

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



知识图谱问答

- 知识图谱问答系统的目标是利用知识图谱中的知识自动回答自然语言问题。
- 知识图谱问答 (Question Answering over Knowledge Graphs, **KGQA**) 也称知识库问答 (Question Answering over Knowledge Bases, **KBQA**)
- 可应用于智能个人助理、聊天机器人、搜索引擎等



知识图谱问答

- 简单问答 (Simple Question Answering)
 - 只涉及一种关系
- 复杂问答 (Complex Question Answering)
 - 涉及多种关系

简单问答

- 问题中只涉及一个实体和一种关系的问答任务

如： 谁导演了流浪地球？

- 方法有：

- 基于模板

What is the population of New York?



<i>what is the</i>	r	<i>of</i>	e	$=$	$r(?, e)$
<i>population</i>				$=$	population
<i>new york</i>				$=$	new-york

- 分解为子任务

- 实体识别、实体链接、关系识别

- 端到端的方法

基于多个子任务的简单知识图谱问答

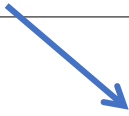
1. 识别问题中的实体提及（简化的NER）

What major cities does **US 2** run through?

识别出其中的“US 2”

2. 找出该实体提及可能指代的实体（利用实体链接）

What major cities does **US 2** run through?



U.S. Route 2 (Highway)
US-2 (Boat)
US 2 (Film)
The Last of Us Part 2 (game)
...

基于多个子任务的简单知识图谱问答

3. 找出候选三元组

- 将所有以候选实体为头实体的三元组作为候选

(U.S. Route 2, major_cities, Kalispell)

(U.S. Route 2, major_cities, Williston)

(US-2, has_use, firefighting)

(US 2, genre, drama)

...

