

基于预训练语言模型的检索-匹配式知识图谱问答系统

张鸿志, 李如霖, 王思睿, 黄江华

美团, 北京市朝阳区 100020

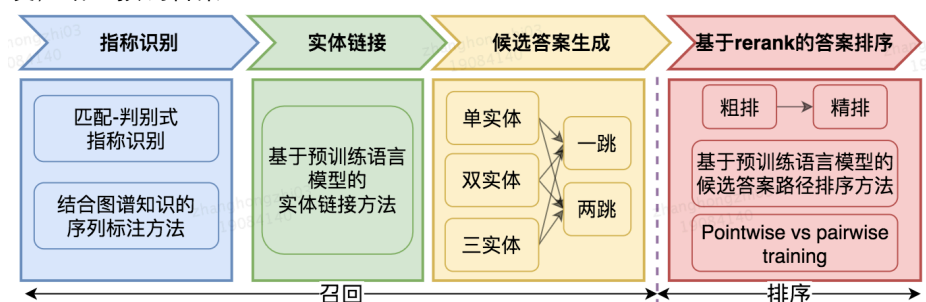
{zhanghongzhi03,lirumei,wangsirui,huangjianghua}@meituan.com

Abstract. 本文介绍了我们在 CCKS-2020 的 KBQA 任务上的技术方案。该系统包括指称识别、实体链接、候选答案生成以及答案排序四个子模块。在指称识别中, 为使识别到的实体指称更容易地被链接到图谱上, 提出了匹配-判别式指称识别方法和融合图谱知识的多粒度序列标注方法。在答案排序模块, 设计了重排序策略以降低计算量、缓解样本不平衡问题, 同时探索了多任务学习、带间隔的 pairwise 损失等训练策略。与传统的以人工特征为主的中文 KBQA 方法不同, 该系统主要基于预训练的语言模型完成指称识别、实体链接和候选答案排序。B(test)榜初次提交时, 本系统的 f1 值显著高于 A(dev)榜分数相当的系统, 该结果证明了所提方法的泛化能力。

Keywords: 知识图谱问答, 指称识别, 语义匹配, 预训练语言模型

1 引言

KBQA 系统基于知识图谱回答用户的自然语言问题。通常采用两种方法, 语义解析的方法直接将用户问题解析为可以在 KB 上执行的查询语句; 检索式方法中, 首先生成候选的答案, 然后对答案进行排序。本文采用检索式方法, 将答案查找过程建模为一个自然语言问句与结构化知识描述的语义匹配过程。整体框架包括召回和排序两大模块。其中召回模块生成候选答案, 该过程保持尽可能高的召回率; 排序模块利用预训练语言模型计算问题和候选答案之间的匹配度, 给出预测答案。



具体地，所提系统的架构如图 1 所示。其中，指称识别和实体链接模块识别出所提及的话题实体并将其映射到知识图谱中的实体节点（称之为话题实体）；然后子图召回模块以话题实体为中心进行漫游，生成候选答案以及从话题实体到候选答案的路径（称之为答案路径）；最终，答案路径排序模块，基于预训练的语言模型计算候选的答案路径与问题之间的语义匹配度、给出预测答案。为避免流水线模型中存在的错误累积问题，在前三个步骤中遵循轻准确、重召回的策略，将排序的压力留给最终的排序模块。实验也初步表明了该模式的合理性，也说明所设计和实现的基于预训练语言模型的答案路径排序模型是相对可靠的。

之前在中文 KBQA 上的工作[2]，主要基于人工特征和 Xgboost 分类模型进行实体链接和候选答案排序。其中也有利用 FastText 和 BERT 等深度学习方法，但只是作为一路语义特征。本文方法中，尝试不再进行手工特征提取，主要基于预训练的语言模型完成指称识别、实体链接和候选答案排序。

