计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.009570]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多层次信息融合的多跳机器阅读理解①

朱海飞[1], 段宗涛[1], 王全伟[2], 曹建荣[1], 席铁钧[1]

[1](长安大学 信息工程学院, 西安 710064)
[2](豫西工业集团 河南北方红阳机电有限公司, 南阳 474679)
通信作者: 朱海飞, E-mail: zhuhaifei@chd.edu.cn

摘　要: 以往机器阅读理解模型中存在文本特征提取单一, 文本和问题的交互信息不全面等问题, 导致模型不能充分对文本进行理解, 本文提出了一种多层次信息融合的机器阅读理解模型. 通过在不同位置使用不同方法, 对文本信息进行多种层次的获取. 使用膨胀卷积网络捕捉文本的全局信息, 采用双向注意力机制和自注意力机制融合文本和问题之间的交互信息, 通过指针网络预测答案及其对应的支撑句. 该模型在 CAIL2019 和 CAIL2020 阅读理解数据集上训练的联合 $F1$ 值分别达到 50.09% 和 58.44%, 相比于其他基线模型取得了明显的性能提升.

关键词: 中国法研杯; 多跳机器阅读理解; 注意力机制; 信息融合

引用格式: 朱海飞,段宗涛,王全伟,曹建荣,席铁钧.基于多层次信息融合的多跳机器阅读理解.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9570.html

Multi-hop Machine Reading Comprehension Based on Multi-level Information Fusion

ZHU Hai-Fei[1], DUAN Zong-Tao[1], WANG Quan-Wei[2], CAO Jian-Rong[1], XI Tie-Jun[1]

[1](School of Information Engineering, Chang'an University, Xi'an 710064, China)
[2](Henan North Hongyang Mechanical and Electrical Co. Ltd., Western Henan Industrial Group, Nanyang 474679, China)

Abstract: In previous machine reading comprehension models, there were some problems, such as single-text feature extraction and incomplete interactive information between text and questions, which led to insufficient text understanding. This study proposes a machine reading understanding model with multi-level information fusion, which can obtain text information at multiple levels by using different methods in different locations. The model uses the dilated convolutional network to capture the global information of the text. Bi-directional attention mechanism and self-attention mechanism are used to fuse the interactive information between text and questions. Finally, the answer and its corresponding supporting sentence are predicted through the pointer network. The joint $F1$ values of the model trained on the CAIL2019 and CAIL2020 reading comprehension datasets reach 50.09% and 58.44% respectively, which achieves significant performance improvement compared with other baseline models.

Key words: China AI law challenge; multi-hop machine reading comprehension; attention mechanism; information fusion

1　引言

作为自然语言处理的研究热点之一, 机器阅读理解 (machine reading comprehension, MRC) 在智能问答、引擎搜索等任务中发挥着重要作用. 机器阅读理解任务试图让机器阅读并理解文本内容, 从而对给定的问题进行回答[1]. 即给定一段文本$P=\{p_1,p_2,\cdots,p_m\}$以及基于文本 P 提出的问题$Q=\{u_1,u_2,\cdots,u_l\}$, 需要机器阅读给定的若干段文本和问题并返回问题对应的答

① 基金项目: 陕西省重点研发计划 (2019ZDLGY17-08); 陕西省特支计划科技创新领军人才项目 (TZ0366)
收稿时间: 2024-01-27; 修改时间: 2024-02-29; 采用时间: 2024-03-11; csa 在线出版时间: 2024-05-31

案$A=\{w_1,w_2,\cdots,w_n\}$. 其中, p_i 、 u_i 、 w_i 分别表示文本、问题和答案中的一个词或一个字, m、l、n 分别表示文本、问题和答案的长度. 根据返回答案的不同类型, 可以将机器阅读理解任务分为完形填空、多项选择、片段抽取和开放式问答 4 类[2], 4 类任务的难度依次递增. 本文研究目标主要针对片段抽取任务, 即从给定的文本 P 中查找出若干段连续的语句作为问题的答案.

预训练模型以及注意力机制的出现极大地促进了机器阅读理解任务的发展, 和传统的深度学习任务相比, 采用预训练模型及注意力机制的阅读理解任务取得了全面性能优势. 传统模型需要人工进行特征工程的构建, 或者采用固定的词向量模型进行文本向量表示, 例如 Word2Vec[3], GloVe[4]等, 这种方式缺乏灵活性且任务性能表现较差. 预训练模型经过大量语料的训练, 对下游任务具有强大的泛化能力, 可以与不同的下游任务进行结合, 极大减少人工工作量, 提高任务精度. 同时, 阅读理解任务中“文本-问题”交互模式, 非常适合使用注意力机制, 对两者的交互信息进行充分挖掘, 以模拟人类阅读习惯, 使模型更关注文本和问题中的重要部分, 提高 MRC 任务的性能[5].

随着近年来裁判文书网相关法律数据的公布, 以及法研杯司法人工智能挑战赛的举行, 越来越多的司法数据得以公开, 为司法大数据相关的研究任务提供了便利条件. 法律机器阅读理解任务的目的是让机器理解法律文书, 并对相应的问题进行回答, 从而辅助司法人员提高案件处理的效率. 因此将 MRC 相关技术应用到司法领域, 对提高司法服务效率和促进 MRC 相关技术的发展具有重要的现实意义.

