DOI：10.19789/j.1004-9398.2025.02.001
文献引用：文勇军，吴金铭，梅硕．基于预训练模型与双向注意力流的抽取式阅读理解模型[J]．首都师范大学学报（自然科学版），2025，46（2）：1-11. WEN Y J，WU J M，MEI S. An extractive reading comprehension model based on pre-trained model with bi-directional attention flow[J]. Journal of Capital Normal University（Natural Science Edition），2025，46（2）：1-11.

基于预训练模型与双向注意力流的抽取式阅读理解模型

文勇军*，吴金铭，梅　硕

（长沙理工大学物理与电子科学学院，湖南 长沙　410114）

摘要：针对服务机器人的抽取式阅读理解任务中出现答案抽取准确度不高的问题，构建了基于预训练模型与双向注意力流的抽取式阅读理解模型。该模型首先采用预训练模型来提取问题与文档上下文的浅层联合语义表征；其次利用双向注意力网络来加强特征交互和信息融合，得到问题与文档上下文的深层联合语义表征；最后结合浅层和深层的联合语义表征，通过排序、滤错和定位操作完成对答案的抽取。在抽取式问答任务的斯坦福英文机器阅读理解数据集SQuAD 1.1和“讯飞杯”中文机器阅读理解数据集CMRC 2018上进行了实验。结果表明：与英文预训练语言模型BERT相比，该模型的性能指标E_{M}和$F1$值分别提高了1.172%和1.194%；与中文预训练语言模型RoBERTa-wwm-ext相比，该模型的E_{M}和$F1$值分别提高了1.336%和0.921%。

关键词：自然语言处理；机器阅读理解；预训练模型；双向注意力流（BERT）；RoBERTa-wwm-ext；答案抽取
中图分类号：TP391　　**文献标志码**：A

An extractive reading comprehension model based on pre-trained model with bi-directional attention flow

WEN Yongjun*，WU Jinming，MEI Shuo

（School of Physical & Electric Science，Changsha University of Science & Technology，Changsha Hunan　410114）

Abstract：An extractive reading comprehension model based on a pre-training model with bidirectional attention flow is constructed for the problem of low accuracy in answer prediction that occurs in extractive reading comprehension tasks of service robots. The model first uses a pre-training model to extract the shallow joint semantic representations of the question and the document context，then uses a bidirectional attention network to enhance feature interaction and information fusion to obtain the deep joint semantic features of the question and the document context. Finally，combines the shallow and deeps joint semantic representations to complete the extraction of answers through ranking，error filtering and localization operations. Experiments were conducted on the Stanford English machine reading comprehension dataset SQuAD 1.1 and the “iFlytek Cup” Chinese machine reading comprehension dataset CMRC 2018 for the extractive question and answer task. The results show that

收稿日期：2023-06-08
*通信作者：micowen@csust.edu.cn

compared with the English pre-trained language model BERT, the performance metrics E_M and $F1$ values of this model are improved by 1.172% and 1.194%, respectively; compared with the Chinese pre-trained language model RoBERTa-wwm-ext, the E_M and $F1$ values are improved by 1.336% and 0.921%, respectively.

Keywords: natural language processing; machine reading comprehension; pre-training model; bidirectional encoder representations from transformers(BERT); RoBERTa-wwm-ext; answer extraction

CLC: TP391 **DC**: A

0 引 言

服务机器人是用来与人进行交互,以自然语言作为传达信息的主要方式。因此,让机器学会理解语义是人机交互的前提,也是制约服务机器人工作效率的一个重要因素[1]。机器阅读理解(machine reading comprehension,MRC)是一种能全面评估机器对自然语言理解能力的方法,即对用户需求进行分析、理解,并从知识库中提取出相应答案,做出准确的回复指令。

早在20世纪70年代,研究者就已经在研究MRC任务,用来测试机器对自然语言的认知与理解。Lehnert[2]给出了一个强有力的语言理解视角,强调了上下文语境对回答问题的重要作用。但当时是以规则为基础的方法,人工生成的数据规模很小,并且仅局限于手动编写的脚本,导致MRC系统性能不佳,要拓展到更广阔的领域是困难的。直到20世纪90年代后期,Hirschman等[3]建立并评估了一个融合了模式匹配(词袋模型)技术和自然语言处理5大语义分析技术的基线系统,使得其对正确语句的检索精度可达30%~40%。但该方法需要依赖大量人工设定的规则或特征,其泛化能力较差,难以在实际应用中得到推广。

目前,大多数MRC模型的答案抽取准确度不高,如何对模型进行优化,以提高服务机器人作答的准确性,是当前MRC领域的热点问题之一[4]。本文的主要工作是提出一种基于预训练模型与双向注意力流的抽取式阅读理解模型,经实验论证该模型可以提高答案抽取的准确性,增强模型的泛化能力,进一步改善服务机器人的用户体验。

1 MRC任务

1.1 框架结构的演变

自2015年以来,随着深度学习的发展,长短期记忆网络[5]和Seq2Seq结构[6]的提出,使得设计更复杂的网络模型成为可能。MRC的基本框架大致可以分为词嵌入层、编码层、交互层和预测层4个网络层[7],如图1(a)所示。

在做MRC任务时,传统的静态词向量具有一定的局限性,如上下文特征信息不足、难以解决一词多义等问题[8]。将预训练模型在文本语料库上,利用大量外部数据进行无监督预训练,可以捕获文本序列之间更深层次的依赖关系,从而解决数据长距离依赖的问题。只需要设计建立高效的模型迁移到具体任务中,再进行有监督训练微调,就能获得相当不错的表现,在某些数据集上甚至超越了人类表现,大大提升了MRC的性能[9]。当前基于预训练语言模型已成为面向MRC任务的主流模型,如图1(b)所示。

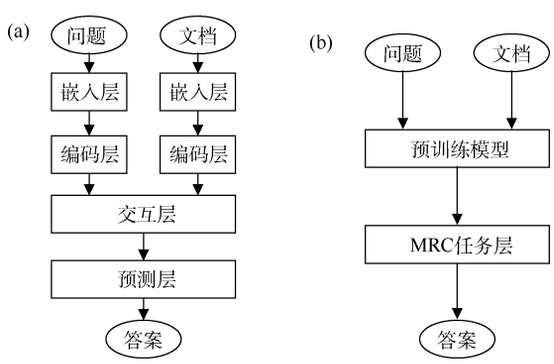


图1 机器阅读理解(MRC)模型结构对比
(a)传统MRC的基本框架;(b)基于预训练模型的MRC

预训练模型取得成功不仅依靠大规模语料库训练的优势,基础模型Transformer[10]的强大也是一个关键因素。Transformer利用多头注意力机制实现快速并行,改进了循环神经网络(recurrent neural networks,RNN)[11]训练速度慢的缺点。Transformer可以扩展到非常深的深度,充分发掘深度神经网络(deep-learning neural network,DNN)[12]的特性,提升模型准确率。注意力机制自从被证明在MRC任务上是有效的,就成为了许多模型改进的重要思路[13]。

