软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(11):5179−5190 [doi: 10.13328/j.cnki.jos.006730]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于多视角图编码的选择式阅读理解方法*

余笑岩[1,2,3], 何世柱[1,2], 宋 燃[2], 刘 康[1,2], 赵 军[1,2], 周永彬[3,4]

[1](中国科学院大学 人工智能学院, 北京 100049)

[2](中国科学院 自动化研究所, 北京 100190)

[3](中国科学院 信息工程研究所, 北京 100093)

[4](南京理工大学 网络空间安全学院, 江苏 南京 210094)

通信作者: 周永彬, E-mail: zhouyongbin@njust.edu.cn

摘 要: 选择式阅读理解通常采用证据抽取和答案预测的两阶段流水线框架, 答案预测的效果非常依赖于证据句抽取的效果. 传统的证据抽取多依赖词段匹配或利用噪声标签监督证据抽取的方法, 准确率不理想, 这极大地影响了答案预测的性能. 针对该问题, 提出一种联合学习框架下基于多视角图编码的选择式阅读理解方法, 从多视角充分挖掘文档句子之间以及文档句子和问句之间的关联关系, 实现证据句及其关系的有效建模; 同时通过联合训练证据抽取和答案预测任务, 利用证据和答案之间强关联关系提升证据抽取与答案预测的性能. 具体来说, 所提方法首先基于多视角图编码模块对文档、问题和候选答案联合编码, 从统计特性、相对距离和深度语义 3 个视角捕捉文档、问题和候选答案之间的关系, 获得问答对感知的文档编码特征; 然后, 构建证据抽取和答案预测的联合学习模块, 通过协同训练强化证据与答案之间的关系, 证据抽取子模块实现证据句的选择, 并将其结果和文档编码特征进行选择性融合, 并用于答案预测子模块完成答案预测. 在选择式阅读理解数据集 ReCO 和 RACE 上的实验结果表明, 所提方法提升了从文档中选择证据句子的能力, 进而提高答案预测的准确率. 同时, 证据抽取与答案预测联合学习很大程度减缓了传统流水线所导致的误差累积问题.

关键词: 选择式阅读理解; 多视角图编码; 证据抽取; 答案预测; 联合学习

中图法分类号: TP18

中文引用格式: 余笑岩, 何世柱, 宋燃, 刘康, 赵军, 周永彬. 基于多视角图编码的选择式阅读理解方法. 软件学报, 2023, 34(11): 5179–5190. http://www.jos.org.cn/1000-9825/6730.htm

英文引用格式: Yu XY, He SZ, Song R, Liu K, Zhao J, Zhou YB. Multiple-choice Reading Comprehension Approach Based on Multi-view Graph Encoding. Ruan Jian Xue Bao/Journal of Software, 2023, 34(11): 5179–5190 (in Chinese). http://www.jos.org.cn/1000-9825/6730.htm

Multiple-choice Reading Comprehension Approach Based on Multi-view Graph Encoding

YU Xiao-Yan[1,2,3], HE Shi-Zhu[1,2], SONG Ran[2], LIU Kang[1,2], ZHAO Jun[1,2], ZHOU Yong-Bin[3,4]

[1](School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China)

[2](Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

[3](Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)

[4](School of Cyber Science and Engineering, Nanjing University of Science & Technology, Nanjing 210094, China)

Abstract: Multiple-choice reading comprehension typically adopts the two-stage pipeline framework of evidence extraction and answer prediction, and the effect of answer prediction highly depends on evidence sentence extraction. Traditional evidence extraction methods mostly rely on phrase matching or supervise evidence extraction with noise labels. The resultant unsatisfactory accuracy significantly

* 基金项目: 国家重点研发计划 (2020AAA0106400); 国家自然科学基金 (61922085, 61976211, U1936209, 62002353); 中国博士后科学基金 (2021M701726); 中国科学院重点研究计划 (ZDBS-SSW-JSC006)
收稿时间: 2022-02-19; 修改时间: 2022-04-13; 采用时间: 2022-06-10; jos 在线出版时间: 2023-05-18
CNKI 网络首发时间: 2023-05-19

reduces the performance of answer prediction. To address the above problem, this study proposes a multiple-choice reading comprehension method based on multi-view graph encoding in a joint learning framework. The correlations among the sentences in the text and those of such sentences with question sentences are fully explored from multiple views to effectively model evidence sentences and their relationships. Moreover, evidence extraction and answer prediction tasks are jointly trained so that the strong correlations of the evidence with the answers can be exploited for joint learning, thereby improving the performance of evidence extraction and answer prediction. Specifically, this method encodes texts, questions, and candidate answers jointly with the multi-view graph encoding module. The relationships among the texts, questions, and candidate answers are captured from the three views of statistical characteristics, relative distance, and deep semantics, thereby obtaining question-answer-aware text encoding features. Then, a joint learning module combining evidence extraction with answer prediction is built to strengthen the relationships of evidence with answers through joint training. The evidence extraction submodule is designed to select evidence sentences and fuse the results with text encoding features selectively. The fusion results are then used by the answer prediction submodule to complete the answer prediction. Experimental results on the multiple-choice reading comprehension datasets ReCO and RACE demonstrate that the proposed method attains a higher ability to select evidence sentences from texts and ultimately achieves higher accuracy of answer prediction. In addition, joint learning combining evidence extraction with answer prediction significantly alleviates the error accumulation problem induced by the traditional pipeline framework.

Key words: multiple-choice reading comprehension; multi-view graph encoding; evidence extraction; answer prediction; joint learning

机器阅读理解 (machine reading comprehension, MRC) 通过对自然语言文档的理解实现对给定问题的正确回答, 常见的机器阅读理解任务根据问答形式的不同可以分为 4 类: 完形填空、多项选择、答案片段抽取以及自由式回答. 其中多项选择即选择式阅读理解需要根据对文档的分析, 自动从多项候选答案中选择一项作为给定问题的正确答案[1−3], 由于其答案不是直接从文档中抽取, 而是从多个候选答案中选择最佳的答案, 因此需要根据问题分析文档与候选答案之间关系来选择答案. 该任务要求机器具备一定的分析推理能力, 相关研究工作难度大, 得到了学术界的广泛关注, 是机器阅读理解任务的重要方向.

