

基于混合注意力机制的中文机器阅读理解

刘高军^{1,2}, 李亚欣^{1,2}, 段建勇^{1,2}

(1.北方工业大学 信息学院,北京 100144; 2.北方工业大学CNONIX 国家标准应用与推广实验室,北京 100144)

摘要: 预训练语言模型在机器阅读理解领域具有较好表现,但相比于英文机器阅读理解,基于预训练语言模型的阅读理解模型在处理中文文本时表现较差,只能学习文本的浅层语义匹配信息。为了提高模型对中文文本的理解能力,提出一种基于混合注意力机制的阅读理解模型。该模型在编码层使用预训练模型得到序列表示,并经过BiLSTM处理进一步加深上下文交互,再通过由两种变体自注意力组成的混合注意力层处理,旨在学习深层语义表示,以加深对文本语义信息的理解,而融合层结合多重融合机制获取多层次的表示,使得输出的序列携带更加丰富的信息,最终使用双层BiLSTM处理输入输出层得到答案位置。在CMRC2018数据集上的实验结果表明,与复现的基线模型相比,该模型的EM值和F1值分别提升了2.05和0.465个百分点,能够学习到文本的深层语义信息,有效改进预训练语言模型。

关键词: 中文机器阅读理解;注意力机制;融合机制;预训练模型;RoBERTa模型

开放科学(资源服务)标志码(OSID):



中文引用格式: 刘高军,李亚欣,段建勇.基于混合注意力机制的中文机器阅读理解[J].计算机工程,2022,48(10):67-72,80.
英文引用格式: LIU G J, LI Y X, DUAN J Y. Chinese machine reading comprehension based on hybrid attention mechanism[J]. Computer Engineering, 2022, 48(10): 67-72, 80.

Chinese Machine Reading Comprehension Based on Hybrid Attention Mechanism

LIU Gaojun^{1,2}, LI Yaxin^{1,2}, DUAN Jianyong^{1,2}

(1.School of Information, North China University of Technology, Beijing 100144, China; 2.CNONIX National Standard Application and Promotion Laboratory, North China University of Technology, Beijing 100144, China)

[Abstract] The pre-training language model performs well in the field of machine reading comprehension. Compared with English machine reading comprehension, the reading comprehension model based on the pre-training language model performs slightly worse in processing Chinese text and can only learn the shallow semantic matching of the text. To improve the ability of the model to understand Chinese text, this paper proposes a Chinese machine reading comprehension model based on hybrid attention mechanism. The model uses the pre-training model to obtain the sequence representation in the coding layer and further deepens the context interaction through BiLSTM processing. Then, this is processed by a hybrid attention layer comprising two variants of self-attention mechanism, which aims to learn the deep semantic representation, to deepen the understanding of the text semantic information. Further, the fusion layer combines multiple fusion mechanisms to obtain the multi-level representation, making the output sequence carry more rich information. Finally, after double BiLSTM processing, input output layer to get the answer position. The experimental results on CMRC2018 dataset show that the EM and F1 values of this model are increased by 2.05 and 0.465 percentage points, respectively, compared with those of the baseline model. This enables to learn the deep semantic information of the text and effectively improve the pre-trained language model.

[Key words] Chinese machine reading comprehension; attention mechanism; fusion mechanism; pre-training model; RoBERTa model

