问答系统项目提案

黄丹禹 [1)] 明静怡 [1)]

[1)](南开大学 计算机学院和网络空间安全学院 , 天津市 中国 300350)

摘 要 随着科技发展，问答系统的应用变得越来越重要，越来越多不同种类的问答系统逐渐出现。本项目计划对基于知识图谱的问答系统展开研究，结合国内外问答系统的发展现状，计划根据基于知识图谱的问答系统发展流程进行四篇论文的复现，分析其运行步骤、原理和结果，根据实验结果和已有研究讨论可能的创新点。最后，对小组的未来工作方向和问答系统的未来进行了讨论。

关键词 问答系统；知识图谱

Question and Answer System Project Proposal

Danyu Huang[1)] Jingyi Ming[1)]

[1)](College of computer science and College of Cyber Science, Nankai University, Tianjin, China)

Abstract

As technology advances, the application of question-answering system becomes increasingly important. A greater variety of question-answering system is gradually emerging. This project plans to conduct research on knowledge graph-based question-answering systems. Combining the current development status of question-answering systems domestically and internationally, the project intends to replicate four papers based on the development of knowledge graph-based question-answering systems. We will analyze their operational steps, principles, and outcomes. Aftert that, we will discuss potential innovative points based on experimental results and existing research. Finally, the project will engage in discussions about the future direction of the team's work and the future of question-answering systems.

Keywords Question answering system; knowledge graph

1 小组成员信息

- 黄丹禹

  学号：2012030

  邮箱：2012030@mail.nankai.edu.cn

- 明静怡

  学号：2013748

  邮箱：2013748@mail.nankai.edu.cn

2 问题定义

2.1 问答系统定义

问答系统（Question Answering System, QA）是信息检索领域的一种高级形式。这类系统的目标是用准确、简洁的自然语言回答用户提出的问题。广义上的问答系统不仅仅局限于回答通过自然语言提出的问题，还可以包括解决通过语音、图片等媒介提出的问题。随着科技的飞速发展，人工智能技术迅速崛起，人机交互的现实场景日益广泛。在互联网和大数据时代，如何快速、准确地获取所需的大量、正确的信息成为各信息检索平台研究的重点。问答系统在这一背景下具有重要意义。

随着技术的发展，多种多样的问答系统纷纷出现，问答系统可以根据不同的标准进行分类，本文主要从两个方面来介绍问答系统的分类方法。首先按照使用模态不同将问答系统分为单模态问答系统和多模态问答系统。单模态问答系统是指只使用一种模态（通常是文本）作为输入和输出的问答系统，而多模态问答系统是指使用多种模态（如语音、图像、视频等）作为输入和输出的问答系统。单模态问答系统又可以按照所涉及知识的领域、范围分为限定域问答系统和开放域问答系统。限定域问答系统是指只针对某一特定领域或主题的知识进行问答的系统，如医疗、法律、体育等，而开放域问答系统是指能够回答任何领域或主题的知识的系统。和单模态只涉及文本信息外，多模态问答系统结合了多种维度信息，比如语音、图像等参与输入输出和知识处理，并基于此对此大类问答系统的应用作简要说明。

其次，按照知识的信息来源进行分类，分为基于文本的问答系统和基于知识图谱的问答系统。

2.2 本文工作

本文主要在基于知识图谱的问答系统方向展开讨论与研究，提出问题定义和总述后对国内外问答系统的发展现状进行讨论。随后讲解小组计划复现的论文所涉及的系统运行步骤、原理和结果，并提出小组经讨论后目前的创新想法和技术路线简述。最后，对于小组的未来工作方向和问答系统的未来进行讨论。

3 目前国内外研究现状

知识图谱这一概念首先由 Google 在 2012 年提出，旨在将现实中接触到的物体、逻辑转为实体、关系成为结构化数据，并将此实体关系数据以图的方式被存储为可以通过计算机进行处理、计算的结构。

与基于非结构化文本数据的问答（例如文档检索、阅读理解）不同，基于知识图谱的问答系统利用知识图谱里高精度、高关联性的结构化知识，对复杂的事实型问题进行准确的语义理解或解析，然后在知识图谱里进行查询推理，得到简洁、正确的答案。

在此类问答系统中，知识图谱主要以三元组的形式被存储，每一条数据被存储为 (head, rel, tail) 这样的一个个三元组，比如在“南开大学位于天津”这一例子中，头实体可以为“南开大学”，尾实体为“天津”，rel 为“位于”。而“天津”这一实体，还存在例如“华北平原”这样的属性，那么便同时存在（“天津”，“属于”，”华北平原”）这一三元组，这边形成了简单的图，众多这样的三元组便构成复杂的知识图谱。