选择式阅读理解的核心是通过建模文档、问句及候选答案之间的关联性, 选择和原文表达最一致的选项作为问题的答案[4−9]. 如图 1 所示, 给定文档, 对于问题“嗓子疼可以吃感康吗?”, 需要从给定的“可以”“不可以”“不确定”这 3 个候选答案中选择正确答案. 按照答案选择方式的不同又可分为排除法、相关度排序法以及匹配法等. Parikh 等人[4]提出了一种基于神经网络的答案排除模型 ElimNet, 通过模仿人类的阅读理解方式, 先计算问题感知的文档表征, 然后通过构建消除门, 计算感知文档表征与候选答案之间关联, 通过关联判断来删除最不相关的选项, 不断重复迭代排除不相关选项, 得到最终答案. Ran 等人[7]提出了一种基于选项比较网络 OCN 的答案选择方法, 该方法在答案选择过程加入了对候选答案之间关系的比较分析, 将每个选项编码成向量, 通过注意力机制实现逐个向量的比较, 利用相关性关系辅助答案选择. 段建勇等人[9]提出一种用于选择式阅读理解的多角度共同匹配模型, 使用多角度匹配机制获得文档、问题和候选答案之间的相关性, 利用相关性对文档表示进行加权, 增强相关性强的文档候选片段, 基于优化的文档表征进一步选出正确答案.

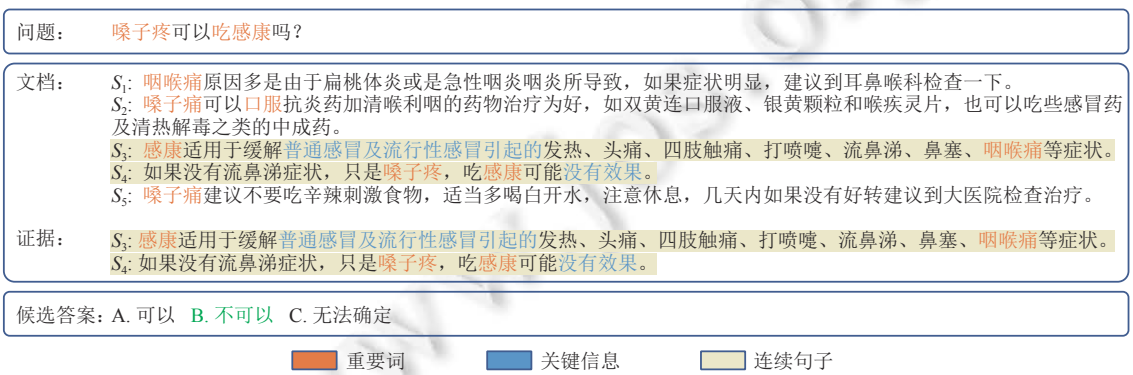

图 1 选择式阅读理解中问题、文档、证据和候选答案示例

从以上分析可以看出, 针对选择式阅读理解任务, 如何更好建模问句和答案约束的原文是保证正确答案选择的关键. 然而文档一般较长, 通常会包含一些和答案选择无关的噪声信息. 如图 1 所示, S_1 和 S_5 对于问题“嗓子痛

可以吃感康吗?”的支撑作用并不明显, 属于冗余信息, 甚至可能影响正确答案的选择. 为了缓解上述问题, 近年来选择式阅读理解的研究聚焦于先从文档中抽取证据句, 在此基础上再进行答案预测的流水线框架[10−12]. Choi 等人[10]首次提出一种从粗到细的框架用于建模基于文档的问答任务, 首先用词袋模型、滑动窗口模型和卷积模型来获得文档句子表示; 然后通过强化学习的方法隐式地从文档中选择一个句子; 最后仅利用选择的句子来完成问答. Wang 等人[11]提出一种远程监督的方法, 着重于抽取能够支撑回答问题和选择答案的证据句, 通过远程监督来生成噪声证据标签, 利用深层概率逻辑学习框架, 将句子级和跨句子信息进行结合, 用于间接监督最终的答案预测. 但是, 上述方法大都通过简单的字段匹配, 或者利用带噪声证据标签来训练证据抽取器, 均难以获得高质量的证据句子, 影响了后续答案预测的准确率.

基于证据的选择式阅读理解的关键是如何有效获取和利用证据信息. 从图 1 可以看出, 直接利用问句和文档中句子之间的相关度来匹配证据可能会得到错误的证据, 如 S_2 表示“嗓子痛可以吃些感冒药”可能会误导模型选择答案“可以”. 因此准确定位和理解证据很关键. 通过观察可以发现: 首先, 在文档中指向同一答案的多个证据句通常具有与问句相同的关键词, 在语义表达上也较为接近, 如作为证据的 S_3 和 S_4 都有问句中的重要词“嗓子痛”和“感康”, 定位这些词在文档中的位置对定位证据句有很大的作用; 第二, 文档中相互靠近的句子往往更相关, 相邻句子之间的关联对挖掘证据信息有促进作用, 如 S_3 和 S_4 和彼此相连, 且两个句子在文档中本就相连, 描述“感康”的用途和其对问题的解释; 第三, 在证据选择和编码过程中不但要考虑问句和候选证据句之间的匹配关系, 同样应该引入候选证据句之间的关系, 促使多个相似证据句在语义上中表达接近, 如 S_3 中“感康适用于缓解普通感冒…引起的…咽喉痛等”, 需要结合该句子和问句的语义信息, 从而推断答案为“不可以”. 另外从图 1 中还可以发现, 答案选择和证据选择两个任务紧密相关, 答案选择对于证据选择具有重要的指导作用. 如图 1 中正确答案为嗓子疼不可以吃感康, 通过答案可以很好的确定证据应该是和不可以吃感康相关, 可以有效排除选择 S_2 作为证据句.