由于抽取式阅读理解更加能够贴近服务机器

人的实际应用,本文主要展开对抽取式阅读理解模型的研究。当前大多数模型的阅读理解行为与人类表现都存在一定的距离,主要体现在答案抽取准确度不高[14-16]。本文利用BERT和RoBERTa预训练语言模型和双向注意力流模型等主流技术,模仿人类在阅读理解中的行为,泛读时得到浅层的字面细节,精读时挖掘深层的内在语义,结合2次预测的片段进行答案推理,可以很好地解决答案抽取准确度不高的问题。

1.2 抽取式任务定义

抽取式阅读理解是MRC任务中一个重要的分支,具体的任务定义为:对于给定的用户问题(Question,$\boldsymbol{Q}$)和相关文档上下文(Context,$\boldsymbol{C}$),根据文档内容给出该问题的答案(Answer,$\boldsymbol{A}$)。数据处理后的每个样本是一个3元组<$\boldsymbol{Q}$,$\boldsymbol{C}$,$\boldsymbol{A}$>形式,可以表示为寻求$P(\boldsymbol{A}|\boldsymbol{Q},\boldsymbol{C})$最优解的过程,如图2所示。答案由相关文档中的1个片段组成,通常也称为片段抽取[17]。

输入 Question = $\{q_1, q_2, \cdots, q_m\}$ 和 Context = $\{c_1, c_2, \cdots, c_n\}$,输出 Answer = $\{a_s, \cdots, a_e\}$。其中:m为用户问题的长度;n为文档上下文的长度;$\boldsymbol{A} \in \boldsymbol{C}$;$a_s$表示答案的开始;$a_e$表示答案结束。

2 抽取式阅读理解模型

采用与人类阅读理解类似的方式,首次阅读时得到浅层的字面细节,再次阅读时挖掘更深层次的内在语义[18]。结合2次预测的片段所在位置进行答案调整,获取最终的预测结果,以提升预测答案的准确性。本文针对抽取式阅读理解任务设计了一种优化模型,模型分为浅层联合语义表征模块、深层联合语义表征模块和答案片段预测模块3个模块,其整体模型结构如图3所示。第1部分是浅层联合语义表征模块,该模块通过大型预训练模型进行词嵌入和编码,完成文本的特征提取,获得$\boldsymbol{A}$和$\boldsymbol{C}$的浅层联合语义表征$\boldsymbol{H} \in R^{d\times(m+n)}$;第2部分是深层联合语义表征模块,该模块将浅层联合语义表征切分成$\boldsymbol{Q} \in R^{d\times m}$和$\boldsymbol{C} \in R^{d\times n}$的单独语义表征,采用双向注意力流进行语义表征交互,融合$\boldsymbol{C}$和$\boldsymbol{A}$之间的特征信息,得到一个问题感知上下文表征$\boldsymbol{G} \in R^{4d\times n}$,降维成$\boldsymbol{G}' \in R^{d\times n}$,再与问题表征$\boldsymbol{Q} \in R^{d\times m}$拼接,得到$\boldsymbol{A}$和$\boldsymbol{C}$的深层联合语义表征$\boldsymbol{D} \in R^{d\times(m+n)}$;第3部分是答案片段预测模块,该模块将前2部分的输出输入至全连接层,分别得到一对答案起止位置的指针概率向量,为了提升预测答案的准确性,将二者预测答案的概率向量叠加进行合并调整,通过Softmax函数归一化获取在整个序列上start和end的概率分布$\boldsymbol{A} \in R^{2\times(m+n)}$,找出$\boldsymbol{A}$的最佳起止位置,并抽取出最终的预测结果。

2.1 浅层联合语义表征模块

2.1.1 重构输入序列

浅层联合语义表征模块将$\boldsymbol{Q} = \{q_1, q_2, \cdots, q_m\}$和$\boldsymbol{C} = \{c_1, c_2, \cdots, c_n\}$表示,作为预训练模型的输入序列,添加[CLS]作为起始标志,[SEP]作为句子间分隔标志,最大序列长度为$L(L = m + n + 3)$。当$\boldsymbol{C}$过长时,如果直接采取截断处理,问题的答案有可能恰巧在被截断部分,因此采用滑动窗口的方法来构造整个模型的输入序列,如图4所示。

假设2个样本构成一个batch,根据指定的最大序列长度和滑动窗口大小,将原始样本进行滑动窗口处理并得到多个子样本,子样本右边的3列数字分别表示每个子样本的起止位置索引(无答案时置0)和原始样本对应的问题编码标识(identity document,ID)。需要注意的是,问题部分不参与滑动,为了方便后一模块的切分,使用[PAD]将长度不同的问题序列补齐到统一长度。

2.1.2 预训练模型的嵌入层

每个子样本都会经过预训练模型的嵌入层,将输入序列中的每个词转换成向量形式,预训练模型的嵌入层如图5所示。嵌入向量是3种不同的向量表示按元素相加[19],分别是:词嵌入将分词后的各个词元转换成固定维度$d(d = 768)$的向量;句子嵌入用以区分一个句子对中的2个句子;位置嵌入带有输入的顺序属性。