DOI:10.19678/j.issn.1000-3428.0062206

0 概述

机器阅读理解是自然语言处理领域的一个极具挑战性的任务,一直受到研究人员的关注。深度学

习技术的成熟以及数据的多样化推动了机器阅读理解技术的快速发展,基于深度学习建立阅读理解模型已成为目前普遍采用的方法。

基金项目: 国家自然科学基金(61972003,61672040)。

作者简介: 刘高军(1962—),男,教授,主研方向为数据处理、软件服务;李亚欣,硕士研究生;段建勇,教授。

收稿日期: 2021-07-29 **修回日期:** 2021-11-03 **E-mail:** duanjy@ncut.edu.cn

机器阅读理解是指让机器通过阅读文本回答相应的问题。机器阅读理解技术通过训练模型帮助用户从大量的文本中快速、准确地找到答案。根据答案类型的不同,机器阅读理解任务可分为4类^[1]:完形填空式任务要求模型从候选答案集合中选择一个正确的词填至问题句,使文章变得完整;抽取式任务要求模型能根据提出的问题从文章中抽取一个连续片段作为答案,输出答案在上下文中的起始位置和结束位置;多项选择式任务需要从候选答案集合中挑选正确答案;在自由作答式任务中,答案的类型不受限制。其中,抽取式阅读理解任务的形式相对灵活,能够适用于现实中大部分场景,如搜索引擎、智能问答等。

预训练语言模型BERT^[2]的出现使得一些模型在阅读理解任务上的表现接近甚至超过了人类,推动了机器阅读理解的研究进入到新的阶段。BERT模型优秀的表现受到了众多专家、学者的高度关注,近年涌现出了很多基于BERT改进的模型,如ALBERT^[3]、RoBERTa^[4]等,使用预训练模型已成为机器阅读理解的发展趋势。由于预训练模型只能学习到文本的浅层语义匹配信息,目前大多数模型都采取了预训练语言模型与注意力机制相结合的方式,即通过预训练模型获取相应表示,再使用注意力机制进行推理,从而捕捉文本的深层语义信息,预测出更加准确的答案。但原始的预训练模型是针对英文语言设计的,无法有效处理中文文本。

本文提出一种基于混合注意力机制的中文机器阅读理解模型。该模型使用混合注意力机制进行推理,并结合多重融合机制丰富序列信息,最终在CMRC2018中文阅读理解数据集上进行了实验。

1 相关工作

1.1 结合注意力机制的机器阅读理解

BAHDAU等^[5]将注意力机制用于机器翻译任务,这是注意力机制第一次应用于自然语言处理领域。引入注意力机制后,不同形式的注意力机制成为基于神经网络模型在阅读理解任务上取得好成绩的一个关键因素。

2015年,HERMANN等^[6]提出The Attentive Reader和The Impatient Reader两个基于神经网络的模型,将注意力机制应用于机器阅读理解的任务中,通过注意力机制得到问题和文章之间的交互信息。随后提出的Attention Sum Reader模型^[7]以及The Stanford Attentive Reader模型^[8]均着重于提升注意力模型中问题和文章的相似度计算能力。

在前期模型中使用的注意力机制大多较为简单,对文本理解能力不足,无法对文章和问题进行有效交互。针对这一问题,研究人员在深层注意力机制方面做了大量的研究。BiDAF模型^[9]同时计算文章到问题和问题到文章两个方向的注意力信息,捕获问题和文章更深层的交互信息。Document

Reader模型^[10]将词性等语法特征融入词嵌入层,经过模型处理得到答案。R-Net模型^[11]在计算问题和文章的注意力之后加入自匹配注意力层,对文章进行自匹配,从而实现文章的有效编码。FusionNet模型^[12]融合多个层次的特征向量作为输入。

2017年,谷歌的研究人员提出了Transformer模型^[13],该模型仅依靠自注意力机制在多个任务上取得了较好结果,证明注意力机制拥有较强的提取文本信息的能力。2018年,谷歌团队提出了基于双向Transformer的预训练语言模型BERT。这种双向的结构能够结合上下文语境进行文本表征,增强了模型的学习能力。BERT的出现刷新了11个自然语言处理任务的最好结果,使得预训练语言模型成为近年来的研究热点。