基于上述思想, 本文提出一种多视角图编码的方法来建模文档、问题和候选答案之间的关系, 以解决传统方法中证据抽取效果不佳的难点问题. 本方法从统计特性、语义关系和相对距离 3 个角度建模文档、问题和候选答案之间的关系, 实现文档中潜在证据信息的精准捕捉, 提升答案预测任务的性能. 另外, 本文提出一种证据抽取和答案预测的联合学习框架, 通过软证据抽取辅助答案预测, 缓解了基于流水线的选择式阅读理解框架中的误差传播问题. 实验结果也表明, 提出的方法能够很好地适应有证据标签和证据标签缺失等场景.

本文的主要贡献总结如下.

(1) 提出多角度图编码网络对编码文档、问题和参考答案进行表征, 从统计特性、相对距离和深度语义 3 个角度捕捉文档、问题和参考答案之间的关系, 深度挖掘文档中潜在证据信息, 在证据标签缺失的场景下有效提高选择式阅读理解性能.

(2) 提出在联合学习的框架下同时训练证据抽取和答案预测, 利用两个任务之间的耦合关系来辅助答案预测的性能, 在有证据标签的场景下进一步提高答案预测的性能, 同时减缓传统流水线框架中的误差传播问题.

(3) 在 ReCO 和 RACE 两个数据集进行验证, 实验结果表明我们的方法明显优于基准方法, 进一步的消融实验也证明了多角度图编码模块和联合训练框架的有效性.

1 问题定义和整体框架

选择式阅读理解可以被形式化为: 给定一个由 n 个句子组成的文档 $D=\{S_1,S_2,\ldots,S_n\}$, 一个问题 Q, 和 m 个(本文举例使用 $m=3$) 候选答案 $C=\{A_1,A_2,\ldots,A_m\}$, 其中有且只有一个为正确答案 y_A.

证据抽取任务要求模型判断文档 $D=\{S_1,S_2,\ldots,S_n\}$ 中每一个句子 S_i 是否为证据, 每一个句子 S_i 都有一个证据标签 $y_i\in\{0,1\}$; 对一个句子判断是否为证据句的过程可描述为:

$$y=\arg\max_{\bar{y}}P(\bar{y}|C,Q,D),$$

其中, $\bar{y}$ 为模型判断证据句结果.

答案预测任务则要求模型从候选答案 $C=\{A_1,A_2,\ldots,A_m\}$ 中选择一个答案 A_j 作为正确答案, 其过程可描述为:

$$y_A = \arg\max_{\overline{A}} P(\overline{A}|C,Q,D),$$

其中, $\overline{A}$ 为模型预测答案.

本文提出的基于多视角图编码的选择式阅读理解方法结构如图 2 所示. 该模型由 3 个模块组成: 多视角图编码模块, 证据抽取模块和答案预测模块. 其中多视角图编码模块从不同视角用图建模文档、问题和候选答案的关系; 证据抽取模块判断文档中句子是否为证据句; 答案预测模块基于问题、候选答案和证据增强后的文档句子表示进行答案预测. 在文本表示方面证据抽取模块和答案预测模块共享一个由多视角图编码模块得到的输入, 最后证据抽取模块和答案预测模块基于多任务联合的方式进行训练.

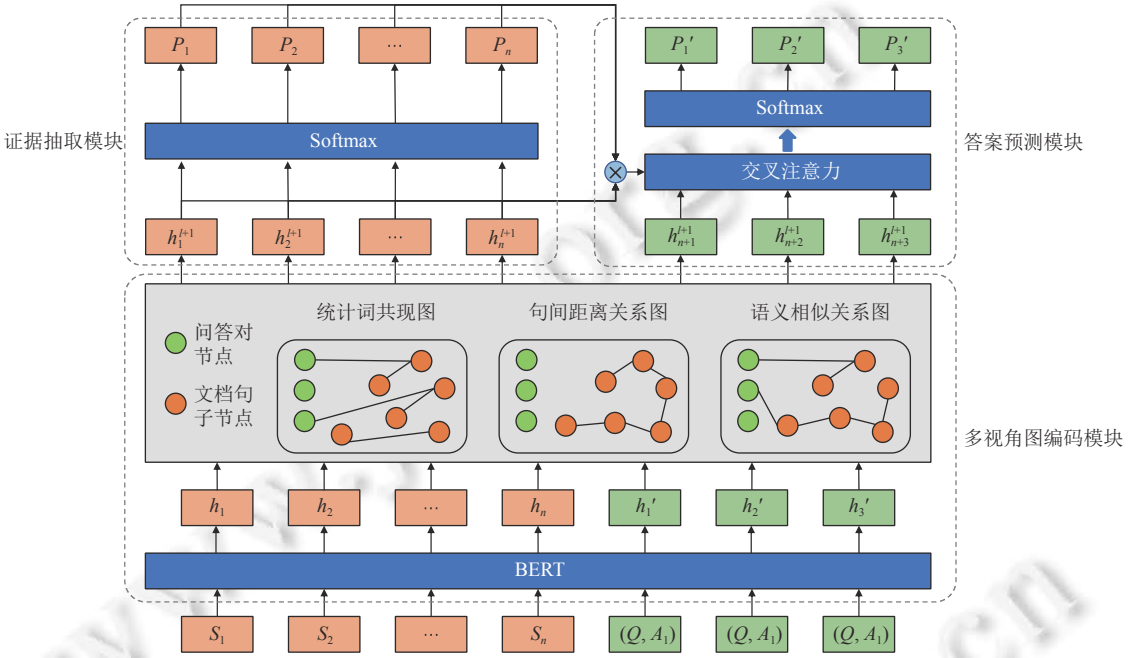

图 2 基于多视角图编码的选择式阅读理解方法整体框架图

2 多视角图编码模块

多视角图编码模块 (multi-aspect graph encoder, MAGE) 由 BERT 编码网络和多视角图编码网络组成. BERT 编码器对文档和问答对进行句子级表征, 多视角图编码网络从 3 个视角挖掘文档中与问答对相关的潜在证据信息, 从而获得问答对和潜在证据感知的文档表示.

