DUTIR中文开放域知识库问答评测报告

曹明宇 李帅驰 王鑫雷 杨志豪 林鸿飞

大连理工大学计算机科学与技术学院, 辽宁大连, 中国

caomingyu1997@mail.dlut.edu.cn

Abstract. 基于知识库的智能问答是自然语言处理领域的热门研究问题。本文使用一套流水线的方法，先对问题进行主语实体识别和属性值识别，将其链接到知识库中的实体，使用逻辑回归对候选实体进行筛选；进一步，抽取其两跳内关系作为候选查询路径，使用基于预训练语言模型的文本匹配模型选择与问题相似度最高的候选路径；最后使用桥接来得到多实体情况的查询路径。本文方法在CCKS2019 CKBQA测试集上达到了67.6%的F值

Keywords: 知识库问答, 文本匹配, 实体识别.

1 研究背景

基于知识库的智能问答（Knowledge Based Question Answering, KBQA）是自然语言处理领域的热门研究方向。知识库是知识的结构化表示，由三元组（主语, 谓词, 宾语）构成，表示实体和实体间存在的语义关系，例如：奥巴马出生在火奴鲁鲁，可以表示为：（侯赛因，出生地，火奴鲁鲁）。知识库问答的主要任务是给定自然语言问题，理解问题中包含的实体、语义关系和逻辑组合，到知识库中检索并返回答案。

目前知识库问答主要方法分为两大类。第一类是基于语义解析的方法，该类方法使用字典、规则和机器学习，直接从问题中解析出实体、关系和逻辑组合。Wang等人[1]使用序列标注模型识别问题中的实体，使用序列到序列模型预测问题中的关系序列，并使用答案验证机制和循环训练方式提升模型的性能，在英文多关系问题数据集WebQuestion上取得了先进水平。Hu等人[2]提出了一种状态转移的框架，设计了四种状态转移动作和限制条件，结合多通道卷积神经网络等多种方法，在英文复杂问题数据集ComplexQuestion上取得了最先进水平。基于语义解析的方法通常使用分类模型进行关系的预测，面临着未登录关系的问题，即训练集未出现的关系难以被预测出来。中文知识库包含数十万种关系，训练集难以覆盖如此庞大规模的数量，使得基于语义解析的方法在中文知识知识库问答（Chinese Knowledge Based Question Answering, CKBQA）上受到限制。

第二类是基于信息检索的方法，该类方法首先根据问题得到若干个候选实体，根据预定义的逻辑形式，从知识库中抽取与候选实体相连的关系作为候选查询路径，再使用文本匹配模型，选择出与问题相似度最高的候选查询路径，到知

识库中检索答案。Yu等人[3]提出了一种增强关系匹配的方法，使用二层BiLSTM与候选关系进行多层次的匹配，并使用关系匹配对实体链接结果进行重排序，在英文多关系问题数据集上取得了最先进水平。这类方法侧重于计算问题和候选关系的相似度，在关系选择上具有更好的泛化能力。近年来，预训练语言模型利用大规模无标注文本语料，学习更充分的上下文信息，在多个自然语言任务上都取得了性能的提升，对文本匹配存在可能的帮助。

在CKBQA任务上，Yang等人[4]提出了一种联合抽取实体的关系的流水线方法，在CCKS2018 COQA任务上取得了第二名的成绩。本文同样使用流水线式的方法：先进行实体和属性值识别，再进行实体链接，进而从知识库中抽取候选查询路径，使用文本匹配模型选择与问题最相似的候选路径，最后使用桥接技术探索多实体情况的可能结果。本文方法在CCKS2019 CKBQA测试集上达到了67.6%的F值。

2 模型与方法

本文方法主要包含以下模块：构建辅助词典、实体提及与属性值提及识别、实体链接与筛选、候选查询路径生成与文本匹配、桥接与答案检索。流程如图1所示。

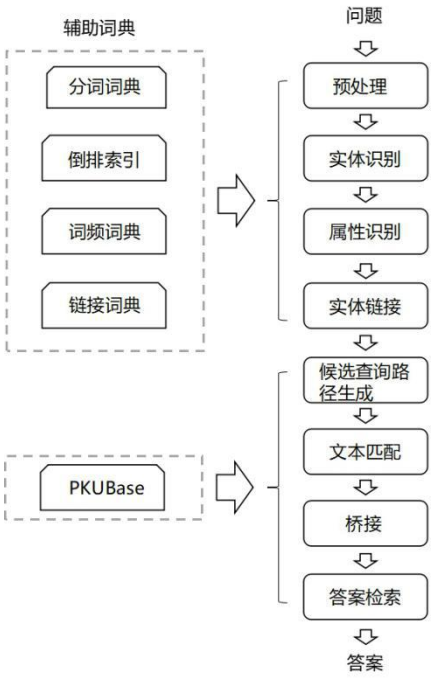

图1 本文方法流程图

2.1 辅助词典构建

本文方法在流程中需要多个词典用于分词、计算词频等，均来自于PKUBase知识库或外部资源，构建方法如下：

（1）实体链接词典：实体链接词典为文本中的实体提及到知识库实体的映射，由CCKS2019 CKBQA主办方提供；

（2）分词词典：分词词典参照Yang等人[4]的方法，通过实体链接词典中的所有实体提及，以及知识库中所有实体的主干成分构建。例如<红楼梦_（中国古典长篇小说四大名著之一）>这个实体，只保留下划线之前的部分“红楼梦”；