图 2 给出了从数据量级视角的系统流程图。整个过程可理解为多次候选生成与候选排序过滤的过程。如实体链接中，将一个实体指称映射为多个候选实体然后再进行判别。由于候选答案生成过程产生大量候选答案，采用了重排序策略。即首先经过一个轻量级的粗排模型进行过滤，然后采用较重的精排模型产生最终答案。如果不考虑计算量的问题，完全依赖精排模型应该可以产生还算满意的结果。在未来的工作中，可以考虑将精排以前的所有步骤，替换为向量哈希召回、简化知识图谱问答工作流程。

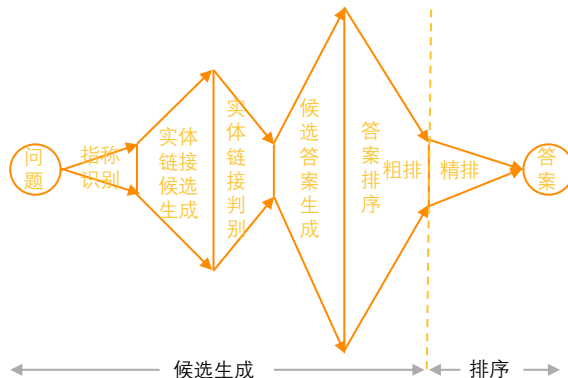


图 2 从数据量来看的系统流程图

2 方案介绍

2.1 实体指称识别模块

指称识别的目标是识别问题中提及的图谱实体节点的文本片段。为了保证召回率，我们实现了两个指称识别系统，取两系统输出的并集作为实体链接的输

入。其中，基于匹配-过滤的方法，主要用于召回实体指称精准提及的情况（即实体指称被包含在问题中的情况），而基于序列标注的方法可召回模糊提及的文本片段。作为 KBQA 系统的子模块，匹配-判别式方法所识别出的实体指称能够很方便的被链接到图谱中的实体，而基于序列标注的方法中，我们也引入图谱中的知识、尽可能识别出跟图谱中数据粒度相一致的实体指称。

1.匹配-判别式指称识别模块

首先采用文本匹配、规则提取和检索式召回等方法生成候选的实体指称。1) 文本匹配根据指称字典，采用最长匹配产出候选的实体指称。2) 将书名号和引号内的文本片段、连续英文和日期片段作为实体指称。仅采用以上方法无法召回模糊提及的实体指称，3)我们通过检索式方法召回一些模糊匹配的实体指称。具体地，构建实体指称的倒排索引，然后将问题作为查询词从而召回匹配度较高的实体指称。然后对于所召回的未被包含在问题中的候选指称，在问题中查找与该指称 Jaccard 相似度最高的文本片段作为候选指称。

由于知识图谱规模可能很大，甚至包含“<那么>”、“<哪些>”以及“<什么>”等实体。候选实体指称数目较多，因此我们通过一个指称判别模型进行过滤，产生多组候选实体指称。对于一个问题中由上述方法提取的 M 个实体指称，我们首先产生 M 个单实体指称的结果，然后从原来 M 个任选两个候选指称产生 C_M^2 个双实体候选集合。指称判别模型借鉴[4]中的实体位置标记方法，在候选指称的开始和结束位置加入“#”号作为实体位置标志位，然后用 BERT 模型做文本分类判断当前的实体指称是否正确。下表中给出了几个指称判别模型的输入示例，以及相应的标签。该模型的训练数据基于给定的标注数据由规则生成。预测时，选取得分前五的五组实体指称识别结果。

实体数	候选指称组合	判别模型输入	标签
单实体	叔本华	#叔本华#信仰什么宗教？	1
双实体	叔本华、宗教	#叔本华#信仰什么#宗教#？	0
单实体	村上春树	#村上春树#的哪部作品出版于 2007 年 7 月？	0
双实体	村上春树、 2007 年 7 月	#村上春树#的哪部作品出版于#2007 年 7 月#？	1

表格 1 候选实体指称判别样本示例

此外，我们训练一个二分类模型，判断当前问句是单实体问句还是多实体问句。如果模型以较高的置信度预测为单实体，则每个句子仅生成单实体指称组合。由于大部分问题是单实体问题，该策略可有效减小候选的数目。