● BERT 编码网络. 本文采用 BERT[13]作为基础编码器以获得文档和问答对句子级的语义表征向量. BERT 是一个多层 Transformer 的编码器, 并在大规模文本上进行预训练. 为了获得 BERT 的输入, 在文本预处理阶段, 首先将文档 D 进行分句, 得到 $S_1,S_2,\ldots,S_n$, 用特殊分隔符“[SEP]”将问句 Q 分别和每个候选答案 $A_1,A_2,\ldots,A_m$ 进行拼接; 最后将拼接好的问答对与分好句的文档组成 BERT 的输入形式 (用特殊分隔符拼接).

$$h'_1,h'_2,\ldots,h'_m,h_1,h_2,\ldots,h_n = \mathrm{BERT}((Q,A_1),(Q,A_2),\ldots,(Q,A_m),\ldots,S_1,S_2,\ldots,S_n),$$

其中, h'_j 表示候选答案 A_j $(i=1,2,\ldots,m)$ 经过 BERT 语义编码后的特征, h_i 表示文档中句子 S_i $(i=1,2,\ldots,n)$ 经过 BERT 语义编码后的特征, (Q,A_j) 表示“[SEP]”拼接的问答对.

● 多视角图编码网络. 为了挖掘不同文档句子和问答对的潜在证据信息, 本文构建了基于统计特征、空间位置和深层语义 3 种关系的图网络对文档、潜在证据信息和问答对进行表征, 以得到与问答对更有关联的篇章句子表征, 以促进答案预测和证据抽取任务.

2.1 统计词共现图

共现词为句子之间有相同的词, 与共现度高的句子往往描述了相同的实体, 当文档句子与问句词共现度高时, 表明句子可能为潜在的证据. 为了挖掘文档中证据信息, 构建基于 TF-IDF 的统计词共现图. TF-IDF 是一种用于信息检索与文本挖掘的加权技术. 其计算为:

$$tfidf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}\times\log\frac{|D|}{1+\left|\left\{j:t_i\in d_j\right\}\right|},$$

其中, $n_{i,j}$ 是词 i 在文档 d_j 中出现的次数, $\sum_k n_{k,j}$ 代表文档 d_j 中所有字词出现次数之和, $|D|$ 是语料库中文档总数, $\left|\left\{j:t_i\in d_j\right\}\right|$ 是含有词 t_i 的文档数量.

构造统计词共现图的具体做法是: 首先计算问句 Q 的 TF-IDF 值, 选择值较高的词作为重要词; 文档中所有句子和问答对作为结点, 当出现重要词共现, 则在两个结点之间连上一条边. 其邻接矩阵定义为:

$$A_{i,j}^{\text{word}}=\begin{cases}1, & \text{if } S_i^{\text{TF-IDF}}\cap S_j^{\text{TF-IDF}}\neq\varnothing\\ 0, & \text{else}\end{cases},$$

其中, $S_i^{\text{TF-IDF}}$ 表示句子 S_i 的 TF-IDF 值较高的词的集合, $S_i^{\text{TF-IDF}}\cap S_j^{\text{TF-IDF}}\neq\varnothing$ 表示第 i 句与第 j 句有重要词共现.

2.2 句间距离关系图

鉴于文档中相互靠近的句子往往更相关, 相邻句子之间的关联对挖掘证据信息有促进作用. 本文在建模句子间关系时, 考虑到句子之间空间上的距离关系. 本文参考 Tian 等人[14]的工作, 采用高斯分布来衡量篇章内句子之间空间距离对相关度的影响. 文档中句子 S_i 与句子 S_j 之间的距离计算为:

$$distance=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(j-i)^2}{2\sigma^2}},$$

其中, σ 为一个超参数.

句间距离关系图的具体构造方法为: 文档中所有句子和问答对作为图中的结点; 计算结点之间的距离值; 设置篇章句间距离阈值, 当距离值超过该阈值时在两个结点间连一条边, 其邻接矩阵定义为:

$$A_{i,j}^{\text{dist}}=\begin{cases}1, & \text{if } distance\geqslant\delta_d\\ 0, & \text{else}\end{cases},$$

其中, δ_d 为句间距离阈值.

2.3 语义相似关系图

为了挖掘文档中的证据信息, 建模文档、问题和候选答案之间的语义关系是必要的. 在语义相似的两个对象之间建立联系对于挖掘证据信息是有利的. 文档和问答对在经过 BERT 的表征后获得了一定的语义信息, 为了进一步计算语义相似度, 我们构造了语义相似关系图. 其具体构造方法为: 文档中所有句子和问答对作为图中的结点; 计算所有结点之间的余弦相似度; 设置语义相似度阈值, 当两个结点 (句子) 之间余弦相似度大于相似度阈值时在两个结点之间连一条边, 其邻接矩阵定义为:

$$A_{i,j}^{\text{simi}}=\begin{cases}1, & \text{if } \mathrm{sim}(\mathrm{S_i},\mathrm{S_j})\geqslant\delta_{\mathrm{s}}\\ 0, & \text{else}\end{cases},$$

其中, $sim(\cdot)$ 表示余弦相似度的计算, $sim(S_i,S_j)$ 表示句子 S_i 和 S_j 之间的语义相似性; δ_s 为相似度阈值.

2.4 多图融合

经过 3 种不同构图方式得到的包含文档、问答对和潜在证据信息的多视角关系图后, 下一步要对 3 个图进行信息融合. 定义图 $G=(H,A)$, 其中 H 为图的结点表征, A 为图的邻接矩阵. 具体的方法为: 首先, 对 3 个视角的图分别进行图卷积, 如下所示:

$$H_u^{l+1}=f_u(H_u^l,A_u),$$

其中, A_u 为邻接矩阵, 取 $A_u=A_{\text{word}},A_{\text{dist}},A_{\text{simi}}$; $f_u(\cdot)$ 为图卷积操作, l 为图卷积层数; H_u^{l+1} 是所有句子的表示.

