doi:10.16104/j.issn.1673-1891.2026.01.010

基于改进型BERT预训练模型的大规模文本语义匹配方法

周晓飞

(丽水职业技术学院电子信息学院,浙江 丽水 323000)

摘要:大规模文本数据具有数据量庞大的特点,且同一词汇在不同语境下可能具有完全不同的含义。仅依赖固定规则或模型,难以适应动态的语义变化,这会导致信息丢失和语义不完整。在这种情况下,无法捕捉到深层次的语义信息和语境关系,进而影响语义匹配的准确性。为解决这一问题,本文提出基于改进型双向编码器表征(bidirectional encoder representations from transformers,BERT)模型的大规模文本语义匹配方法。该改进的BERT预训练模型通过文本词向量的位置编码来增强文本的语境信息特征,从而有效捕捉文本的语境信息。此外,采用注意力机制动态计算特征融合权重,并通过加权融合方法生成文本的融合语义特征。通过文本特征信息提取、多维知识编码、融合语义标签生成以及语义匹配关系预测4个步骤,评估待匹配文本之间的语义一致性。本文设定一致性阈值为0.8,即当预测值超过0.8时,认为待匹配文本具有较高的语义一致性,从而实现准确的文本语义匹配。实验结果表明,基于大规模文本样本数据得到的平均倒数排名(mean reciprocal rank,MRR)高于0.7,且与对比方法相比,匹配结果更加准确。

关键词:改进型BERT预训练模型;融合特征;位置编码;文本向量化;注意力机制;语义匹配

中图分类号:TP391　　**文献标志码:**A　　**文章编号:**1673-1891(2026)01-0084-09

A Large Scale Text Semantic Matching Method Based on an Improved BERT Pre-Training Model

ZHOU Xiaofei

(School of Electronic Information, Lishui Vocational & Technical College, Lishui 323000, Zhejiang, China)

Abstract: Large-scale text data is characterized by an enormous volume, and the same vocabulary may carry completely different meanings in diverse contexts. Relying solely on fixed rules or models makes it difficult to adapt to dynamic semantic changes, which leads to information loss and semantic incompleteness. In such cases, deep semantic information and contextual relationships cannot be captured, thereby impairing the accuracy of semantic matching. To address this issue, a large-scale text semantic matching method based on an improved bidirectional encoder representations from transformers (BERT) pre-training model is proposed in this paper. Therefore, a large-scale text semantic matching method based on an improved BERT pretrained model is proposed." The improved BERT pre-training model is applied to enhance the contextual information features of text via positional encoding of text word vectors, thus capturing the contextual information of text effectively. Furthermore, the attention mechanism is adopted to dynamically calculate the feature fusion weights, and a weighted fusion method is used to generate the fused semantic features of text. The semantic consis-

收稿日期:2025-10-14
基金项目:浙江省高职教育"十四五"教学改革项目(jg20240348)。
作者简介:周晓飞(1980—),女,浙江丽水人,助理研究员,博士,主要研究方向为信息技术、教学管理。E-mail:vvbb090@sina.com。

tency between texts to be matched is evaluated through four steps: text feature information extraction, multi-dimensional knowledge encoding, fused semantic tag generation, and semantic matching relationship prediction. A consistency threshold of 0.8 is set, meaning that the texts to be matched are considered to have high semantic consistency when the predicted value exceeds 0.8, thus achieving accurate text semantic matching. Test results show that the Mean Reciprocal Rank (MRR) obtained based on large-scale text sample data is higher than 0.7, and the matching results are more accurate compared with the contrast methods.

Keywords: improved BERT pre-training model; fused features; positional encoding; text vectorization; attention mechanism; semantic matching

0 引言

在计算机与互联网技术的快速发展中,海量文本数据以前所未有的速度呈指数式增长,广泛分布于新闻资讯、社交媒体言论、学术论文等领域,蕴含巨大信息价值。然而,如何从海量数据中快速、准确地获取与用户需求高度匹配的信息,成为自然语言处理领域亟待解决的关键问题。文本语义匹配作为解决该问题的核心手段,旨在精准判断2个文本语义是否高度相似,在数据库信息搜索、人工客服智能回复等众多场景中至关重要,对大规模文本高效语义匹配意义重大且前景广阔。但传统的文本匹配方法多聚焦于基于符号和关键词的匹配策略,虽能实现简单匹配任务,却存在明显局限,其仅停留在文本表面层次,无法挖掘深层次特征如语义内涵、上下文语境等,处理复杂语义文本时匹配效果不佳,且为提升匹配效果常需借助外部知识库补充信息,这增加了系统复杂度与计算成本,还面临知识库更新不及时、覆盖范围有限等问题,限制了其在大规模文本匹配场景中的应用。因此,对大规模文本进行语义匹配具有重要意义。

