计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(3):212−219 [doi: 10.15888/j.cnki.csa.008417]

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

面向机器阅读理解的多任务层次微调模型①

丁美荣, 刘鸿业, 徐马一, 龚思雨, 陈晓敏, 曾碧卿

(华南师范大学 软件学院, 佛山 528225)
通信作者: 刘鸿业, E-mail: a470141216@163.com

摘　要: 机器阅读理解与问答一直以来被认为是自然语言理解的核心问题之一, 要求模型通过给定的文章与问题去挑选出最佳答案. 随着 BERT 等预训练模型的兴起, 众多的自然语言处理任务取得了重大突破, 然而在复杂的阅读理解任务方面仍然存在一些不足, 针对该任务, 提出了一个基于回顾式阅读器的机器阅读理解模型. 模型使用 RoBERTa 预训练模型对问题与文章进行编码, 并将阅读理解部分分为词级别的精读模块与句子级别的泛读模块两个模块. 这两个模块以两种不同的粒度来获取文章和问题的语义信息, 最终结合两个模块的预测答案合并输出. 该模型在 CAIL2020 的数据集上综合 *F*1 值达到了 66.15%, 相较于 RoBERTa 模型提升了 5.38%, 并通过消融实验证明了本模型的有效性.

关键词: 自然语言处理; 机器阅读理解; 多任务学习; 预训练语言模型; 层次微调

引用格式: 丁美荣,刘鸿业,徐马一,龚思雨,陈晓敏,曾碧卿.面向机器阅读理解的多任务层次微调模型.计算机系统应用,2022,31(3):212–219. http://www.c-s-a.org.cn/1003-3254/8417.html

Multi-task Hierarchical Fine-tuning Model Toward Machine Reading Comprehension

DING Mei-Rong, LIU Hong-Ye, XU Ma-Yi, GONG Si-Yu, CHEN Xiao-Min, ZENG Bi-Qing

(School of Software, South China Normal University, Foshan 528225, China)

Abstract: Machine reading comprehension and question answering has long been considered as one of the core problems of natural language understanding, which requires models to select the best answer from a given text and question. With the rise of pre-trained language models such as BERT, great breakthroughs have been made in natural language processing (NLP) tasks. However, there are still some shortcomings in complex reading comprehension tasks. To solve this problem, this study proposes a machine reading comprehension model based on retrospective readers. The proposed model uses the pre-trained model RoBERTa to encode questions and articles and divides the reading comprehension section into two modules: an intensive reading module at the word level and a comprehensive reading module at the sentence level. These two modules capture the semantic information in articles and problems at two different granularity levels. Finally, the prediction results of the two modules are combined to produce the answer with the highest probability. The model accuracy is improved in the CAIL2020 dataset and the *Joint_F*1 value of the model reaches 66.15%, which is 5.38% higher than that of the RoBERTa model. The effectiveness of this model is proved by ablation experiments.

Key words: natural language processing; machine reading comprehension; multi-task learning; pre-trained language model; hierarchical fine-tuning

1　引言

机器阅读理解 (machine reading comprehension, MRC) 是通过计算机理解文章语义并回答相关问题的一项重要研究任务. 让计算机系统能够理解文本的含

① 基金项目: 国家自然科学基金 (61876067); 广东省普通高校人工智能重点领域专项 (2019KZDZX1033); 广东省信息物理融合系统重点实验室建设专项 (2020B1212060069)
收稿时间: 2021-05-31; 修改时间: 2021-07-07, 2021-07-20; 采用时间: 2021-07-30; csa 在线出版时间: 2022-01-24

义, 并且做出正确的反馈, 是自然语言处理的长期目标. 机器阅读理解的研究对提升机器的自然语言理解能力具有重要促进作用, 已受到学术界和工业界的广泛关注.

将机器阅读理解技术运用到智能客服问答系统中, 能够更进一步的提高问答系统的扩展性和效率. 例如, Long 等[1] 提出了基于机器阅读理解的智能审判模型 AutoJudge. Xu 等[2] 推出了医疗阅读理解数据集 CliCR. Šuster 等[3] 提出了一种深度学习模型 StockNet, 可以同时分析股价历史以及 Twitter 上相关的新闻.