最后采用一个两层感知机对 3 个不同关系图的卷积结果进行融合, 得到一个最终表示:

$$H' = \tanh\left(W_a \tanh\left(W\left(H_{\text{word}}^{l+1} \| H_{\text{dist}}^{l+1} \| H_{\text{simi}}^{l+1}\right) + b\right) + b_a\right),$$

其中, W_a, b_a, W, b 为模型的可训练参数, $\|$为特征拼接操作.

分别从 3 个视角建模文档、问题和候选答案之间的关系后, 对 3 个视角关系图分别卷积, 然后进行融合可以得到了文档和问答对的特征向量. 这些特征向量包含丰富的信息, 将这些表示作为后续模块的输入能有效改善答案预测的效果.

3 联合学习框架

为了减缓传统流水线框架中误差传播的问题, 即抽取到错误证据必然导致答案预测的错误, 本文在联合学习的框架下训练证据抽取模块和答案预测模块. 具体做法是, 使用证据抽取模块获得的每个文档句子作为证据的概率对多视角图编码模块得到的表征进行加权, 从而实现证据句权重更高, 其他句子权重相对较低, 而辅助答案预测任务. 通过联合优化参数的方法, 使得两个耦合的任务相互帮助, 而达到更好的性能.

3.1 证据抽取模块

证据预测任务能够有效辅助答案预测任务, 本文设计一个证据抽取模块 (evidence extractor, EE), 对多视角图编码模块得到的文档表示进一步的处理, 目的为确定文档中哪些句子为证据信息, 并且根据确定的证据信息辅助后续的答案预测任务.

对于每个文档中的句子都预测一个是否为证据句的表示, 其过程可描述为:

$$P_1, P_2, \ldots, P_n = Softmax\left(W_d \cdot \tanh(W_c\left(h_1^{l+1} \| h_2^{l+1} \| \ldots \| h_n^{l+1}\right) + b_c) + b_d\right),$$

其中, P_i 为文档中句子 S_i 作为证据的概率; h_i^{l+1} 是图卷积后句子 S_i 的表示; $\|$为特征拼接操作; W_c, b_c, W_d, b_d 是模型的可训练参数, 其中 P 既能够作为证据概率, 又可以加权到句子的表示上以提高答案预测的效果.

3.2 答案预测模块

答案预测模块 (answer prediction, AP) 旨在利用多视角图编码器得到的表示, 在证据抽取模块获得的证据信息的加权辅助下, 实现根据文档选择一个候选答案作为给定问题的答案.

为了提高证据句子在后续答案预测中的重要性, 同时不将非证据句子的影响降为无, 从而减弱流水线框架中错误传播问题带来的负面影响, 答案预测模块的输入由证据抽取模块得到的文档中每个句子作为证据的概率与每个句子的表示加权相乘和多视角图编码器得到的问答对表示拼接而成, 如图 2 所示. 通过线性操作, 进行交叉注意力, 其过程为:

$$Q = W_e\left(h_{n+1}^{l+1} \| h_{n+2}^{l+1} \| \ldots \| h_{n+m}^{l+1}\right) + b_e,$$

$$K, V = W_f(P_i \cdot H_i) + b_f,$$

$$C = Softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V,$$

其中, $Q \in \mathbb{R}^{m \times d}$; $K, V \in \mathbb{R}^{n \times d}$; $C \in \mathbb{R}^{m \times d}$; h_{n+j}^{l+1} 表示图卷积后候选答案 A_j 的表示; C 为交叉注意力后得到的特征表示; W_e, b_e, W_f, b_f 为模型可训练参数.

最后得到答案预测模块输出:

$$P_1', P_2', \ldots, P_m' = Softmax(W_h \cdot C + b_h),$$

其中, P_j' 为候选答案 A_j 作为答案的概率; C 为交叉注意力后得到的特征表示; W_h, b_h 为模型可训练参数.

3.3 联合训练

在联合训练的过程中, 首先分别计算证据抽取和答案预测模块的损失值, 再计算两个任务的总损失值对任务同时进行优化.

证据抽取模块的损失计算如下:

$$\mathcal{L}_1 = \sum_i -[y_i \cdot \log(P_i) + (1-y_i) \cdot \log(1-P_i)],$$

其中, y_i 是句子 S_i 的证据标签, P_i 是模型分类句子 S_i 为证据的概率.

答案预测模块的损失计算如下:

$$\mathcal{L}_2 = -\sum_j^m y_A \cdot \log\left(P'_j\right),$$

其中, P'_j 是模型判断候选答案 A_j 为正确答案的概率, y_A 是正确答案.

证据抽取和答案预测两个任务在联合学习的框架下进行训练, 两个任务的参数都通过联合优化来获得. 联合损失的计算公式如下:

$$\mathcal{L} = \alpha\mathcal{L}_1 + \mathcal{L}_2,$$

其中, α 是超参数.

4 实验分析

4.1 实验数据

我们在公开选择式阅读理解数据集 ReCO[3]和 RACE[2]上进行实验来评估所提模型的有效性.

ReCO 数据集包含 30 万个数据样本, 每个数据样本由文档、问题、候选答案、证据句和答案 (以及参考文档来源 URL 等) 组成. 模型需要从 3 个候选答案中选择一个作为正确答案, 这要求模型具备较强的推理能力. 其划分为 28 万训练数据和 2 万测试数据. 表 1 所示为该数据集样本中问题领域分布和文档来源. 值得注意的是, 在 ReCO 数据集中, 由于人为标注的原因, 部分证据标注是标注员自行总结或复述的, 与原文表述不完全一致. 因此, 在数据处理过程中, 我们预先将那些与原文表述不一致的证据通过相似度比对的方式找到其在原文表述中所对应的句子.

