



一只小狐狸带你解锁 炼丹术&NLP 秘籍

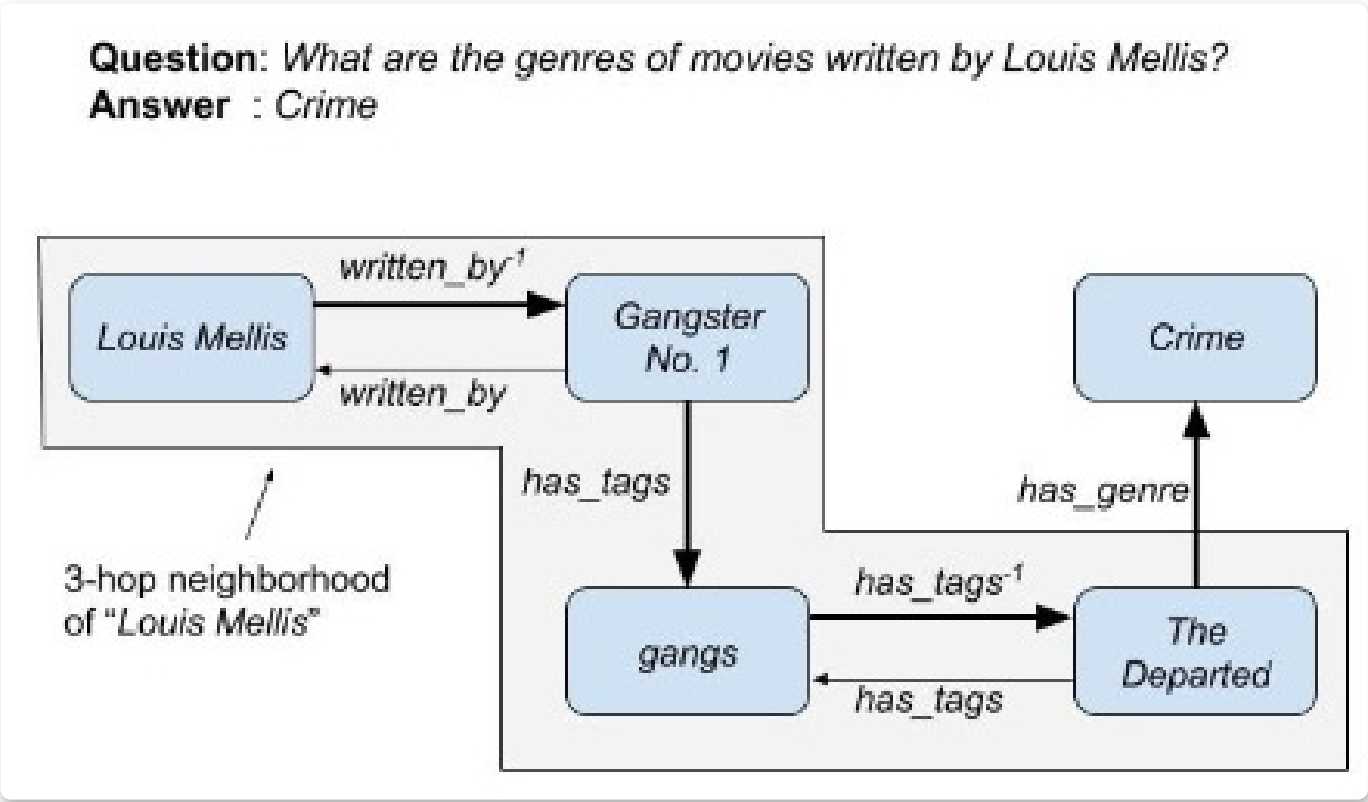
作者：舒意恒（南京大学硕士生，知识图谱方向）

背景

什么是知识图谱问答？

知识图谱 (KG) 是一个多关系图，其中包含数以百万计的实体，以及连接实体的关系。知识图谱问答 (Question Answering over Knowledge Graph, KGQA) 是利用知识图谱信息的一项研究领域。给定一个自然语言问题和一个知识图谱，通过分析问题和 KG 中包含的信息，KGQA 系统尝试给出正确的答案。

多跳知识图谱问答指的是，该问答系统需要通过知识图谱上的多条边执行推理，以获得正确答案。



一般而言，针对一个简单的事实类问题，KGQA 尝试找到一个三元组来回答这一问题。具体而言，KGQA 需要分析自然语言问

题，将自然语言问题中的实体描述和关系描述分别链接到知识图谱中的实体 h 和关系 r 。如果知识图谱中存在三元组 (h, r, t) ，则 t 是潜在的答案。这一过程可以称作单跳问答，一个问题的查询通过找到独立的三元组完成，而没有涉及多个有关联的三元组。