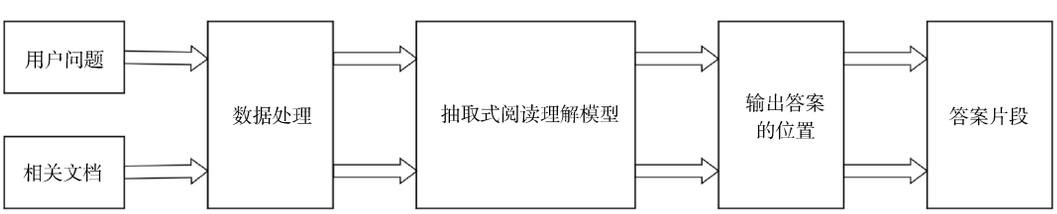
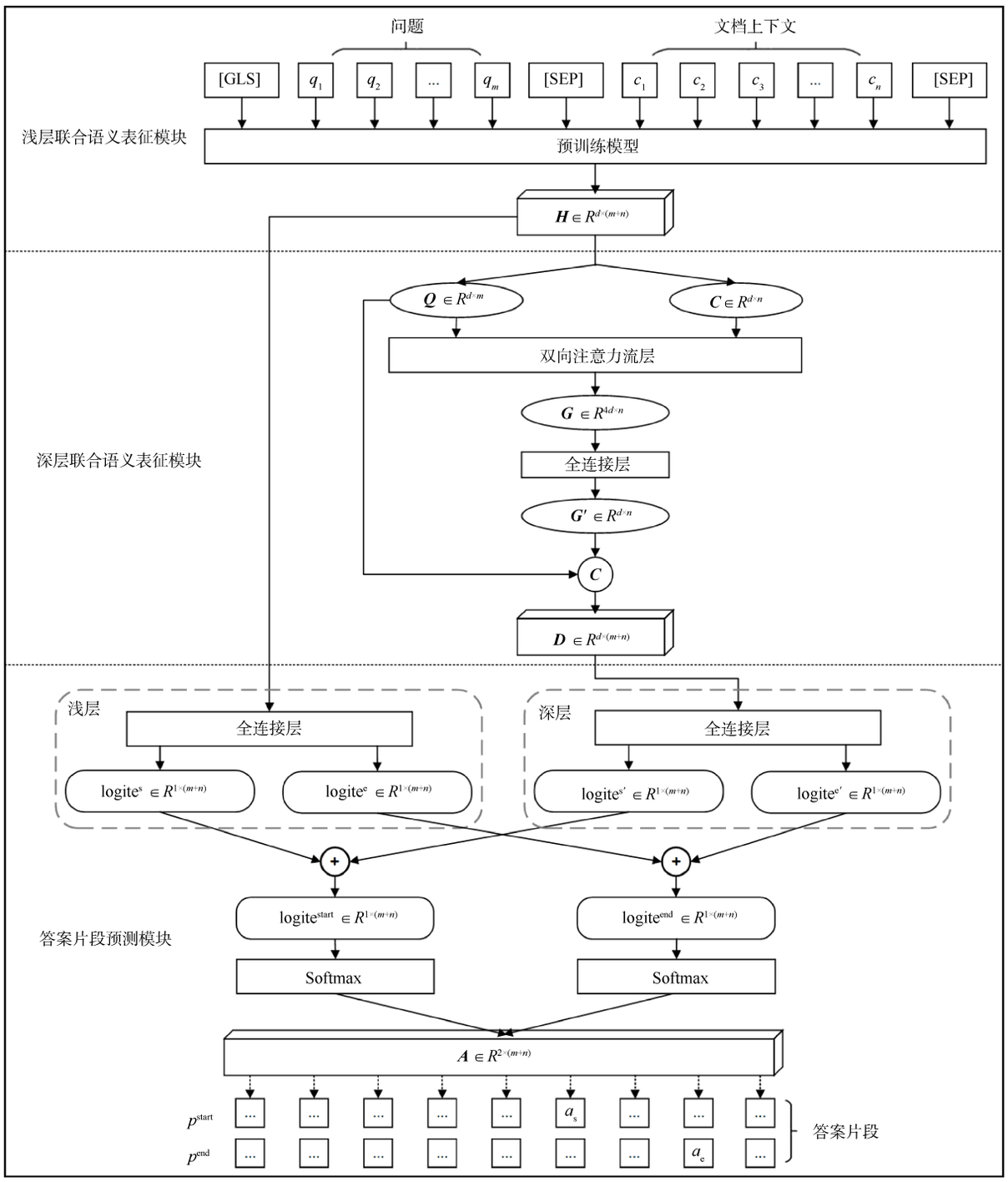


图2 抽取式阅读理解任务



注:q_m为问题子序列,表示有m个问题词;c_n为文档上下文子序列,表示有n个文档词;$\boldsymbol{Q} \in R^{d\times m}$是问题表征;$\boldsymbol{C} \in R^{d\times n}$是文档上下文表征;$\boldsymbol{H} \in R^{d\times(m+n)}$是浅层联合语义表征;$\boldsymbol{D} \in R^{d\times(m+n)}$是深层联合语义表征;$\boldsymbol{G} \in R^{4d\times n}$是问题感知上下文表征;$R$代表实数,其指数为矩阵大小;[CLS]为起始标志,[SEP]为句子间分隔标志;logits为指针概率向量。

图3　抽取式阅读理解模型结构

2.1.3　预训练模型的编码层

随后嵌入层的输出将被输入至预训练模型的编码阶段,编码器是由N个相同的Transformer[10]中Encoder层堆叠在一起[19]。每个Encoder层都是由一个多头自注意力机制,外加全连接的前馈神经网络组成,还使用了残差连接和归一化操作。多头自注意力机制的头数用h(一般$h=8$)表示,其输出由多个自注意力层的输出拼接后进行线性变换得到,如图6所示。