表 1 ReCO 数据集问题领域分布

分组	领域	例子	比例 (%)
问题领域	健康	晒太阳真的能补钙吗?	44.5
	科技	U盘进水还能用吗?	12.5
	社会	湖南高考分数线高吗?	17.5
	生活	移动营业厅周日开门吗?	20.5
	文化	西方人吃米饭吗?	5.0
参考文档来源	问答平台	FDA、妈妈帮	40.0
	论坛	Quora、搜狗问问	44.5
	其他	搜狗百科、人民日报	15.5

RACE 是一个从英语考试中收集而来的大规模选择式阅读理解数据集. 其包含 27933 个文档和 97687 个问题, 每个问题对应 4 个给定的候选答案, 其中只有一个为正确答案. RACE 数据集中的样本覆盖多种问题类型, 如总结、推理、演绎和上下文匹配, 大多数问题需要推理, 而不是词汇层面的匹配.

4.2 基线模型

为了评估所提方法的有效性, 我们使用了几个在很多机器阅读理解任务上都表现很好的模型作为基线模型.

● BiDAF[15]: 一个采用多阶段、层次化处理, 使其可以捕获原文不同粒度的特征的模型. 使用长短期记忆 (long short-term memory, LSTM) 作为其编码器, 并通过双向注意力流机制建模问题和答案之间的联系.

● BiDAF*: 用 ELMO[16] (一种在无监督数据上训练的语言模型) 取代了 BiDAF 中的传统词嵌入, 从而得到更好的效果.

● BERT_doc[13]: 一个多层的 Transformer 编码器, 在大量的无标签数据上进行了预训练, 并在许多自然语言处

理任务上超过了最先进 (state-of-the-art) 的模型, BERT_doc 表示在进行阅读理解时, 仅依靠文档作为参考信息的输入, 没有证据标签.

● BERT_evi: 使用 BERT 对文档、问题和候选答案进行编码, BERT_evi 表示在进行阅读理解时, 依靠标注的证据句作为参考信息的输入, 有证据标签 (仅 ReCO 数据集有证据标签, RACE 数据集没有该基线模型实验).

对于流水线框架的模型, 我们分别训练一个基于证据抽取的流水线机器阅读理解模型和一个基于证据生成的机器阅读理解模型.

● Pipeline_Ex: 为了对比证据抽取的流水线框架与证据抽取辅助的答案预测的联合学习框架, 设计一个使用 MAGE 对文档编码, 证据抽取模块与答案预测模块相互独立, 答案预测在证据抽取的基础上.

● Pipeline_Gen[17]: 设计了一个基于编码器-解码器框架的流水线框架. 为了对比流水线框架与联合学习框架的性能, 该基线方法通过 MAGE 对文档进行编码, 然后用 LSTM 解码器生成证据, 在生成的证据上进行答案预测.

4.3 实验设置和评估方法

对于文本编码部分, 本文使用 Wolf 等人[18]提出的 Transformers 框架, 并使用隐向量长度为 768 的 BERT_base 为作为初始模型, 优化超参数与原模型一致. 在构图方面, 每个句子最大长度为 64, 若超过 64 则拆分为两句相同长度的句子. 句子个数最大为 20, 丢弃掉大于 20 的部分, 此部分数据少于 1.8%. 对于伪证据句构建, 使用标注的证据句子与文本中每个句子计算相似度, 当相似度大于 0.8 时, 标注该文档句子为证据句, 标签为 1, 其余句子则为非证据句, 标签为 0. 对于语义图使用相似度阈值 $\delta_s = 0.8$, 对于统计图使用阈值 0.4, 对于空间图使用句间距离阈值 $\delta_d = 0.6$. 本文使用 DGL[19]库中的 GCN 模块对图进行编码, 其中 GCN 层数为 2, 隐状态长度为 256. 模型训练使用 AdamW[20]优化器, 联合损失超参数设置 $\alpha = 0.25$, 学习率为 4E–5, 训练批次大小为 8, 共训练 6 轮, 每 15 000 步进行一次验证并选择准确度最好的模型保存. 所有超参数都是通过网格搜索的方式获得的.

跟之前的机器阅读理解任务相同, 本文使用准确度作为衡量模型是否将每个样本准确分类的标准.

$$\text{准确度} = \frac{\text{正确分类的样本数量}}{\text{总样本数量}} \times 100\%.$$

4.4 有证据标签场景下 MAGE-JF 实验

本文在 ReCO 数据集上评估了有证据标签场景下基于多视角图编码的选择式阅读理解方法 (下文简称为 MAGE-JF) 的有效性, 实验结果如表 2 所示. 从中可以看到 MAGE-JF 相比于基线方法 BiDAF 提升了 9.3 个百分点, 说明了仅从语义角度建模文档和问题对于回答问题是远远不足够的, 多视角建模文档、问题和候选答案对于选择式阅读理解的答案预测有很大的促进作用; 相比于基于 BERT 使用文档作为参考信息输入的基线方法, 提升了 3.7 个百分点, 说明本文提出的从多视角建模文档、问题和候选答案的方法的有效性, 其中每个视角的有效性在第 4.6 节中证明; MAGE-JF 相比于流水线框架的方法提升了 2.6 个百分点, 在证据抽取方法和答案预测方法相同的情况下, MAGE-JF 在联合学习的框架下进行训练, 联合优化证据抽取和答案预测, 而流水线框架中两个任务分别训练, 因此, 可以看出联合学习的框架可以减缓流水线框架中误差传播的问题.