4. 三元组选择

- 结合问题对三元组中的头实体和关系打分，选出得分最高的形成答案

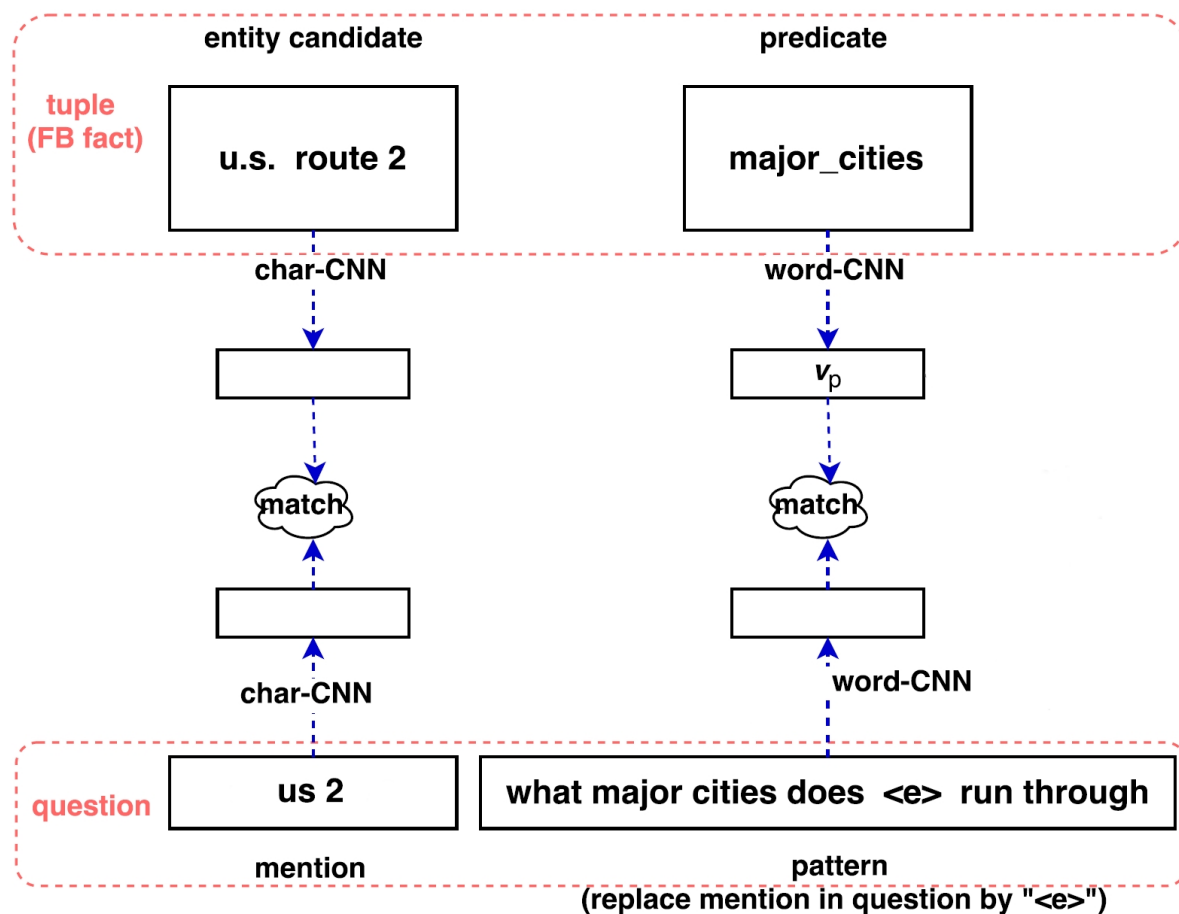
基于多个子任务的简单知识图谱问答

- 识别实体提及（方法举例）
 - 可通过序列标注实现；可使用BERT、BERT-CRF、LSTM-CRF等模型

词	What	major	cities	does	US	2	runs	through	?
标签	c	c	c	c	e	e	c	c	c

基于多个子任务的简单知识图谱问答

• 三元组选择（方法举例）



- 对问题中的实体提及和候选三元组中的实体分别获取向量表示后计算匹配程度: m_e
- 对问题上下文和候选三元组中的关系（即 predicate）分别获取向量表示后计算匹配程度: m_r

最后分数:

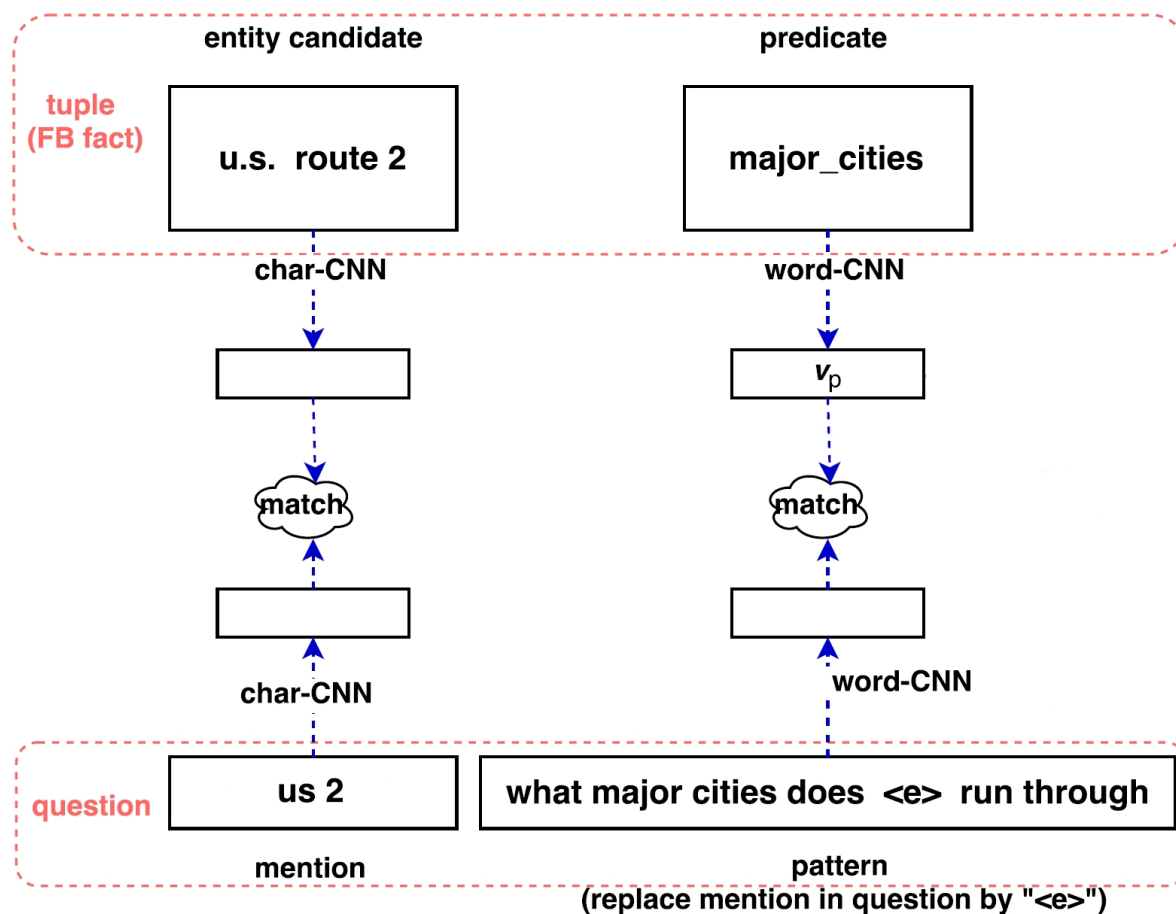
$$f(q, t) = m_e + m_r + s_e$$

其中 s_e 为实体链接分数

在已有实体链接分数 s_e 的情况下, m_e 并不是必须的

基于多个子任务的简单知识图谱问答

- 三元组选择



最后分数:

$$f(q, t) = m_e + m_r + s_e$$

训练loss:

$$l(q, t^+, t^-) = \max(0, \lambda + s_t(q, t^-) - s_t(q, t^+))$$

Yin, Wenpeng, et al. "Simple question answering by attentive convolutional neural network." COLING 2016.

复杂问答

- 无法只基于知识图谱中的一个实体和一种关系回答的问题

江湖儿女的导演的老婆是谁？

(江湖儿女, 导演, 贾樟柯) (贾樟柯, 妻子, 赵涛)

