两阶段文档筛选和异步多粒度图多跳问答

张雪松[1,2],李冠君[2],聂士佳[1,2],张大伟[2],吕　钊[1],陶建华[3]
(1. 安徽大学 计算机科学与技术学院,安徽 合肥 230601;
2. 中国科学院自动化研究所 模式识别国家重点实验室,北京 100190;
3. 清华大学 自动化系,北京 100084)

摘　要:多跳问答旨在通过对多篇文档内容进行推理,来预测问题答案以及针对答案的支撑事实。然而当前的多跳问答方法在文档筛选任务中旨在找到与问题相关的所有文档,未考虑到这些文档是否都对找到答案有所帮助。因此,该文提出一种两阶段的文档筛选方法。第一阶段通过对文档进行评分且设置较小的阈值来获取尽可能多的与问题相关文档,保证文档的高召回率;第二阶段对问题答案的推理路径进行建模,在第一阶段的基础上再次提取文档,保证文档的高精确率。此外,针对由文档构成的多粒度图,提出一种新颖的异步更新机制来进行答案预测以及支撑事实预测。提出的异步更新机制将多粒度图分为异质图和同质图来进行异步更新以更好地进行多跳推理。该方法在性能上优于目前主流的多跳问答方法,验证了该方法的有效性。

关键词:多跳问答;文档筛选;多粒度图;异步更新;答案预测
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2024)01-0121-07
doi:10.3969/j.issn.1673-629X.2024.01.018

Two-stage Document Filtering and Asynchronous Multi-granularity Graph Multi-hop Question Answering

ZHANG Xue-song[1,2],LI Guan-jun[2],NIE Shi-jia[1,2],ZHANG Da-wei[2],LYU Zhao[1],TAO Jian-hua[3]
(1. School of Computer Science and Technology,Anhui University,Hefei 230601,China;
2. State Key Laboratory of Pattern Recognition,Institute of Automation,Chinese Academy of Sciences,
Beijing 100190,China;
3. Department of Automation,Tsinghua University,Beijing 100084,China)

Abstract:Multi-hop question answering aims to predict the answer to a question and the supporting facts for the answer by reasoning over the content of multiple documents. However,current multi-hop question answering methods aim to find all documents related to the question in the document filtering task,without considering whether all these documents are useful for finding the answer. Therefore,we propose a two-stage document filtering approach. In the first stage,the documents are scored and a small threshold is set to obtain as many relevant documents as possible to ensure a high recall of documents. In the second stage,the inference path of the question answer is modeled,and the documents are extracted again based on the first stage to ensure high accuracy. In addition,we propose a novel asynchronous update mechanism for answer prediction and supporting fact prediction for multi-granularity graph composed of documents. The proposed asynchronous update mechanism divides the multi-grain graph into heterogeneous and homogeneous graphs to perform asynchronous updates for better multi-hop inference. The performance of the proposed method is better than that of the current mainstream multi hop question answering method,and the effectiveness of the proposed method is verified.

Key words:multi-hop question answering;document filtering;multi-granularity graph;asynchronous update;answer prediction

0　引　言

问答(Question Answering,QA)是自然语言处理中的一个热门话题。随着深度学习的蓬勃发展,QA模型已经取得了重大进展,甚至在简单 QA 基准测试

收稿日期:2023-02-25　　**修回日期**:2023-06-27
基金项目:国家重点研发计划(2020AAA0140003);浙江实验室开放研究项目(2021KH0AB06);北京市科委、中关村管委会计划(Z211100004821013)
作者简介:张雪松(1997-),男,硕士研究生,CCF 会员(N5869G),研究方向为多跳问答;通讯作者:陶建华(1972-),男,研究员,CCF 常务理事,研究方向为人机交互、模式识别。

中超过了人类[1]。然而大部分 QA 模型为单跳 QA,主要聚焦从单篇文档中寻找答案。当单篇文档不足以获得正确答案时,单跳 QA 通常缺乏从多篇文档中推理答案的能力。

为了提高 QA 模型在多篇文档中的推理能力,近年来,学者们提出了多跳 QA 模型,并且设计了多个专门用于评估多跳推理能力的多跳 QA 数据集。例如目前流行的 WikiOmnia[2], HotpotQA[3] 和 NarrativeQA[4]。这些数据集对应的 QA 任务十分具有挑战性,因为它们要求 QA 模型能够在多篇文档和文档噪声干扰下进行多跳推理,以获得问题答案。尤其是在 HotpotQA 数据集中,QA 模型除了需要预测答案,还需提供答案的支撑事实。图 1 显示了来自 HotpotQA 数据集的一个示例。除了“3,677 seated”这个答案外,HotpotQA 数据集还在文档中标注了答案的支撑事实句子来解释答案。