因此, 本文提出了基于多层次信息融合的机器阅读理解模型. 本文的主要贡献包括数据重构和多层次信息融合两方面. 数据重构主要使用改进的 BM25 算法[6]对 CAIL2019 机器阅读理解赛道的数据进行重构以适用于多跳 MRC 的训练任务, 增加训练数据量, 以及用来对模型的泛化性能进行验证. 多层次信息融合主要是对文本和问题, 以及两者的交互信息进行充分提取和融合, 以提高模型性能.

2 相关工作

2.1 机器阅读理解

早期的 MRC 任务是基于规则的或者基于机器学习技术的. Hirschman 等[7]于 1999 年开始探索 MRC 技术的研究, 设计出第 1 个以小学年级故事为语料库的自动阅读理解测试系统 Deep Read, 该系统采用词袋模型进行向量编码, 利用人工编写的规则进行模式匹配, 达到了 40% 的正确率. 这种传统的 MRC 任务大多采用模式匹配来提取特征, 其鲁棒性差、耗时长、泛化能力差. 2015 年 Hermann 等[8]提出使用深度学习神经网络模型, 其提出的 Deep LSTM Reader、Attentive Reader 和 Impatient Reader 这 3 种神经网络模型, 极大地促进了 MRC 相关任务的发展, 但其无法解决复杂推理和长文档答案推理的问题. Wang 等[9]提出的 Match-LSTM 模型, 首次将 Pointer Net 中指针的思想应用于 MRC 任务中, 提高了答案预测的准确性. Seo 等[10]提出的 BiDAF 模型奠定了 MRC 任务的基本框架, 其将 MRC 任务分为 5 层, 分别为文本嵌入层、上下文嵌入层、注意力流层、模型层和输出层, 并且首次提出了文本到问题和问题到文本的双向注意力计算方法. Yu 等[11]提出的 QANet 模型是预训练模型发布之前排名最优的一个 MRC 模型, 与之前的模型明显不同的是, QANet 模型抛弃了循环神经网络 (recurrent neural network, RNN), 只使用卷积神经网络 (convolutional neural network, CNN) 和自注意力机制完成编码工作, 但是此模型使用 CNN 进行编码不能保留长文档的信息. 随着谷歌团队的 Devlin 等[12]提出 BERT (bidirectional encoder representations from Transformers) 预训练模型后, 其在 11 个自然语言处理任务中均取得了最好效果. 李芳芳等[13]提出了一种基于多任务联合训练的法律文本 MRC 模型, 使用 RoBERTa 预训练模型进行文本编码, 对答案分类、答案预测、答案支持句子判别 3 个子任务进行联合训练, 有效地提升了 MRC 性能. 朱斯琪等[14]使用实体图结构和注意力机制对文本进行建模, 通过注意力机制融合不同细粒度的实体图结构的信息. Zhang 等[15]以 BiDAF 模型为基础, 提出了 ELMo+Gated self-attention 模型, 该模型通过引入 ELMo 预训练模型[16]和自注意力机制提高了任务性能. 深度学习模型具有强大的语义特征捕捉能力, 但是其本身缺乏可解释性, 并不能给出问题对应的答案的推理过程, 而图结构中实体和关系的网络连接关系则适合于进行逻辑推理. Ding 等[17]提出了一种基于认知图的多跳 MRC 模型, 通过抽取文本中的实体和关系建立图网络, 并加入线索增强推理时各个答案间的联系. Qiu 等[18]

通过设计一个图融合层来研究模型推理过程中的每一个推理步对应的子图变化. 但是这种通过实体和实体间的关系建立的图结构很难捕获文本的全局信息, 尤其是当问题的答案需要融合多个文档的信息时[19].

2.2 多跳机器阅读理解

根据答案抽取时是否需要参考多个文章 P 进行多步推理, 可以将机器阅读理解任务分为单跳机器阅读理解和多跳机器阅读理解[20]. 单跳机器阅读理解任务只需要根据给定的相关文章, 给出问题的答案即可, 不需要在相关文章中找出和答案对应的支持句. 因此, 大多数问题的答案趋向于查找文章 P 中与问题 Q 最相似的句子, 一般通过构建问题感知上下文表示来解决[21], 并不涉及复杂的推理过程. 使用这种阅读理解模型, 会导致模型过度关注答案的生成而并非理解文章, 使模型失去泛化能力.

多跳机器阅读理解任务与传统的单跳机器阅读理解任务不同, 其不仅需要查找相关问题的答案, 还需要给定答案在文本中的支持句. 具体而言, 给定一个问题, 模型只通过一个文章无法找到问题的正确答案, 而是需要模型根据多篇文章、经过多次理解推理才能找到答案, 相比于单跳机器阅读理解具有更好的泛化能力. 这就要求机器不仅只关注寻找相关问题的答案, 而是需要真正对文本进行理解, 从而回答问题和查找答案相关的支持句.