目前机器阅读理解任务仍然面临着许多挑战, 例如推理性不强, 缺乏可解释性等问题[4]. 而将机器阅读理解应用于法律等其他领域时, 对知识推理能力与可解释性有着更加巨大的需求和挑战. 缺乏大量的训练数据是当前阅读理解领域急需解决的问题之一. 现有的阅读理解模型和相关数据集存在一定限制, 限定输出只有单个答案. 而实际应用中, 存在大量场景不仅需要答案, 还需要这个答案具有一定的解释性, 需要同时找出能够支撑该答案的相关句子. 目前在司法领域的机器阅读理解研究还较少, 因此也需要我们更多地围绕该领域对以上问题展开对有效模型和方法的研究探索.

在本文 CAIL2020 比赛提供的中文法律机器阅读理解数据集就是能够解决这一问题的数据集. 该数据集与传统机器阅读理解的数据集有所不同, 传统机器阅读理解任务中, 只需要给定文章和问题, 通过模型得到答案. 而这里的数据集参考了 HotpotQA 数据集, 使得该机器阅读理解任务具有更大的挑战, 要求实现观点类问题作答的同时, 每个问题都需给出答案的依据, 如图 1 所示.

针对以上问题, 本文结合预训练模型, 提出一种层次微调模型 (hierarchical fine-tuning reader model, HF-Reader Model), 与机器阅读理解中主流的几种预训练模型进行实验对比和探索性研究分析, 主要贡献包括 3 个方面.

(1) 提出一种机器阅读理解模型, 能够以文章作为依据回答问题, 并且能够有效定位多个支持答案的句子.

(2) 参考 Retrospective reader 方法并进行修改, 将模型分为精读模块与泛读模块两个部分. 精读模块使用字级别的向量对问题和文章进行处理, 用作抽取以及分类, 泛读模块使用句子级别的向量对问题与文章进行处理, 用作句子二分类. 通过实验证明该思路的有效性.

(3) 提出层次微调方法. 使用 RoBERTa 预训练模型, 在数据集上先使用多任务学习方法进行微调, 再用于精读模块与泛读模块进行二次微调, 使模型得到进一步的提升, 验证该方法是较优的提升方式.

[0] 本院经审理认定事实如下: 被告系呼和浩特市赛罕区风华园小区23号楼5单元2楼中户房屋的业主, [1] 房屋建筑面积为81.79平方米。[2] 014年3月20日,[3] 原告与风华园小区业主委员会签订《物业管理委托合同》, [4] 合同期限为2014年3月20日至2016年3月20日,[5] 合同约定原告为风华园小区提供物业服务, [6] 住宅房屋业主按建筑面积每月每平方米0.45元支付物业管理服务费。[7] 2016年3月1日,[8] 原告与风华园小区业主委员会续签《物业管理委托合同》,[9] 合同期限为2016年3月20日至2019年3月20日, [10] 住宅房屋的物业费变更为按建筑面积每月每平方米0.5元。[11] 2014年3月21日至2017年3月20日被告未向原告交纳物业服务费。[12] 另查明, [13] 呼和浩特市欣民物业服务有限责任公司于2016年1月14日将公司名称变更为内Xxxxxxxxxx0。
问题: 续签的《物业管理委托合同》中是否变更了物业费?
答案: yes
答案支撑句: 6, 8

图 1 CAIL2020 阅读理解数据样例

2 研究进展

机器阅读理解是一种利用算法使计算机理解文章语义并回答相应问题的技术. 本节从当前国内外经典的抽取式机器阅读理解数据集以及相关模型两个角度来对机器阅读理解进行表述.

抽取式机器阅读理解的主要任务是给定一段文本和问题, 通过模型从文中抽取相应的内容来回答问题.

Rajpurkar 等[5] 提出了 SQuAD 数据集, 该数据集是通过人工方式获取的基于维基百科构造的数据集, 许多经典的模型都是基于此数据集提出的, 例如, Seo 等[6] 提出了 BiDAF 模型, 使用双向注意力流的方法获取上下文表示, 使用多层次处理数据在 SQuAD 上取得了较好的效果; Wang 等[7] 提出了 R-Net 模型, 一种基于门控与自注意力的模型, 该模型在某些指标上已经接近人类水平. Huang 等[8] 提出了 FusionNet, 基于对以往工作中注意力方法的分析, 提出单词历史与全关注注意力使模型结合了不同语义层的信息流.