在如何提高语义匹配准确度方面,相关研究者做了一些尝试[1]。赵云肖等[2]通过信息编码层对汉字的形音义多元知识进行编码,结合编码分类标签和信号监督标签完成文本语义关系的判别。该方法充分利用了汉字的形音义多元知识,增强了模型对文本语义的理解能力。然而,模型在编码和整合多元知识时可能存在信息丢失或误判的情况,导致匹配精度降低。王奥等[3]增强农业问句文本的语义推断特征和文本距离特征后,将其嵌入多方位匹配函数中,实现语义匹配。此方法能够从多个角度进行相似度对比,进一步增强了匹配的鲁棒性,但对于一些罕见或复杂的农业术语,模型难以准确捕捉到所有关键特征,影响了匹配精度。张文慧等[4]采用差异特征提取器提取文本的关联特征,并结合门控和语义融合方法实现匹配,该方法可以较好地解决语义匹配多元性问题,提升匹配效率。然而,短文本的长度限制会导致模型在解析和理解文本时出现信息不足的情况,从而影响匹配精度。刘萌等[5]基于全局特征和局部特征,利用深度学习算法训练多模态模型,从而达到语义匹配的目的。但图像和文本的质量对匹配结果具有较大影响,无法保证匹配准确度。

上述方法虽在提升语义匹配准确度上各有探索,但均因存在如信息丢失误判、难捕捉关键特征、信息不足、受数据质量影响等问题,导致在大规模文本语义匹配中匹配精度受限。为此,本文提出基于改进型双向编码器表征(bidirectional encoder representations from transformers,BERT)模型的大规模文本语义匹配方法。该方法利用预训练词嵌入矩阵实现文本词汇位置编码的向量化,为文本处理提供良好基础;借助改进型 BERT 模型强化上下文特征提取,提升对文本语义的理解能力;采用注意力机制加权融合生成语义特征,有效突出关键信息;通过多步骤评估待匹配文本语义一致性,并以预测值超过0.8为阈值实现精准匹配。

1 大规模文本语义匹配方法设计

针对大规模文本上下文特征难捕获导致匹配不准确问题,提出基于改进型BERT预训练模型的大规模文本语义匹配方法,其流程图如图1所示。

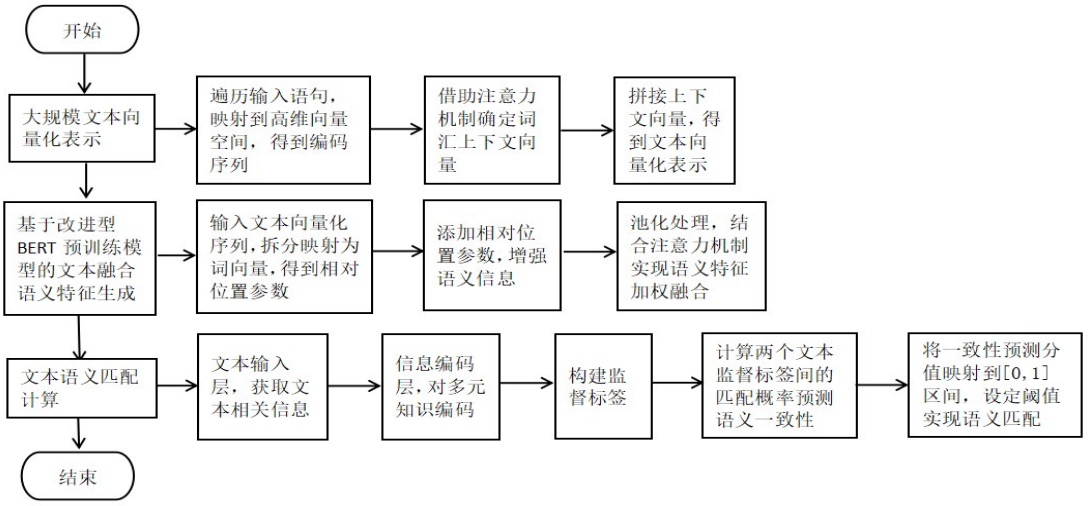


图1 基于改进型BERT预训练模型的大规模文本语义匹配流程

由图1可知,先利用预训练词嵌入矩阵进行位置编码获取文本向量化表示。再用改进型BERT模型增强上下文特征,用注意力机制生成融合语义特征。最后经过一系列步骤计算语义一致性并设阈值实现匹配。