1.2 中文机器阅读理解

中文机器阅读理解由于起步较晚,缺少优质中文数据集,发展相对缓慢。在近年来发布的各种中文机器阅读理解数据集的影响下,越来越多的研究人员致力于中文领域的探索。

2016年,CUI等^[14]发布了大规模填空型中文机器阅读理解数据集People Daily and Children's Fairy Tale,填补了大规模中文阅读理解数据集的空白。2017年,CUI等^[15]在此数据集的基础上提出了CMRC2017数据集,作为第一届“讯飞杯”中文机器阅读理解评测比赛的数据集。

2018年,CUI等^[16]发布了抽取型中文机器阅读理解数据集CMRC2018,该数据集作为第二届“讯飞杯”中文机器阅读理解评测比赛使用的数据集,也是本文实验使用的数据集。该数据集由近两万个人工标注的问题构成,同时发布了一个需要多句推理答案的挑战集。

HE等^[17]于2018年提出DuReader数据集,该数据集共包含20万个问题、100万篇文章和超过42万个人工总结的答案,数据来源更贴近实际,问题类型丰富,是目前最大的中文机器阅读理解数据集。

徐丽丽等^[18]搜集全国各省近10年高考题及高考模拟题中的981篇科技文章语料,构建了4905个问题,同时搜集5万篇新闻语料,构造10万个补写句子类选择题语料。SHAO等^[19]提出了繁体中文机器阅读理解数据集DRCD,该数据集包含从2108篇维基百科文章中摘取的10014篇段落以及超过3万个问题。中文机器阅读理解领域受到研究人员越来越多的关注,不断有优秀的方法与模型被提出,呈现较好的发展趋势。

2 本文模型结构

为了提高模型对中文文本的理解能力,本文提出一种基于混合注意力机制的中文机器阅读理解模型。首先经过编码层得到序列表示,使用混合注意力机制提取文本中可能与答案有关的关键信息,然后结合多

重融合机制融合多层次的序列信息,经过双层 BiLSTM 建模后传入输出层,最终输出正确答案所在位置。

本文模型包含编码层、混合注意力层、融合层、建模层以及输出层,其结构如图1所示。

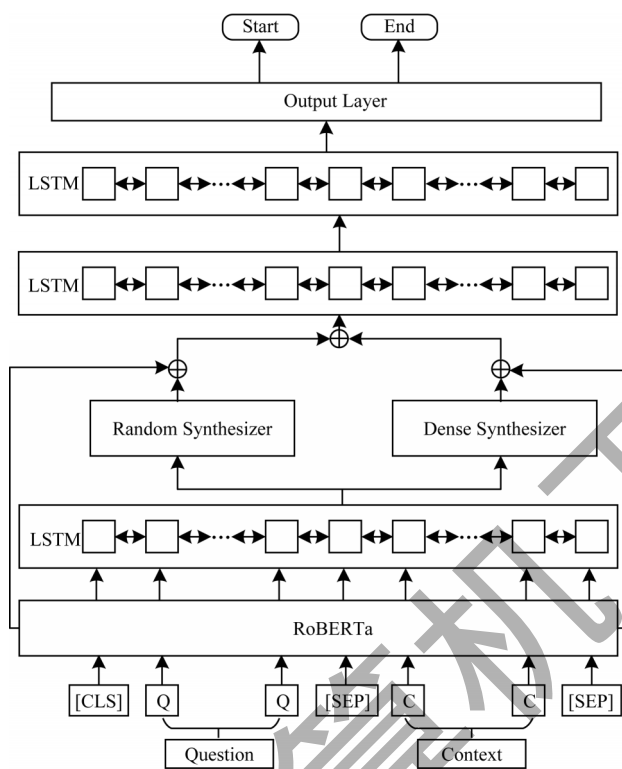


图1 本文模型结构

Fig.1 The structure of the proposed model

2.1 编码层

编码层通过中文预训练语言模型 RoBERTa^[18]对问题和文章进行编码。RoBERTa模型仍使用BERT的基本结构,在模型训练时有以下4个方面的差异:

- 1)使用动态掩码机制。
- 2)移除BERT中采用的下一句预测训练任务。
- 3)使用更大Byte级别的文本编码方式。
- 4)使用更大批次以及更大规模的数据进行训练。

可以看出,RoBERTa模型在多个任务上的表现优于BERT。

编码层将问题和文章拼接后的文本输入到RoBERTa模型中,经过分词器处理后的每一个词称为 token,最终 RoBERTa 模型输入的编码向量为 token 嵌入、位置特征嵌入以及用以区分问题和文章的分割特征嵌入之和。本文使用的 RoBERTa 模型由12层Transformer编码器组成,该模型取最后一层编码输出作为文本嵌入表示,得到的向量表示 H 如式(1)所示:

$$H=[h_1, h_2, \dots, h_N] \quad (1)$$

其中: h_i 为序列中第 i 个 token 经过 RoBERTa 编码后的向量表示; N 为序列长度。

利用 BiLSTM 进一步加深文本的上下文交互,

捕捉文本序列的局部关系,如式(2)所示:

$$H^1 = \text{BiLSTM}(H) \quad (2)$$

2.2 混合注意力层

混合注意力层基于混合注意力机制处理编码层得到的上下文向量 H^1 ,进而学习文本中更深层次的语义信息,该层是模型的核心部分。该层的混合注意力机制由文献[10]中提出的两种自注意力机制的变体注意力 Random Synthesizer 和 Dense Synthesizer 组成。传统的自注意力机制通过计算序列中每一个 token 与序列中其他 token 的相关度得到权重矩阵 R ,再将归一化后的权重和相应的键值进行加权求和,得到最终的注意力表示。这里的相关度一般通过点积得分矩阵体现,点积自注意力的主要作用是学习自对齐信息,即 token 对的交互信息。自注意力机制通过比较序列本身捕捉序列和全局的联系,获取文本特征的内部相关性,其简化结构如图2所示。

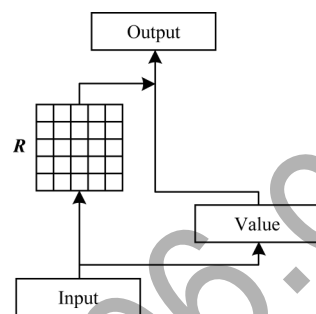


图2 自注意力机制结构

Fig.2 Structure of self-attention mechanism

这种从 token-token 交互中学习到的注意力权重有一定的作用,但也存在缺点。传统自注意力机制中的权重包含实例中 token 对的交互信息,通过计算点积的方式得到每个 token 与序列其他 token 的相对重要度。这种方式过度依赖特定实例,仅通过 token 对之间的相关度决定答案的概率是不稳定的,缺乏一致的上下文联系,很大程度上会受不同实例影响,不能学习到更多的泛化特征。文献[19]的实验结果表明,与传统自注意力机制相比, Synthesizer 注意力机制得到的权重曲线更加平滑。受其启发,本文认为这种合成权重矩阵的自注意力机制不会从特定的 token 中获益,可以在提取序列关键信息的同时减小因不同实例产生的影响,因此该层使用这种合成注意力来提取文本深层信息。这种合成矩阵的注意力与点积注意力或考虑上下文的注意力不同,它不依赖于 token-token 交互的方式生成权重矩阵,受特定样本的影响较小,能够学习到较为稳定的权重值。

1) Random Synthesizer 使用随机值初始化权重矩阵 R ,并随着模型一起训练这些值。这种方式下所有实例均使用相同的对齐模式,不依赖输入的 token,不会因特定实例而影响权重矩阵,因此 Random 关注的是全局的注意力权重。