Devlin 等[9] 提出的双向语言理解模型 BERT, 在 11 种不同的自然语言处理任务中达到了最佳成绩. 该模型使用了 Vaswani 等[10] 提出的多层 Transformers 结构, 并且使用掩码机制对大量文本进行无监督训练, 再将训练好的模型用于下游任务. 在这之后如 CoQA, SQuAD2.0, HotpotQA 等更具有挑战性的数据集也逐

渐被发表出来. Reddy 等[11] 提出的 CoQA 数据集将抽取式阅读理解引入对话场景中, 通过多轮对话问答的方式进行问答, 该数据集还额外包含了答案为 Yes/No 的问题, 使得模型不仅仅能从文中抽取答案, 还能根据原文信息判断 Yes/No 的问题. Zhu 等[12] 提出了一种全新的基于上下文注意力机制的深度神经网络 SDNet 来解决对话问答任务, 并将前几轮的问题和对当前问题的回答加入上下文从而解决多轮对话问题.

Rajpurkar 等[13] 提出的 SQuAD2.0 数据集则是包含了不可回答问题, 使得模型不仅仅需要回答原文中可以找到答案的问题, 还要避免回答原文中找不到答案的问题, 从而达到更深层次的理解. Zhang 等[14] 首次提出回顾式阅读方法 Retro-Reader 模型, 将其化为两个模块分别进行训练, 第一个模块先进行判断是否是可回答问题, 第二个模块来产生答案候选, 最后综合两个模块来得到最后的答案.

Yang 等[15] 提出的 HotpotQA 数据集的挑战在于, 该数据集是基于多文档以及推理的数据集. Ding 等[16] 提出了 CogQA 模型, 使用 BERT 模型输出答案信息以及多跳信息在 GNN 上生成新的节点和下一跳的关系, 以此方法进行推理计算. Qiu 等[17] 提出了 DFGN 模型, 使用动态融合图网络来解决多跳推理问题, 设计了融合模块来提高实体图和文章之间的交互性. Tu 等[18] 提出了 HDE 模型, 通过互注意力学习候选答案、问题、文档以及实体之间的关系, 同时利用这些关系构建了一个异构图, 并通过图卷积神经网络进行推理寻找答案的支撑句. Nishida 等[19] 提出了 QFE 模型, 将片段抽取任务与多跳推理任务进行联合学习, 使用 RNN 来提取答案的支撑句. Tu 等[20] 提出了 SAENet 模型, 提出了 learning-to-rank 算法过滤冗余文档信息, 结合多任务学习以及图神经网络对答案以及证据共同预测, 增强了模型的可解释性. Shao 等[21] 提出了 C2F Reader 模型, 通过实验证明了 Transformers 有能力学习从一个实体到另一个实体的注意力从而替代图结构, 认为多跳推理并不一定需要图结构.

法律的智能化在近些年来成为一个热点研究方向, 其中一项重要的任务就是将机器阅读理解技术应用在司法领域, 让人工智能自动地阅读和分析海量的法律文书, 以提高司法人员在案件处理环节的效率. 但目前关于中文司法领域的机器阅读理解的数据集相对匮乏. 因此, Duan 等[22] 提出了一个中文司法阅读理解 (CJRC) 数据集, 它包含了大约一万个文档和近五万个问题和答案. 文件来源于判决书, 问题由法律专家注释. CJRC 数据集可以帮助研究人员通过阅读理解技术提取元素. 在该数据集上, 谭红叶等[23] 对阅读理解中问题类型多样性的解答展开研究, 提出一种基于 BERT 的多任务阅读理解模型, 利用注意力机制获得丰富的问题与篇章的表示, 并对问题进行分类, 然后将分类结果用于任务解答, 实现问题的多样性解答.

3 模型与方法

3.1 任务定义

机器阅读理解任务主要分为 3 大类型: 填空式阅读理解, 选择式阅读理解, 抽取式阅读理解. 本文的主要研究就是基于抽取式阅读理解的数据集 CAIL2020, 与传统的抽取式阅读理解任务不同, 传统的抽取式阅读理解任务定义为:

给定一段文本 $c=\{w_1^c, w_2^c, \cdots, w_n^c\}$和问题 $q=\{w_1^q, w_2^q, \cdots, w_n^q\}$, 通过模型从文本中抽取出对应的答案 a.

而本文研究的数据集, 除了需要从文本中抽取对应的答案部分之外还需要额外抽取支撑该答案的句子, 即支撑答案的证据 S_k, k 为证据的句子数目. 对于上述任务, 其定义如式 (1) 所示:

$$f(c,q)=\arg\max P(a,S_k|c,q) \tag{1}$$

3.2 模型构建

层次微调模型分为两层, 第一层我们先使用 RoBERTa 模型进行编码分别用线性层进行输出计算损失值, 并保存最优权重, 进行第一次微调, 如图 2 中 (a) 所示. 第二层再结合 Retro-Reader 方法, 将模型分为精读模块和泛读模块分别读取模型 (a) 微调后的权重进行二次微调, 最后进行预测, 如图 2 中 (b) 所示.