1.1 大规模文本向量化表示

由于原始文本数据一般以非结构化状态存在,直接进行语义匹配运算不仅会增加计算复杂度,还无法保证匹配结果的准确性,因此,在进行匹配计算之前,应对文本数据进行向量转化,即在不改变原始文本语义特征的同时,将原始的文本数据转换为便于处理和计算的数值向量形式,从而使得语义匹配变得更加高效[6]。

遍历所有输入语句,对于其中2个语句$\boldsymbol{X}$和Y,基于位置映射转换思想,将语句映射到高维向量空间中,且保证映射前后语句在空间中的位置一致,得到语句对应的编码序列[7],公式如式(1)所示。

$$\begin{cases} a_X = \sum_{i=1}^{n} b_i h_0 \\ a_Y = \dfrac{\exp(e_o)}{\sum_{i=1}^{n} y_t} \end{cases} \tag{1}$$

式中:n为词汇表中词汇个数,b_i为第i个词汇的词长序列,h_0为嵌入矩阵维度,a_X为语句X的编码序列,e_o为嵌入矩阵阶数,y_t为词汇表中语句序列的最大长度,a_Y为语句Y的编码序列。

为了不破坏文本的原始语义,有利于改进型BERT预训练模型的梯度传播,借助多头注意力机制确定各个词汇对应的上下文向量[8]。设定注意力头的数量为M,对于每个注意力头,计算注意力分数并聚合多头注意力结果。对于单个注意力头,具体计算公式如式(2)所示。

$$\boldsymbol{c}_i = -\frac{1}{N_i}\sum\left[p(x)\log\left(q(x)\right) + \left(a_X - a_Y\right)r(x)\right] \tag{2}$$

式中:N_i为文本中词汇i出现的次数,$p(x)$为索引位置为x时隐藏层的注意力分数,$\boldsymbol{q}(x)$为索引位置x随机失活后的句向量,$\boldsymbol{r}(x)$为标签空间中的概率向量;$\boldsymbol{c}_i$为词汇i的上下文向量。在得到M个注意力头计算的上下文向量$\boldsymbol{c}_i^{\,1}, \boldsymbol{c}_i^{\,2}, \boldsymbol{c}_i^{\,3}, ..., \boldsymbol{c}_i^{\,M}$后,采用平均聚合的方式将多头注意力结果进行聚合,得到词汇i最终的上下文向量$\boldsymbol{C}_i = \dfrac{1}{M}\sum_{i=1}^{M}\boldsymbol{c}_i$。

对各个词汇的上下文向量进行拼接,固定词向

量与编码向量，得到文本向量化表示[9]，表达式如式(3)所示。

$$\boldsymbol{D}=\frac{C_i E_i}{\sqrt{\boldsymbol{\eta}_x \times u_i}} \tag{3}$$

式中：E_i为词汇i在高维映射空间中的维度；$\boldsymbol{\eta}_x$为索引x位置处的偏置矩阵；u_i为词汇i的最大化对数似然函数；$\boldsymbol{D}$为向量序列。

通过向量拼接获取大规模文本向量化表示，便于融合语义特征生成。

1.2 基于改进型BERT预训练模型的文本融合语义特征生成

利用改进型BERT预训练模型通过掩码语言和下一句预测学习文本向量的语言知识和语义信息，最终生成文本融合语义特征。

改进型BERT预训练模型以多层Transformer编码器作为基础网络，首先将文本向量化序列输入改进型BERT预训练模型中，对文本按照字进行拆分并映射为词向量，通过位置编码器描述同一句子内不同位置之间的关系，由此得到语句相对位置参数[10]，如式(4)所示。

$$P_j=\frac{\left(\boldsymbol{v}_i W_e\right)\left(\iota_j W_u+\boldsymbol{d}_t\right)}{\sqrt{\boldsymbol{D}}} \tag{4}$$

式中：$\boldsymbol{v}_i$为词汇i的嵌入向量，W_e为指示函数，ι_j为第j个语句的拆分个数，W_u为转换函数，$\boldsymbol{d}_t$为参数矩阵，$\boldsymbol{D}$为文本向量序列，P_j为语句j的相对位置参数。

在编码器中添加相对位置参数，并通过点积对不同位置的字进行交互，得到绝对位置编码产生的位置向量，进而把位置向量与词向量进行相加来增强语义信息[11]，公式如式(5)所示。

$$\begin{cases} g_{ik}=\operatorname{sgn}\left(\dfrac{\phi_i \phi_k}{\varphi_l}\right) F_b \\ H_j=\dfrac{g_{ik}}{P_j}\boldsymbol{\beta}_t \times \boldsymbol{f}' \\ s=H_j \oplus \boldsymbol{\varpi}_s \end{cases} \tag{5}$$