目前针对多跳机器阅读理解任务的研究方法可分为以下 3 类: ① 基于图结构的方法: 通过抽取文本中的实体和实体间的关系建立图网络, 通过实体间的关系对相关问题进行逻辑推理. 该方法具备一定的可解释性, 符合人的直觉思维, 但是这种图结构只具备推理实体间关系的能力, 对于整体文本的复杂逻辑关系难以推理, 尤其是跨语句间的答案查找, 图网络无法进行推理[22]. ② 将多跳任务转换为多个单跳任务, 再按照传统的单跳问题进行回答. 该方法可以分别对单跳任务进行回答, 再进行答案综合, 可以取得一定的效果. 但是多个单跳任务间缺乏逻辑联系和可解释性, 对于复杂的多跳任务难以有效模拟[23,24]. ③ 使用深层次的神经网络模拟多跳任务[25]: 使用深层次的神经网络强大的特征捕捉能力对文本信息进行挖掘, 使模型对文本信息进行充分理解, 其在多跳机器阅读理解任务上表现较好. 但是随着神经网络层次的加深, 网络训练面临着梯度消失和梯度爆炸的问题, 需要采取各种措施进行缓解. 本文采用基于预训练模型的深度神经网络来进行建模, 以充分利用预训练模型强大的泛化能力和深度神经网络的特征捕捉能力.

针对以上问题, 本文提出了一种机器阅读理解模型, 我们的主要工作包括 3 个方面.

(1) 提出了一种基于多层次信息融合的机器阅读理解模型, 模型可对文本和问题的交互信息进行有效融合, 可较为准确地查找问题的答案和相应的支持句.

(2) 提出了一种基于 BM25 算法的改进算法, 对 CAIL2019 数据集进行重构, 使其适用于多跳 MRC 任务, 在增加数据量的同时, 可以对模型的泛化性能进行验证.

(3) 在 CAIL2020 数据集和重构的 CAIL2019 数据集上进行实验, 验证模型性能.

3 模型设计

本文提出的机器阅读理解模型, 使用多种方法在多个层次上对文本信息进行获取. 使用膨胀卷积网络 (dilated convolution network, DCN)[26]对输入文本序列的全局特征进行捕捉, 膨胀卷积网络可以在保持参数不变的情况下, 增大卷积核的感受野, 更好的捕捉文本特征. 使用双向注意力机制和自注意力机制, 捕获全局文本交互信息以及问题和文本之间的交互信息.

3.1 模型结构

为了充分提取文本序列特征以及对文本和问题包含的信息进行交互, 本文提出了基于多层次信息融合的多跳阅读理解模型, 模型结构分为 4 个子模块: 文本嵌入模块、上下文交互模块、信息融合模块、答案预测模块, 模型结构如图 1.

3.1.1 文本嵌入模块

文本嵌入模块的作用是对问题和文档进行向量化编码以输入到下层模块中. 本文将多个文档和问题进行拼接作为输入文本, 文本拼接方式为 *Inputs*=[CLS]+*Q*+[SEP]+*D*+[SEP], 其中 Q 表示问题, D 表示多篇拼接的文档, [CLS]表示文本的起始位置, [SEP]表示问题和文档的分隔符. 将 *Inputs* 输入到 RoBERTa-wwm-ext[27] 预训练模型进行编码得到编码向量表示, 再将编码结果输入到高速公路网络 (highway network)[28], 得到文本的向量化表示 $P \in R^{H \times d}$, 对应于图 1 中的 Passage_encoding. 其中, H 为文本的长度, d 为向量编码的维度. 高速公路网络的核心是将网络的输出跳跃着连接到后

面的网络层, 在保证模型深度的同时减少梯度消失和梯度爆炸的问题. 根据文档和问题的长度, 分别获取文档和问题的编码表示, 分别记为 $Q\in R^{Hq\times d}$ 和 $C\in R^{Hc\times d}$, 分别对应于图 1 中的 Question_encoding 和 Context_encoding, 其中 Hq 和 Hc 分别为问题和文档的长度.

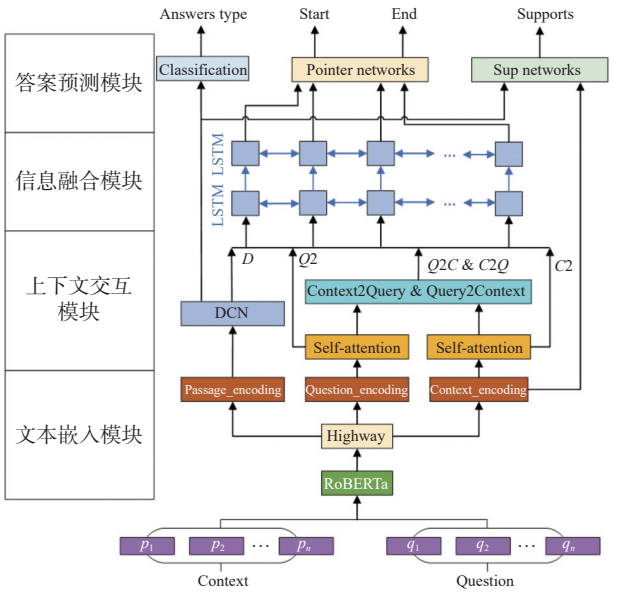

图 1 多层次信息融合的机器阅读理解模型结构图

3.1.2 上下文交互模块

上下文交互模块包含两个子任务: 全局特征捕捉和文本注意力计算, 其中全局特征捕捉使用膨胀卷积网络 DCN:

$$D=\mathrm{DCN}(P)\in R^{H\times 2d} \tag{1}$$

文本注意力计算部分包含 4 个部分: 问题自注意力、文档自注意力、文档到问题的交互注意力、问题到文档的交互注意力, 注意力计算如下.