2) Dense Synthesizer 通过对输入序列进行线性变换得到权重矩阵 R , 序列中的每个 token 为自己相应位置的 token 独立预测权重, 即按序列顺序处理每个向量。Dense 学习的是局部的注意力权重, 权重矩阵的生成需要依赖样本的每一个 token, 因此它能关注到序列中每一个 token 携带的信息。线性变换方式如式(3)所示:

$$R_N = \text{softmax}(F_N(H_{i,N})) \quad (3)$$

其中: 参数化函数 $F_N(\cdot)$ 由两层前馈层和 ReLU 激活函数组成; i 为 $H_{i,N}$ 的第 i 个 token; N 为序列长度。

这两种自注意力使用不同方法合成权重矩阵, 分别从不同角度提升获得信息的质量。因此, 该层采取两种注意力混合使用的策略, 能够结合全局与局部注意力的优势, 不会过度依赖输入样本, 既能从原始序列中获取特征信息, 又能减弱不同实例对模型的影响, 可以更加有效地处理相关任务。

将上一层得到的向量 H^r 分别输入到 Random Synthesizer 和 Dense Synthesizer 中, 与权重矩阵 R 加权求和, 得到两组具有深层语义的向量表示 H_N^r 和 H_N^d , 如式(4)、式(5)所示:

$$H_N^r = R_N^r L_N(H_N^r) \quad (4)$$

$$H_N^d = R_N^d L_N(H_N^d) \quad (5)$$

其中: H_N^r 和 H_N^d 分别为 Random Synthesizer 和 Dense Synthesizer 输出的表示; R_N^r 和 R_N^d 分别表示 Random Synthesizer 和 Dense Synthesizer 的权重矩阵; $L_N(\cdot)$ 为一层线性层; $L_N(H_N)$ 等同于注意力机制中的 V 矩阵。

2.3 融合层

为防止模型过于关注某一部分而过滤掉文本其他特征信息, 融合层结合多重融合机制丰富序列表示。

首先, 将上层得到的两组注意力 H^r 和 H^d 分别与 RoBERTa 模型得到的序列 H 进行融合, 如式(6)、式(7)所示, 实现在不丢失原始信息的基础上更加关注关键信息。

$$\overline{H}^r = \alpha_1 H + (1 - \alpha_1) H^r \quad (6)$$

$$\overline{H}^d = \alpha_2 H + (1 - \alpha_2) H^d \quad (7)$$

其次, 对处理后的两组序列进行融合, 得到混合语义表示, 如式(8)所示:

$$\overline{H} = \alpha_3 \overline{H}^r + (1 - \alpha_3) \overline{H}^d \quad (8)$$

在式(6)~式(8)中: $\alpha_1, \alpha_2, \alpha_3$ 均为模型训练参数; \overline{H}^r 和 \overline{H}^d 分别为两组注意力与序列 H 融合后的表示; \overline{H} 为最终融合后的输出表示。

最后, 输出结合全局和局部的注意力信息, 融入一定比例的全局上下文信息, 能够有效降低实例不同对信息造成的影响。以上3次均融合采用同一种策略。

2.4 建模层

建模层使用双层 BiLSTM 对融合多重信息的序列 \overline{H} 进行整体上的建模, 整合每个向量关于整个序列的上下文信息, 得到具有前后位置信息的新序列

H^f , 如式(9)和式(10)所示:

$$H^h = \text{BiLSTM}(\overline{H}) \quad (9)$$

$$H^f = \text{BiLSTM}(H^h) \quad (10)$$

2.5 输出层

输出层将建模后的序列 H^f 输入到线性层, 得到针对答案开始位置和结束位置预测的两个输出, 由 softmax 函数计算概率得到最终预测答案在文章中的起止位置 s 和 e , 如式(11)所示:

$$s, e = \text{softmax}(\text{Linear}(H^f)) \quad (11)$$

3 实验结果与分析

3.1 数据集

本文使用 CMRC2018 评测任务数据集以及 DRCD 数据集进行实验。两个数据集的格式相同, 均用于抽取式阅读理解任务。其中, CMRC2018 数据集为简体中文数据集, DRCD 数据集为繁体中文数据集。除对比实验外, 其余几组实验均使用 CMRC2018 数据集。以 CMRC2018 数据集为例, 数据集实例如下:

[Document] 白荡湖位于中国安徽枞阳县境内, 紧邻长江北岸, 系由长江古河床摆动废弃的洼地积水而成。湖盆位置介于北纬 $30^\circ 47' \sim 30^\circ 51'$ 、东经 $117^\circ 19' \sim 117^\circ 27'$ 。白荡湖原有面积近 100 km^2 , 经过近五十年的围垦, 目前面积缩小为 39.67 km^2 , 平均水深 3.06 m , 蓄水量 $1.21 \times 10^9 \text{ m}^3$ 。通过白荡闸与长江连通, 是长江重要的蓄洪湖之一。湖水补给主要依赖降水与长江倒灌, 入流的罗昌河、钱桥河等均为季节性溪流, 入水量较小。白荡湖是重要的水产养殖基地, 盛产各种淡水鱼类与水禽, 其中以大闸蟹产量最大。每年冬季开启白荡闸排干湖水捕鱼, 次年5月左右再引长江水倒灌, 水位至7月、8月份达到最高。

[Question] 白荡湖是怎样形成的?

[Answer] 系由长江古河床摆动废弃的洼地积水而成。

CMRC2018 数据集和 DRCD 数据集由几万个真实问题组成, 篇章均来自中文维基百科, 问题由人工编写。两个数据集规模分别如表1、表2所示。

表1 CMRC2018数据集规模

Table 1 CMRC2018 dataset size

集合	短文数	问题数	答案数
训练集	2 659	11 144	1
开发集	848	3 219	3
测试集	1 718	6 895	3

表2 DRCD数据集规模

Table 2 DRCD dataset size

集合	短文数	问题数	答案数
训练集	8 016	26 936	1
开发集	1 000	3 524	2
测试集	1 000	3 493	2

3.2 实验配置

本文实验采用 GPU 进行训练,开发语言为 Python,深度学习框架为 Pytorch。由于本文模型加入注意力层以及 BiLSTM,增加了序列之间的交互过程,因此相比基线模型,本文模型的训练速度更加缓慢。实验参数如表 3 所示。

表 3 实验参数

Table 3 Experimental parameters

参数	参数值
Epoch	3
Batch size	12
学习率	4e-5
Dropout	0.1
最大答案长度	50
最长输入序列	512

3.3 评价指标

本文采用 EM 值和 F1 值作为评价指标。EM 值为精确匹配度,计算预测答案与真实答案是否完全匹配。F1 值为模糊匹配度,计算预测答案与标准答案之间的匹配程度。这两个指标通常作为抽取式机器阅读理解的评价指标。

3.4 结果分析

3.4.1 对比实验

为验证本文提出的模型在中文机器阅读理解任务的有效性,将本文模型与以下模型进行实验对比:

1) BERT-base (Chinese) 和 BERT-base (Multi-lingual)为 CMRC2018 评测任务选用的基线模型。

2) RoBERTa-wwm-ext^[21]为本文选取的基线模型,该模型针对中文改进预训练模型中的全词掩码训练方法。

3) MacBERT-base 为文献[22]提出的预训练模型,该模型主要针对 mask 策略对 RoBERTa 进行改进。

表4,表5所示为本文模型与其他模型在CMRC2018数据集与DRCD数据集上的EM值和F1值。其中RoBERTa-wwm-ext(*)为本文复现的结果。

表 4 不同模型在 CMRC2018数据集上的实验结果

Table 4 Experimental results of different models on the CMRC2018 dataset

%

模型	EM	F1
BERT-base(Chinese)	63.600	83.900
BERT-base(Multi-lingual)	64.100	84.400
RoBERTa-wwm-ext	67.400	87.200
MacBERT-base	68.500	87.900
RoBERTa-wwm-ext(*)	67.785	87.572
本文模型	69.835	88.037