Question: The arena where the Lewiston Maineiacs played their home games can seat how many people?
Document1,Lewiston Maineiacs: [0]The Lewiston Maineiacs were a junior ice hockey team of the Quebec Major Junior Hockey League based in Lewiston, Maine.[1]The team played its home games at the Androscoggin Bank Colis.[2]They were the second QMJHL team in the United States, and the only one to play a full season.[3]They won the President's Cup in 2007.
Document2,Androscoggin Bank Colis: [0]The Androscoggin Bank Colis\u00e9e (formerly Central Maine Civic Center and Lewiston Colisee) is a 4,000 capacity (3,677 seated) multi-purpose arena, in Lewiston, Maine, that opened in 1958.[1]In 1965 it was the location of the World Heavyweight Title fight during which one of the mostfamous sports photographs of the century was taken of Muhammed Ali standing over Sonny Liston.
Answer: 3,677 seated
Supporting facts: [["Lewiston Maineiacs",1],["Androscoggin Bank Colis",0]]

图 1 HotpotQA 数据集的示例(只显示了两篇文档)

直观地说,如果一个问题需要通过多篇文档才能得到正确答案,通常模型需要两个步骤:(1)筛选文档;(2)在筛选出的文档中预测出问题答案并找到支撑事实。

在步骤 1 筛选文档任务中,大多数早期的工作要么将所有文档作为输入,要么单独处理文档,而不管大多数文档是否与问题有关,或者对找到答案是否有帮助。一个准确的文档筛选模块可以提高多跳 QA 模型的可伸缩性,且不会降低性能[5]。最近的工作包括 Tu 等[6]和 Wu 等[7],他们附加一个多头自注意力层(MHSA)来鼓励文档间的交互。Fang 等[8]设计一个级联文档检索模块,但其词法匹配引入太多的噪声,使得检索性能不佳。以上的工作旨在找到所有与问题相关的文档,然而并非所有的相关文档都对找到答案有所帮助。

为此,该文在相关文档的基础上进一步提取答案所需的支持文档。将筛选文档任务分为两个阶段。第 1 阶段通过处理每对文档间的信息来对文档进行评分并选出文档得分超过阈值 g 的文档作为相关文档,此外通过设置较小的 g 来尽可能多地获取文档,保证文档的高召回率;第 2 阶段训练一个递归神经网络(Recurrent Neural Network,RNN)对问题答案的推理路径进行建模,在第 1 阶段获取相关文档的基础上再次提取文档,保证文档的高精确率。

在步骤 2 找出问题答案和支撑事实任务中,已有工作证明,图神经网络(Graph Neural Networks,GNN)由于其关系表示能力和归纳偏差,非常适用于多跳 QA。根据 HotpotQA 数据集的特点,大多数工作从构建实体图的实体中选择答案,或通过将实体图的信息传播回文档表示从而在文档中选择答案。然而所构建的图大多仅用于答案预测,不足以发现支持事实。此外,上面的方法在图更新的步骤中同步更新所有节点,忽略了不同节点具有不同优先级以及有序逻辑推理。

因此,在步骤 2 中将筛选后的文档构建成一个包含问题、实体以及句子的多粒度图,并使用不同的粒度节点执行不同的任务(例如答案预测、支撑事实预测)。此外,该文提出了一种基于多粒度图的异步更新机制来更好地进行多跳推理。具体来说,将该更新分为两个阶段,首先是不同粒度节点之间的更新(例如问题-句子、问题-实体、句子-实体),使节点捕获与其不同粒度节点的线索完善自身信息;其次是相同粒度节点之间的更新(例如句子-句子、实体-实体),比较相同粒度节点之间的描述性信息以更好定位答案和支撑事实,最后将更新后的节点表示传递到预测模块,该模块预测问题的答案、答案类型以及支撑事实。

该文的贡献如下:

(1)提出一种新颖的两阶段筛选文档方法。第 1 阶段保证文档的高召回率,第 2 阶段保证文档的高精确率。

(2)针对由文档构成的多粒度图,提出一种新颖的异步更新机制来进行答案预测以及支撑事实预测,以更好地进行多跳推理。

(3)在 HotpotQA 数据集上进行了对比试验,验证了所提方法的有效性。

1 相关工作

1.1 文档筛选

对于多跳 QA 数据集来说,一个问题通常提供多篇文档,其中包含许多冗余的文档。当前的预训练语言模型一次接受所有文档作为输入通常是不可行的,

因为类似于 BERT,其输入的最大长度限制为 512。因此,文档筛选是必不可少的。文档筛选的目的是减少噪声信息,为下游任务生成高质量的上下文,即高召回率和高精确率的上下文。Qiu 等[9]使用 BERT 模型单独计算每个文档的相关性,忽略了文档之间的语义关系,因此会在筛选的文档中引入噪声,导致阅读理解任务的表现不佳。因此 Tu 等[6]和 Wu 等[7]附加了多头自注意力层(Multi-Head Self-Attention,MHSA)以鼓励文档交互,为了获得更优的结果,Wu 等[7]进一步采用一个级联文档筛选模块,该模块将所选的文档作为输入,随后探索它们之间更深层次的关系。此外,由于他们试图同时定位所有相关文档,导致简单的二分类器无法很好地执行。因此,Tu 等[6]和 Wu 等[7]将分类目标重新制定为排名和评分目标,以符合选择器的排名性质。与上面的方法不同的是,该文在相关文档的基础上进一步提取文档用于下游任务。

1.2　多跳 QA

多跳 QA 是机器阅读理解的一个特殊任务,是近年来自然语言理解领域的一个极具挑战性的课题,它更接近真实场景。HotpotQA 是最具代表性的多跳 QA 数据集,因为它不仅要从上下文中提取正确的答案,还需要提供答案的支撑事实。现有的多跳 QA 工作主要分为两大类:基于记忆检索的递归推理和基于 GNN 的多跳推理。

第一类专注于多跳问题分解[10],并在循环网络中通过问题和上下文的相互作用来更新潜在特征。Qi 等[11]通过迭代重排序检索系统查询缺失的实体。Asai 等[12]构建了一个带有超链接的离线维基百科图,构建推理链来进行多跳 QA。周展朝等[13]将问题分解视作一个阅读理解任务,更好地捕捉了多跳问题和文档之间的交互语义信息,以此指导多跳问题分解。杨玉倩等[14]提出了一种融合事实文本的问解分解式语义解析方法,对复杂问题的处理分为分解-抽取-解析三个阶段来进行多跳问题分解。

由于 GNN 具有固有的消息传递机制,可以通过图传播传递多跳信息,因此在找出问题答案和支撑事实方面有着巨大潜力[15]。对于多跳 QA 来说,基于 GNN 的多跳推理方法占主导地位。大多数的工作集中在单一层次的粒度表示上。Qiu 等[9]提出了一种动态融合图网络,沿着实体图动态探索,从上下文中寻找支撑实体。邵震等[16]在表征抽取层的神经网络隐藏部分使用参数共享和矩阵分解技术来降低模型的空间复杂度,使用点积计算方式进行答案预测。龙欣等[17]提出了一种多视图语义推理网络,该网络利用全局和局部两种视图的信息共同进行推理。与上述方法不同的是,该文侧重于多粒度图的异步更新。

2　两阶段的文档筛选和异步多粒度图更新多跳问答

在本章节中,将详细介绍文中方法。如图 2 所示,提出的模型由 5 个主要模块组成。

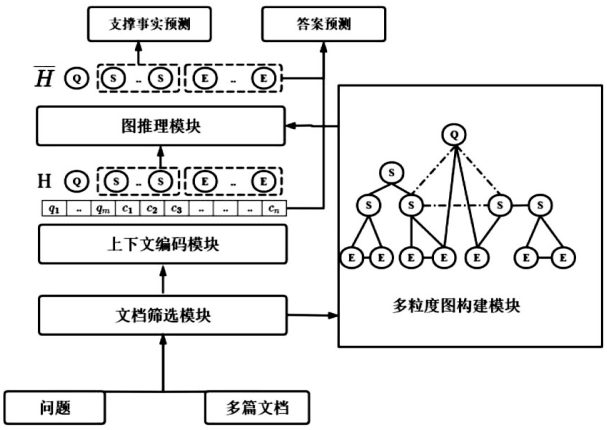

图 2　模型框架

(1)文档筛选模块(第 2.1 节),通过该模块筛选出高召回率和高精确率的文档,然后将其传递给下游任务;

(2)上下文编码模块(第 2.2 节),通过基于 RoBERTa 的编码器获得图形节点的初始表示;

(3)多粒度图构建模块(第 2.3 节),通过该模块构建多粒度图以连接来自不同信息源的线索;

(4)图推理模块(第 2.4 节),基于 GNN 的消息传递算法和异步更新机制用于更新节点表示;

