基于预训练语言模型的中文知识图谱问答系统

王鑫雷，李帅驰，杨志豪，林鸿飞，王健（大连理工大学，计算机科学与技术学院）

论文摘要

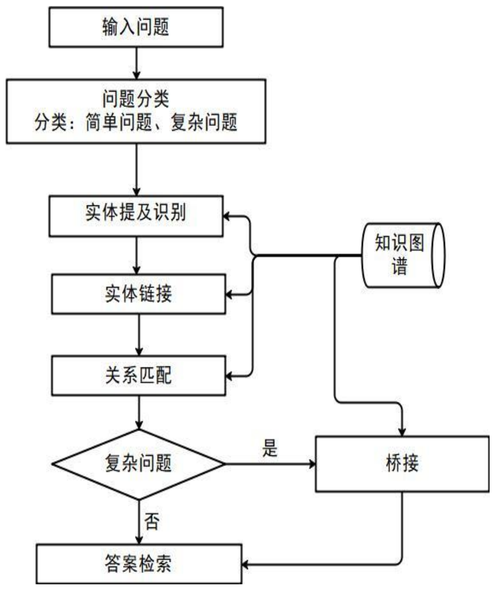
近年来，预训练语言模型在英文知识图谱问答研究中取得了令人瞩目的成绩。该文将预训练语言模型应用到中文知识图谱问答研究中，并通过实验结果分析不同模型及不同预训练语言模型的性能，验证了 ERNIE(Enhanced Representation from Knowledge Integration)语言模型更适合完成中文问答任务。同时，该文提出一套高效的流水线方法，在实体提及识别、实体链接、关系匹配子任务上提出新的框架来提升识别匹配结果，并在 CCKS2019-CKBQA 测试集上达到了 69.9%的 F1 值。最终基于该文方法在 web 端实现了知识图谱的问答系统展示，可回答大量开放域问题。

论文简介

该文基于不同预训练模型，设计了一套知识问答的流水线系统。将系统模块化为实体提及识别、实体链接、关系匹配模块。同时基于问句类型分类，将复杂问句的两种类型迭代化回答。

系统模型

模型如图，包含 以下五个模块：问题分类、实体提及识别、实体链接、关系匹配、桥接与答案检索。



算法原理

1.实体提及识别：分词（实体链接词典、分词词典、领域词典）
基于语言模型的实体识别
属性识别（正则式、时间属性、模糊匹配）
2.实体链接：实体提及特征
实体特征
3.关系匹配：基于语言模型的关系匹配模型
4.桥接及答案检索：多实体桥接结合字数重叠解决部分多实体复杂问题

实验仿真

表1 数据集划分

问题类型	训练集（条）	验证集（条）	测试集（条）
单实体单关系	1159	476	455
单实体多关系	682	156	140
多实体	357	134	171
总数	2298	766	766

表2 问答系统的F1值

模型	F1 值
Siamase	49.7%
BERT-match	64.6%
XLNet-match	64.9%
RoBERTa-match	65.0%
ERNIE-match	65.5%
ERNIE-match+桥接	67.8%
ERNIE-match+桥接+字面匹配	**69.9%**

图1 系统网站页面

论文结论

该文通过实验验证ERNIE语言模型更适合应用在中文图谱问答任务中。同时在实体提及识别、实体链接、关系匹配子任务上提出的新框架有助于高效精确的识别匹配结果。并通过在 CCKS2019-CKBQA 测试集上的结果验证方法的有效性，最后基于本文方法实现了问答系统展示。后续工作考虑对复杂问题进行语义解析，并融入知识图谱的全局信息增强问答性能。

西安电子科技大学 XIDIAN UNIVERSITY