表 5 不同模型在 DRCD数据集上的实验结果

Table 5 Experimental results of different models on the DRCD dataset

%

模型	EM	F1
BERT	83.100	89.900
RoBERTa-wwm-ext	86.600	92.500
MacBERT-base	88.300	93.500
RoBERTa-wwm-ext(*)	88.791	94.025
本文模型	89.047	94.138

本文模型在CMRC2018数据集的EM值和F1值分别达到69.835%和88.037%,相比复现的基线模型分别提高了2.05和0.465个百分点,在DRCD数据集上的EM值和F1值分别达到89.049%和94.138%,相比基线模型分别提高了0.256和0.113个百分点,在两个数据集上的表现均优于其他对比模型。实验结果表明,本文模型在性能上有显著提升,能够学习到文本的深层语义信息,有效改进了预训练语言模型。

3.4.2 消融实验

为研究混合注意力以及多重融合机制对模型的贡献,设计消融实验进一步分析本文模型。由于多重融合机制需要混合注意力的输出信息,因此本节实验考虑两部分共同作用的影响,实验结果如表6所示。

表 6 消融实验结果

Table 6 Ablation experiment results

%

模型	EM	F1	AVG
本文模型	69.835	88.037	78.936
未使用混合注意力和多重融合机制的模型	67.847	87.973	77.910

从表6可以看出,当模型未使用混合注意力和多重融合机制时,EM值和F1值分别下降了1.988和0.064个百分点。结果表明,使用混合注意力机制以及多重融合机制能够加深对文本的理解,防止模型随着训练遗失原有信息,使模型更好地预测答案。

3.4.3 不同注意力策略实验分析

为了验证变体注意力以及混合策略对模型的影响,本文针对传统自注意力机制以及单一注意力机制两个方面设计对比实验,结果如表7所示。

表 7 不同注意力策略的实验结果

Table 7 Experiment results of different attention strategies

%

模型	EM	F1	AVG
Random+Dense	69.835	88.037	78.936
Random+Self-Attention	67.505	87.218	77.362
Dense+Self-Attention	67.630	87.133	77.382
Random	68.500	87.471	77.986
Dense	68.593	88.505	78.549
Self-Attention	68.437	87.639	78.038

表7所示为使用不同注意力方法对模型的影响,其中,Random和Dense分别表示Random Synthesizer注意力和Dense Synthesizer注意力,“+”表示混合使用两种注意力,Self-Attention表示使用传统自注意力机制。实验结果分析如下:

1)传统自注意力的表现略低于Dense Synthesizer,证明以往利用token对生成权重矩阵的方式并没有合成矩阵有竞争力,使用合成注意力能够降低过多关注局部注意力的影响,提升模型性能。

2)综合比较EM值和F1值,混合使用Random Synthesizer注意力和Dense Synthesizer注意力的方法效果最好。Random Synthesizer与Dense Synthesizer两种注意力在合成权重矩阵时输入的信息不同,因此联合使用这两种方法可以学习到综合注意力权重,能够进一步提升模型性能。对比结果发现,使用单一Dense Synthesizer注意力的F1值最高,混合注意力加入一定比例的全局注意力,减少样本不同导致的权重波动,因此会在一定程度上影响个别样本的准确度。

3.4.4 注意力层和融合层位置实验分析

为了研究混合注意力层和融合层加入位置的不同对模型的影响,本文设置了注意力层和融合层位置对比实验。RoBERTa+Att+BiLSTM对应于将注意力层和融合层加在RoBERTa模型之后,实验结果如表8所示。

表8 注意力层和融合层不同位置的实验结果

Table 8 Experiment results of different positions of attention layer and fusion layer

模型	EM	F1	AVG
本文模型	69.835	88.037	78.936
RoBERTa+Att+BiLSTM	68.872	87.596	78.234

通过实验发现,混合注意力层和融合层的位置在第1个BiLSTM和序列建模层之间表现更好,表明对使用BiLSTM建模后的序列进行自注意力处理,能较好地理解文章,更有效地预测答案。