问题自注意力 $Q2$ 和文档自注意力 $C2$ 使用自注意力进行计算, self-attention 为自注意力计算函数, 计算公式如下所示:

$$Q2=\text{self-attention}(Q) \tag{2}$$

$$C2=\text{self-attention}(C) \tag{3}$$

$C2Q$ 是文档到问题的注意力, 首先通过问题和文档的向量表示 C 和 Q 计算两者的相似度概率矩阵 $S\in R^{Hc\times Hq}$, 即文档和问题中各个词之间的相似度得分, 然后对 S 按行做归一化处理, 将归一化的矩阵与 Q 相乘得到融合注意力权重的问题向量:

$$S_{ij}=\alpha\left(C_{:i},Q_{:j}\right) \tag{4}$$

$$\alpha(c,q)=u[c;q;c\cdot q] \tag{5}$$

$$\delta_i=Softmax(S_{i:}) \tag{6}$$

$$Q'=\sum_j\delta_{ij}Q_{:j}\in R^{H\times 2d} \tag{7}$$

其中, S_{ij} 表示文档中第 i 个词和问题中第 j 个词的相似度, α为可训练的缩放函数, u 为可训练参数矩阵, 式中[;]表示向量拼接, [·]表示向量点积. δ_i是 S 每一行归一化后的结果, 将δ_i与 Q 中的每一列加权求和得到新的向量 Q', 表示文档到问题的注意力.

$Q2C$ 是问题到文档的注意力, 计算出每一个问题的单词在文档中相似度最高的词, 对相似度矩阵 S 的列方向进行归一化, 然后计算文档向量加权和, 计算公式如下:

$$\alpha_j=Softmax\left(S_{:j}\right) \tag{8}$$

$$C'_{:j}=\sum_i\alpha_{ij}C_{:j} \tag{9}$$

$$I=\omega(C,Q',C')\in R^{H\times 2d} \tag{10}$$

$$\omega(c,q',c')=[c;q';c\cdot q';c\cdot c'] \tag{11}$$

其中, α_j是 S 每一列归一化后的结果, 将α_j与 C中的每一行加权求和得到新的向量 C', 对以上计算的向量进行线性变换, 其中ω为可训练函数, 最终的输出向量为 I, 表示问题到文档的注意力.

3.1.3 信息融合模块

信息融合模块主要负责将上下文交互模块产生的独立的信息进行进一步的交互融合, 此模块的输入有 5 部分, 分别为文本全局特征 D、问题自注意力 $Q2$、文档自注意力 $C2$、问题到文档的注意力 $Q2C$、文档到问题的注意力 $C2Q$, 将以上信息进行拼接后输入到前馈网络层进行线性变换, 然后将其输入到双向长短期记忆网络 (bi-directional long short-term memory, BiLSTM) 中进行融合, 计算公式如式 (12)、式 (13) 所示:

$$in=w(D;Q2;Q2C;C2Q) \tag{12}$$

$$out=\mathrm{BiLSTM}(in)\in R^{H\times 4d} \tag{13}$$

其中, w 为可训练参数矩阵, 拼接的向量经过维度转换之后维度为 $in\in R^{H\times 2d}$, 将其输入 BiLSTM 得到上下文交

互模块的最终输出 *out*.

3.1.4 答案预测模块

答案预测模块主要完成 3 个子任务: 答案类型判断、答案片段抽取、答案支持句判别.

答案类型判断: CAIL2020 机器阅读理解任务的答案类型有 4 种: Span、YES、NO、Unknown. 其中, Span 类型答案是从文档中抽取出的连续片段作为答案, YES 和 NO 类型答案表示答案类型为是或否, Unknown 类型答案是文档中不存在答案的情况. 答案类型判断可视为分类任务, 本文使用 RoBERTa-wwm-ext 预训练模型编码中的分类标记 [CLS]位置的编码向量进行文本分类, 将 [CLS]位置的编码向量 $P \in R^{H \times d}$ 送入全连接层进行四分类任务, 得到答案类型的概率为 p_i, 计算公式如式 (14) 所示:

$$p_i = Softmax(\text{Linear}(P)) \tag{14}$$

在答案类型预测子任务中, 使用实际答案类型 q_i 和预测答案类型 p_i 的交叉熵作为损失函数, 计算公式如式 (15) 所示:

$$L_1 = \text{CrossEntropy}(q_i, p_i) \tag{15}$$

答案片段抽取: 本文使用指针网络 pointer network[29] 预测答案在文档中的开始位置 q_s 和结束位置 q_e, 计算公式如式 (16)、式 (17) 所示:

$$q^s = \tanh(W^s out) \tag{16}$$

$$q^e = \tanh(W^e out) \tag{17}$$

其中, W^s 和 W^e 是可训练的参数矩阵.