（3）词频词典：该词典用于计算实体提及和属性值提及的词频特征，使用搜狗开源的中文词频词典构建；

（4）倒排索引字典：该词典用于识别属性值的模糊匹配，使用知识库中所有属性值，构建字到词的映射。

2.2 实体提及和属性值提及识别

实体提及识别. 本文采用词典分词和神经网络模型结合进行实体识别。首先将分词词典导入分词工具，对自然语言问题进行分词，在分词词典中的词语都加入候选实体提及中。由于分词存在一定错误，且嵌套实体通常只保留较长的情况，比如问题“大连理工大学校歌是什么？”，正确的分词结果应当为“大连理工大学|校歌|是|什么|？”，但由于分词词典中存在更长的实体“大连理工大学校歌”，所以实际的分词结果为“大连理工大学校歌|是|什么|？”，进而得到错误的实体提及。针对这样的问题，本文基于预训练语言模型bert[5]，将训练集的标注实体还原为实体提及，训练了一个实体识别模型。将实体识别模型抽取出的实体提及同样加入到候选实体中。

属性值提及识别. 问题中包含的属性值规范性较低，可能是很长的字序列，也可能无法直接与知识库对齐，因此上述基于词典分词的方式不适用。本文针对不同类型的属性值，使用不同方式进行识别：

（1）书名、称号或数字：构建正则表达式，判断匹配结果是否在知识库的属性值中，在则加入候选属性值提及；

（2）时间属性：构建正则表达式，将其还原为知识库中规范的时间表达，如“1989年九月”还原为“1989.09”，加入候选属性值提及；

（3）模糊匹配属性：得到问题中每个字对应的所有属性值，统计每个属性值的次数，选择top3的属性加入候选属性值提及

2.3 实体链接及筛选

对于候选实体提及中的每个提及，首先判断其词性，过滤掉词性为语气词、副词等的提及。使用实体链接词典，将其对应的所有实体都加入候选实体中。对于候选属性值提及中的每个属性，由于抽取时已经与知识库对齐，将其直接加入候选实体中。平均每个问题初步得到的候选实体数量为13.6，多余的候选实体会引入干扰，同时增加后续步骤的时间成本。因此，参考Yang等人[4]的方法，对每个实体计算一些特征：

（1）实体提及的长度：该实体对应的实体提及的字数；
（2）实体提及的词频：该实体对应的实体提及的词频；
（3）实体提及的位置：该实体对应的实体提及距离句首的距离；
（4）实体两跳内关系和问题重叠词的数量；
（5）实体两跳内关系和问题重叠字的数量；

在训练集上，令标注的实体标签为1，其余候选实体标签为0，使用逻辑回归对上述特征进行拟合。在验证集和测试集上，使用训练好的逻辑回归模型对每个实体打分，保留分数排名前n的候选实体。

2.4 候选查询路径生成及文本匹配

在CCKS2019 CKBQA任务中，70%以上的问题只包含一个主语实体且最多包含两个语义关系，更复杂的情况也可以由简单问题桥接得到。因此，对于每个候选实体，抽取与其相连的单跳关系和两跳关系作为候选的查询路径，形式如（entity，relation）或（entity，relation1，relation2）。

传统上，文本匹配模型（如ESIM[6]）被用来学习自然语言问题和候选查询路径间的相似度，更侧重于学习同一语义不同表达间的相似性，需要大规模的语料作为支撑，模型的性能受到语料规模的约束。预训练语言模型可以依靠大规模的无标注语料，使用mask词预测、下文单词预测、句子对分类等无需人工标注的监督学习任务，学习到词级别、句子级别的信息。将预训练语言模型迁移到下游自然语言处理任务，作用类似于扩大了语料，增加了模型的性能和泛化能力。因此，本文基于预训练的bert[5]模型，使用训练集进行文本匹配的微调，在验证集和测试集上，使用该模型计算问题和候选查询路径的相似度。

在微调过程中，由于预训练模型是基于自然语言训练的，而将生成的候选查询路径是不符合自然语言逻辑的。因此，本文将候选查询路径还原为人工问题。例如，（侯赛因，出生地）被还原为“侯赛因的出生地？”。在训练集上，对于每个问题，随机选择三个候选查询路径作为负例，令标注的候选查询路径标签为1，负例的标签为0，将自然语言问题和人工问题拼接，训练一个文本分类模型。在验证集和测试集上，使用该模型对所有的自然语言问题-人工问题对进行打分。