2.融合图谱知识的多粒度序列标注方法

在指称识别中，由于指称类型丰富、粗细粒度混杂。指称识别不仅需要足够的泛化能力，而且需要使指称识别结果粒度与图谱中实体粒度尽可能一致，因此我们提出了结合图谱知识的多粒度指称识别模型。

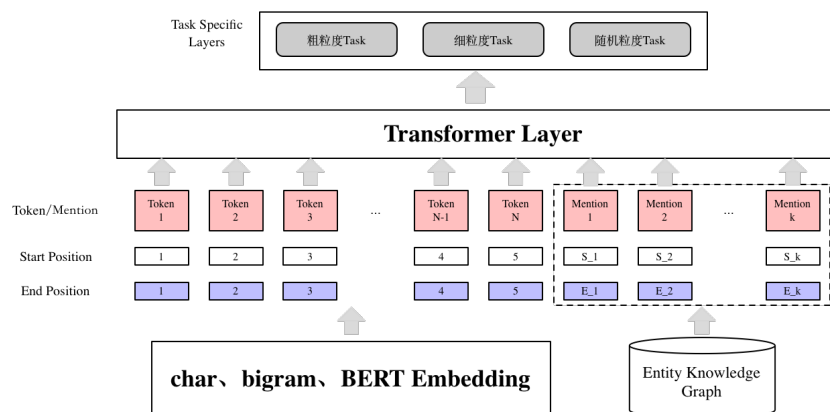


图3 结合知识图谱的NER方法框架

基于Transformer的序列标注。将指称(mention)识别建模为序列标注任务，利用Transformer作为问题的序列建模层。Transformer不仅接收token级别的embedding特征、BERT embedding，也整合了图谱实体匹配特征和位置特征，改变attention的计算方式以便于模型适应这种不同维度的信息。Transformer的输出提供给不同粒度的任务做序列标注建模。

图谱知识的融合。1) 候选指称的获取：将query分成n-gram片段，与mention-entity词典做匹配，保留匹配到结果的n-gram作为候选指称，这部分指称能链接到图谱中的实体，能为指称的识别提供特征。2) Lattice特征的融合：由于多个候选指称之间可能相互包含、交叉、覆盖，Transformer模型无法直接使用这种lattice的指称特征，我们采用FLAT-Lattice NER[1]模型的做法，提取每个指称、token的起始位置和终点位置，利用相对位置编码建模指称的position embedding，连同指称本身的word embedding一起输入Transformer。3) Token特征的增强：每个token，我们同时结合了char embedding、bigram-embedding和预训练的BERT-wwm embedding（使用的embedding和FLAT-Lattice NER一致），再用一层全连接映射确保三种embedding融合后的维度和lattice特征相一致。

多粒度弱监督训练数据生成与基于多任务学习的多粒度指称识别。由于比赛数据集中没有给出问题中的指称标注，我们基于规则生成了弱监督数据。多粒度弱监督数据生成步骤如下：一个实体可以被多个指称召回，如实体“<清华大学>”的实体指称包括“清华”和“清华大学”等。我们生成了三种粒度的训练数据：最长粒度、最短粒度、随机粒度。最长粒度就是选择实体对应最长的指称，最短粒度就是选择实体对应最短的指称，随机粒度会在实体中对应的指称中随机选择一个。最后我们通过BMEOS对token进行打标。

基于多任务学习的多粒度指称识别模型。为了让模型输出多粒度的指称，我们采用多任务学习框架训练不同粒度的指称识别模型，包括三个子任务：粗粒度、细粒度和随机粒度的指称识别。三个子任务共享特征提取层的参数，在Transformer层提取的语义特征上添加不同的线性任务层。在训练过程中，每种