在答案片段抽取子任务中, 损失函数定义为预测答案开始、结束位置和答案实际的开始、结束位置的交叉熵, 计算公式如式 (18) 所示:

$$L_2 = \text{CrossEntropy}\left(q_j^s, P_j^s; q_j^e, P_j^e\right) \tag{18}$$

其中, P_j^s和P_j^e表示第 j 个样本答案的真正开始和结束位置, q_j^s和q_j^e表示第 j 个样本预测答案的开始和结束位置.

答案支持句判别: 此子任务主要用来查找答案对应的支持句子, 可将其视为分类任务. 首先计算文档编码 C 和文本编码 P 的点积, 获取每个句子的表示 z_{sup}, 并对其列方向进行归一化, 然后输入到全连接层进行支持句预测, 得到支持句的分布概率为 p_{sup}, 计算公式如式 (19)、式 (20) 所示:

$$z_{\text{sup}} = [C \cdot P] \tag{19}$$

$$p_{\text{sup}} = \text{Linear}\left(Softmax\left(z_{\text{sup}}\right)\right) \tag{20}$$

在支持句判别子任务中, 损失函数定义为预测的支持句 p_{sup} 和实际的支持句 q_{sup} 的交叉熵, 计算公式如式 (21) 所示:

$$L_3 = \text{CrossEntropy}\left(p_{\text{sup}}, q_{\text{sup}}\right) \tag{21}$$

最终结果通过共享模型底层参数设计多任务学习函数[30], 共同训练上述 3 个子任务, 计算公式如式 (22)、式 (23) 所示:

$$L = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 \tag{22}$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1 \tag{23}$$

3.2 数据集样例

本文所使用的数据集为法研杯机器阅读理解赛道发布的数据集, 包括 CAIL2019 数据集和 CAIL2020 数据集, CAIL2020 数据集样例如图 2 所示, 每一个数据样例包含若干个句子级的文档、一个根据文档提出的问题、问题对应的答案和答案对应的支持句子.

句子1: 经审理查明,
句子2: 2016年6月初至同月21日,
句子3: 被告人汪某1在经营位于绍兴市柯桥区漓渚镇永安路16号邱某凤熟食摊期间,
句子4: 在制作熟烤麸、熟腐竹的过程中添加依法禁止添加的防腐剂山梨酸,
句子5: 并销售供人食用.
句子6: 经检测汪某1销售的熟烤麸、熟腐竹中山梨酸含量为0.2 g/kg,
句子7: 不符合《食品安全国家标准食品添加剂使用标准》的要求(熟烤麸、熟腐竹不在山梨酸添加范围内).
句子8: 案发后被告人汪某1经民警电话联系主动到案,
句子9: 如实供述了上述事实.
句子10: 上述事实,
句子11: 被告人汪某1在开庭审理过程中亦无异议,
句子12: 并有营业执照、抓获经过、绍兴市柯桥区市场监督管理局涉嫌犯罪移送材料,
句子13: 被告人汪某1的违法行为,
句子14: 足以认定.

问题: 汪某1在哪些食物中添加了防腐剂山梨酸?
答案: 熟烤麸、熟腐竹
支持句: 3, 4

图 2 CAIL2020 数据集样例

CAIL2019 数据集样例如图 3 所示, 每一段文本对应着若干个问题, 每一个问题对应的答案仅有答案文本, 并没有支撑答案所需要的支持句, 不符合多跳 MRC 任务的数据要求.

文章: 经审理查明, 被告人张某3、张某1系张广庙乡九龙村村民, 系父子. 2013年至2014年期间, 张某3、张某1以非法占有为目的, 强行索要他人钱财. 具体事实如下: 2013年2月1日, 被告人张某3利用固始县张广庙乡九龙村开发建设九龙新型农村社区之机, 以阻止施工相威胁, 强行索要施工方张某2、李某5人民币20万. 其中7万被其送给李某2, 后李某2将此款退还张某3、张某1.

问题: 被告人利用什么阻止施工进行?

答案: 九龙新型农村社区之机

图 3 CAIL2019 数据集样例

3.3 数据重构

目前 MRC 任务面临的困难之一是相关数据集的缺乏, 尤其是针对中文 MRC 任务的相关数据集. CAIL2019 的数据集是针对单跳 MRC 任务的, 需要对其进行重构以适应多跳 MRC 任务, 用来增加训练数据量以及验证模型泛化性能. 因此, 本文提出了一种基于 BM25 (best matching) 的改进算法对 CAIL2019 数据集进行重构, 将文章拆分为多个句子级的文档, 以寻找答案对应的支持句子.