3.3 编码层

编码层将问题和文章的离散符号转换为连续向量表示序列. 本文使用 RoBERTa-large-wwm 来实现编码层. 该方法主要更改了原预训练阶段的训练样本生成策略, 将全词掩码的方法应用在了中文中, 即对组成的同一个词的汉字全部进行 [MASK].

首先本文使用预训练模型 RoBERTa 的词表先将输入的文本转化 Input=[CLS]Q[SEP]T[SEP] 的方式. 其中 [CLS] 用于分类任务, Q 为问题, T 为文章, [SEP] 作为分隔符来对问题 Q 和文章 T 进行分割. Input 长度为 512, 如果总长度未达到 512, 空余部分以 0 补全. 之

后将离散的符号映射至高维的向量, 通过 24 层 Transformers, 隐藏层维度为 1 024, 激活函数使用高斯误差线性单 (GELU), 并采用 16 头注意力机制进行编码. 经过编码层得到的向量将用于后续的交互层进行处理.

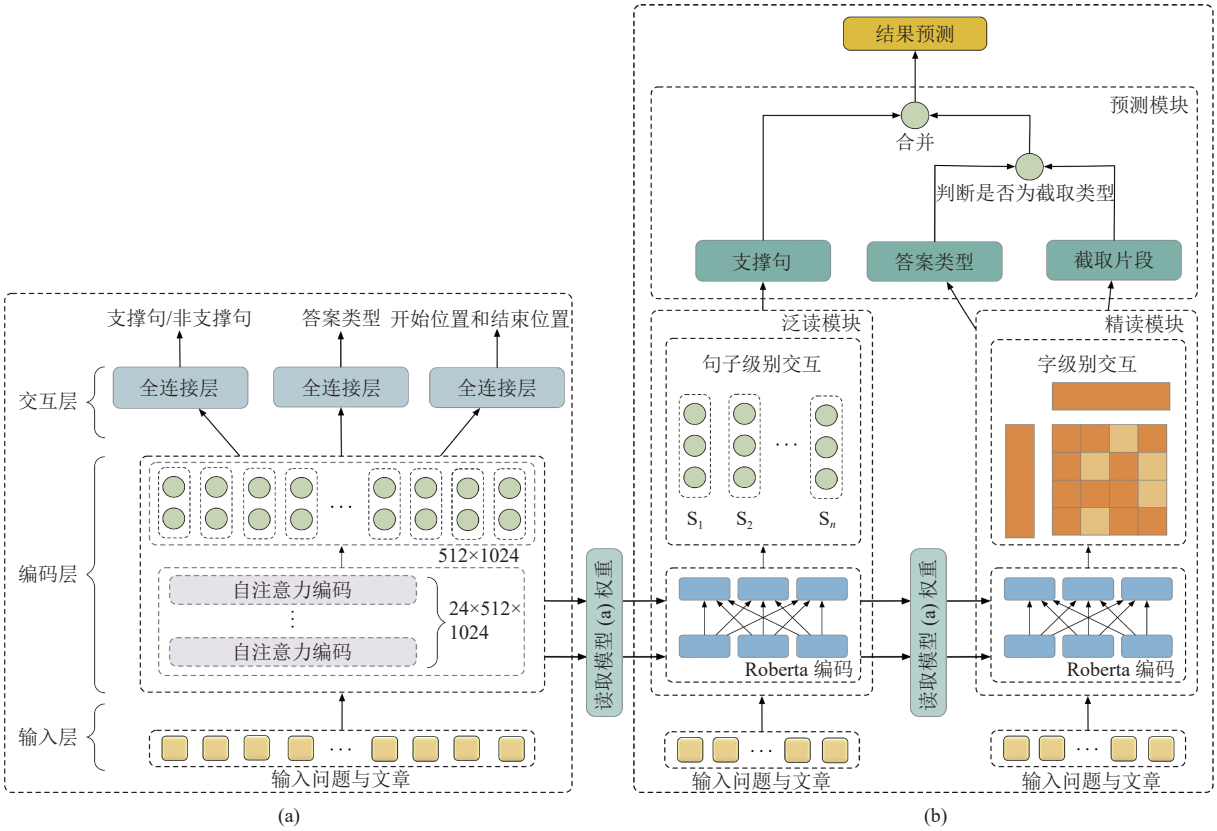

图 2 HF-Reader 模型架构图

3.4 层次微调方法

层次微调方法的主要作用是将图 2 中 (a) 的模型结合更多的任务获取更广泛的相关领域知识.