(5)预测模块(第 2.5 节),该模块用来执行寻找支撑事实和预测答案任务。

2.1　文档筛选模块

在这个模块,目标是过滤干扰信息,为下游任务生成高召回率和高精确率的上下文。

如图 3 所示,在第 1 阶段中,对于每一篇文档,通过连接“[CLS]+问题+[SEP]+文档+[SEP]”来构建一个输入,供 BERT 使用。通过 BERT 对每个问题/文档对进行编码,提取代表全局表示的[CLS]标记作为每个问题/文档对的总结向量。该向量只包含了各个文档自身的特征,但是文档间存在一定的关系,所以通过 MHSA 让文档间的信息得到交互,再利用一个双线性层来输出每对文档的相关概率。其二元交叉熵损失如下:

$$\text{loss} = \sum_{i=0}^{n}\sum_{j=0}^{i} l_{i,j}\log P(D_i,D_j) + (1-l_{i,j})\log(1-P(D_i,D_j)) \tag{1}$$

其中,n 表示文档数量,i,j 表示第 i,j 篇文档,$l_{i,j}$ 为文档(D_i,D_j)的标签,若 D_i 是支持文档 $l_{i,j}$ 为 1 反之为 0。$P(D_i,D_j)$ 表示 D_i 比 D_j 更相关的预测概率。

最后利用评分器 SCORE 处理每对文档的信息从

而获到每篇文档的评分。为了后续 g 选取操作,将每篇文档评分控制在 0 到 100 分之间。

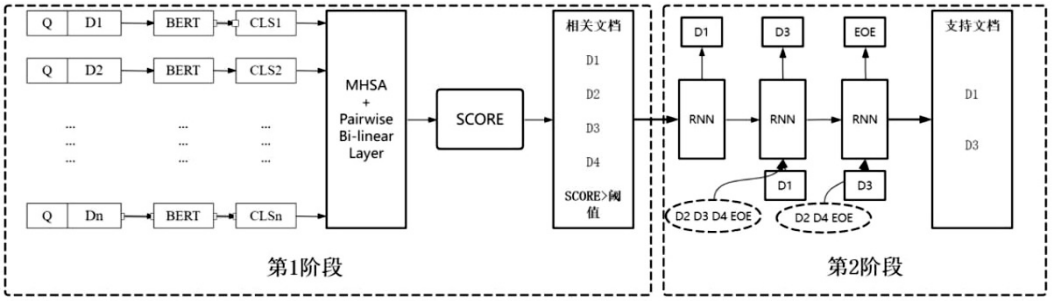

图 3 文档筛选模块示意图

$$\mathrm{Score}[D_i] = \frac{\sum_{j,j\neq i}^{n} P(D_i, D_j)}{\sum_{i}^{n}\sum_{j,j\neq i}^{n} P(D_i, D_j)} * 100$$

其中, $D = \{D_1, D_2, \cdots, D_n\}$, n 是文档的数量。

$$D_t = \{D_{t1}, D_{t2}, \cdots, D_{tM}\}$$

$$\mathrm{Score}[D_m] > g, m = \{1, 2, \cdots, M\} \tag{2}$$

选取集合 D_t 作为相关文档,其中 D_t 为 D 的子集($D_t \subseteq D$),集合 D_t 中每篇文档评分均大于 g 。

第 2 阶段主要是从第 1 阶段选取的文档中检索出获得问题答案所需的支持文档 y_{SD} 。受 Asai 等[12]的启发,该文在 D_t 中利用 RNN 和波束搜索来寻找最佳的支持文档路径。选择 y_{SD} 的过程如下所示:

$$w_i = \mathrm{BERT}_{[CLS]}(Q, D_i)$$

$$h_t = \sigma(Wh_{t-1} + Uw_i + b_h)$$

$$O_t = Vh_t + b_o \tag{3}$$

其中, h_t 是 RNN 在第 t 个推理步骤的隐藏状态, σ 是激活函数, W, U, V, b_h, b_o 为参数。使用波束搜索在 D_t 中进行检索,当选择到结束符号(EOE)时过程终止。最后,输出推理路径,选择得分最高的路径上的文档作为 y_{SD} 。

$$y_{SD} = \mathrm{argmax}\prod_{1\leqslant t} O_t \tag{4}$$

2.2 上下文编码模块

首先,将文档筛选模块中筛选出的文档合并成一个上下文,然后将上下文与问题相连并输入到预先训练过的 RoBERTa 模型中,得到编码的问题表示 $Q = \{q_1, q_2, \cdots, q_m\} \in R^{m*d}$ 的上下文表示为 $C = \{c_1, c_2, \cdots, c_n\} \in R^{n*d}$,其中 m, n 分别是问题和上下文的长度, d 为隐藏状态的大小。紧接着 C 需要再经过一层 BiLSTM,从 BiLSTM 的输出 $M \in R^{n*2d}$ 中获得不同节点(句子 S ,实体 E)的表示。整张图可以表示为 $H = \mathrm{diagram}\{Q, S, E\} \in R^{z*d}$,其中 z 为问题节点、句子节点、实体节点数量之和。

