基于特征融合的中文知识库问答方法

汪洲 侯依宁 汪美玲 李长亮 *

AI Lab, KingSoft Corp, Beijing, China
{wangzhou1, houyining, wangmeiling1, lichangliang }@kingsoft.com

Abstract. 知识库问答即 KBQA 是自然语言处理领域的热点、难点问题。本文提出一种基于特征融合的中文知识库问答方法。此方法的 pipeline 主要由 mention 识别、实体链接、问句分类、路径生成、路径排序、答案检索六个部分组成，其创新点在于采用了多种特征融合策略，通过充分融合各个阶段挖掘的语义信息和覆盖浅层、深层的多级特征，从知识库中更加精准地召回答案。本文方法在 CCKS 2020 CKBQA 测试集 TestB 达到了 86.078% 的 F1 值。

Keywords: 知识库问答，实体链接，路径排序，特征融合.

1 引言

基于知识库的问答（Konwledge Based Question Answering, KBQA）[1] 是自然语言处理领域的热门研究方向。知识库问答的主要任务是接收一个自然语言问句为输入，识别问句中的实体、理解问句的语义关系、构建关于实体和关系的查询语句，进而从既有知识库中检索答案。例如针对问句 “龙卷风的英文名是什么?”，基于知识库中 [< 龙卷风 _（一种自然天气现象）> < 外文名 > “Tornado”] 三元组形式的知识，知识库问答给出答案 “Tornado”。

知识库问答的主要方法包括基于语义解析 (Semantic Parsing, SP) 的方法和基于信息检索 (Information Retrieval, IR) 的方法两大类。基于语义解析的方法直接从自然语言问句中解析出实体、关系及逻辑组合，转化为知识库上的查询语句并从知识库查询返回答案。例如 Wang 等人 [6] 利用序列标注模型解析问句中的实体、利用端到端模型解析问句中的关系序列。基于语义解析的方法通常依赖大量人力进行关系分类的标注，难以预测训练集中未出现的关系。基于信息检索的方法在问句实体识别与实体链接的基础上，从知识库中召回候选实体相关路径，并依据与问句的语义匹配进行路径排序，进而选择最可能的路径从知识库中检索答案。例如 Yu 等人 [8] 提出增强路径匹配的方法，实现问句与候选路径的多层次匹配。相比于基于语义解析的方法，基于信息检索的方法在路径选择方面具有更好的泛化能力，能够应用在较大的知识库中。

2020 年，新型冠状病毒疫情爆发，OpenKG 搭建了以新冠为核心的高质量知识图谱，并提出新冠知识图谱构建与问答相关的四个测评任务。其中，任务四 “新冠知识图谱问答评测” 要求针对输入的中文问句从给定知识库中选择若干实体或属性值作为答案返回，问句覆盖简单类型与复杂类型，例如单实体多度问句、多实体问句。本文针对任务四提出了一种基于特征融合的中文知识库问答方法，此方法采用基于信息检索的方法实现，该方法的 pipeline 主要由 mention 识别、实体链接、问句分类、路径生成、路径排序、答案检索六个阶段组成。为了从知识库中更加精准召回答案，本方法采用了多种特征融合策略，具体是在路

径排序阶段融合了实体链接和路径生成阶段挖掘的语义信息，在实体链接、路径排序等阶段融合了词法、句法等浅层特征和基于 Bert[3] 的深层特征。本文方法在 CCKS 2020 CKBQA 测试集 TestB 达到了 86.078% 的 F1 值。

2 方法

本文方法的 pipeline 主要包括 mention 识别、实体链接、问句分类、路径生成、路径排序、答案检索六个阶段，如图 1所示。其中，mention 识别阶段识别问句中出现的实体和属性 mention，实体链接阶段将 mention 链接到知识库中的候选实体并排序，句子分类阶段判断问句的类型，例如单实体问句或多实体问句，路径生成阶段从知识库中召回候选路径并通过路径相似度计算进行筛选，路径排序阶段针对单实体问题的候选路径进行重排序，答案检索阶段构造查询语句并从知识库中检索答案。