该方法是通过 RoBERTa 预训练模型与多任务学习方法进行训练, 并保存最优的 RoBERTa 模型权重. 然后再将其权重值传递给图 2 中 (b) 部分的精读模块和泛读模块分别进行读取, 继续在对应的任务上进行训练, 从而达到一个较好的效果.

3.5 精读模块

精读模块主要包含编码层, 交互层以及输出层. 而编码层均采用 3.3 节方法进行编码故本节主要描述交互层以及输出层.

采用多任务学习方法, 任务 1 为序列标注问题从文章中标注答案开始位置以及结束位置, 任务 2 为分类问题, 结合问题以及文章进行判断得到答案的类型. 由编码层可以得到输入层的 1 024 维向量表示, n 为 512.

任务 1 中定义 BERT 的输入为 [CLS]Q[SEP]P. 定义 P 中 n 个单词的 BERT 编码为 $[h_1,h_2,\ldots,h_n]$, $h_i \in R^d$. 在 BERT 的模型上加入一个前向网络 $W^s \in R^{d*l}$ 来获取分数 $s_i \in h_i W^s$. 经过 *Softmax* 计算得到模型预测的答案在文本中每个位置开始的概率 $P_i{}^S$, 如式 (2) 所示:

$$P_1{}^S, P_2{}^S, \cdots, P_n{}^S = Softmax(s_1, s_2, \cdots, s_n) \tag{2}$$

同理, 加入另一个的前向网络 $W^E \in R^{d*l}$ 来获取分数 $e_i \in h_i W^E$. 经过 *Softmax* 计算得到模型预测的答案在文本中每个位置结束的概率 $P_i{}^E$, 如式 (3) 所示:

$$P_1{}^E, P_2{}^E, \cdots, P_n{}^E = Softmax(e_1, e_2, \cdots, e_n) \tag{3}$$

使用 n 维度的矩阵 W, 将每个开始位置与每个结束位置的概率相加, 若开始位置大于结束位置, 则置为 0. 从中 W 寻找概率最大的位置. 矩阵 W 的定义如式 (4) 所示.

$$\boldsymbol{W} = \begin{bmatrix} P_1{}^S + P_1{}^E & P_1{}^S + P_2{}^E & \cdots & P_1{}^S + P_n{}^E \\ 0 & P_2{}^S + P_2{}^E & \cdots & P_2{}^S + P_n{}^E \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_n{}^S + P_n{}^E \end{bmatrix} \tag{4}$$

任务 2 中 BERT 的输入同样为 [CLS]Q[SEP]P. 定义 [CLS] 的 BERT 编码为$h_0 \in R^d$. 在 BERT 模型上加入一个前向网络$W^s \in R^{d*4}$用来获取分数得到 $[t_1,t_2,t_3,t_4]$. 经过 Softmax 计算得到模型预测的每个答案类型的概率P_i, 并得到概率最大的答案类别为P_t. 答案类型为 4 种分别为 SPAN 类型, YES 类型, NO 类型以及 UNKNOWN 类型. 对于类型的分类定义, 如式 (5) :

$$P_1,P_2,P_3,P_4 = Softmax(t_1,t_2,t_3,t_4) \tag{5}$$

将任务 1 中, W 矩阵中概率最大的开始位置与结束位置与标准答案的位置输入交叉熵损失函数得到 $loss_1$.

同样将任务 2 中概率最大的类型与标准答案的类型进行交叉熵函数得到 $loss_2$ 并将两者的 $loss$ 相加返回总损失值L进行训练, 具体定义如式 (6)–式 (8) 所示:

$$loss_1 = -y^s \log P_s - y^E \log P_E \tag{6}$$

$$loss_2 = -y^t \log P_t \tag{7}$$

$$L = a \cdot loss_1 + b \cdot loss_2 \tag{8}$$

其中, y^S 与 y^E 分别代表开始位置的标准值与结束位置的标准值, y^t代表答案类别的标准值, P_S 与 P_E 则是分别代表模型预测的开始位置与结束位置, P_t代表模型预测的答案类别, a 与 b 为系数.

3.6 泛读模块

泛读模块也同样包含编码层, 交互层以及输出层. 编码层均采用 3.3 节方法进行编码故本节主要描述交互层以及输出层.

本模块单独进行编码不与精读模块进行共享权重参数, 该模块先将 3.3 节得到 512×1024 维度向量按照句子的位置信息, 转化为 m×1024 维的句子向量, 其中 m 为文章中句子的数目, 句子向量表示为 $S=\{s_1, s_2, \cdots, s_n\}$. 然后再经过全连接层与 *Sigmoid* 函数输出预测值 x, 判断该句子是否为答案的证据.