基于知识图谱的问答系统主要采用基于循环神经网络的端到端深度神经网络模型。该模型包括三个关键步骤：词向量转化、编码层和解码层。最初，基于简单的 RNN 网络，为了提高语义信息的表达，优化后的模型引入了注意力机制。注意力机制能够加强模型对上下文信息的处理，从而更好地理解输入问题的语义。

随后，在问答系统中引入外部知识库（知识图谱），在知识图谱中引入多跳机制进行逻辑推。多跳机制使得问答系统能够处理更复杂的问题，通过多步逻辑推理在知识图谱上获取答案。然而，多跳机制存在遍历三元组的劣势，限制了跳跃次数，可能导致准确率降低。

为解决多跳机制的劣势，可以通过设计得分函数对所有结点进行打分，最终选择最高分对应实体作为结果。这扩大了推理候选范围，但需要使用矩阵对实体、知识图谱和问题进行表达，引入了 embedding。

至此，基于知识图谱的问答系统已经可以进行简单的对话了，但是现有知识图谱往往对于时序复杂的推理问题表现不佳，故可以将时间这一属性单独提出，为模型和数据集引入时间属性，在涉及设

计时序问题时进行时间敏感的检索和推理从而提高模型的问答效果。

若将现有世界上的全部知识全部用一张知识图谱来表示，则难以想象此知识图谱的规模和大小，也具有极高的构建难度，所以现有的基于知识图谱的问答系统多用于特定领域，如医疗领域 [5]、金融领域等。

4 计划复现论文理解阐述

4.1 计划复现论文

本项目计划复现以下四篇论文：

(1) Zhou H, Young T, Huang M, et al. Commonsense knowledge aware conversation generation with graph attention[C]//IJCAI. 2018: 4623-4629.
(2) Madotto A, Wu C S, Fung P. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems[J]. arXiv preprint arXiv:1804.08217, 2018.
(3) Saxena A, Tripathi A, Talukdar P. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 4498-4507.
(4) Saxena A, Chakrabarti S, Talukdar P. Question answering over temporal knowledge graphs[J]. arXiv preprint arXiv:2106.01515, 2021.

4.2 论文理解阐述

在基于知识图谱的问答系统中，目前最主流的模型是基于循环神经网络并以此为基准进行变形形成端到端的深度神经网络模型，其主要包含三个步骤：词向量转化、编码层、解码层。基于循环神经网络的记忆功能，此类模型最开始仅基于简单的 RNN 网络，后续优化中 Zhou H[1] 等人于 2018 年年提出加入了注意力机制的 CCM 模型，该模型对于知识的记忆主要通过 GRU 网络，并使用了编码器-解码器模型。CCM 模型中主要原理为，通过对已有生成图（并非知识图谱，而是将问题中实体作为头结点生成的知识生成图）进行检索，即通过在问题中选取关键词后作为一次检索的头结点，将对现有知识图谱（各个三元组）得到的所有三元组和子图输入 GRU 网络，通过 GRU 的记忆功能拥有语义信息，最后在每个编码位置根据被检索到的图和每个图中的实体，在提前设置存储好的词汇表中生辰各一个实体作为答案。其加入的注意力机制对网络中的上下文信息给予权重，能够更好的反应语义信息。

随后 Madotto A[2] 等人基于此思路引入外部知识库，即知识图谱，提出 Mem2Seq 模型，并在模型中引入了多跳机制，即在知识图谱上进行多步的逻辑推理，从而得到问题的答案。和（“南开大学”，“位于”，”天津”），（“天津”，“属于”，”华北平原”）的例子类似，倘若用户提出“南开大学处于高原还是平原？”这一问题时，需要将“南开大学”作为推理的头结点，在检索到“位于”这条关系时，得到“天津”结点后，并不能就此结束，还需要将中间结果“天津”再次作为头结点进行推理，最后得到“华北平原”这一结点，此过程需要进行起码 2 次跳跃（实际需要华北平原属于平原这一属性，应该进行 3 次跳跃），结合外部知识图谱和引入多跳机制促使问答系统可以处理更加复杂的问题，给出更丰富准确的推理结果。

但多跳机制存在一些先天劣势，由于多跳这一动作需要进行对三元组进行大规模遍历，所以在实际推理过程中不能进行次数过多的跳跃，最常出现的跳跃次数多为 2-5，但倘若正确答案未在多跳结点范围内的子图中，则导致准确率降低。对此，Saxena A[3] 等人在 2020 年提出了 EmbedKGQA，其思想为为实体、知识图谱、问题都创建相应的 Embedding，并在推理过程中将所有实体都视为候选，通过对 embedding 的计算完成对每个结点的评价（通过得分函数），最后选出最高分对应实体作为结果，该方法大幅扩大了推理候选范围，但同时也存在一些问题，比如面对超大规模知识图谱，实体和关系过于复杂多样，导致计算压力过大，论文中也对此提及了一些解决办法，比如使用特定的模型或算法进行知识图谱剪枝等进行解决。