表 2 有证据标签场景下 MAGE-JF 实验 (%)

模型	ReCO_Dev	ReCO_Test
随机	33.3	33.3
BiDAF_doc	55.8	56.4
BiDAF*_doc	57.5	58.9
BERT_evi	76.3	77.3
BERT_doc	61.4	61.1
Pipeline_ex	62.5	62.9
Pipeline_gen	61.9	62.1
MAGE	**64.3**	**64.7**
MAGE-JF	**65.1**	**65.4**

表 2 中 MAGE-JF 联合学习框架下的基于多视角图编码的选择阅读理解方法, 即在有监督场景 (有证据标签场景) 下; BERT_doc 和 BERT_evi 分别为基于文档的机器阅读理解和基于证据句的机器阅读理解; 流水线抽取和流水线生成分别表示流水线框架下基于证据抽取和基于证据生成的方法. 总的来说, 本文所提出的多视角图能够有效地捕捉分散在文档中的潜在证据; 联合学习框架下的证据抽取模块可以能够增强模型对证据的捕捉能力, 同时不过度干预模型, 让模型能够感知潜在证据信息, 整合全文信息进行答案预测.

4.5 无证据标签场景下 MAGE 实验

本文在 RACE 数据集上评估了无证据标签场景下基于多视角图编码的选择式阅读理解方法 (下文简称为 MAGE) 的有效性, 实验结果如表 3 所示. 从表 3 可以看出, MAGE 在 RACE-M、RACE-H 和 RACE 上均取得了最佳性能, 说明了 MAGE 的有效性. 由于本实验仅通过多视角图编码模块对文档、问题和候选答案进行编码, 编码后直接进行答案预测, 没有额外证据抽取部分, 即通过多视角编码上下文来进行阅读理解, 与其他基线方法思路一致. 从实验数据可以看出, 与 BERT_doc 相比, MAGE 在性能上提升了 2.2 个百分点, 充分说明了从多个角度建模文档、问题和候选答案对答案预测的有效性, 即通过多个视角编码文档, 可以有效地挖掘文档中的潜在证据信息, 证明本文提出方法能够在没有证据标注的情况下可以有效建模文档间句子的关系, 并提高答案预测的准确度. 实验证明本文方法能够适用于选择式阅读理解任务, 本文方法具有泛化性.

表 3 无证据标签场景下 MAGE 实验 (%)

模型	RACE	RACE-M	RACE-H
随机	25	25	25
BiDAF_doc	38.2	40.2	41.8
BiDAF*_doc	40.9	43.6	43.3
BERT_doc	65.0	71.7	62.3
MAGE	**66.1**	**72.1**	**64.5**

由于 RACE 数据集没有证据句标注, 实验都是使用基于文档的阅读理解方法; 同样的, 没有证据句标注的场景下流水线框架中的证据抽取部分无法评估其性能, 因此本组实验可以很好地证明多视角图编码模块的有效性. 从表 3 可以看出 RACE-H 提升的效果高于 RACE-M, 但根据数据集的构建规则可知, RACE-H 中的文本预测难度比 RACE-M 中的要高, 从这一点可以看出, 多视角图编码适用于编码更复杂的文本, 更能从复杂的文本中挖掘出证据信息来辅助答案预测.

4.6 多视角图编码模块消融实验

本小结通过消融实验来验证多视角图编码中每一种构图方式的有效性, 实验结果如表 4 所示. 表 4 中“+”表示在 BERT 基础编码器的基础上加上指定视角的构图方法. 第 1 行到最后一行分别代表使用文档作为阅读理解参考信息输入的 BERT、仅构建空间图、仅构建统计图、仅构建语义图、构建空间图和统计图、构建空间图和语义图、构建统计图和语义图、MAGE 的性能.

表 4 多视角图编码模块消融实验 (%)

模型	ReCO	RACE-H
BERT_doc	61.1	62.3
+空间图	61.4	62.3
+统计图	61.9	62.4
+语义图	62.1	62.6
+空间 & 统计图	62.4	62.9
+空间 & 语义图	63.3	63.1
+统计 & 语义图	63.9	63.4
MAGE	**64.3**	**64.7**

从表 4 中第 2 行到第 4 行可以看出, 3 个不同视角图对最终答案预测性能排序如下: 语义图> 统计图> 空间图. 经过分析, 本文认为造成空间图共现最小的原因是, 在多视角图编码模块中, BERT 是基础编码器, 其中有位置嵌入操作, 因此 BERT 编码器已经考虑到文档句子之间的空间位置信息. 同时, 语义信息是判断文档和问题相关度的重要部分, 故语义图对答案预测性能的贡献最大, 由于 BERT 与训练语言模型能够有效捕捉语义信息, 因此基于此构建的语义图具有较好的表征能力.

4.7 案例分析

BERT 对文档编码时, 句子距离对于有着较大的影响. 由于问题离句子 S_3 更近, 所以模型更加趋向于将 S_3 作为证据句, 因此, 根据句子 S_3 模型将回答“多”. 但对于表 5 所示例子可以看出: “叶子发黄”的原因有很多, 多视角图编码模块能够建模问题、文档句子间的关系, 有效地捕捉信息. 这些句子被从不同的角度链接起来, 能够相互作为补充, 此时信息则显示“叶子发黄, 有很多原因”, 其中多浇水或少浇水都会使叶子变黄, 因此, 对于“叶子发黄”的原因, 应该综合不同情况来看, 结合文档中给的信息, 根据图编码后得到的信息可以回答“不确定”.

表 5 案例分析

问题	玫瑰花叶子发黄是水浇多还是少了?		
文档	S_1: 家庭养花最容易碰到的问题是叶子发黄. S_2: 花的叶子黄, 一般有以下几种原因: S_3: 一为水黄, 就是水浇得太多导致土壤积水久湿, 透气性差甚至部分须根腐烂, 表现出嫩叶暗黄无光泽, 新梢萎缩. S_4: 二为肥黄, 就是肥多, 表现为老叶枝尖变黄脱落, 新叶虽肥厚有光泽但一般凹凸不舒展. S_5: 三为旱黄, 由于长期浇水少导致脱水, 新叶虽叶色正常, 但下部叶片渐向上干黄脱落老化. S_6: 四为碱黄, 尤其一些南方花卉喜酸性土壤, 而北方水质偏碱则出现叶子渐退色变黄甚至脱落.		
候选答案	A. 多了	B. 少了	C. 无法确定
BERT模型预测	**0.6**	0.3	0.1
MAGE模型预测	0.07	0.03	**0.9**
正确答案	×	×	√

5 总 结

本文提出一种基于多视角图编码的选择式阅读理解方法, 通过从多视角有效建模证据句以及问句答案之间的关系, 很好地缓解了选择式阅读理解任务中证据句抽取难和不准确的问题, 同时通过联合优化答案选择和证据句选择的方法缓解了传统流水线框架中误差传播的问题. 在有证据标签的 ReCO 和无证据标签的 RACE 两个数据集上的实验结果也证明了提出方法的有效性.