最后损失函数使用二元交叉熵函数, 计算过程如式 (9) 和式 (10) 所示, 其中 y 为目标值, x 为预测值.

$$x_1,x_2,\cdots,x_m = Sigmoid(s_1,s_2,\cdots,s_m) \tag{9}$$

$$BCELoss = -\frac{1}{n}\sum (y_m \log x_m + (1-y_m)\log(1-x_m)) \tag{10}$$

3.7 预测模块

精读模块的预测结果有两部分, 第 1 部分为答案类型一共有 4 种分类. 分别为截取类型 (SPAN 类型), 是否类型 (YES/NO 类型) 以及不可回答类型 (UNKNOWN 类型). 第 2 部分为根据问题在原文中所截取的片段.

我们先取第 1 部分答案类型分类结果, 如果分类为是否类型与不可回答类型, 则直接输出答案类型作为精读模块结果. 如果为截取类型, 则取第 2 部分输出内容为结果.

最后, 将精读模块预测内容与泛读模块中输出的答案支撑句进行拼接, 进行最终答案的输出, 得到如图 1 的结果方便后续评估.

4 实验

4.1 数据集

本文的实验数据集为 CAIL2020, 训练集包括重新标注的约 5 100 个问答对, 其中民事、刑事、行政各约 1 700 个问答对, 均为需要多步推理的问题类型. 验证集和测试集各分别约为 1 900 和 2 600 个问答对, 同样均为需要多步推理的问题类型. 其中训练集中按照答案类型进行划分, 如图 3 截取类型 (SPAN 类型) 数据包含 2748 个问答对, 是否类型 (YES/NO 类) 包含 1 512 个问答对, 不可回答类型 (UNKNOWN 类型) 包含 758 个问答对.

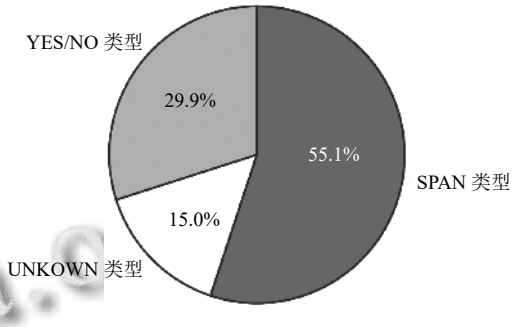

图 3 答案类型图

4.2 实验参数

实验训练过程中模型采用 Loshchilov 等[24] 提出的带有权重衰减的自适应动量估计算法 (AdamW) 作为优化算法, 学习率为 1e–5. 使用预热学习率 (warmup step), 即先用较小的学习率训练, 然后每步逐渐增大, 直到达到最初设置的学习率, 批处理大小为 1, 一共训练 10 轮.

4.3 实验结果

由于预测结果主要包含两个字段, 分别为“answer”和“sup”. “answer”对应的是模型预测的答案, “sup”则是模型预测的答案的依据. 本文使用 $F1$ 作为评分的主要评分标准, 分别对“answer”和“sup”的值进行计算精确率 (*precision*) 和召回率 (*recall*), 得到 p^{Ans}, r^{Ans}, p^{Sup} ,

r^{Sup}. 再计算综合的 Joint F1 值作为最终评价标. 其中 *precision* 用来描述所有预测的答案文本中与真正答案文本的相同字数所占比率, *recall* 用来描述所有真正答案中与预测答案文本的相同字数所占比率. *F*1 综合了 *precision* 和 *recall* 两个指标, 其定义如式 (11) 和 式 (12) .

$$precision = \frac{TP}{TP+FP} \tag{11}$$

$$recall = \frac{TP}{TP+FN} \tag{12}$$

TP 表示真实样本与预测样本中令牌 (token) 相同的部分, *TP*+*FP* 表示预测样本中所有的令牌数目, *TP*+*FN* 表示真实样本中所有的令牌数目.