粒度的数据训练对应的任务，共同调节共享层参数。预测时，解码出三组不同粒度的预测结果。

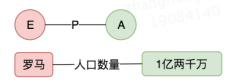
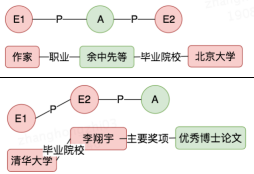
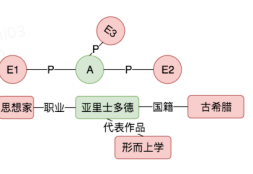
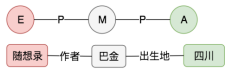
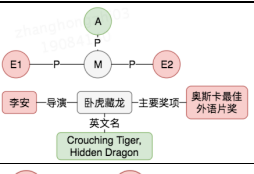
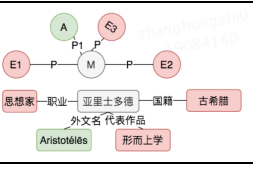

2.2 实体链接

给定指称识别的结果，基于词典和倒排索引将实体指称映射到知识图谱中的节点，生成实体链接的候选。此外，我们对日期字符串进行了特殊处理，将日期类指称串格式化成标准日期格式进行候选实体召回。

采用实体链接模型对候选实体进行排序过滤。为简化实现、避免手工特征提取，我们同样采用了基于预训练的语言模型进行语义匹配度计算。实体链接模型的输入为问句、识别到的指称（一个或多个）以及指称对应的候选实体拼接成 Question[SEP]Mention1\$Entity1#Mention2\$Entity2 的形式，其中多个指称用“#”号分隔。然后保留得分大于一定数值的所有实体链接候选。以往工作，如[2]中仅保留 top2 实体，而我们这里尽可能的保证召回率，将排序工作留给最后的答案排序模型。

2.3 候选答案和答案路径生成

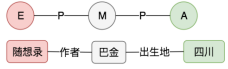
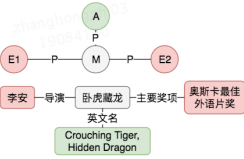
给定所识别到的实体，以实体为中心“画圈”召回候选答案，同时产生话题实体到答案实体的路径作为答案排序的依据。下表给出了具体的单跳和两跳候选答案与关系路径的生成策略。

	单实体查询	双实体查询	三实体查询
一跳			
两跳			
关系查询			

表格 2 候选答案生成策略，其中红色为输入实体，绿色为候选答案

我们采用了一些简单的剪枝策略避免候选答案爆炸，如果某一跳节点 M 的第二跳节点数目超过 500 时，不再对该节点进行二跳遍历。如“<人物>”、“<中华人民共和国>”等，其中有太多实体的“<类型>”和“<国籍>”指向该节点，对该节点进行二跳遍历，将产生大量候选。此外，在两跳查询中，我们删掉了两关系名相同但反向的候选答案路径，避免将话题实体本身作为答案。

最终，我们基于规则将召回的路径展开为一个文本序列。用\$表示答案节点，用@表示中间节点，下表中给出了部分示例。

	单实体查询	双实体查询
两跳		
展开序列	<随想录>#<作者><出生地>\$	@<导演><李安>@<主要奖项><奥斯卡>@<英文名>\$

表格 3 候选答案路径展开为文本序列示例

2.4 答案路径排序模块

答案排序问题建模

答案排序模块计算用户问题 q 与答案路径 p_i 的语义匹配度 $s(q, p_i)$ ，选择得分最高的答案路径 p_i 对应的问题作为答案。我们采用 BERT 进行语义相似度的计算。将 q 和 p_i 拼接后 $[CLS]q[SEP]p_i$ 输入到 BERT 中提取语义相似度特征，最后基于输出特征计算相似度值。