2.5 桥接及答案检索

2.4节描述的方法只适用于单实体的情况，实际上，仍然有一部分问题包含两个及以上的主语实体，例如“北京大学出了哪些哲学家”。因此，本文采用桥接的方式，探索每个问题作为双实体问题的候选答案。

对于每个问题，首先对2.4节打分后的候选查询路径进行排序，保留前30个单关系的查询路径（entity1，relation1）。对于这些查询路径，到知识库中进行检索，验证其是否能和其他候选实体组成多实体情况的查询路径（entity1，relation1，ANSWER，relation2，entity2），将其加入候选查询路径中。最后，本文将2.4节单实体情况排名前三的候选查询路径和本节得到的双实体情况查询路径同时和问题计算重叠的字数，选择重叠字数最多的作为最终的查询路径，认为其在语义和表达上最与问题相似。

3 实验与分析

3.1 实验设置

本文实验基于CCKS2019 CKBQA数据集，是北京大学和恒生电子有限公司共同发布的中文开放域知识库问答任务。该任务中问题的标注SQL语句均来自于PKUBase知识库（http://pkubase.gstore-pku.com/）。数据集的数据统计如表1所示。

表1 语料集数据统计

问题类型	训练集	验证集	测试集
单实体单关系	1159	476	-
单实体多关系	682	156	-
多实体	356	133	-
总数	2297	765	765

3.2 实体链接结果

对于实体链接环节，由于测试集暂未提供标注的SQL语句，本文在验证集上针对5种特征进行了消融实验，并且记录了保留不同数量的候选实体的召回率。实验结果如表2所示，Recall@n表示在保留前n个候选实体情况下，所有问题标注实体的召回率。可以看出：（1）实体提及的特征和实体的特征对候选实体的筛选性能均有促进作用；（2）保留分数前5的候选实体，可以得到与保留全部候选实体接近的召回率，并且可以有效降低噪音和后续计算量。

表2 验证集上实体链接结果

特征	保留数量	Recall@n
mention+实体特征	全部（平均13.6）	92.3%
mention+实体特征	1	73.1%
mention+实体特征	3	84.6%
mention+实体特征	5	**89.7%**
mention+实体特征	10	90.5%
仅实体特征	5	81.1%
仅mention特征	5	53.4%

3.3 知识库问答结果

进一步，本文在验证集上，计算了文本匹配环节使用不同数量负例及不同检索方案答案的F值。本文对比了三种方案的性能：（1）直接选择文本匹配后相似度最高的查询路径；（2）对所有问题使用桥接获得可能的多实体情况查询路径，对于可以获得多实体查询路径的问题，直接覆盖方案一的路径；（3）对文本匹配排名前3的路径和多实体路径和问题重新进行重叠字数的匹配，选择字面上最相近的作为最终查询路径。

从表3的实验结果及分析可以得到：在文本匹配环节上，合适数量的负例可以获得更好学习文本相似性，本任务上3个负例效果最佳；桥接可以考虑多实体的情况，但会引入一些错误，即一些实际为单实体的问题得到了多实体情况的查询路径，而重叠字数匹配可以有效缓解该问题。

表3 验证集上知识库问答结果

检索方案	负例数量	F-score
仅文本匹配结果	3	56.7%
+桥接	3	58.6%
+桥接和字面匹配	3	61.5%
+桥接和字面匹配	1	61.1%
+桥接和字面匹配	5	59.4%
+桥接和字面匹配	10	55.3%

4 结论

本文提出了一套流水线式的模型，依次对问题进行实体及属性识别、实体链接及筛选、文本匹配和桥接等流程，并验证了预训练语言模型在知识库问答上的性能，在CCKS2019 CKBQA测试集上取得了67.6%的F值。

References

1. Wang, Y., Zhang, R., Xu, C. and Mao, Y., 2018, August. The APVA-TURBO Approach To Question Answering in Knowledge Base. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1998-2009).
2. Hu, S., Zou, L. and Zhang, X., 2018. A State-transition Framework to Answer Complex Questions over Knowledge Base. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2098-2108).
3. Yu, M., Yin, W., Hasan, K.S., Santos, C.D., Xiang, B. and Zhou, B., 2017. Improved neural relation detection for knowledge base question answering. arXiv preprint arXiv:1704.06194.
4. Li, Y., Miao, Q., Yin, C., Huo, C., Mao, W., Hu, C. and Xu, F., 2018. A Joint Model of Entity Linking and Predicate Recognition for Knowledge Base Question Answering. In CCKS Tasks (pp. 95-100).
5. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
6. Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H. and Inkpen, D., 2016. Enhanced lstm for natural language inference. arXiv preprint arXiv:1609.06038.