BM25 算法是信息检索领域计算查询和文档的相似度得分的经典算法, 该算法包含两个重要指标, 分别是查询中每个词的权重以及查询中每个词与文档的相关性得分, 计算公式如下:

$$Score(Q,d)=\sum_{i}^{n}W_iR(q_i,d) \tag{24}$$

其中, Q 表示一条查询语句, q_i 表示对 Q 分词后的词语; d 表示查询文档; W_i 表示查询 Q 中单词 q_i 的权重; $R(q_i, d)$ 表示单词 q_i 与文档 d 的相关性得分. 单词权重 W_i 一般使用逆文档词频 (inverse document frequency, IDF) 进行计算:

$$IDF(q_i)=\log\frac{n-n(q_i)+0.5}{n(q_i)+0.5} \tag{25}$$

其中, n 为检索中的所有文档数, $n(q_i)$ 为包含了 q_i 的文档数. 由式 (25) 可以看出, 包含 q_i 的文档数越多, 则 q_i 的权重越低. 查询中的单词与文档的相关性得分计算公式如下所示:

$$R(q_i,d)=\frac{f_i(k_1+1)}{f_i+K}\frac{F_i(k_2+1)}{F_i+k_2} \tag{26}$$

$$K=k_1\left(1-b+b\frac{l}{avg(l)}\right) \tag{27}$$

其中, k_1、k_2、b 为调节因子, 通常根据经验设置, 一般为 k_1=1.2, b=0.75; f_i 为 q_i 在文档 d 中的出现频率, F_i 为 q_i 在查询 Q 中的出现频率; l 为文档的长度, $avg(l)$ 为所有文档的平均长度. 由于绝大多数情况下, q_i 在查询 Q 中只会出现一次, 即 F_i=1. 因此式 (26) 可化简为:

$$R(q_i,d)=\frac{f_i(k_1+1)}{f_i+K} \tag{28}$$

由 BM25 算法的式 (24) 可知, BM25 算法计算时考虑了查询 Q 中每个单词 q_i 和文档 d 的相似度, 但没有考虑查询 Q 和文档 d 的整体相似度, 因此本文提出了一种改进的 BM25 算法, 将考虑计算查询 Q 和文档 d 的整体相似度, 改进的 BM25 算法的计算公式如下:

$$Score(Q,d)=R(Q,d)\sum_{i}^{n}W_iR(q_i,d) \tag{29}$$

其中, $R(Q,d)$ 表示查询 Q 和文档 d 的整体相似度得分, 其计算步骤如下.

使用 RoBERTa-wwm-ext 预训练模型对文本和查询分别进行向量编码, 再使用余弦相似度方法计算两者编码向量的相似度, 得到两者的相似度得分为$R(Q,d)$. 通过在具有答案支持句标记的 CAIL2020 数据集上进行改进 BM25 算法的验证实验, 验证了改进算法的有效性. 因此, 本文使用改进的 BM25 算法对 CAIL2019 数据集进行重构. 将问题 Q 作为查询, 将文章拆分后的句子作为文档, 计算两者的相似度得分并进行排序. 实验结果表明, 选取和得分最高的句子分差小于 20% 的句子作为支持句, 可以获得最佳的支持句判定精度.

4 实验

4.1 数据集

使用改进的 BM25 算法重新标记的 CAIL2019 数据集包含 37 600 条数据, CAIL2020 数据集包含 5 000 条数据. 将两个数据集按照 7:2:1 的比例划分为训练集、验证集、测试集, 划分结果如表 1 所示.

表 1 实验数据集

数据集	train	dev	test
CAIL2019	26 320	7 520	3 760
CAIL2020	3 500	1 000	500

CAIL2019 和 CAIL2020 数据集的文章长度统计如图 4、图 5 所示, 由于数据集中文本长度超过 512 的部分较少, 所以不对输入数据进行截断或滑动窗口处理, 仅对文本长度超过 512 的数据进行适当删减, 使其长度小于预训练模型输入长度限制.

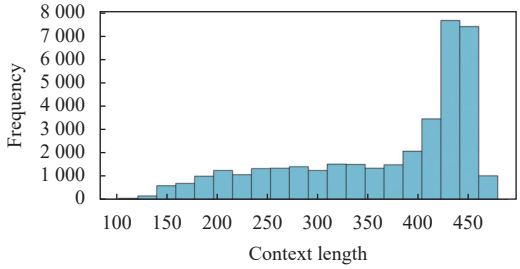

图 4 CAIL2019 数据集文本长度分布

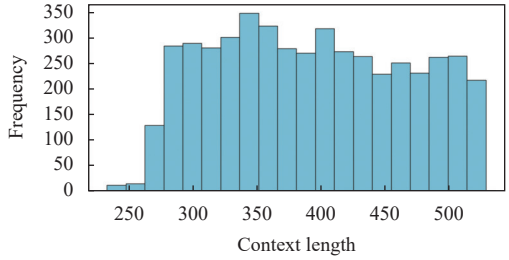

图 5 CAIL2020 数据集文本长度分布

4.2 评估指标

片段抽取式机器阅读理解任务的性能评价, 主要是根据对比预测答案和真实答案的匹配度来进行的, 评价指标主要为精确匹配 (exact match, *EM*) 值和模糊匹配 *F*1 值, 同时, 片段抽取阅读理解任务中还需要考虑答案支持句, 所以本文评价指标增加了联合 *F*1 值进行评价.

EM 用来计算预测答案与真实答案完全相同的比例, 当两个答案完全一致时得分为 1, 否则为 0, 计算公式如式 (30) 所示:

$$\begin{cases} EM = 1, \text{ if pred} = \text{gold} \\ EM = 0, \text{ otherwise} \end{cases} \tag{30}$$

其中, pred 为预测答案, gold 为真实答案.