*Ans_F*1 以及 *Sup_F*1 的计算方法则是将其对应的精确率与召回率分别相乘如式 (13) 和 式 (14) 所示:

$$Ans\_F1 = \frac{2\times p^{Ans}\times r^{Ans}}{p^{Ans}+r^{Ans}} \tag{13}$$

$$Sup\_F1 = \frac{2\times p^{Sup}\times r^{Sup}}{p^{Sup}+r^{Sup}} \tag{14}$$

*Joint_F*1 的计算方法则是将“answer”与“sup”的精确率与召回率分别相乘, 再计算 Joint 最终评价指标. 具体如下式所示:

$$p^{joint} = p^{Ans}\times p^{Sup} \tag{15}$$

$$r^{joint} = r^{Ans}\times r^{Sup} \tag{16}$$

$$Joint\_F1 = \frac{2\times p^{joint}\times r^{joint}}{p^{joint}+r^{joint}} \tag{17}$$

本文使用机器阅读理解中流行的主流预训练模型进行对比实验. 并且选取了 4 种不同规模的模型进行对比分析, 主要为:

(1) BERT 模型, 在预训练模型之后使用多任务学习方法进行预测.

(2) ALBERT 模型, 采用矩阵分解等方法为轻量级的 BERT 模型[25], 并在预训练模型之后采用多任务学习方法进行预测.

(3) RoBERTa 模型, 相较于 BERT, 训练时去除了下一句预测部分. 其中对比实验中用到了其 base 版本和 large-ext 版本[26]. 同样使用多任务学习方法进行预测.

(4) MJL-DPCNN 模型, 使用句法关系增强的关系要素图构建方法在 DFGN 模型上进行支撑句挖掘并使用 DPCNN 进行观点类问题分类[27,28].

(5) DFGN_CAIL 模型, 按照 CAIL2020 的数据格式, 修改了 DFGN 的数据处理部分.

(6) Cola 模型, CAIL2020 阅读理解比赛第 4 名所用模型.

由于本文的数据集为中文数据集, 因此使用中文全词覆盖的方式取代原本英文的 WordPiece 方法. 本文采用 RoBERTa 作为预训练模型, 实验的评价指标与 HotpotQA 一致. 以上模型在测试集上的实验结果如表 1 所示, 可以看出所有的基线模型中 RoBERTa-large 模型效果最好. 本文的模型方法相较于 RoBERTa-large 基线模型综合评分提升了 3.38%.

表 1 CAIL2020 实验结果 (%)

模型	*Ans_F*1	*Sup_F*1	*Joint_F*1
BERT	68.61	66.11	49.02
ALBERT	62.97	66.70	46.91
RoBERTa	71.09	73.63	57.27
RoBERTa-large	74.86	77.46	62.77
MJL-DPCNN	77.43	75.07	61.80
DFGN_CAIL	68.79	72.34	53.82
Cola	74.63	73.68	59.62
HF-Reader	78.48	80.33	66.15

4.4 消融实验

消融实验部分将训练集按照 9:1 进行分割分别作为新的训练集和验证集. 为了进一步评估模型各个模块的贡献, 本文进行了如下消融实验:

(1) -Retro Reader: 去掉精读模块与泛读模块, 并使用多任务方法取代进行预测.

(2) -HF: 去掉层次微调方法.

表 2 实验结果显示, 去掉层次微调方法后, *Ans_F*1 的值下降了 1.24, *Sup_F*1 下降了 1.99, *Joint_F*1 下降了 2.24. 去精读模块与泛读模块 , 并使用多任务方法取代后, *Ans_F*1 的值下降了 2.01, *Sup_F*1 下降了 2.48, *Joint_F*1 下降了 3.7. 通过消融实验分析, 证明了本文所提模型的有效性.

5 结束语

本文提出了一种多任务层次微调模型, 灵活使用预训练与多任务学习方法获取到更广泛的语义信息, 并将答案类型预测和答案预测两个任务放入精读模块. 使用多任务学习方法进行预测, 再将证据抽取任务单独分放进泛读模块, 单独进行训练预测, 最后综合两个

模块的预测结果进行输出, 并且选取了 BERT、ALBERT、RoBERTa、RoBERTa-large 四种不同规模的预训练模型对输出的实验结果进行了对比分析. 实验表明, 本文所提方法可以有效提高机器阅读理解的答案抽取以及证据抽取的效果. 为了进一步评估模型各个模块的贡献, 本文又通过消融实验将训练集按照 9:1 进行了分割, 用分割后新的训练集和验证集进行实验, 分别去掉 Retro-Reader 部分以及层次微调部分, 并使用多任务方法取代进行预测后, *F*1 的值均显示下降, 进一步证明了本文所提模型的有效性. 未来将把侧重点放在精读模块与泛读模块上, 以进一步对模型进行改进优化, 例如结合当前热门的图卷积网络技术以句子作为节点进训练并行预测, 进一步提升机器阅读理解的效果以及可解释性.