多跳知识图谱问答面临的挑战

知识图谱作为一种知识存储的形式，其中最重要的缺陷之一是它们通常都是不完整的，而这给 KGQA 提出了额外的挑战，尤其是多跳 KGQA。如上图所示，多跳 QA 需要一个长路径，而该路径上任意三元组的缺失都将导致真正的答案无法被搜索到。因此，采取某种方式预测知识图谱中缺失的链接，对于提升多跳 QA 的表现是有帮助的。当前缓解知识图谱不完整性的方法主要有：将 KG 与外部文本语料库结合，或者对知识图谱内的三元组进行补全等。

链接预测

链接预测的任务即预测知识图谱中缺失的链接，以减缓知识图谱的稀疏性。知识图谱嵌入是一种常见的链接预测方法，它为知识图谱中的实体和关系学习高维向量表示，但作者发现它尚未应用于多跳 KGQA 中。作者首次将嵌入用于多跳 KGQA，其目的在于充分利用嵌入方法在应对知识图谱稀疏性上的良好表现，增强模型的多跳推理能力。

论文题目：Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings

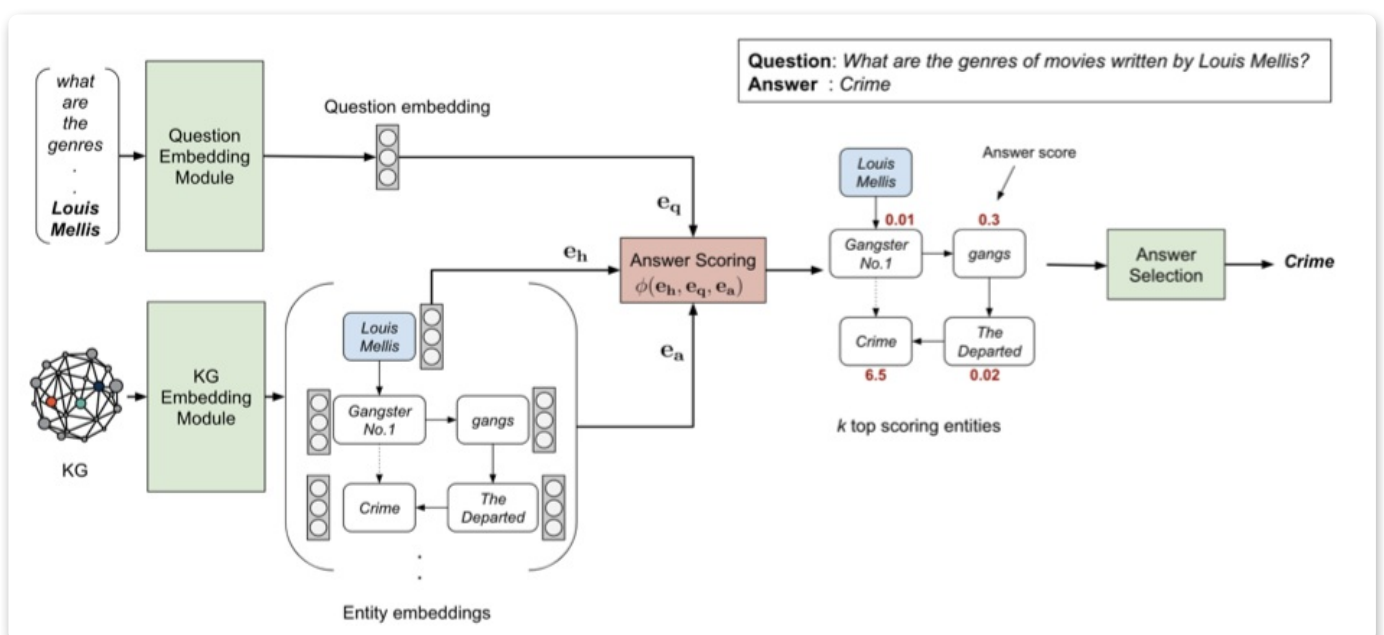
论文链接：<https://arxiv.org/abs/1910.03262v1>