自注意力机制实际上是一种特殊的注意力机

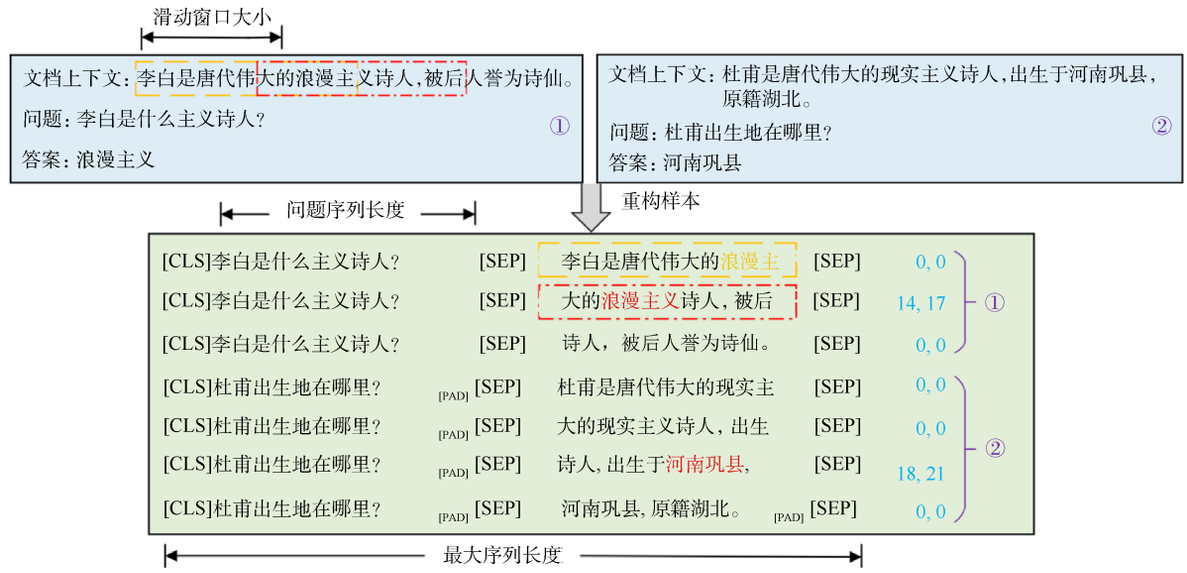


图 4 滑动窗口原理

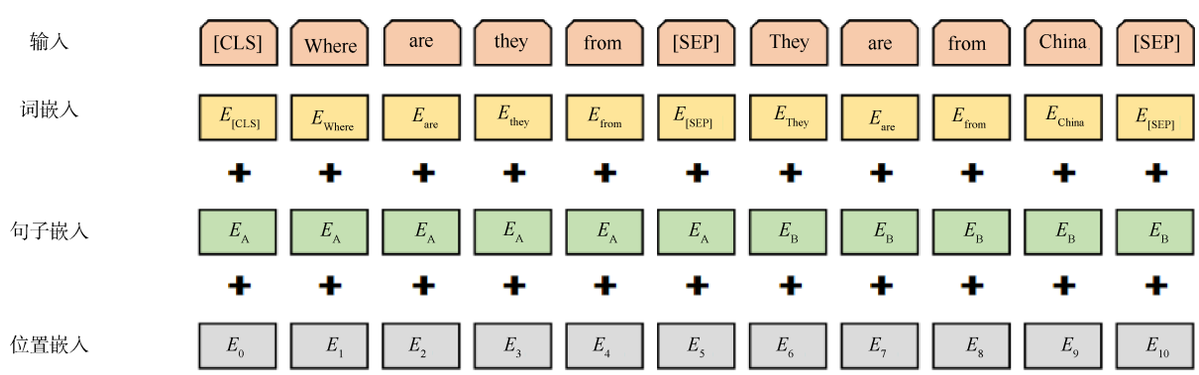


图 5 预训练模型的嵌入层

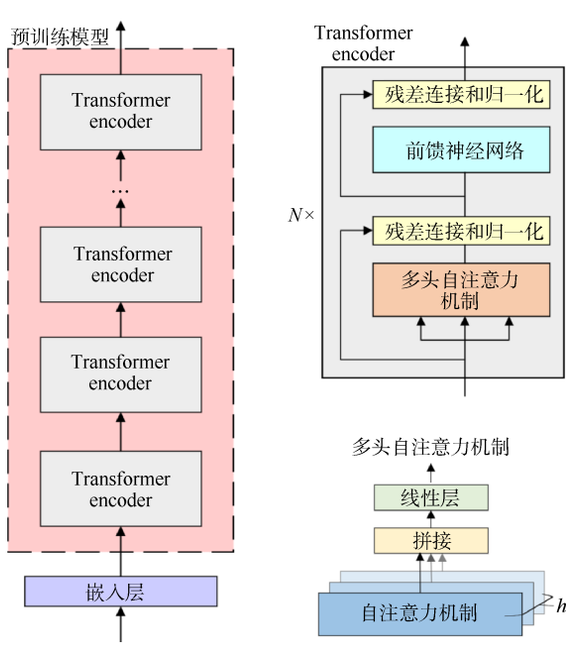


图 6 预训练模型的编码过程

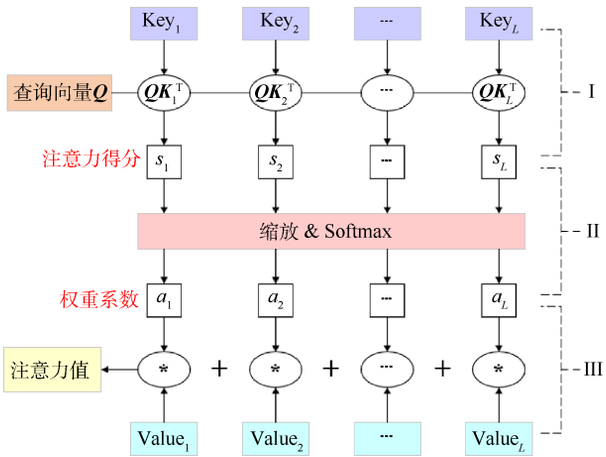
制,主要分为 3 个部分:第 1 部分先介绍注意力机制的公式原理,第 2 部分介绍自注意力机制和注意力机制的区别,第 3 部分给出多头自注意力机制的演变示例图。

(1)注意力机制[20]能使研究者从大量信息中,聚焦在少量重要的信息上,忽略大多数不重要的信息。注意力机制的计算公式为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Soft max}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_K}}\right)\boldsymbol{V}。 \quad (1)$$

式中:Attention 为注意力机制;$\boldsymbol{Q}$(Query)、$\boldsymbol{K}$(Key)和 $\boldsymbol{V}$(Value)分别为查询向量、键向量和值向量;T 表示转置运算,d_k 为向量 $\boldsymbol{K}$ 的维度;Softmax 为归一化函数。

注意力机制的计算流程具体实现过程如图 7 所示,第 1 步将 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 做点积计算二者之间的相似性,得到注意力得分;第 2 步对注意力得分进行缩放(除以维度的开方 $\sqrt{d_k}$)来确保稳定的梯度,再进行 Softmax 归一化,得到权重系数;第 3 步根据权重系数对 $\boldsymbol{V}$ 进行加权求和,得到的注意值具有注意力信息。



注:$\boldsymbol{Q}$代表查询向量;Key表示键向量$\boldsymbol{K}$;Value表示值向量$\boldsymbol{V}$;$\boldsymbol{QK}^{\mathrm{T}}$表示$\boldsymbol{Q}$与$\boldsymbol{K}$的转置做点积运算;I为做点积计算相似度得到注意力得分;II为$\sqrt{d_k}$缩放Softmax()归一化;III为加权求和。

图7　注意力机制的计算流程

(2)自注意力机制[10]是注意力机制的一个具体做法,不同点在于$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$来自同一个输入向量$\boldsymbol{X}$,由向量序列$\boldsymbol{X}$和3个可训练的转换矩阵$\boldsymbol{W}^{K}$、$\boldsymbol{W}^{Q}$和$\boldsymbol{W}^{V}$相乘得到,如图8所示。按照注意力机制的步骤进行计算,其中$\sqrt{d_k}=\sqrt{L/h}=8$。通过自注意力模型,给定向量序列$\boldsymbol{X}$被编码成新的向量序列$\boldsymbol{Z}$,$\boldsymbol{Z}$相较于$\boldsymbol{X}$拥有了句法特征和语法特征。

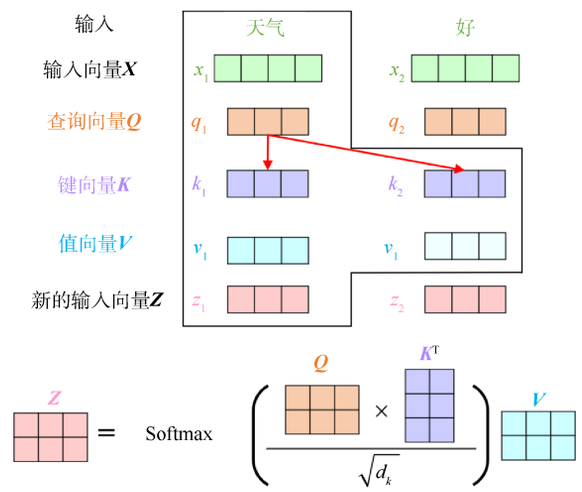


图8　自注意力机制的计算流程

(3)多头自注意力机制。自注意力机制在对当前位置的信息进行编码时,会过度地将注意力集中于自身的位置,因此希望基于相同的注意力机制学习到不同的行为,然后将不同的行为作为知识组合起来,这种设计被称为多头注意力机制[8]。多头自注意力机制的计算公式为:

$$\begin{aligned}\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})&=\mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_h)\boldsymbol{W}^{\mathrm{O}},\\ \mathrm{head}_i&=\mathrm{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i),\\ \boldsymbol{Q}_i&=\boldsymbol{X}_i^{Q},\boldsymbol{K}_i=\boldsymbol{X}_i^{K},\boldsymbol{V}_i=\boldsymbol{X}_i^{V},\end{aligned}\tag{2}$$

式中:MultiHead为多头注意力机制;Concat为拼接操作;$\boldsymbol{W}^{\mathrm{O}}$为输出权重;$\boldsymbol{X}$为输入向量。