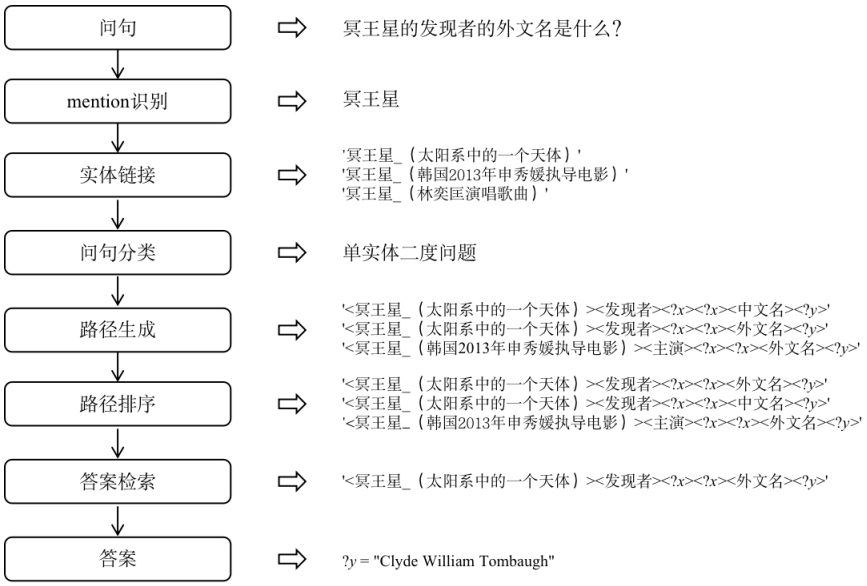

Fig. 1. 方法流程图.

2.1 Mention 识别

Mention 是指出现在问句原文中的命名实体和属性，本文的 mention 识别融合了模型、词典、规则三种策略。

(1) 模型识别：基于 Bert+CRF[7] 训练 2 个 mention 识别模型，其一是在评测的训练数据上训练识别实体边界的模型，其二是在百度 lic2019 数据[1]上训练

[1] 2019 语言与智能技术竞赛信息抽取任务数据集：http://lic2019.ccf.org.cn/kg

识别人名和机构名的模型，召回人名和机构名 mention，以提升 mention 的召回率。

(2) 词典识别：基于链接词典和知识库构建实体词典、基于知识库构建属性值词典，在此基础上通过与词典最大匹配识别问句中的实体 mention 和属性值 mention。

(3) 规则识别：针对数字、日期、书名等特殊实体和属性值，通过正则匹配进行 mention 识别。

预测时，针对每个输入问句，首先分别进行模型识别、词典识别、正则识别，之后以模型识别结果为主，利用词典识别结果和正则识别结果加以补充。

2.2 实体链接

实体链接将 mention 识别结果链接到知识库中的实体。本文的实体链接通过两个步骤实现：

(1) 候选实体召回：针对 mention 识别结果中的每个实体 mention，基于实体链接词典召回其所有链接结果作为候选实体，属性 mention 因已与知识库对齐直接加入候选实体中；

(2) 候选实体排序：考虑到噪音实体会干扰 pipeline 后续阶段的结果，而且候选实体过多会导致后续计算耗时过长，因而对所有候选实体排序并选择 top N_l 候选实体用于后续的计算。候选实体排序融合了基于 Bert 的深层特征和词法、句法等浅层特征，具体：

- Bert 模型：将每个候选实体与其全部一度关系拼接，使用 Bert-1 模型计算与问句的相似度得分，为候选实体排序；
- LightGBM[4] 模型：融合 mention 的浅层特征包括 mention 长度、与疑问词距离、在问句中的位置、流行度等，候选实体的浅层特征包括候选实体长度、与问句的重合度、在知识库中的流行度、两度内关系与问句的重合度等，以及通过 Bert-2 模型计算的候选实体与问句的深层语义相似度特征；
- 最后以 Bert 模型得分与 LightGBM 模型得分相加的结果排序候选实体。

2.3 问句分类

经统计，在评测训练数据中包含 57% 的单实体一度问句、17% 的单实体二度问句、15% 的多实体问句。因而本文在问句分类阶段基于 mention 识别结果、实体链接结果和问句分类模型对问句类型进行如下判断：

(1) 首先根据 mention 数量和实体链接结果的 mention 数量判断问句是单实体问句、二实体问句或三实体问句；

(2) 之后针对判断为单实体类型的问句，基于 Bert 分类模型进一步判断问句的度数，即判断问句类型具体为单实体一度问句或单实体多度问句。

2.4 路径生成

在判定了问句类型之后，路径生成阶段根据候选实体从知识库中召回候选路径并进行筛选，通过路径召回和路径筛选两个步骤实现。

在路径召回步骤，训练数据的问句类型及相关头尾实体的统计结果如下：

(1) 单实体一度问句共 2301 条，其中作为头实体的问句 2131 条、作为尾实体的问句 153 条；
(2) 单实体二度问句共 695 条，其中实体在一度关系头 402 条、实体在一度关系尾 263 条；
(3) 多实体问句共 944 条，其中二实体一度 596 条、三实体一度 184 条。