答案路径排序与实体链接联合模型基于所提取的语义交互特征，一方面判断关系路径是否正确，一方面判断当前话题实体是否正确。

pointwise 和 pairwise 模型训练

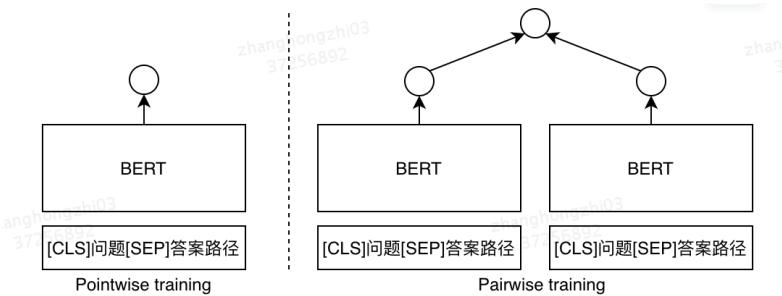


图 4 pointwise 和 pairwise 模型训练

采用了不同的训练策略进行模型训练。对于问题“乐山大佛始建于哪一年？”的正确 path 是“<开凿时间>”。但是采用 pointwise 损失训练出来的模型，会给该 path 一个很低的分数 $10e-6$ 量级，候选路径“<乐山大佛>#<大佛通高>”的得分约为 10^{-3} 量级。Pointwise 策略训练出来的模型，对排序偏后的样本难以给出合适的排序分数。分析原因，pointwise 模型训练中，模型主要

损失来自相似但非正确答案样本的得分。即因为训练时，如果问题是问“出生时间”，模型会花很多时间“学习”给“逝世时间”一个极低的分数。而我们希望，top1 之后的，看似不匹配的情况下，能给出一个合理的排序分，即学到“开凿时间”作为正确答案的概率应高于“大佛通高”。因此，也采用了间隔化的成对损失函数来进行模型训练。具体地，我们采用带有软间隔的损失函数，只要某负样本与正样本的间隔大于 γ 时，就不再惩罚

$$\text{loss} = \max(0, \gamma - s(q, p^+) + s(q, p^-))$$

重排序策略

对一个问题，将生成数千条候选的答案路径。如果直接训练，一方面正负样本比极不均衡，另一方面模型训练将极其耗时、无法训练 BERT-large 等大模型。因此，我们首先训练一个小模型（BERT-base 6 层裁剪），采用该模型为每个问题选出 topk 的候选答案路径。将正负样本比例控制在 k:1，同时降低了训练数据的规模。然后在新的训练数据上进行 Bert-large 等模型的训练。

预测过程与之类似，先采用一个轻量级模型进行粗排，选择 topN 样本输入到一个或几个集成的大模型中进行重排序。

3 实验

3.1 实验数据与实验设置

在 CCKS 新冠知识图谱构建与问答评测任务(四)的数据集上进行了实验验证。数据集包括全领域、金融和医疗领域的训练数据。训练集 4000 条、验证集 1529 条、测试集 1599 条。所给出的知识图谱中包含 6600 万三元组，两千多万实体。验证集中，单实体、双实体和三实体问题的大约分别有 1140、227 和 108 个。

如非特殊说明，本文采用了该仓库中的 BERT 模型¹。在实体链接句对判别任务上，首先将 BERT 模型在开源的 LCQMC 数据集上进行预精调，然后在实际任务上进行模型训练。采用 ElasticSearch 搭建的倒排索引。

¹ https://github.com/brightmart/roberta_zh

3.2 实验结果与分析

总体结果，B(test)榜初次提交时，本系统的 f1 值显著高于 A(dev)榜分数相当的系统，该结果证明了我们所提出方法的泛化能力。

	dev(2020.09.10)	test(2020.09.18)	test(2020.09.30)
see	0.91004	0.78269	0.85474
Artemis	0.90863	0.69448	0.86078
MiQA	0.88569	0.69143	0.85453
SAIL-PZLZ	0.89574	0.68169	0.74396

表格 4 与其他参赛系统的结果对比

3.3 各模块实验分析

指称判别模块

指称判别模型的准确率相对尚可，但发现一类较为明显的问题。当候选的实体指称较长时，如超过 8 个字符时，正样本给分较低会导致一些漏召，需通过规则进行调整得分。总体而言，在实体的开始和结束位置加标志位总体上是一种简单可行的方法，但是在长文本片段的标记中也面临一些待解决的问题。