多头注意力机制的计算流程具体实现如图9所示,把输入向量序列$\boldsymbol{X}$映射到8个不同的子空间上,对应$h(h=8)$组$\boldsymbol{W}^{Q}$、$\boldsymbol{W}^{K}$和$\boldsymbol{W}^{V}$,通过输出8组不同的查询/键/值($\boldsymbol{Q}/\boldsymbol{K}/\boldsymbol{V}$)计算注意力得到$\boldsymbol{Z}_0$~$\boldsymbol{Z}_7$,拼接后乘以权重矩阵$\boldsymbol{W}^{\mathrm{O}}$,进行线性变换(降维)得到$\boldsymbol{Z}$。这样可以捕捉更加丰富的特征信息,模型效果更好。

2.2　深层联合语义表征模块

本模块将浅层联合语义表征$\boldsymbol{H}\in R^{d\times(m+n)}$切分成$\boldsymbol{Q}\in R^{d\times m}$和$\boldsymbol{C}\in R^{d\times n}$,并参考BiDAF模型[21]的双向注意力流,计算2个方向上的注意力特征信息,既能根据问题词找到文档中最相关部分,又能找到每个文档词对回答问题的重要性,起到相互补充的作用[22]。

2.2.1　双向注意力流层

首先,计算$\boldsymbol{Q}$和$\boldsymbol{C}$的上下文嵌入之间的共享相似矩阵$\boldsymbol{S}\in R^{m\times n}$,使用的相似函数是三线性函数

$$\begin{aligned}S_{ij}&=\alpha(\boldsymbol{Q}_{:i},\boldsymbol{C}_{:j}),\\ \alpha(q,c)&=\boldsymbol{w}_{(s)}^{\mathrm{T}}[q;c;q\circ c],\end{aligned}\tag{3}$$

式中:S_{ij}表示第i个问题词与第j个文档词之间的相似度;α是可训练的标量函数,编码其2个输入向量之间的相似性;$\boldsymbol{Q}_{:i}$是$\boldsymbol{Q}$的第i列向量,$\boldsymbol{C}_{:j}$是$\boldsymbol{C}$的第j列向量;$\boldsymbol{w}_{(s)}\in R^{3d\times 1}$是可训练的权重向量;[;]是向量在行上的拼接;∘是对应元素相乘。

使用相似矩阵$\boldsymbol{S}$来获得2个方向的注意力权重和参与向量,下面将详细阐述2个方向的注意力流,即文档上下文-问题注意力流(context-to-question,C2Q)和问题-文档上下文注意力流(question-to-context,Q2C)。

(1)文档上下文-问题注意力流。表示哪些问题词与每个文档词最相关,如图10(a)所示。让$a_j=\mathrm{Softmax}(\boldsymbol{S}_{:j})\in R^{m\times 1}$表示第$j$个文档词对每个问题词的注意力权重,对于所有$j$,$\sum_i^m a_{ij}=1$。随后每个参与的问题向量为$\tilde{\boldsymbol{Q}}_{:j}=\sum_i^m a_{ij}\boldsymbol{Q}_{:i}\in R^{d\times 1}$,因此,$\tilde{\boldsymbol{Q}}\in R^{d\times n}$包含整个文档的参与问题向量。

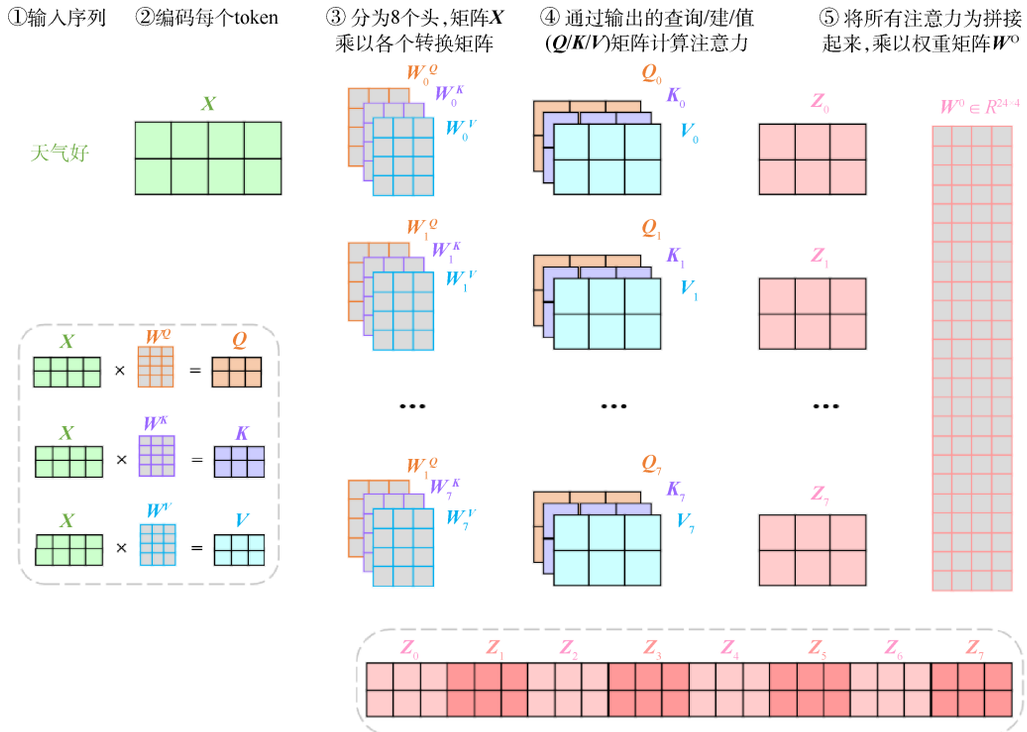


图 9　多头注意力机制的计算流程

(2)问题-文档上下文注意力流。表示哪些文档词与问题词最相似,这对于回答问题至关重要。如图 10(b)所示,通过 $b = \mathrm{Softmax}(\max_{\mathrm{col}}(S)) \in R^{1\times n}$ 获得文档词的注意力权重,其中最大函数($\max_{\mathrm{col}}$)跨列执行。参与的文档向量是 $\tilde{\boldsymbol{c}} = \sum_{j}^{n} b_j \boldsymbol{C}_{:j} \in R^{d\times 1}$,该向量表示文档中关于问题的最重要单词的加权和。$\tilde{\boldsymbol{c}}$在列上扩展 n 次,然后得到 $\tilde{\boldsymbol{C}} \in R^{d\times n}$。

将文档上下文表征 $\boldsymbol{C}$、注意力向量 $\tilde{\boldsymbol{Q}}$ 和 $\tilde{\boldsymbol{C}}$ 组合在一起产生 $\boldsymbol{G} \in R^{d_G\times n}$,其中每个列向量可以被视为每个文档词的问题感知表示。定义 $\boldsymbol{G}$

$$\boldsymbol{G}_{:j} = \beta(\boldsymbol{C}_{:j}, \tilde{\boldsymbol{Q}}_{:j}, \tilde{\boldsymbol{C}}_{:j}) \in R^{d_G},$$
$$\beta(c, \tilde{q}, \tilde{c}) = [c;\ \tilde{q};\ c\circ\tilde{q};\ c\circ\tilde{c}] \in R^{4d\times 1},\quad (4)$$
$$(d_G = 4d),$$