但现有问答系统在处理复杂的时序问题时往往出现时间偏差，导致回答准确率降低。比如“美国总统是谁？”这一问题在不同的聊天中会得到不同的答案，倘若我们的聊天基于 2012 年，则答案应该是“奥巴马”，但如果基于 2018 年提出问题，则答案应该为“特朗普”。Saxena A[4] 等人在 2021 年提出时序知识图谱数据集 CRONQUESTIONS，通过新增时间属性，在涉及时序推理时，对于知识图谱进行涉及时间依赖的查询检索，并基于此修改了 EmbedKGQA 模型，将时间敏感特性加入，提出新的模型 CRONKGQA，从而推理出准确时间的正确答案。

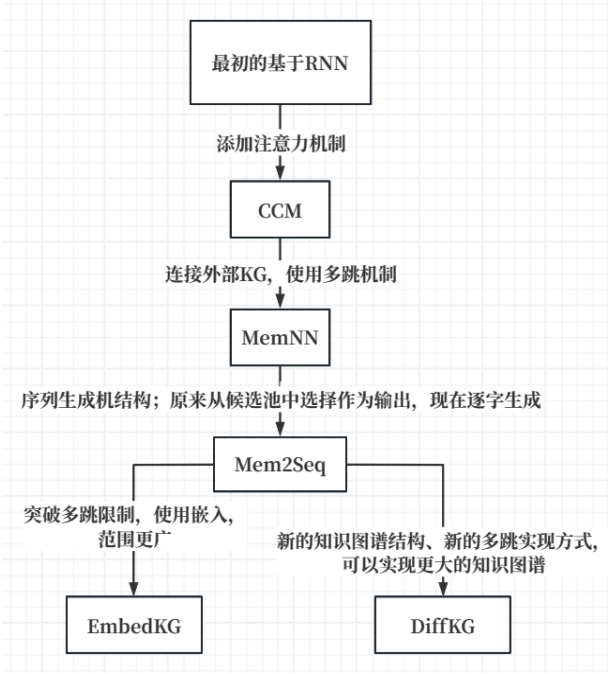

图 3 基于知识图谱的模型演变过程

5 创新想法和技术路线描述

小组计划在复现以上几篇论文的基础上，对最后的时序模型可以进行修改，如在编码层添加一些新的注意力机制，或使用更先进的预训练模型进行改进，如果进行编码操作。除此之外，值得注意的是，时序知识图谱的构建以及模型的训练需要高成本和高投入，在一般的普适问答系统中，涉及到的基于时序的复杂推理的问题其实占比并不高，所以可以针对不同的问题采用不同的知识图谱进行推理，考虑模型在不同时间段、不同领域之间选用不同知识库进行检索、推理。

6 未来工作及展望

现有的交互式问答系统已进入飞速发展阶段，这类问答系统可以与用户进行多轮的对话，以获取更多的信息或反馈，从而提供更准确和满意的答案。这类系统在 OpenAI 发布的 chatgpt4.0 中已经可以近乎完全做到，其可以根据用户的问题和上下文来提出适当的追问或澄清，例如询问用户的意图、偏好、需求等。

而自适应问答系统建立在交互式问答系统的基础上，这类问答系统可以根据不同的用户、任务和环境来动态地调整问答的策略和方式，以提高问答的效率和效果。这些系统可以根据用户的特征和行为来个性化问答，例如考虑用户的知识水平、兴趣爱好、情感状态等。可以尝试在历史对话中对用户生成用户画像并进行存储以在接下来甚至下次对话中进行参考、调整参数。

这些系统也可以根据任务的目标和难度来优化问答，例如选择合适的数据源、模型和算法等。这些系统还可以根据环境的变化和不确定性来适应问答，例如处理噪声、异常、冲突等。这些系统需要具备强大的自我学习和自我调整能力，以及灵活的决策和执行能力。

7 组内成员 Final Project 分工

项目计划由明静怡完成 CCM 模型和 Mem2Seq 模型的复现，对比分析两种模型的优缺点，黄丹禹完成 EmbedKGQA 模型和 CronKGQA 模型的复现，对比分析两种模型的优缺点。之后通过对复现结果进行讨论分析，共同思考可能的优化方法，结合已有文献探索可能的创新点。

参 考 文 献

[1] Zhou H, Young T, Huang M, et al. Commonsense knowledge aware conversation generation with graph attention[C]//IJCAI. 2018: 4623-4629.

[2] Madotto A, Wu C S, Fung P. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems[J]. arXiv preprint arXiv:1804.08217, 2018.

[3] Saxena A, Tripathi A, Talukdar P. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 4498-4507.

[4] Saxena A, Chakrabarti S, Talukdar P. Question answering over temporal knowledge graphs[J]. arXiv preprint arXiv:2106.01515, 2021.

[5] https://github.com/liuhuanyong/QASystemOnMedicalKG