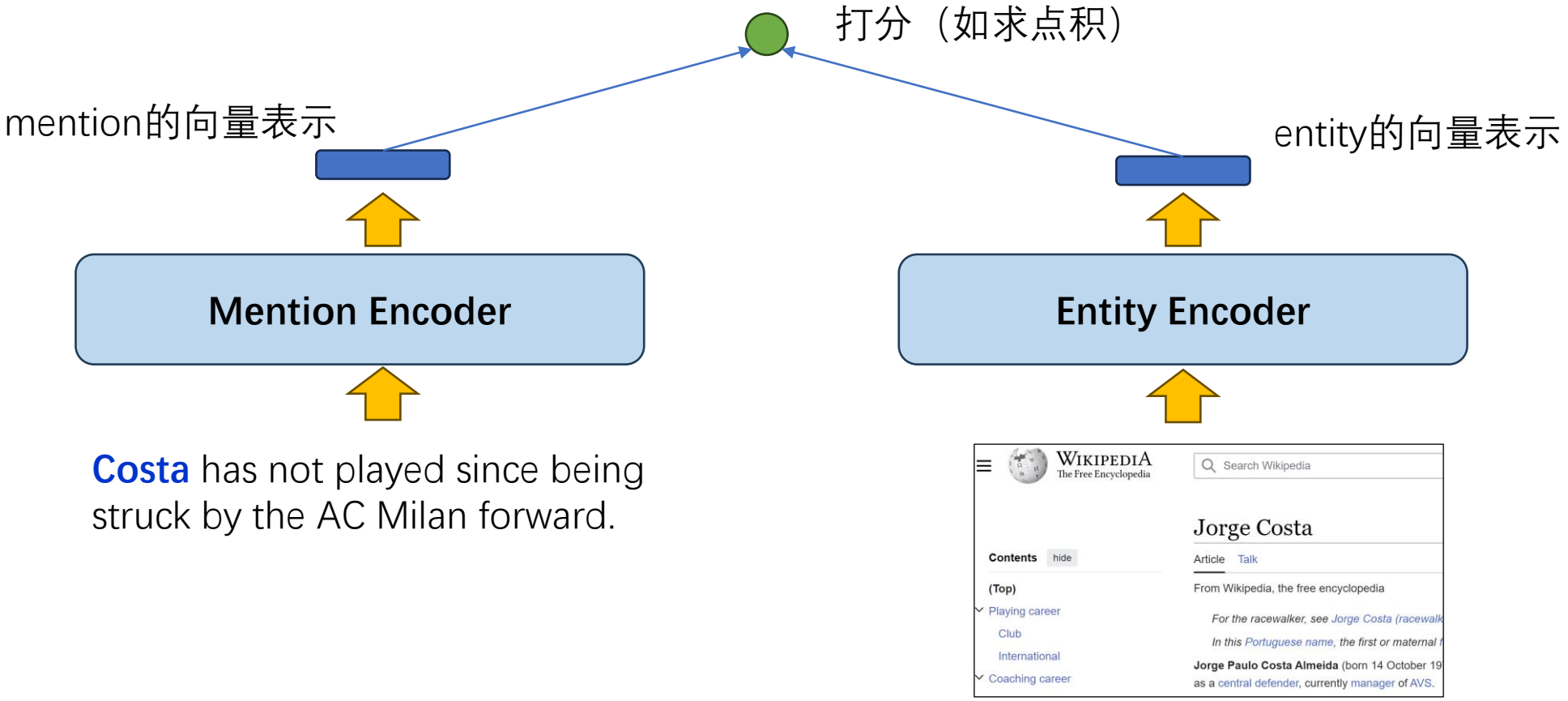
- 构建别名表
  - 即 实体名->实体 的词典
- 可利用资源：Wikipedia的消歧页、重定向页、内部超链接等