式中:$\boldsymbol{G} \in R^{d_G\times n}$ 是第 j 个列向量(对应于第 j 个文档词);β 是融合其(3 个)输入向量的可训练向量函数;d_G 是 β 函数的输出维度。

2.2.2　降维拼接操作

为了后一模块的合并,将其与编码后的文档上下文表征 $\boldsymbol{C} \in R^{d\times n}$ 保持维度一致,将双向注意力流输出的问题感知上下文表征 $\boldsymbol{G} \in R^{4d\times n}$ 经过全连接层[23],进行线性变换(降维)得到 $\boldsymbol{G}' \in R^{d\times n}$,再用原始问题表征 $\boldsymbol{Q} \in R^{d\times m}$ 与之拼接[24],如图 11 所示。

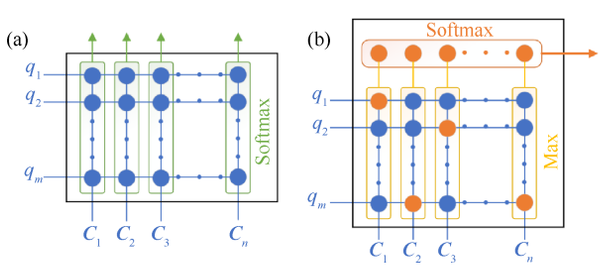


图 10　双向注意力流

(a)文档上下文-问题注意力流;(b)问题-文档上下文注意力流

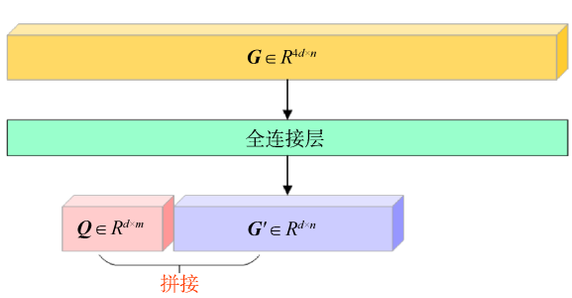


图 11　降维拼接操作

2.3　答案片段预测模块

抽取式阅读理解模型的最后一个模块,将浅层联合语义表征输入至全连接层,对整个序列中的每个位置进行二分类,判断其是否为答案的起始位置(或者终止位置),并输出 start/end 的概率值,得到答案起始的指针概率向量(log its^s)和答案终止的指针

概率向量($\log its^{e}$)。同样将深层联合语义表征输入至全连接层后,获得答案开始位置的输出$\log its^{s'}$和结束位置的输出$\log its^{e'}$。

2.3.1　合并预测

合并调整[25]得到起止位置的答案指针概率向量$\log its^{start}$和$\log its^{end}$,表示整个序列中每个位置上属于start/end的概率值,按照式(5)和(6)进行计算:

$$\log its^{start} = \log its^{s} + \log its^{s'}, \tag{5}$$

$$\log its^{end} = \log its^{e} + \log its^{e'}, \tag{6}$$

式中:$\log its^{s}$和$\log its^{e}$由浅层联合语义表征计算得出;$\log its^{s'}$和$\log its^{e'}$由深层联合语义表征计算得出。

通过Softmax函数计算输入序列中每个位置上$\log its^{start}$和$\log its^{end}$的概率分布

$$\begin{cases} p^{start} = \text{Soft}\max(\log its^{start}), \\ p^{end} = \text{Soft}\max(\log its^{end}), \end{cases} \tag{7}$$

式中:p^{start}为答案开始位置的概率;p^{end}为答案结束位置的概率。

2.3.2　答案推理

通过上述合并预测,对整个序列中每个词进行二分类,判断是否是答案的起始位置或者结束位置[26],并输出start/end的概率值。答案推理过程如图12所示,第1步取前K(K=10)个概率值最大的结果作为候选[27]。第2步需依据2个条件[18]:(1)起止索引肯定在文档部分;(2)起止索引肯定是小于等于终止索引,排除明显的错误结果。第3步起止位置概率值相加,排序取概率最大列,定位答案并完成抽取。

最小化训练损失采用交叉熵损失函数,定义为预测概率分布中真实开始和结束索引的负对数概率之和,再对所有样本取平均,即

$$L(\theta) = -\frac{1}{N}\sum_{i}^{N}[\log(p_{y_i^s}^{start}) + \log(p_{y_i^e}^{end})], \tag{8}$$

式中:θ是模型中所有可训练权重的集合;N是数据集中样本的数量;y_i^s和y_i^e分别是第i个样本的真实开始和结束索引。

3　实验与结果分析

3.1　实验数据

英文数据集SQuAD 1.1[28]:斯坦福大学研究人员提出第1个真正意义上的大规模抽取式阅读理解数据集SQuAD 1.1,直接推动了当时抽取式阅读理解的发展。该数据集以原文、问题和答案三元组形式,目前已经成为MRC任务中的基准数据集。

中文数据集CMRC 2018[29]:第二届"讯飞杯"MRC评测任务中发布了首个人工标注的中文片段抽取式阅读理解数据集CMRC 2018,该数据集由哈工大讯飞联合实验室推出,不仅弥补了抽取式阅读理解模型在中文语料的缺失,也推动了中文MRC的发展。

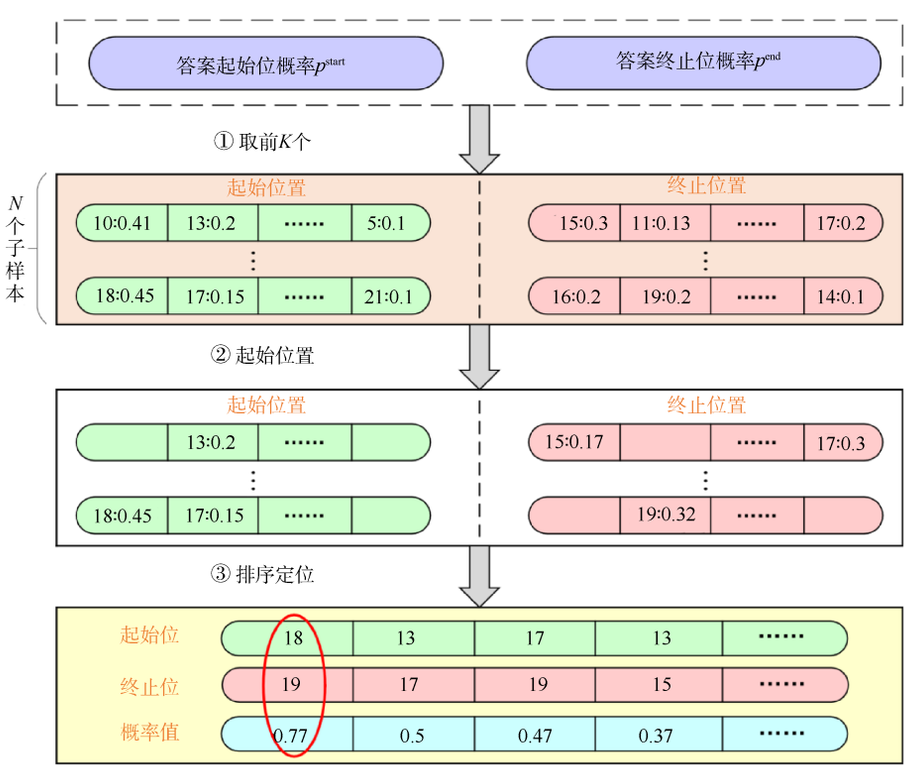


图12　答案推理处理流程

分别在英文数据集SQuAD 1.1和中文数据集CMRC 2018上进行实验,用以验证模型在不同语言上的性能和有效性。训练集用来训练模型,划分30%的训练集作为开发集,来优化模型参数;采用测试集进行推理来评估模型性能。预处理数据统计如表1所示。