结合图谱知识的 NER 方法

下表给出了不同策略时，在 dev 集合上结合知识图谱的 NER 方法 F1 值的变化。

模型	F1
BERT	81.92%
FLAT-Lattice+char+bigram	83.43%
FLAT-Lattice+char+bigram+BERT	88.54%

表格 5 结合图谱知识的 NER 方法

引入基于训练标注的方法将验证集上实体链接的召回率从 94.6%提高到 97.8%。

答案路径排序模型

对答案路径排序模型的各个机制进行了消融实验，不同配置下 dev 集合 F1 值结果如下表所示。我们首先测试了重排序策略对实验结果的影响，直接在原始的训练数据上进行模型训练，F1 降低近 3%，且训练过程极为耗时。在预测阶段，直接采用大模型先进排序，能够带来约 0.1%的提高，但是预测时间约为原来的数倍。采用带有间隔损失的 pairwise 排序函数，能带来约 2%的提高。同时也观察到模型对相似度较低的候选，能够给出更为合适的分数。如“林建华毕

业于哪个学校？”这个问题，采用 pointwise 和 pairwise 损失训练出的模型对三个候选关系的排序分别为“毕业院校>出生日期>任教院校”和“毕业院校>任教院校>出生日期”。这里“任教院校”虽然不是正确答案，但是其作为正确答案的概率应高于“出生日期”。将答案路径排序任务与实体链接任务进行联合训练能够带来小幅提高。也观察到，BERT-large 模型显著优于 Bert-base 模型，该结果表明预训练过程中学习到的语言知识能够被迁移到答案路径排序任务中。

模型	Dev 集 F1
最佳单模型结果	88.51%
- w/o rerank (训练阶段)	85.60%
- w/o rerank (预测阶段)	88.63%
- w/o pairwise 训练	86.43%
- w/o 实体链接多任务训练	87.65%
- w/o BERT-Large (w BERT-base)	83.68%

表格 6 答案路径排序消融实验结果

3.4 错例分析

从验证集随机抽取了 62 个错误样本进行归类分析，错误归因如下表所示。

错误类型	数目	细分类型	举例说明
实体错误	2	—	赵薇林心如范冰冰主演的经常 重播的神剧 是？
子图召回错误	11	三跳查询	慕容云海的女朋友的初恋的心上人是？
		union 操作	著名的亚历山大大帝的亲人有哪些？
rank 错误	27	实体错误	五指峰景区是在哪里呢？ <五指峰_ (湖南张家界五指峰) > <五指峰_ (江西省上饶县五指峰) > 带鱼科的动物一般分布在哪里？ 带鱼科被链接到了<带鱼>
		关系错误	小儿腹泻散属于什么药种？ <类型> <药品类型> 谁的徒弟自创了太极拳？ \$<徒弟>#<自创>#"太极拳" @<自创>#"太极拳"@<徒弟>\$
		实体&关系错误	张翠山在冰火岛认识了谁，后来有了儿子？ 维力医疗哪个高管曾任韦士泰董事长
后处理错误	14	—	防风通圣丸是中药还是西药？

			过滤了答案中非中药和西药的节点。
多个正确答案	8	—	"负荆请罪"说的是谁的故事？ ?x <典故> <负荆请罪> <负荆请罪>#<相关人物>\$

表格 7 验证集错例归类分析

参考文献

1. Li X, Yan H, Qiu X, et al. FLAT: Chinese NER Using Flat-Lattice Transformer[J]. arXiv preprint arXiv:2004.11795, 2020.
2. 骆金昌, 尹存祥等, 混合语义相似度的中文知识图谱问答系统
3. Liu X, Chen Q, Deng C, et al. Lcqm: A large-scale chinese question matching corpus[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 1952-1962.
4. Wu S, He Y. Enriching pre-trained language model with entity information for relation classification[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 2361-2364.