*F*1 值用来计算预测答案和真实答案之间单词的重合程度, 它不要求两者完全一致, 因此比 *EM* 评价指标更为宽松, 计算公式为准确率和召回率的加权平均值, 计算公式如式 (31) 所示:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{31}$$

其中, *precision* 为准确率, *recall* 为召回率.

联合 *F*1 根据答案和支持句的精度和召回率进行联合计算, 计算公式如式 (32)–式 (34) 所示:

$$precision^{\text{joint}} = precision^{\text{ans}} \times precision^{\text{sup}} \tag{32}$$

$$recall^{\text{joint}} = recall^{\text{ans}} \times recall^{\text{sup}} \tag{33}$$

$$Joint\ F1 = \frac{2 \times precision^{\text{joint}} \times recall^{\text{joint}}}{precision^{\text{joint}} + recall^{\text{joint}}} \tag{34}$$

其中, $precision^{\text{ans}}$、$precision^{\text{sup}}$ 和 $recall^{\text{ans}}$、$recall^{\text{sup}}$ 分别为答案和支持句的精度和召回率. $precision^{\text{joint}}$ 和 $recall^{\text{joint}}$ 为联合精度和召回率.

4.3 实验设置

本文选用哈工大讯飞联合实验室 (HFL) 发布的中文预训练模型 Chinese-RoBERTa-wwm-ext[27]来进行向量编码, 向量编码维度为 768, 实验设置文档最大输入长度为 512, 问题最大输入长度为 50, 训练集 batch_size 大小为 2, 测试集和验证集 batch_size 为 4, dropout 设置为 0.1, DCN 网络的膨胀率设置为 2. 训练中使用 BERTAdam 优化器进行优化, 学习率设置为 2E–4, 模型训练 10 个 epoch 后停止.

4.4 实验结果

本文选择了 6 个基准模型进行对比实验, 分别为: BiDAF[10]模型、QANet[11]模型、BERT[12]模型、比赛官方的 RoBERTa[27]基准模型、RAiO[31]模型、R-NET[32]模型. 由于本文针对的是中文数据集, 所以使用 RoBERTa 对文本进行向量编码, 代替基准模型中针对英文数据集的字符级和词级别的向量编码. 各个基准模型和本文模型在 CAIL2019 和 CAIL2020 数据集上的实验结果如表 2、表 3 所示.

表 2 不同模型在 CAIL2019 数据集上的实验结果 (%)

实验模型	*EM*	*F*1	*Joint F*1
BiDAF	41.30	60.20	41.67
R-NET	42.34	58.44	41.50
QANet	47.72	63.81	48.04
RAiO	46.54	67.52	47.95
BERT	47.21	68.40	49.11
RoBERTa	47.91	69.64	49.27
Ours	48.14	69.33	50.09

表 3 不同模型在 CAIL2020 数据集上的实验结果 (%)

实验模型	*EM*	*F*1	*Joint F*1
BiDAF	44.32	61.83	49.10
R-NET	46.28	60.95	48.74
QANet	49.10	66.32	53.22
RAiO	51.67	70.90	56.17
BERT	52.31	69.12	53.06
RoBERTa	52.70	68.05	54.89
Ours	66.64	76.22	58.44

本文提出的模型在两个数据集上均表现较好, 体现了模型良好的泛化性能. 相比于比赛官方的基准模型和其他基准模型, 本文所提出的模型的表现优于基

准模型. 同时能够发现各个模型在 CAIL2020 数据集上的训练结果明显优于在 CAIL2019 数据集上的训练结果, 这可能是由于改进的 BM25 算法对 CAIL2019 数据集进行标记的过程中, 并没有找到问题所对应的所有支持句或将不相关的句子作为支持句而引起的.

4.5 消融实验

为了评估模型中每个部分的贡献, 在本文所提出的模型中分别去掉双向注意力模块、自注意力模块、DCN 网络、Highway 网络, 在 CAIL 2020 数据集上验证各个模块对模型性能的影响, 实验结果如表 4 所示.

表 4 不同模块在 CAIL2020 数据集上的消融实验结果 (%)

模型	*EM*	*F*1	*Joint F*1
Ours	66.64	76.22	58.44
双向注意力	49.19	67.51	55.20
自注意力	50.06	71.28	57.11
DCN网络	51.30	68.69	54.39
Highway网络	51.90	72.10	57.52

由表 4 可以看出, 各个模块均会对模型性能产生影响. 其中双向注意力模块对模型性能影响最大, 可见文本和问题的信息交互对 MRC 模型充分理解文本有明显作用. 自注意力模块加强了文本和问题对自身的理解, 两者的信息相对独立, 从而对模型性能影响较小. DCN 网络加强了整个文本的特征捕捉能力, 对 MRC 模型寻找正确答案及支持句有较大贡献. Highway 网络对模型计算稳定性和收敛速度有贡献, 而对模型性能影响相对较小.

5 结论

本文提出了一种基于多层次信息融合的机器阅读理解模型, 该模型能够较为全面地综合文档和问题的特征以及两者的交互信息. 同时, 针对 CAIL2019 数据集只能适用于单跳 MRC 任务的情况, 提出了一种改进的 BM25 算法对 CAIL2019 的数据集进行重构, 使其适应于多跳 MRC 任务, 增加了数据量, 同时更好地验证模型的泛化性能.