基于如上统计结果，针对每个候选实体召回如下 4 种路径，训练数据的召回率可达 94.6%：

(1) 作为头实体的一度路径 <实体><关系><?x>；
(2) 作为尾实体的一度路径 <?x><关系><实体>；
(3) 对 (1) 扩展至二度路径 <实体><关系1><?x><?x><关系2><?y>；
(4) 对 (2) 扩展至二度路径 <?x><关系1><实体><?x><关系2><?y>。

在路径筛选步骤，利用 Bert 模型计算候选路径（按邻接关系拼接实体与关系）与问句的相似度分数，并对召回的路径进行筛选：

(1) 针对单实体一度问句和多实体问句，根据候选路径与问句的相似度分数，对单实体一度问句的 top 1 候选实体筛选出 top N_{selr} 候选一度路径，对多实体问句每个 mention 的 top 1 候选实体筛选出 top N_{me} 候选一度路径；
(2) 针对单实体多度问句，利用多模型分步筛选候选路径；例如对于单实体二度问句，针对 top 1 候选实体，首先利用模型 1 计算其一度路径与问句的相似度分数，筛选出 top N_{semr1} 候选一度路径用于扩展至二度路径，并利用模型 2 计算扩展出的所有第二度路径与问句的相似度分数并筛选出 N_{semr2} 个候选路径，之后将筛选出的 N_{semr1} 个第一度路径与 N_{semr2} 个第二度路径拼接出最多 $N_{semr1} \times N_{semr2}$ 个二度路径，使用模型 3 计算与问句的相似度分数并筛选出 top N_{semr} 为最终候选路径。

2.5 路径排序

路径排序阶段针对单实体类型问句的候选路径进行重新排序，并在重排序中融合了实体链接阶段的候选实体得分和路径生成阶段的路径相似度得分，此外还利用 LightGBM 模型融合如下特征：

(1) 候选路径的长度；
(2) 候选路径与问句的相同字数及字级别向量相似度；
(3) 候选路径与问句的相同词数及词级别向量相似度；
(4) 候选路径在知识库中的流行度，即在知识库中出现的次数；
(5) 候选路径是否直接出现在问句原文中；
(6) 候选路径与问句的 Bert 相似度分数；
(7) 候选实体与候选路径的相对位置，即出现在三元组头还是三元组尾；

融合三种得分后重新对候选路径排序，保留 top N_p 为最终候选路径。

2.6 答案检索

针对多实体问句，使用桥接的手段从知识库中检索答案。对于所有候选实体的 top N_{me} 个候选一度路径，依据候选路径的排序进行路径桥接，并剔除知识库检索答案交集为空的情况。

针对所有类型问句，通过扩充 top N_{es} 候选实体并依据 2.4 节扩充候选路径，以扩大答案检索的范围，缓解实体链接错误带来的的影响。

3 实验

CCKS 2020 新冠知识图谱问答评测任务组织方提供开放领域中文知识图谱 PKU-base 作为知识库，该知识图谱包含 66,499,920 条三元组、25,574,536 个实体、408,690 条关系，发布训练集 4000 条、测试集 TestA 1529 条。在模型训练过程中，本文从 4000 条训练数据中随机抽取 90% 样本作为训练集、10% 样本作为验证集用于模型调优以及超参数选择。本文实验基于 TestA 进行评价，最终答案的测评指标为 F1 值。本方法集成了 bert-wwm[2][2]、roberta[3][5]、roberta-large-pair[4] ，ernie[5][9] 等模型。本方法的实验参数设置如表 1所示。

Table 1. 实验参数设置.