表1 实验数据集统计信息

数据集	语言类型	训练集/个	开发集/个	测试集/个
SQuAD 1.1	英文	89 947	26 985	11 015
CMRC 2018	中文	19 071	5 722	6 181

3.2 评估指标

对于抽取式阅读理解任务,需要模型对预测所得答案和真实答案进行对比。常用精准匹配(exact match,E_M)和模糊匹配$F1$值2个度量标准来评估模型[28]。E_M指的是模型输出的预测答案与实际的真实答案相符合所占的比例。$F1$值是精确率和召回率的调和平均数

$$F1=\frac{2\times P\times R}{P+R},\tag{9}$$

式中:P是精确率,表示正确预测正样本占实际预测为正样本的比例;R是召回率,表示正确预测正样本占正样本的比例。

3.3 实验配置

实验环境配置为64位的Windows操作系统,Intel Xeon E5-2620六核处理器,32 GB Nvidia TITAN RTX显卡。使用PyTorch深度学习框架来完成模型的构建,首先将已训练好的预训练模型(英文BERT-base和中文RoBERTa-wwm-ext)参数进行下载,然后设置适当的循环轮数、批量大小、问题最大长度、序列最大长度与学习率进行微调,具体参数设置如表2所示。

表2 参数设置

参数	循环轮数	批量大小	问题最大长度	序列最大长度	学习率/10^{-5}
BERT-base	2	24	64	384	3.5
RoBERTa-wwm-ext	2	32	64	512	3.0

注:BERT-base为英文预训练模型;RoBERTa-wwm-ext为中文预训练模型。

3.4 实验结果及分析

3.4.1 模型性能对比

将微软亚洲研究院提出的多层网络结构R-NET[30]作为传统模型的代表,基准模型分别选择谷歌AI研究院提出的英文预训练模型BERT[17]和哈工大讯飞实验室基于全词mask[31]技术发布的中文预训练模型RoBERTa-wwm-ext[32],与本文模型在中英文2种数据集上进行了性能对比实验。性能结果如表3所示,R-NET通过设计巧妙的结构来搭建,性能指标却被只接1个全连接层的预训练模型反超,这不仅说明预训练语言模型具有很强的文本编码能力,也正是传统模型难以提升的瓶颈所在。本文模型在英文数据集SQuAD 1.1上E_M达到了81.972%,$F1$值达到了89.694%,相较于基准模型$\text{BERT}_{\text{base}}$分别提升了1.172%和1.194%;在中文数据集CMRC 2018上E_M达到了68.717%,$F1$值达到了88.121%,相较于基准模型$\text{RoBERTa-wwm-ext}_{\text{base}}$也分别提升了1.336%和0.921%,这说明本文模型适用于中英文2种不同语言的同时,还能提高模型的性能。

表3 不同数据集上的模型性能对比 单位:%

模型	SQuAD 1.1		CMRC 2018	
	E_M	$F1$	E_M	$F1$
人类表现	80.300	90.500	91.083	97.348
R-NET	71.100	79.500	45.418	69.825
基准模型	80.800	88.500	67.381	87.200
本文模型	81.972	89.694	68.717	88.121

注:R-NET表示传统的机器阅读理解模型;E_M表示精准匹配,$F1$表示模糊匹配。

3.4.2 超参数验证

由表3的实验结果可知,在中文数据集上预训练模型选择的是比BERT(Chinese)效果更佳的RoBERTa-wwm-ext,但其性能指标还是不如英文数据集上的好,尤其是完全匹配E_M数值相差较大。因此,本文在模型优化过程中调整超参数大小,试图找到模型的最优解。实验统计了中文数据集在不同步长值下,基准模型和本文模型的答案不匹配个数。由图13可知,本文模型相较于基准模型,答案不匹配个数都有所下降,且步长取596时,答案抽取的错误率最低,与实验参数设置相符合($S=\frac{N}{B}=\frac{19\,071}{32}=595.968\,75\approx 596$,$S$表示步长,$N$表示训练样本数,$B$表示批量大小)。

3.4.3 预测片段对比

为了更直观地对比观察模型对答案片段的预测效果,仍然以中文数据集CMRC 2018为例,选择

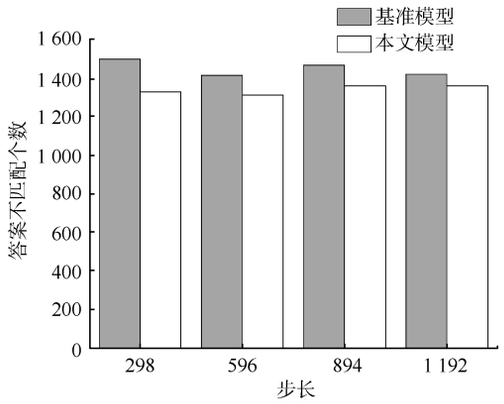


图13 中文数据集在不同步长值下的答案不匹配个数

模型最优参数,将基准模型和本文模型的预测答案片段,与人工标注的三元组中的正确答案片段进行对比,总结出存在的常见3类问题并举例说明二者之间存在的差异(表4)。基准模型预测答案片段时:(1)答案不全导致与正确答案存在偏差,出现似是而非的答案;(2)答案过于冗长,导致出现答案指代不明的现象;(3)对于错综复杂的相关文档,甚至出现了答案错误的情况,可见中文是一种更依赖语境与语义的语种。反观本文模型与正确答案的匹配度和准确性会相对较高,这也说明本文设计的模型结构对提升性能是有所帮助的。