未来的研究工作中, 我们希望将知识图谱引入到模型训练过程中, 通过构建不同层级的知识图谱来查找支持句子, 同时能够将答案的推理过程进行可视化分析, 这对于认识 MRC 模型对文本的理解过程具有重要意义.

参考文献

1 Cui YM, Zhang WN, Che WX, *et al*. Multilingual multi-aspect explainability analyses on machine reading comprehension models. Iscience, 2022, 25(5): 104176. [doi: 10.1016/j.isci.2022.104176]

2 Liu SS, Zhang X, Zhang S, *et al*. Neural machine reading comprehension: Methods and trends. Applied Sciences, 2019, 9(18): 3698. [doi: 10.3390/app9183698]

3 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. Proceedings of the 2013 International Conference on Learning Representations. Scottsdale: ICLR, 2013.

4 Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: ACL, 2014. 1532–1543.

5 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

6 Wang JJ, Wang S, Cui Q, *et al*. Local-based active classification of test report to assist crowdsourced testing. Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering. Singapore: IEEE, 2016. 190–201.

7 Hirschman L, Light M, Breck E, *et al*. Deep Read: A reading comprehension system. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. College Park: ACL, 1999. 325–332.

8 Hermann KM, Kočiský T, Grefenstette E, *et al*. Teaching machines to read and comprehend. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 1693–1701.

9 Wang SH, Jiang J. Machine comprehension using match-lstm and answer pointer. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

10 Seo MJ, Kembhavi A, Farhadi A, *et al*. Bidirectional attention flow for machine comprehension. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview. net, 2017.

11 Yu AW, Dohan D, Luong MT, *et al*. QANet: Combining local convolution with global self-attention for reading comprehension. Proceedings of the 6th International Conference on Learning Representations. Vancouver: OpenReview.net, 2018.

12 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding.

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL, 2018. 4171–4186.

13 李芳芳, 任星凯, 毛星亮, 等. 基于多任务联合训练的法律文本机器阅读理解模型. 中文信息学报, 2021, 35(7): 109–117, 125. [doi: 10.3969/j.issn.1003-0077.2021.07.013]

14 朱斯琪, 过弋, 王业相, 等. TransformerG: 基于层级图结构与文本注意力机制的法律文本多跳阅读理解. 中文信息学报, 2022, 36(11): 148–155, 168. [doi: 10.3969/j.issn.1003-0077.2022.11.015]

15 Zhang WW, Ren FJ. ELMo+Gated self-attention network based on BiDAF for machine reading comprehension. Proceedings of the 11th IEEE International Conference on Software Engineering and Service Science (ICSESS). Beijing: IEEE, 2020. 1–6.

16 Ilić S, Marrese-Taylor E, Balazs JA, *et al*. Deep contextualized word representations for detecting sarcasm and irony. Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Brussels: ACL, 2018. 2–7.

17 Ding M, Zhou C, Chen QB, *et al*. Cognitive graph for multi-hop reading comprehension at scale. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 2694–2703.

18 Qiu L, Xiao YX, Qu YR, *et al*. Dynamically fused graph network for multi-hop reasoning. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 6140–6150.

19 Shao N, Cui YM, Liu T, *et al*. Is graph structure necessary for multi-hop question answering? Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2020. 7187–7192.

20 Cui Y, Liu T, Che W, *et al*. A span-extraction dataset for Chinese machine reading comprehension. arXiv:1810.07366, 2018.

21 Jiang YC, Bansal M. Self-assembling modular networks for interpretable multi-hop reasoning. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 4474–4484.

22 Wang WY, Pan S. Deep inductive logic reasoning for multi-hop reading comprehension. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin: ACL, 2022. 4999–5009.

23 Min S, Zhong V, Zettlemoyer L, *et al*. Multi-hop reading comprehension through question decomposition and rescoring. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 6097–6109.

24 Perez E, Lewis P, Yih WT, *et al*. Unsupervised question decomposition for question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2020. 8864–8880.

25 Mohammadi A, Ramezani R, Baraani A. A comprehensive survey on multi-hop machine reading comprehension approaches. arXiv:2212.04072, 2022.

26 Yu F, Koltun V, Funkhouser T. Dilated residual networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 636–644.

27 Cui YM, Che WX, Liu T, *et al*. Pre-training with whole word masking for Chinese BERT. arXiv:1906.08101, 2019.

28 Srivastava RK, Greff K, Schmidhuber J. Highway networks. arXiv:1505.00387, 2015.

29 Vinyals O, Fortunato M, Jaitly N. Pointer networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 2692–2700.

30 Fan Y, Liu QW. A retriever-reasoner method for multi-document reading comprehension. Proceedings of the 19th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). Chengdu: IEEE, 2022. 1–5.

31 Phan TA, Jung JJ, Bui KHN. Read-all-in-once (RAiO): Multi-layer contextual architecture for long-text machine reading comprehension. IEEE Access, 2023, 11: 77873–77879. [doi: 10.1109/ACCESS.2023.3298100]

32 Microsoft Asia Natural Language Computing Group. R-NET: Machine reading comprehension with self-matching networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Beijing: Microsoft Reserach Asia, 2017.

(校对责编: 孙君艳)