2.3 多粒度图构建模块

不同粒度的节点可以从不同的信息源捕获语义,因此与同质节点的简单图相比,它可以更准确地定位支撑事实和答案。为了将分散在多篇文档中的线索汇总起来,构建一个包含问题、句子以及实体的多粒度图。不同粒度的节点针对不同的下游任务。句子节点主要用于事实预测,此外由于答案可能不在实体节点中,因此将实体节点信息融合到上下文表示中来共同预测答案。

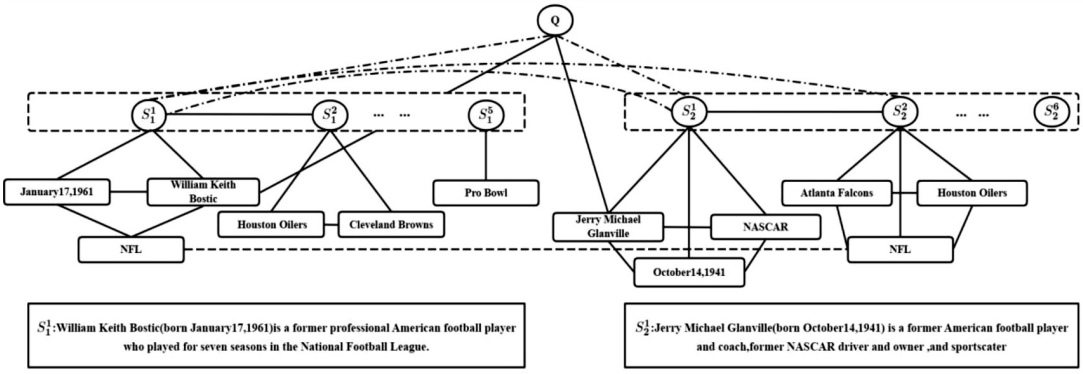

图 4 多粒度图的一个示例

图 4 显示了多粒度图的一个示例。定义了不同类型的边,如下所示:

(1)问题节点与其对应实体节点(问题里的实体)之间的边;

(2)句子节点与其对应实体节点(句子里的实体)之间的边;

(3)出现在同一篇文档中的句子节点之间的边;

(4)出现在同一个句子中实体之间的边;

(5)具有相同实体的句子之间的边;

(6)具有问题中的实体的句子之间的边(问题实体可以不同);

(7)具有相同实体的问题节点和句子节点之间的边;

(8)相同实体之间的边。

设计前 4 种类型的边使 GNN 能够掌握每个文档中呈现的全局信息。此外,跨文档推理是通过从问题中的实体跳到未知的桥接实体或比较问题中两个实体的属性来实现的。因此,设计了后 4 种类型的边,以更好地捕获跨文档推理路径。最后多粒度图由这 8 种类型的边和 3 种类型的节点组成。

2.4　图推理模块

为了实现显式和可解释的图推理,使用基于图注意网络[18](Graph Attention Network,GAT)的两阶段图推理。对图中的节点先进行异质更新再进行同质更新来进行多跳推理。例如句子节点,首先进行异质更新使其捕获不同粒度节点的线索完善自身信息,再通过同质更新,比较句子间的描述性信息以确定支撑事实。具体来说,首先将多粒度图 H 分为异质更新图 H_1 和同质更新图 H_2。异质/同质更新图通过屏蔽节点之间相应的连接得到,如图 5 所示。

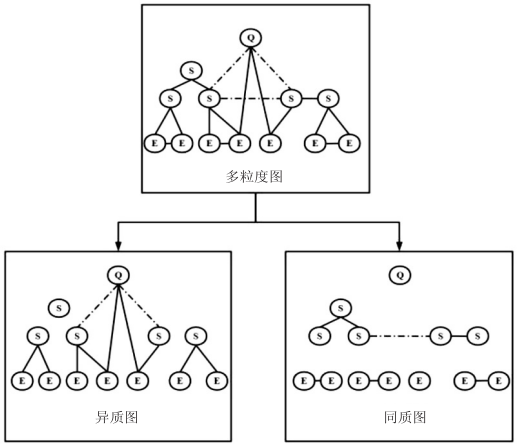

图 5　多粒度图的分解

对于图节点的更新表示,具体来说,首先从图中某个节点开始推理,关注在图上与该节点有连接的其他节点。然后通过计算它们之间的注意力分数,更新节点的特征表示。假设对于任意节点 i,其相邻节点集合为 N_i,则节点 i($i \in \{Q,S,E\}$)的注意力权重由下面公式得出:

$$S_{i,j} = \text{LeakyRelu}(\boldsymbol{W}^{\mathrm{T}}[h_i;h_j]),\ j \in N_i$$

$$a_{i,j} = \frac{\exp(S_{i,j})}{\sum_{k \in N_i}\exp(S_{i,k})}$$

其中,$\boldsymbol{W}$ 为可训练的线性变化矩阵,$S_{i,j}$ 表示两个节点之间的相关度分数,$a_{i,j}$ 表示节点 i 相比于其相邻节点的注意力权重系数。最后通过公式计算出节点 i 的最终的特征表示:

$$\overline{h_i} = \text{Relu}(\sum a_{i,j}h_j) \tag{5}$$

经过两阶段图推理后,得到图的更新表示 $\bar{H} = \{\overline{h_0},\overline{h_1},\cdots,\overline{h_z}\} \in R^{z*d}$。之前的工作通常将文档信息聚合到实体图中,然后直接在实体图的实体上选择答案。然而在 HotpotQA 数据集中,答案可能不存在于提取的实体图的实体中。如果答案不是实体图中的实体,则需要进一步处理以定位最终答案。

因此,使用门控注意力机制[19](Gated Attention)来融合图信息 $\bar{H}$ 和上下文表示 M,生成新的上下文表示 $\bar{M}$ 用于最终答案预测步骤。

$$\bar{M} = \text{Gated Attention}(\bar{H},M) \tag{6}$$

2.5　预测模块

使用更新后的句子节点表示和新的上下文表示 $\bar{M}$ 分别进行支撑事实预测和答案预测。遵循 Fang 等[8]的预测模块设计,预测模块有 5 个输出,分别为支撑事实预测、实体预测、答案的开始和结束位置以及答案类型预测。

$$\begin{aligned}
y_{\text{type}} &= \text{MLP}_0(\bar{M}[0])\\
y_{\text{start}} &= \text{MLP}_1(\bar{M})\\
y_{\text{end}} &= \text{MLP}_2(\bar{M})\\
y_{\text{sent}} &= \text{MLP}_3(\bar{S})\\
y_{\text{entity}} &= \text{MLP}_4(\bar{E})
\end{aligned} \tag{7}$$

其中,$\bar{M}[0]$ 是 $\bar{M}$ 的第一个隐藏表示,$\bar{S}$ 和 $\bar{E}$ 分别是更新后的句子节点表示和实体节点表示,$\bar{S}$ 和 $\bar{E}$ 从 $\bar{H}$ 中获得。每一个 MLP_i 都是一个用于不同的输出感知器(Multi-Layer Perceptron,MLP)。

模型的训练损失是答案跨度预测、支撑句预测、实体预测和答案类型预测损失的总和。

$$\begin{aligned}
L = {} & (y_{\text{start}},\bar{y}_{\text{start}}) + (y_{\text{end}},\bar{y}_{\text{end}}) +\\
& \gamma_1(y_{\text{sent}} + \bar{y}_{\text{sent}}) + \gamma_2(y_{\text{entity}} + \bar{y}_{\text{entity}}) +\\
& \gamma_3(y_{\text{type}} + \bar{y}_{\text{type}})
\end{aligned} \tag{8}$$

其中,$\bar{y}_{\text{start}}$,$\bar{y}_{\text{end}}$,$\bar{y}_{\text{sent}}$,$\bar{y}_{\text{entity}}$ 和 $\bar{y}_{\text{type}}$ 分别为问题答案、支撑句、支持实体、答案类型的真实标签。此外,跨度损失中添加了 3 个权重,以解释不同损失的比例差异。

3　实验及结果分析

3.1　实验数据集

使用 HotpotQA 数据集,这是第一个考虑模型解释能力的多跳 QA 数据集,也是一个多跳 QA 任务的流行基准。具体来说,该数据集中包括两个子任务,答

案预测(Ans)和支撑事实预测(Sup)。对于每个子任务都有两个官方评估,分别为精确匹配(EM)和部分匹配(F1)。EM 表示模型预测的标签中与真实标签完全匹配的百分比。Fl 表示模型预测的标签中与真实标签重叠的百分比。EM 和 F1 的联合得分用作整体指标(Joint)。在 HotpotQA Distractor 验证集上评估模型,该数据集使用整个英文维基百科转储作为数据集的语料库,约有 11 万个基于英文维基百科的问答对。对于该数据集上的每个问题,提供了 2 篇相关文档和 8 篇干扰文档,这些文档是由英文维基百科的高质量 TF-IDF 检索器收集的。

3.2 基线模型

将所提方案与以下方法进行了对比:

(1) Baseline:将 HotpotQA 数据集[3]中自带的方法作为基线;