References:

[1] Richardson M, Burges CJC, Renshaw E. MCTest: A challenge dataset for the open-domain machine comprehension of text. In: Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing. Seattle: Association for Computational Linguistics, 2013. 193–203.

[2] Lai GK, Xie QZ, Liu HX, Yang YM, Hovy E. RACE: Large-scale ReAding comprehension dataset from examinations. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 785–794. [doi: 10.18653/v1/D17-1082]

[3] Wang BN, Yao T, Zhang Q, Xu JF, Wang XC. ReCO: A large scale Chinese reading comprehension dataset on opinion. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 9146–9153. [doi: 10.1609/aaai.v34i05.6450]

[4] Parikh S, Sai A, Nema P, Khapra MM. ElimiNet: A model for eliminating options for reading comprehension with multiple choice questions. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: Morgan Kaufmann, 2018. 4272–4278.

[5] Tang M, Cai JR, Zhuo HH. Multi-matching network for multiple choice reading comprehension. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence and the 31st Innovative Applications of Artificial Intelligence Conf. and the 9th AAAI Symp. on Educational

Advances in Artificial Intelligence. Hawaii: AAAI Press, 2019. 870.

[6] Zhu HC, Wei FR, Qin B, Liu T. Hierarchical attention flow for multiple-choice reading comprehension. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence and the 30th Innovative Applications of Artificial Intelligence Conf. and the 8th AAAI Symp. on Educational Advances in Artificial Intelligence. Louisiana: AAAI Press, 2018. 746.

[7] Ran Q, Li P, Hu WW, Zhou J. Option comparison network for multiple-choice reading comprehension. arXiv:1903.03033, 2019.

[8] Duan QW, Huang J, Wu HY. Contextual and semantic fusion network for multiple-choice reading comprehension. IEEE Access, 2021, 9: 51669–51678. [doi: 10.1109/ACCESS.2021.3068993]

[9] Duan JY, Wei XP, Wang H. A multi-perspective co-matching model for machine reading comprehension. Data Analysis and Knowledge Discovery, 2021, 5(4): 134–141 (in Chinese with English abstract). [doi: 10.11925/infotech.2096-3467.2020.0714]

[10] Choi E, Hewlett D, Uszkoreit J, Polosukhin I, Lacoste A, Berant J. Coarse-to-fine question answering for long documents. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017. 209–220. [doi: 10.18653/v1/P17-1020]

[11] Wang H, Yu D, Sun K, Chen JS, Yu D, McAllester D, Roth D. Evidence sentence extraction for machine reading comprehension. In: Proc. of the 23rd Conf. on Computational Natural Language Learning. Hong Kong: Association for Computational Linguistics, 2019. 696–707. [doi: 10.18653/v1/K19-1065]

[12] Niu YL, Jiao FK, Zhou MT, Yao T, Xu JF, Huang ML. A self-training method for machine reading comprehension with soft evidence extraction. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020. 3916–3927. [doi: 10.18653/v1/2020.acl-main.361]

[13] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[14] Tian ZX, Zhang YZ, Feng XW, Jiang WB, Lyu YJ, Liu, K, Zhao J. Capturing sentence relations for answer sentence selection with multi-perspective graph encoding. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 9032–9039. [doi: 10.1609/aaai.v34i05.6436]

[15] Seo MJ, Kembhavi A, Farhadi A, Hajishirzi H. Bidirectional attention flow for machine comprehension. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: ICLR, 2017.

[16] Ilić S, Marrese-Taylor E, Balazs J, Matsuo Y. Deep contextualized word representations for detecting sarcasm and irony. In: Proc. of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Brussels: Association for Computational Linguistics, 2018. 2–7. [doi: 10.18653/v1/W18-6202]

[17] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 3104–3112.

[18] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu CW, Le Scao T, Gugger S, Drame M, Lhoest Q, Rush AM. Transformers: State-of-the-art natural language processing. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, 2020. 38–45. [doi: 10.18653/v1/2020.emnlp-demos.6]

[19] Wang MJ, Yu LF, Zheng D, Gan Q, Gai Y, Ye ZH, Li MF, Zhou JJ, Huang Q, Ma C, Huang ZY, Guo QP, Zhang H, Lin HB, Zhao JB, Li JY, Smola A, Zhang Z. Deep graph library: Towards efficient and scalable deep learning on graphs. In: Proc. of the 2019 ICLR Workshop on Representation Learning on Graphs and Manifolds, 2019: 10311680

[20] Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv:1711.05101, 2019.

附中文参考文献:

[9] 段建勇, 魏晓鹏, 王昊. 基于多角度共同匹配的多项选择机器阅读理解模型. 数据分析与知识发现, 2021, 5(4): 134–141. [doi: 10.11925/infotech.2096-3467.2020.0714]

余笑岩(1998－), 女, 硕士生, 主要研究领域为自然语言处理, 机器阅读理解.

刘康(1981－), 男, 博士, 研究员, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理.

何世柱(1987－), 男, 博士, 副研究员, CCF 专业会员, 主要研究领域为知识图谱, 问答系统.

赵军(1966－), 男, 博士, 研究员, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理, 知识工程.

宋燃(1996－), 男, 博士生, 主要研究领域为自然语言处理, 信息检索, 知识图谱.

周永彬(1973－), 男, 博士, 研究员, 博士生导师, CCF 高级会员, 主要研究领域为网络信息安全理论与技术.