参数	名称	值
N_l	实体链接候选实体数	10
N_{selr}	单实体一度候选路径数	5
N_{me}	多实体候选路径数	5
N_{semr1}	单实体二度的候选一度路径数	5
N_{semr2}	单实体二度的候选二度路径数	5
N_{semr}	单实体二度候选路径数	5
N_p	最终候选路径数	5
N_{es}	候选实体扩充数	3

3.1 实体链接评估

在实体链接阶段，评估 Bert-1 模型与 Bert-2 模型对候选实体排序结果的影响，实验结果如表 2所示，其中 $Recall@n$ 表示 top n 候选实体的召回率。从表 2中结果可知：

(1) Bert-2 模型相似度特征可以显著提升实体链接的排序结果，说明融合深层的语义特征有助于方法对问句和实体的理解；

[2] https://github.com/ymcui/Chinese-BERT-wwm
[3] https://github.com/brightmart/roberta_zh
[4] https://github.com/CLUEbenchmark/CLUEPretrainedModels
[5] https://github.com/PaddlePaddle/ERNIE

(2) Bert-1 模型得分可以提升实体链接结果，并且将 Bert-1 模型得分与融合 Bert-2 特征的 LightGBM 模型得分求和后可以得到最好的候选实体排序效果。

Table 2. 实体链接阶段实验结果.

Model	Recall@1	Recall@3	Recall@5
Baseline	62.7%	93.0%	93.9%
Baseline+Bert-2	79.0%	93.9%	94.5%
Baseline+Bert-2+Bert-1	80.4%	94.3%	94.7%

3.2 路径生成、路径排序与答案检索评估

本文从三个方面评估所提方法中路径生成、路径排序和答案检索阶段的有效性：

(1) 针对单实体二度问题，评估路径生成阶段利用多模型分步筛选候选路径对方法最终结果的影响，其中：
   - Our Model w/o Multi-model Filtering 表示所提方法去掉多模型分步筛选候选路径；

(2) 评估路径排序阶段实体链接的候选实体得分、路径生成的路径相似度得分、LightGBM 得分对方法最终结果的影响，其中：
   - Our Model w/o EL score 表示所提方法去掉实体链接阶段的候选实体得分；
   - Our Model w/o PG score 表示所提方法去掉路径生成阶段的路径相似度得分；
   - Our Model w/o LGBM score 表示所提方法去掉 LightGBM 得分；

(3) 评估答案检索阶段候选实体扩充数量 N_{es} 对方法最终结果的影响，其中：
   - Our Model with N_{es}=0 表示所提方法的候选实体扩充数量 N_{es}=0；
   - Our Model with N_{es}=1 表示所提方法的候选实体扩充数量 N_{es}=1；

实验结果如表 3所示：

(1) 针对单实体二度问句，在路径生成阶段去掉多模型分步筛选候选路径使方法的最终结果降低 0.6%，影响了 TestA 所有单实体二度问句的 5%；
(2) 针对所有问句，实体链接阶段的候选实体得分和 LightGBM 得分对最终结果的影响较大，分别为 1.1% 和 1.3%。
(3) 针对所有问句，候选实体扩充数量为 0 的 F1 分数降低 2.4%，说明加入候选实体扩充能够有效缓解实体链接错误带来的的影响。

Table 3. 路径生成、路径排序与答案检索实验结果.

Model	F1 score	∧
Our Model	90.8%	
Our Model w/o Multi-model Filtering	90.2%	-0.6%
Our Model w/o PG score	90.5%	-0.3%
Our Model w/o EL score	89.7%	-1.1%
Our Model w/o LGBM score	89.5%	-1.3%
Our Model with $N_{es} = 1$	90.5%	-0.3%
Our Model with $N_{es} = 0$	88.4%	-2.4%

4 总结

本文针对新冠知识图谱问答评测任务提出了一种基于特征融合的中文知识库问答方法，该方法采用了多种特征融合策略，经过 mention 识别、实体链接、问句分类、路径生成、路径排序、答案检索六个阶段，从知识库中返回问句的答案。本文方法在 CCKS 2020 CKBQA 测试集 TestB 达到了 86.078% 的 F1 值。

References

1. Cui, W., Xiao, Y., Wang, H., Song, Y., Hwang, S.W., Wang, W.: Kbqa: Learning question answering over qa corpora and knowledge bases (2019)
2. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., Hu, G.: Pre-training with whole word masking for chinese bert (2019)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in neural information processing systems. pp. 3146–3154 (2017)
5. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
6. Wang, Y., Zhang, R., Xu, C., Mao, Y.: The apva-turbo approach to question answering in knowledge base. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1998–2009 (2018)
7. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models. arXiv preprint arXiv:1910.11470 (2019)
8. Yu, M., Yin, W., Hasan, K.S., Santos, C.d., Xiang, B., Zhou, B.: Improved neural relation detection for knowledge base question answering. arXiv preprint arXiv:1704.06194 (2017)
9. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: Ernie: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129 (2019)