表 2 消融实验 (%)

模型	答案				支撑句				综合
	precision	*recall*	*F*1	EM	*precision*	*recall*	*F*1	EM	*F*1
HF-Reader	**82.26**	**81.38**	**80.81**	**71.68**	68.10	**69.15**	**66.97**	**46.73**	**55.51**
-HF	80.42	80.65	79.57	69.70	**71.21**	63.21	64.98	44.15	53.27
-Retro Reader	81.51	78.72	78.80	69.30	67.42	65.49	64.49	44.35	51.81

参考文献

1 Long SB, Tu CC, Liu ZY, *et al*. Automatic judgment prediction via legal reading comprehension. China National Conference on Chinese Computational Linguistics. Kunming: Springer, 2019. 558–572.

2 Xu YM, Cohen SB. Stock movement prediction from tweets and historical prices. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. (Volume 1: Long Papers). Melbourne: ACL, 2018. 1970–1979.

3 Šuster S, Daelemans W. CliCR: A dataset of clinical case reports for machine reading comprehension. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans: ACL, 2018. 1551–1563.

4 李舟军, 王昌宝. 基于深度学习的机器阅读理解综述. 计算机科学, 2019, 46(7): 7–12. [doi: 10.11896/j.issn.1002-137X.2019.07.002]

5 Rajpurkar P, Zhang J, Lopyrev K, *et al*. SQuAD: 100, 000+ questions for machine comprehension of text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: ACL, 2016. 2383–2392.

6 Seo M, Kembhavi A, Farhadi A, Hajishirzi H. Bidirectional attention flow for machine comprehension. 5th International Conference on Learning Representations. Toulon: ICLR, 2017.

7 Wang WH, Yang N, Wei FR, *et al*. Gated self-matching networks for reading comprehension and question answering. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. (Volume 1: Long Papers). Vancouver: ACL, 2017. 189–198.

8 Huang HY, Zhu C, Shen Y, Chen W. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. 6th International Conference on Learning Representations. Vancouver: ICLR, 2018.

9 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, 2019. 4171–4186.

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

11 Reddy S, Chen DQ, Manning CD. Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 2019, 7: 249–266. [doi: 10.1162/tacl_a_00266]

12 Zhu CG, Zeng M, Huang XD. Sdnet: Contextualized attention-based deep network for conversational question answering. arXiv: 1812.03593, 2018.

13 Rajpurkar P, Jia RB, Liang P. Know what you don't know: Unanswerable questions for SQuAD. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne: ACL, 2018. 784–789.

14 Zhang ZS, Yang JJ, Zhao H. Retrospective reader for machine reading comprehension. Proceedings of the 35th AAAI Conference on Artificial Intelligence, Virtual Event: AAAI Press, 2021. 14506–14514.

15 Yang ZL, Qi P, Zhang SS, *et al*. HotpotQA: A dataset for

diverse, explainable multi-hop question answering. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 2369–2380.

16 Ding M, Zhou C, Chen Q, *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.

17 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.

18 Tu M, Wang GT, Huang J, *et al*. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 2704–2713.

19 Nishida K, Nishida K, Nagata M, *et al*. Answering while summarizing: Multi-task learning for multi-hop QA with evidence extraction. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 2335–2345.

20 Tu M, Huang K, Wang G T, *et al*. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020. 9073–9080.

21 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. ACL, 2020. 7187–7192.

22 Duan XY, Wang BX, Wang ZY, *et al*. Cjrc: A reliable human-annotated benchmark dataset for Chinese judicial reading comprehension. China National Conference on Chinese Computational Linguistics. Kunming: Springer, 2019. 439–451.

23 谭红叶, 屈保兴. 面向多类型问题的阅读理解方法研究. 中文信息学报, 2020, 34(6): 81–88. [doi: 10.3969/j.issn.1003-0077.2020.06.011]

24 Loshchilov I, Hutter F. Fixing weight decay regularization in Adam. arXiv: 1711.05101v1, 2018.

25 Lan Z, Chen M, Goodman S, *et al*. Albert: A lite bert for self-supervised learning of language representations. 8th International Conference on Learning Representations. Addis Ababa: ICLR, 2020. 344–350.

26 Liu YH, Ott M, Goyal N, *et al*. RoBERTa: A robustly optimized BERT pretraining approach. arXiv: 1907.11692v1, 2019.

27 Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. (Volume 1: Long Papers). Vancouver: ACL, 2017. 562–570.

28 张虎, 王宇杰, 谭红叶, 等. 基于 MHSA 和句法关系增强的机器阅读理解方法研究. 自动化学报, 2021: 1–11. [doi: 10.16383/j.aas.c200951]