候选生成

- 基于向量表示相似度的方法
- 对bi-encoder训练好后，可把实体的向量表示预先计算好



候选生成

- 基于向量表示相似度的方法
- 基于BERT的mention和entity encoder

Mention encoder: $[\mathrm{CLS}]\ \mathrm{ctxt}_l\ [\mathrm{M}_s]\ \mathrm{mention}\ [\mathrm{M}_e]\ \mathrm{ctxt}_r\ [\mathrm{SEP}]$

Entity encoder: $[\mathrm{CLS}]\ \mathrm{title}\ [\mathrm{ENT}]\ \mathrm{description}\ [\mathrm{SEP}]$

训练损失函数： $\mathcal{L}(m_i, e_i) = -s(m_i, e_i) + \log \sum_{j=1}^{B} \exp\left(s(m_i, e_j)\right)$

其中 $s(m, e_i) = \boldsymbol{y_m} \cdot \boldsymbol{y_{e_i}}$

Wu, Ledell, et al. "Scalable Zero-shot Entity Linking with Dense Entity Retrieval." Proceedings of EMNLP. 2020.

实体消歧

- 实体消歧阶段可有两类做法
  - 每次只考虑一个实体提及，对其候选实体打分后选出得分最高的（local方法）
  - 一次性考虑文档中所有实体提及，希望总的链接结果最好（global方法）

实体消歧

- 单独考虑一个提及（mention）的Local方法
- 传统的基于手工设计特征的方法

特征例：

实体受欢迎度：

（基于维基百科内部超链接计算）

$$Pop(e_i) = \frac{count_m(e_i)}{\sum_{e_j \in E_m} count_m(e_j)}$$

$count_m(e)$：文本为m，且指向实体e的维基百科页面的超链接数

上下文与实体描述相似度： 上下文与实体对应Wikipedia文章的TF-IDF表示相似度

实体消歧

- 单独考虑一个提及（mention）的Local方法
  - 与基于向量表示相似度的候选生成方法不同，得到一定数目的候选实体后后，可采用需计算量更大的方法对它们打分

- Cross-encoder模型：

为BERT构建输入：

[CLS] ctxt_l [M_s] mention [M_e] ctxt_r [SEP] entity_title [SEP] entity_description [SEP]

输入BERT后获取[CLS]对应的向量表示，记为$\boldsymbol{y_{m,e}}$

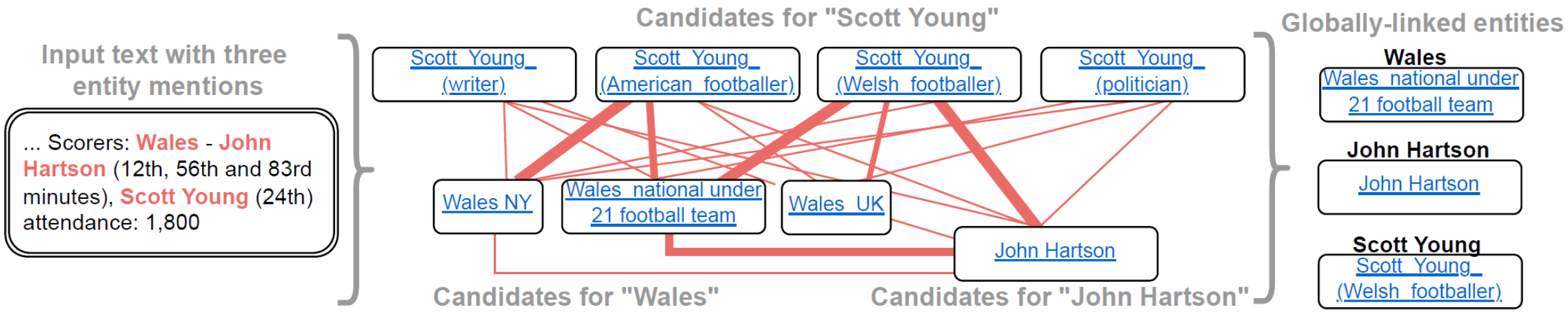
计算实体得分： $s_{cross}(m,e) = y_{m,e} \cdot w$

可采用与候选生成阶段类似的loss训练或Margin-based loss

Wu, Ledell, et al. "Scalable Zero-shot Entity Linking with Dense Entity Retrieval." Proceedings of EMNLP. 2020.