Arxiv访问慢的小伙伴也可以在订阅号后台回复关键词【0616】下载论文PDF。

技术简介

如下图所示，该文将其 KGQA 方法称为 EmbedKGQA。其中包含三个关键模块。

- KG 嵌入模块：为 KG 中所有实体构建嵌入。
- 问题嵌入模块：为问题找到嵌入。
- 答案选择模块：减小候选答案实体的集合，并选择最终的答案。



KG 嵌入模块

作者直接选择了 ComplEx [1] 作为嵌入模型。实体嵌入被用于学习头实体、问题和答案实体构成的三元组评分函数。作者没有在 KG 嵌入模块进行过多关注和额外设计。

问题嵌入模块

该模块的目标是将一个自然语言问题 q 嵌入到一个固定维度的复向量 $e_q \in \mathbb{C}^d$ 。使用复向量的原因是匹配 KG 嵌入的 ComplEx 方法。作者使用了现有方法 RoBERTa [2] 将自然语言问题 q 嵌入到 768 维的向量中, 其中包含 4 个全连接线性层, 使用 ReLU 激活函数。

给定一个自然语言问题 q , 一个查询实体 $h \in \mathcal{E}$, 和一个答案实体集合 $\mathcal{A} \in \mathcal{E}$, 该模块学习问题嵌入以达到以下目标:

$$\begin{aligned}\phi(e_h, e_q, e_a) &> 0 \quad \forall a \in \mathcal{A} \\ \phi(e_h, e_q, e_{\bar{a}}) &< 0 \quad \forall \bar{a} \notin \mathcal{A}\end{aligned}$$

其中, ϕ 是 ComplEx 的评分函数, e 是之前学习到的实体嵌入。对于每个问题, 评分函数 $\phi(\cdot)$ 通过所有候选答案实体 $a' \in \mathcal{E}$ 计算得到。但哪些实体可以被认为是候选答案实体, 作者在此处并没有明确阐述, 而是在答案选择模块中讲解了如何对候选答案实体进行修剪操作。

读到这里, 不禁疑惑该文对多跳问题的处理有什么特别之处? 和单跳问题的处理是否有不同? 我们发现问题嵌入和 KG 嵌入并没有特殊的设计, 而实际上多跳 QA 的能力就来源于嵌入方法本身的性质。作者在实验部分对此进行了解释,

答案选择模块

在进行推断时, 模型根据 (head, question) 头实体-问题对, 对所有可能的答案 $a' \in \mathcal{E}$ 计算分数。对于较小的知识图谱, 例如后文实验部分所述的 MetaQA, 该模块直接选择具有最高分数的实体作为答案, 即从整个知识图谱的所有实体中选择最高分数的实体。此处实际上已经通过嵌入的方式, 完全避免了任何形式的检查某个结点的邻居结点的过程, 或者说克服了生成某种局部子图的方法的弊端, 实现了多跳。

然而对于较大的知识图谱, 作者认为**需要对候选实体进行修剪以提升性能**。修剪的方式即**关系匹配**。

关系匹配

为了在众多实体中选择符合期望的候选答案实体并计算分数, 作者利用了知识图谱中实体之外的信息, 即关系。作者利用类似于 PullNet [3] 的方法, 学习一个评分函数, 它能够在**给定一个问题时对所有关系进行排序**。

首先, 对于一个自然语言问题 q' , 将它作为输入得到它在 RoBERTa 最后一个隐藏层的输出:

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

然后, 根据 KG 嵌入模块学习到的关系嵌入 h_r , 我们可以计算一个度量关系和问题二者的评分:

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

在所有的关系中, 作者选择分数 $S(r, q) > 0.5$ 的关系, 将这个关系集合记为 \mathcal{R}_a 。这些关系是我们认为的与问题比较相关的关系。

然后, 对于每个候选实体 a' , 我们找到头实体 h 和 a' 之间的最短路径的关系, 将这个关系集合称为 $\mathcal{R}_{a'}$ 。这些关系是我们认为的与头实体最相关的关系。每个候选答案实体在给定问题时的关系分数可以通过这两个集合的交集的大小来计算:

$$\text{RelScore}_{a'} = |\mathcal{R}_a \cap \mathcal{R}_{a'}|$$

通过线性组合关系分数和 ComplEx 分数, 我们就可以找到答案实体。

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

值得注意的是，这里的候选答案实体仅仅是 $a' \in \mathcal{N}_h$ ，而其具体含义并未被作者直接说明。

效果

数据集

实验所用数据集是 MetaQA 和 WebQuestionsSP。

MetaQA 是一个大规模多跳 KGQA 数据集, 包含电影领域中超过四十万个问题。在 QA 之外, 该数据集包含 13.5 万项三元组和 4.3 万个实体, 以及 9 种关系。