4 结束语

本文对抽取式中文机器阅读理解任务进行研究,提出一种基于混合注意力机制的阅读理解模型。该模型使用两种自注意力机制的变体模型对序列进行处理,加深对文本语义信息的理解,并对输出的注意力进行多层次的融合,使得输出的序列携带更加丰富的信息。实验结果表明,本文方法提升了模型的理解能力,改进了模型对语义的获取方法,同时保留了原序列的信息特征,提高了预测答案的准确率。目前的中文机器阅读理解模型多数存在答案边界不准确的问题,下一步通过使用分词器优化模型输入,将分词结果作为输入特征加入到序列中,从而优化答案边界。此外,结合双向注意力机制,融合文章到问题以及问题到文章双向的注意力优化模型结构,加深对文本的理解。

参考文献

- [1] CHEN D Q. Neural reading comprehension and beyond[D]. Palo Alto, USA: Stanford University, 2018.
- [2] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1810.04805>.
- [3] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1909.11942>.
- [4] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1907.11692>.
- [5] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1409.0473>.
- [6] HERMANN K M, KOČISKÝ T, GREFFENSTETTE E, et al. Teaching machines to read and comprehend[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1506.03340>.
- [7] KADLEC R, SCHMID M, BAJGAR O, et al. Text understanding with the attention sum reader network[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1603.01547>.
- [8] CHEN D Q, BOLTON J, MANNING C D. A thorough examination of the CNN/daily mail reading comprehension task[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1606.02858>.
- [9] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1611.01603>.
- [10] CHEN D Q, FISCH A, WESTON J, et al. Reading Wikipedia to answer open-domain questions[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1704.00051v2>.
- [11] WANG W H, YANG N, WEI F R, et al. Gated self-matching networks for reading comprehension and question answering[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Washington D. C., USA: IEEE Press, 2017: 189-198.
- [12] HUANG H Y, ZHU C G, SHEN Y L, et al. FusionNet: fusing via fully-aware attention with application to machine comprehension[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1711.07341>.
- [13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of NIPS'17. Cambridge, USA: MIT Press, 2017: 5998-6008.
- [14] CUI Y M, LIU T, CHEN Z, et al. Consensus attention-based neural networks for Chinese reading comprehension[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1607.02250>.
- [15] CUI Y M, LIU T, CHE W X, et al. A span-extraction dataset for Chinese machine reading comprehension[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1810.07366>.
- [16] CUI Y M, LIU T, YANG Z Q, et al. A sentence cloze dataset for Chinese machine reading comprehension[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/2004.03116>.
- [17] HE W, LIU K, LIU J, et al. Dureader: a chinese machine reading comprehension dataset from real-world applications[C]//Proceedings of Workshop on Machine Reading for Question Answering. Washington D. C., USA: IEEE Press, 2018: 37-46.

(上接第 72 页)

- [18] 徐丽丽,李茹,李月香,等. 面向机器阅读理解的语句填补答案选择方法[J]. 计算机工程, 2018, 44(7): 183-187, 192.
- XU L L, LI R, LI Y X, et al. Answer selection method of sentence filling for machine reading comprehension[J]. Computer Engineering, 2018, 44(7): 183-187, 192. (in Chinese)
- [19] SHAO C C, LIU T, LAI Y T, et al. DRCD: a Chinese machine reading comprehension dataset[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1806.00920>.
- [20] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for Chinese BERT[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/1906.08101>.
- [21] TAY Y, BAHRI D, METZLER D, et al. Synthesizer: rethinking self-attention in transformer models[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/2005.00743>.
- [22] CUI Y M, CHE W X, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[EB/OL]. [2021-06-20]. <https://arxiv.org/abs/2004.13922>.

编辑 索书志