表4 CMRC2018数据集上预测的答案片段对比

类别	文档上下文	问题	基准模型	正确模型	本文模型
答案不全	国际人类基因组单体型图(HapMap)计划是一个组织,其目标是开发人类基因组的单体型图,来描述人类遗传变异的常见模式	国际人类基因组单体型图计划的目标是什么?	开发人类基因组的单体型图	开发人类基因组的单体型图,来描述人类遗传变异的常见模式	开发人类基因组的单体型图,来描述人类遗传变异的常见模式
答案冗余	乍浦镇是隶属于中国浙江省嘉兴平湖市的一个镇,位于杭州湾北岸,毗邻上海市金山区。面积54 km^2,常住人口5万余人。乍浦镇地处杭嘉湖平原,宋、元两代亦曾开港贸易	乍浦镇在哪里?	中国浙江省嘉兴平湖市的一个镇,位于杭州湾北岸,毗邻上海市金山区	中国浙江省嘉兴平湖市	中国浙江省嘉兴平湖市
答案错误	海山汽车客运股份有限公司的前身为三峡汽车运输合作社,也是今日中兴巴士集团旗下之淡水客运前身。1974年3月,三峡汽车运输合作社由于经营困难,让售于台北客运,并更名为三峡汽车客运股份有限公司	1974年3月三峡汽车运输合作社更名成什么?	海山汽车客运股份有限公司	三峡汽车客运股份有限公司	三峡汽车客运股份有限公司

4 结束语

服务机器人可以将人们从繁杂且重复性高的工作中解放出来,实现“机器助人”。本文对抽取式阅读理解任务进行了研究,提出了基于预训练模型与双向注意力流的抽取式阅读理解模型,使得答案不匹配个数有所下降,答案抽取的错误率低,能很好地解决答案抽取准确度不高的问题。在中英文2种数据集上验证了本文模型的有效性,有利于增强模型的泛化能力,改善服务机器人的用户体验。

在未来的研究工作中,将从以下几个方面进一步完善:(1)抽取式阅读理解任务中答案是文档的某个子片段,而现实生活中人类阅读后可能需要组织新的文字表达,采用生成式与抽取式阅读理解相结合的方式,既能避免模型产生毫无依据的答案,也能进一步优化模型性能;(2)抽取式阅读理解任务中公开数据集是人工标注的数据集,利用未标注的数据来构建完整的问答系统,更加符合真实的应用场景;(3)服务机器人还涵盖多跳推理和多轮对话等任务,仅使用抽取式阅读理解模型的实验结果,来证明能提升MRC的整体水平还不够充分,后续可以进一步探索本文模型在其他类型任务中的可行性和实际效果。

参 考 文 献

[1] 潘磊,代翔,黄细凤,等.抽取式机器智能阅读理解问答系统:CN111611361B[P].2022-06-14.

[2] LEHNERT W G. The process of question answering[R]. New Haven: Yale University New Haven Department of Computer Science, 1977: 186-190.

[3] HIRSCHMAN L, LIGHT M, BRECK E, et al. Deep read: a reading comprehension system[C]//Association for Computational Linguistics. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. Maryland: Association for Computational Linguistics, 1999: 325-332.

[4] CHEN D Q. Neural reading comprehension and beyond[D]. Stanford: Stanford University, 2018.

[5] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[6] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Advances in Neural Information Processing Systems, 2014, 27: 3104-3112.

[7] QIU B,CHEN X,XU J,et al. A survey on neural machine reading comprehension[EB/OL].(2019-07-10)[2023-10-17]. https://doi.org/10.48550/arXiv.1906.03824.pdf.

[8] 张超然,裘杭萍,孙毅,等.基于预训练模型的机器阅读理解研究综述[J].计算机工程与应用,2020(11):17-25.

[9] 包玥,李艳玲,林民.抽取式机器阅读理解研究综述[J].计算机工程与应用,2021(12):25-36.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems,2017,30:5998-6008.

[11] ZAREMBA W,SUTSKEVER I,VINYALS O. Recurrent neural network regularization[EB/OL].(2015-02-19)[2023-10-17].https://doi.org/10.48550/arXiv.1409.2329.pdf.

[12] COVINGTON P, ADAMS J, SARGIN E. Deep neural networks for youtube recommendations[C]//Association for Computing Machinery. Proceedings of the 10th ACM Conference on Recommender Systems. San Francisco: Conference on Recommender Systems,2016:191-198.

[13] HERMANN K M, KOCISKY T, GREFENSTETTE E, et al. Teaching machines to read and comprehend [J]. Advances in Neural Information Processing Systems, 2015,28:1693-1701.

[14] 刘溪.基于预训练模型的阅读理解式问答技术研究及应用[D].哈尔滨:哈尔滨工业大学,2021.

[15] 李鑫.问题生成与答案抽取关键技术研究[D].长沙:国防科技大学,2020.

[16] 王洁,夏晓明.基于机器阅读理解的BiLSTM-BiDAF命名实体识别[J].华南理工大学学报(自然科学版),2022,50(12):80-88.

[17] LIU S, ZHANG X, ZHANG S, et al. Neural machine reading comprehension: methods and trends[J]. Applied Sciences,2019,9(18):3698.

[18] 彭宇,李晓瑜,胡世杰,等.基于BERT的三阶段式问答模型[J].计算机应用,2022(1):64-70.

[19] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Association for Computational Linguistics. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics,2019:4171-4186.

[20] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL].(2016-05-19)[2023-10-17].https://doi.org/10.48550/arXiv.1409.0473.pdf.

[21] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension [EB/OL]. (2018-06-21) [2023-10-17]. https://doi. org/10. 48550/arXiv.1611.01603.pdf.

[22] WU D,MA N,WAN F. Application research of attention mechanism in machine reading comprehension [C]// Association for Computing Machinery. Proceedings of the 3rd Asia-Pacific Conference on Image Processing, Electronics and Computers. Asia-Pacific: Conference on Image Processing,2022:314-318.

[23] LIN M,CHEN Q,YAN S. Network in network[EB/OL]. (2014-03-04) [2023-10-17]. https://doi. org/10.48550/arXiv.1312.4400.pdf.

[24] 叶俊民,赵晓丽,杜翔,等.片段抽取型机器阅读理解算法研究[J].计算机应用研究,2021(11):3268-3273.

[25] LIU Q, MAO R, GENG X, et al. Semantic matching in machine reading comprehension: an empirical study[J]. Information Processing & Management, 2023, 60(2): 103145.

[26] 苏立新,郭嘉丰,范意兴,等.面向多片段答案的抽取式阅读理解模型[J].计算机学报,2020(5):856-867.

[27] WANG Y, LEI C. QANet-based candidate answer rethink model for machine reading comprehension[J]. International Journal of Wireless and Mobile Computing,2021,20(3):246-254.

[28] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD:100,000+ questions for machine comprehension of text[C]//Association for Computational Linguistics. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA:Association for Computational Linguistics,2016.

[29] CUI Y,LIU T,CHE W,et al. A span-extraction dataset for Chinese machine reading comprehension [C]// Association for Computational Linguistics. Proceedings of Conference on Empirical Methods in Natural Language Processing and the 19th International Joint Conference on Natural Language Processing. Hong Kong: Conference on Natural Language Processing, 2019.

[30] WANG W. R-NET: machine reading comprehension with self-matching networks. Natural Language Computer Group,Microsoft Research. Asia[R]. Beijing: Technical Report,2017.

[31] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT [J]. IEEE/ACM Transactions on Audio,Speech,and Language Processing, 2021,29:3504-3514.

[32] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized BERT pretraining approach[EB/OL].(2019-07-26)[2023-10-17].https://doi.org/10.48550/arXiv.1907.11692.pdf.

(责任编辑:兰丽丽)