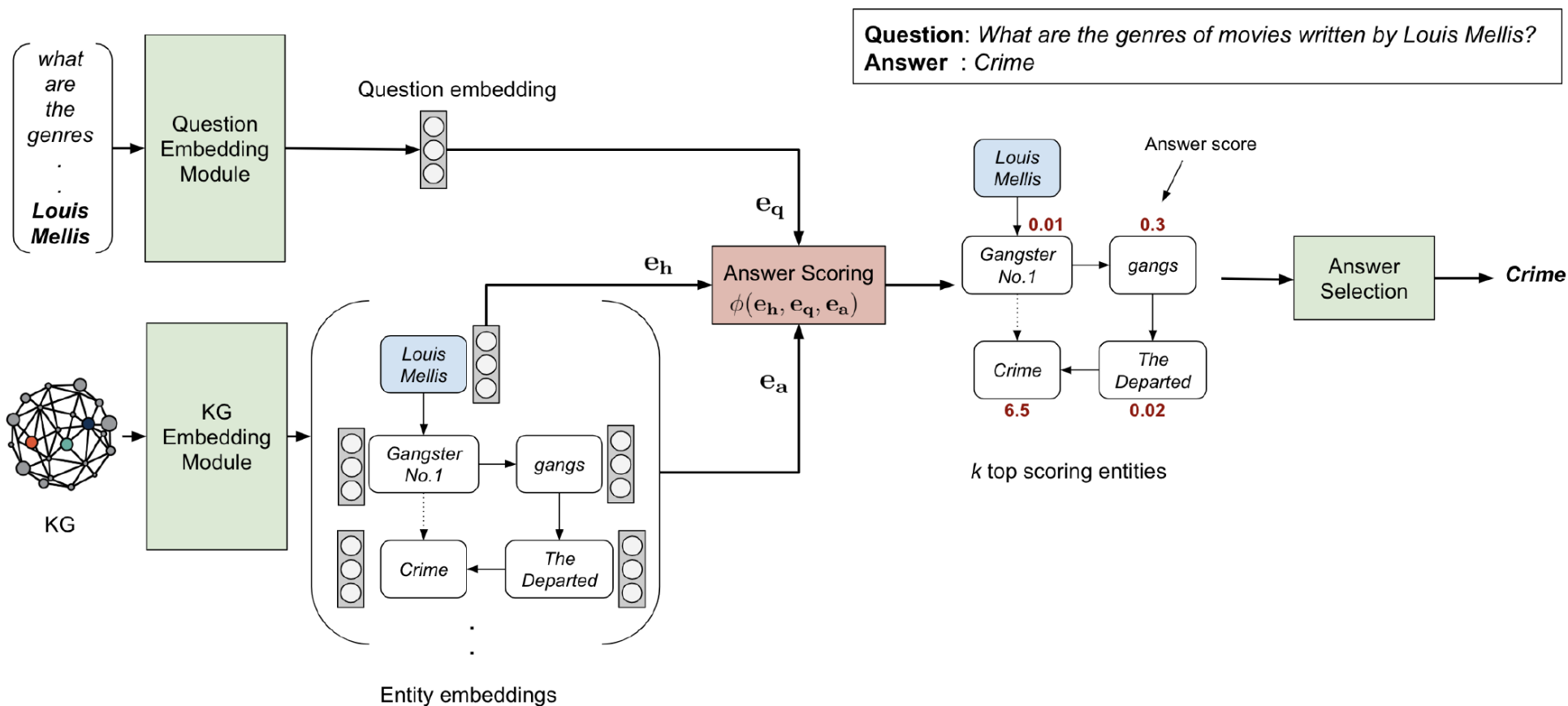
- 主要方法
 - Embedding-based (基于嵌入向量的方法)
 - 将问题和候选答案 (候选实体) 表示为向量后计算候选答案的分数
 - Semantic-parsing-based (基于语义解析的方法)
 - 将问题转为形式化查询语句 (如SPARQL) 后, 对知识图谱执行查询

基于嵌入向量的方法

- 一种基于知识图谱嵌入的方法 (Saxena et al., 2020)
 1. 用知识图谱嵌入方法获得知识图谱中的实体向量表示
 2. 用神经网络模型获得问题的向量表示
 3. 根据向量表示为可能是答案的实体打分，得分最高的作为最终答案

基于嵌入向量的方法

- 一种基于知识图谱嵌入的方法 (Saxena et al., 2020)



基于嵌入向量的方法

- 一种基于知识图谱嵌入的方法 (Saxena et al., 2020)
 - 知识图谱嵌入: ComplEx
 - 为问题获取向量表示: RoBERTa模型

基于实体和问题的向量表示对候选答案实体 a' 打分:

$$\phi(e_h, e_q, e_{a'})$$

e_h : 问题中的实体的向量表示

e_q : 问题的向量表示

$e_{a'}$: 候选答案实体 a' 的向量表示

ϕ : 知识图谱嵌入方法ComplEx的打分函数

基于嵌入向量的方法

- 一种基于知识图谱嵌入的方法 (Saxena et al., 2020)
- 关系匹配
 - 该方法还包括一个判断问题实体与候选答案实体间的路径中的关系是否与问题匹配的模块

$$h_q = \text{RoBERTa}(q)$$

关系 r 的分数: $S(r, q) = \text{sigmoid}(h_q^T h_r)$

$S(r, q)$ 反应问题 q 中是否出现（涉及）了 r 这种关系，其相关参数可以预先训练好

候选答案 a' 的关系分数: $\text{RelScore}_{a'} = |R_q \cap R_{a'}|$

其中 R_q 为得分高于0.5的关系集合， $R_{a'}$ 为问题实体到候选答案实体的最短路径中的关系集合

基于嵌入向量的方法

- 一种基于知识图谱嵌入的方法 (Saxena et al., 2020)
- 候选答案的最终选取:

$$e_{ans} = \arg \max_{a' \in \mathcal{N}_h} \phi(e_h, e_q, e_{a'}) + \gamma * \text{RelScore}_{a'}$$