(2) QFE:Nishida 等[20]通过考虑支撑事实之间的依赖关系来进行预测;

(3) DFGN:Qiu 等[9]根据实体间的关系构造动态实体图,在实体图上进行多跳推理;

(4) GRR:Asai 等[12]提出了一种基于图的递归检索查找支持文档,然后扩展现有的阅读理解模型回答问题;

(5) SAE:Tu 等[6]提出了一个管道系统,首先选择出相关文档,然后使用所选文档预测答案和事实;

(6) C2F:Shao 等[21]认为图结构不是多跳问答所必需的,提出了一种新的自注意力机制来进行预测答案和事实。

3.3 实验结果

表 1 显示了模型在 HotpotQA Distractor 验证集上的实验结果,可以看到文中模型超过了大部分已经发表的结果。在正确答案的预测上,文中模型得到的精确匹配(EM)为 69.3%,F1 为 82.3%。所提方法与基线相比在答案预测 EM 上获得 24.9% 的绝对改进,在答案预测 F1 上获得 24% 的绝对改进。在下面实验分析小节中,将详细分析模型性能增益的来源。

3.4 实验分析

文中的实现基于 Transformer 库[9]。在阈值 g 的选取上,人为设置阈值 g =0.000 001(接近于 0),使得文档筛选模块第 1 阶段尽可能多地选择文档。在多粒度图构建阶段,使用 Manning 等[22]提出的预训练实体模型来提取命名实体。图中的句子节点数量设置为 20,实体节点数量设置为 30。

在文档筛选模块中,当设置阈值 g = 0.000 001 时,文档筛选模块第 1 阶段文档召回率达到 99.9%。在第 1 阶段的基础上进行第 2 阶段检索再次获取文档,最后将文档筛选结果和 DFGN,HGN 和 SAE 中的文档筛选结果进行比较,如表 2 所示。所提方法在文档精确率和召回率方面分别为 96.7% 和 96.8%,相比于 SAE 在各个指标上均有提升。

表 1 在 HotpotQA Distractor 验证集上的结果 %

Model	Ans		Sup		Joint	
	EM	F1	EM	F1	EM	F1
Baseline	44.4	58.3	22.0	66.7	11.6	40.9
QFE	53.7	68.7	58.8	84.7	35.4	60.6
DFGN	55.7	69.3	53.1	82.2	33.7	59.9
GRR-large	68.0	81.2	58.6	85.2	/	/
SAE-large	67.6	80.7	63.3	87.4	46.8	72.7
C2F	68.0	81.2	60.8	87.6	44.7	72.7
所提出的方法	69.3	82.3	62.5	88.0	47.2	74.2

表 2 在 HotpotQA Distractor 验证集上的文档选择结果 %

Model	精确率	召回率	EM	F1
DFGN	83.3	97.5	63.1	88.2
SAE	96.0	96.0	92.3	96.0
HGN	49.8	99.3	0	66.3
所提出的方法	96.7	96.8	92.5	96.6

3.5 消融实验

消融实验结果如表 3 所示。为了证明提出文档筛选方法的有效性,对文档筛选两个阶段进行了消融研究。在文档筛选模块仅使用第 1 阶段或仅使用第 2 阶段中筛选出的文档用于下游任务。从表 3 中可以看到,仅使用第 1 阶段筛选出的文档用于下游任务会使模型在 Joint F1 指标上下降 7.8 百分点,仅使用第 2 阶段检索出的文档用于下游任务会使模型在 Joint F1 指标上下降 0.7 百分点。在两阶段图推理方面,当去除实体节点时,模型在 ANS F1 和 Sup F1 指标上分别下降 0.1 百分点和 0.2 百分点。在图节点更新顺序上,比较了 3 种不同的顺序。观察到从同质到异质的顺序更新节点实验结果较差,低于同步更新的结果。但提出的从异质到同质的方法更新图中节点相比于同步更

新在 3 个指标上均有提升。通过对消融实验结果的分析,证明了所提方法的有效性。

表 3　消融实验　%

Model	Ans F1	Sup F1	Joint F1
所提出的方法	82.3	88.0	74.2
文档选择仅用第 1 阶段	77.4	82.2	66.4
文档选择仅用第 2 阶段	82.0	87.4	73.5
图构建去除实体节点	82.2	87.8	74.0
图推理节点同步更新	82.1	87.9	74.0
图推理节点同质到异质更新	81.9	87.9	73.7

4　结束语

该文提出一种两阶段的文档筛选方法。第 1 阶段通过对文档进行评分和设置较小的阈值来尽可能多地获取文档,保证支持文档的高召回率。第 2 阶段利用递归神经网络对问题答案的推理路径进行建模,结合波束搜索在第 1 阶段的基础上再次提取文档,保证支持文档的高精确率。最后将支持文档构建成一个图节点为问题、句子和实体的多粒度图,并利用一种新颖的异步更新机制从多粒度图上进行答案预测以及支撑事实预测。实验结果证明了该模型的有效性。在未来工作中,希望结合文本构建图形的新进展来解决更困难的推理问题。此外,希望在其他多跳问答数据集上评估该模型。

参考文献:

[1] MAVI V,JANGRA A,JATOWT A. A survey on multi-hop question answering and generation[J]. arXiv:2204. 09140, 2022.