实体消歧

- 同时考虑文档中所有mention，希望总的链接结果最好（global方法）



实体消歧

- 简单的global做法

e为对文档中所有mention的一种链接结果，对其打分为：

$$g(e,m,c)=\sum_{i=1}^{n}s(e_i,m_i,c_i)+\sum_{i,j,i\neq j}r(e_i,e_j)$$

$s(e_i,m_i,c_i)$：第i个mention对应的候选实体与该mention、该mention的上下文的匹配程度分

$r(e_i,e_j)$：第i和第j个mention对应的候选实体e_i和e_j的关联程度

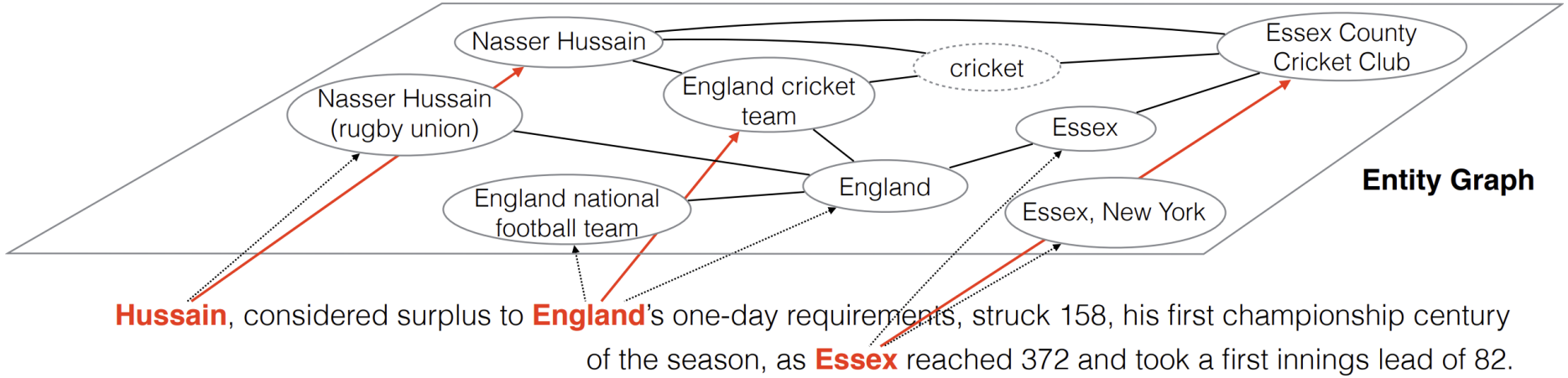
两个实体间关联程度的一种计算方法：

$$r(e_1,e_2)=\frac{\log(\max(|P_1|,|P_2|))-\log(|P_1\cap P_2|)}{\log(|W|)-\log(\min(|P_1|,|P_2|))}$$

其中P_1（P_2）为有链接指向e_1（e_2）的维基百科页面集合；W为所有维基百科页面的集合

实体消歧

- 使用更先进技术的global做法
  - (Cao et al., 2018)基于知识图谱为候选实体构建图
  - 使用图神经网络获取不同实体间的关联程度



Cao, Yixin, et al. "Neural Collective Entity Linking." Proceedings of the 27th International Conference on Computational Linguistics. 2018.

无法链接到实体的情况

- 有些提及所指代的实体在知识库中不存在
  - 此时实体链接系统应为该提及输出NIL，表示找不到对应实体

热心市民**小明**向记者描述了当时的情况

↓

NIL

- 处理NIL的一些方法：
  - 有时可依靠候选生成部分，没有对应的候选实体则输出NIL
  - 实体消歧阶段设置分数阈值，最高分低于阈值则不链接
  - 设置一个NIL假实体，将它与其他实体相同对待
  - 训练一个额外的二分类器，判断是否应该链接

END