可只将与问题实体 h 在知识图谱中较相近的实体作为候选答案
如：只选从问题实体出发经3跳可达到的实体作为候选

基于嵌入向量的方法

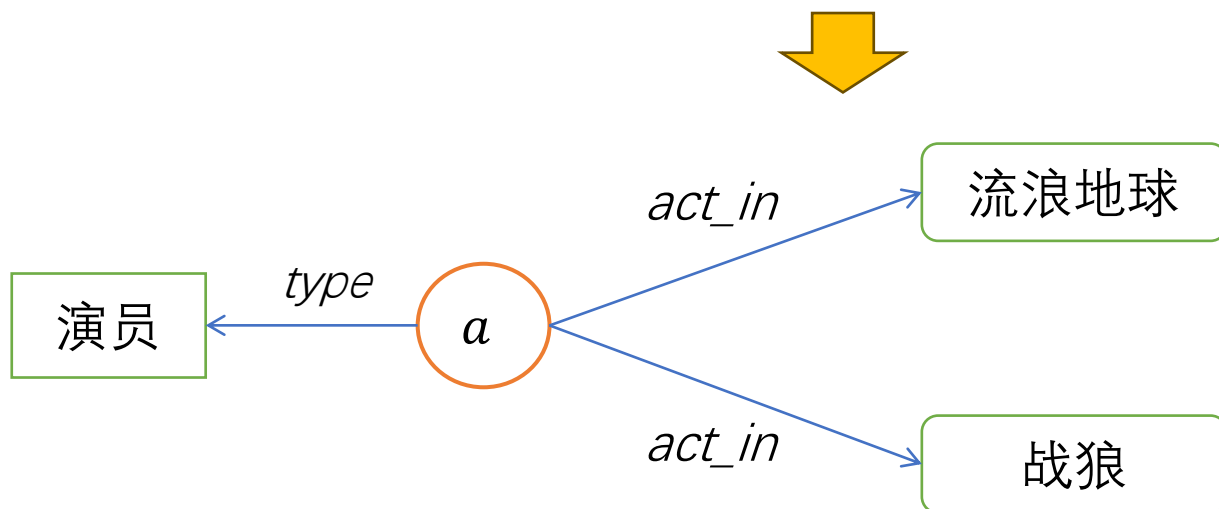
- 一种基于知识图谱嵌入的方法 (Saxena et al., 2020)
- 该方法的缺点:
 - 只能选出一个答案，但有些问题会有多个正确答案
 - 只考虑了问题中只有一个实体的情况

哪个演员既演了流浪地球又演了战狼？

基于语义解析的方法

- 将问题转为形式化查询语句（如SPARQL）后，对知识图谱执行查询
- （Luo et al., 2018）的方法将问题转化为查询图，再根据查询图对知识图谱查询，得到答案

哪个演员既演了流浪地球又演了战狼？



基于语义解析的方法

- 查询图生成

1. 生成候选查询图

2. 对候选查询图打分，选出最高分的查询图

- 生成候选查询图

- 如：（Luo et al., 2018）先生成主要关系路径，再基于该路径添加限制
 - 主要关系路径：通过将 a 节点与问题中的实体以一跳或两跳关系链接得到
 - 限制：类别限制、时间限制等

基于语义解析的方法

- 查询图生成
 1. 生成候选查询图
 2. 对候选查询图打分，选出最高分的查询图
- 对候选查询图打分
 - 分别获得问题的向量表示以及候选查询图的向量表示，作点积或计算相似度打分
 - 问题的向量表示可用如CNN、LSTM、BERT等模型得到
 - 查询图的向量表示可用图神经网络得到

$$loss = \max\{0, \lambda - S(q, G^+) + S(q, G^-)\}$$

基于语义解析的方法

- 除了先生成候选查询再rank，也可以直接以翻译的形式转换问题
- 使用encoder-decoder结构，以sequence to sequence的形式，输入原问题，输入问题的形式化表达

Where is Carew Cross?



```
SELECT ?x WHERE { dbr:Carew_Cross dbo:location ?x . }
```

END