[2] PISAREVSKAYA D,SHAVRINA T. WikiOmnia:generative QA corpus on the whole Russian Wikipedia[J]. Information Retrieval Journal,2022,15(12):60-73.

[3] YANG Z,QI P,ZHANG S,et al. HotpotQA:a dataset for diverse, explainable multi-hop question answering[C]//Conference on empirical methods in natural language processing. Brussels:[s. n.],2018:2369-2380.

[4] KOČISKY T,SCHWARZ J,BLUNSOM P,et al. The narrativeqa reading comprehension challenge[J]. Transactions of the Association for Computational Linguistics,2018,14(6): 317-328.

[5] MOHAMED T, CHEN Zhiyu, BRIAN D D, et al. Neural ranking models for document retrieval[J]. Information Retrieval Journal,2021,24(6):73-85.

[6] TU M,HUANG K,WANG G,et al. Select,answer and explain:interpretable multi-hop reading comprehension over multiple documents[J]//AAAI conference on artificial intelligence. New York:AAAI,2020:9073-9080.

[7] WU B,ZHANG Z,ZHAO H. Graph-free multi-hop reading comprehension:a select-to-guide strategy[J]. arXiv:2107. 11823v1,2021.

[8] FANG Y,SUN S,GAN Z,et al. Hierarchical graph network for multi-hop question answering[C]//Conference on empirical methods in natural language processing. [s. l.]:[s. n.],2020:8723-8838.

[9] QIU L,XIAO Y,QU Y,et al. Dynamically fused graph network for multi-hop reasoning[C]//Annual meeting of the association for computational linguistics. Florence: ACL, 2019:6140-6145.

[10] 冯　钧,李　艳,杭婷婷. 问答系统中复杂问题分解方法研究综述[J]. 计算机工程与应用,2022,58(17):23-33.

[11] QI P,LEE H. Retrieve,rerank,read,then iterate:answering open-domain questions of arbitrary complexity from text [J]. arXiv:2010. 12527v1,2010.

[12] ASAI A,HASHIMOTO K,HAJISHIRZI H,et al. Learning to retrieve reasoning paths over wikipedia graph for question answering[C]//8th international conference on learning representations. [s. l.]:[s. n.],2019:2240-2252.

[13] 周展朝,刘茂福,胡慧君. 基于问题分解的多跳机器阅读理解模型[J]. 计算机工程与科学,2022,44(8):1506-1513.

[14] 杨玉倩,高盛祥,余正涛,等. 融合事实文本的问句分解式语义解析方法[J]. 小型微型计算机系统,2023,44(9): 1932-1939.

[15] 陈雨龙,付乾坤,张　岳. 图神经网络在自然语言处理中的应用[J]. 中文信息学报,2021,35(3):1-23.

[16] 邵　霭,许彩娥,万　健,等. 针对机器问答中多跳问题的深度学习网络模型[J]. 浙江科技学院学报,2022,34(5): 419-425.

[17] 龙　欣,赵容梅,孙界平,等. 面向多跳问答的多视图语义推理网络[J]. 工程科学与技术,2023,55(2):285-297.

[18] VELIČKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]//International conference on learning representations. [s. l.]:[s. n.],2018:45-55.

[19] DU Yongping,LIU Yang,PENG Zhi,et al. Gated attention fusion network for multimodal sentiment classification[J]. Knowledge-Based Systems,2022,240:108107.

[20] NISHIDA K, NISHIDA K, NAGATA M, et al. Answering while summarizing: multi-task learning for multi-hop QA with evidence extraction[C]//Annual meeting of the association for computational linguistics. Florence: ACL, 2019: 2335-2345.

[21] WOLF T,DEBUT L,SANH V,et al. Transformers:state-of-the-art natural language processing[C]//Conference on empirical methods in natural language processing. Boston:Kluwer,2020:38-45.

[22] MANNING C D, SURDEANU M, BAUER J, et al. The Stanford CoreNLP natural language processing toolkit[C]//Annual meeting of the association for computational linguistics. Baltimore:ACL,2014:55-60.