WebQuestionSP 是一个较小的 QA 数据集, 包含 4737 个问题, 问题为 1 跳或 2 跳, 可通过 Freebase 回答。作者选取了 Freebase 的子集用于该数据集的实验, 其中包含 180 万个以上的实体和 570 万项以上的三元组。

显然, 后者所用的知识图谱规模要远大于前者。在实验部分中重点关注后者, 可以窥见该模型能否良好地适应较大规模的知识图谱。

竞争算法

作者将 EmbedKGQA 与 Key-Value Memory Network [4]、VRN [5]、GraftNet [6]、PullNet [7] 进行了对比。这些算法都实现了多跳 KGQA。其中, PullNet 限制答案实体在抽取的问题子图中, 这种抽取问题子图的方法在本质上限制了长路径的多跳推理能力。而 EmbedKGQA 本质上可以应对头实体和答案实体不连通的情况, 统一的嵌入空间在某种程度上已经包含了任意实体之间的关系, 这超越了 PullNet 局部子图的思路。

实验结果

下图是 MetaQA 数据集上的实验结果, KG-50 表示知识图谱的 50%。评价指标是 hit@1。括号中的数字表示文本被用于增强不完整的 QA 时的实验结果。

Model	MetaQA KG-Full			MetaQA KG-50		
	1-hop	2-hop	3-hop	1-hop	2-hop	3-hop
VRN	97.5	89.9	62.5	-	-	-
GraftNet	97.0	94.8	77.7	64.0 (91.5)	52.6 (69.5)	59.2 (66.4)
PullNet	97.0	99.9	91.4	65.1 (92.4)	52.1 (90.4)	59.7 (85.2)
KV-Mem	96.2	82.7	48.9	63.6 (75.7)	41.8 (48.4)	37.6 (35.2)
EmbedKGQA (Ours)	97.5	98.8	94.8	83.9	91.8	70.3

作者使用 50% 的知识图谱进行测试, 其目标是测试模型在不完整知识图谱上进行链接预测的能力。MetaQA KG-50 实际上很稀疏, 导致结点间的路径很可能变得更长。

下图是 WebQSP 数据集上的实验。EmbedKGQA 在链接预测上展现出非常好的表现, 但 KG-Full 上尚未取得超过 SOTA。

Model	WebQSP KG-Full	WebQSP KG-50
KV-Mem	46.7	32.7 (31.6)
GraftNet	66.4	48.2 (49.7)
PullNet	68.1	50.1 (51.9)
EmbedKGQA	66.6	53.2

总结

EmbedKGQA 构建方法简单且有效，充分利用了现有嵌入方法在应对知识图谱稀疏性与实现链接预测方面的良好表现，实现了多跳 QA。

不过，个人认为该文的瑕疵是，由于关系匹配部分中，候选答案实体选择的方法细节并未在文中被阐述清楚，在大型知识图谱上的多跳 QA 效果可能值得深究。



重磅惊喜:卖萌屋小可爱们苦心经营的 **自然语言处理讨论群** 成立三群啦!扫描下方二维码, 后台回复「**入群**」即可加入。众多顶会审稿人、大厂研究员、知乎大V以及美丽小姐姐(划掉 ♀)等你来撩噢~(手慢无

可能喜欢

- ICLR2020满分论文 | 为什么梯度裁剪能加速模型训练?
- 卖萌屋算法工程师思维导图part3—深度学习篇
- 万能的BERT连文本纠错也不放过
- 面试必备! 卖萌屋算法工程师思维导图—统计机器学习篇
- 告别自注意力, 谷歌为Transformer打造新内核Synthesizer



夕小瑶的卖萌屋

关注&星标小夕, 带你解锁AI秘籍
订阅号主页下方「撩一下」有惊喜哦



参考文献

[1] The 6 Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In International Conference on Machine Learning, pages 2071–2080.

- [2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
- [3] Haitian Sun, Tania Bedrax-Weiss, and William W Cohen. 2019a. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. arXiv preprint arXiv:1904.09537.
- [4] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. arXiv preprint arXiv:1606.03126.
- [5] Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In Thirty-Second AAAI Conference on Artificial Intelligence.
- [6] Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. arXiv preprint arXiv:1809.00782.
- [7] Haitian Sun, Tania Bedrax-Weiss, and William W Cohen. 2019a. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. arXiv preprint arXiv:1904.09537.

声明：pdf仅供学习使用，一切版权归原创公众号所有；建议持续关注原创公众号获取最新文章，学习愉快！