式中：ϕ_i、ϕ_k分别为词汇i和k的列向量维度，φ_l为缩放因子，sgn(.)为符号运算函数，F_b为非线性映射层，$\boldsymbol{\beta}_t$为一个随机向量，$\boldsymbol{f}'$为权重矩阵，H_j为语句j的全连接参数，$\oplus$为绝对位置向量按位相加符号，$\boldsymbol{\varpi}_s$为激活函数，s为增强后的文本语义信息。

对增强语义信息进行池化处理，结合注意力机制的动态计算获得融合权重，进而实现语义特征的加权融合[12]，表达式如式(6)所示。

$$B=\frac{s \times n}{\left\|\boldsymbol{\varpi}_f\right\| \times \left\|\boldsymbol{\varsigma}_v\right\|} \tag{6}$$

式中：n为词向量映射后的均方误差值；$\boldsymbol{\varpi}_f$为语义特征融合权重；$\boldsymbol{\varsigma}_v$为类别标签嵌入函数；B为融合语义特征。

利用改进型BERT预训练模型增强文本语义特征信息，通过融合权重计算生成融合语义特征，便于语义匹配的计算。

1.3 文本语义匹配计算

本文所提方法通过文本特征信息提取、多元知识编码、融合语义标签生成以及语义匹配关系预测这4个步骤实现，流程如图2所示。

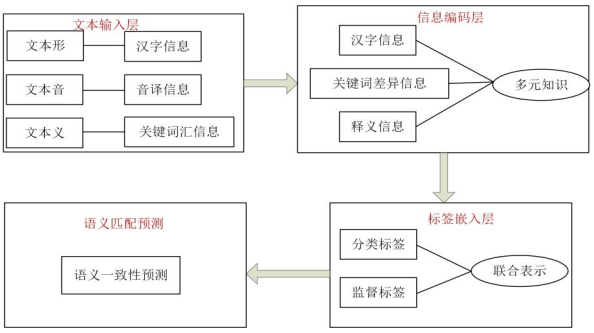


图2　文本语义匹配流程

在文本输入层，从语句的形音义3个角度获取文本的相关信息，并在信息编码层对多元知识进行编码[13]，公式如式(7)所示。

$$U=\frac{\exp\left(\kappa_f\right)}{\sum B \exp\left(\sigma_t\right)} \tag{7}$$

式中：κ_f为输入字符的长度，B为融合语义特征，σ_t为向量范数，U为编码后的多元知识。

基于编码后的分类标签，将标签信息融合至文本特征信息的联合表示，从而构建监督标签[14]，公式如式(8)所示。

$$R_\lambda = \text{soft}\max(Um' + \boldsymbol{L}) \tag{8}$$

式中:m'为可训练参数,$\boldsymbol{L}$为偏置矩阵,R_λ表示文本λ的监督标签。

为了更准确地衡量2个文本之间的相似度,在计算2个文本监督标签之间的匹配概率预测语义一致性时,引入相似度计算公式。设2个文本的监督标签分别为U_1和U_2,它们的相似度K^ξ计算公式如式(9)所示。

$$K^\xi = \frac{U_1 U_2}{\|U_1\| \times \|U_2\|} \tag{9}$$

式中:‖ ‖为向量的范数。

通过计算2个文本监督标签之间的匹配概率预测语义一致性[15],表达式如式(10)所示。

$$y^* = \arg\max \boldsymbol{T}^\gamma \boldsymbol{S}^\gamma \otimes \arg\max K^\xi R_\lambda \tag{10}$$

式中:γ为分类匹配标签,$\boldsymbol{T}$、$\boldsymbol{S}$分别为预测属于和不属于分类匹配类别的概率分布向量,⊗为同或运算,K^ξ为隐含输出的余弦相似度,y^*为2个文本之间语义一致性预测值。

将一致性预测分值映射到[0,1]区间,设定0.8为相似度阈值,即预测值超过0.8时,认为这2个文本语义存在高度一致性,实现语义匹配。

2 实例论证分析

为了验证基于改进型BERT预训练模型的大规模文本语义匹配方法的性能,设计对比实验进行分析。

2.1 实验数据集介绍

实验使用CCKS2018数据集(数据集1)、PKU-Paraphrase-Bank数据集(数据集2)、Chinese-MNLI数据集(数据集3)及OCNLI数据集(数据集4)进行测试;同时,为增强模型在跨领域场景下的验证效果,额外引入法律领域的CAIL2019-Small语义匹配数据集(数据集5)和医学领域的CMeIE语义匹配数据集(数据集6)。其中,CCKS2018数据集是智能客服问答匹配数据,提供了12万对训练文本对和2万测试文本对,每对问题标注为“语义相同”或“语义不同”,其文本序列具有很强的口语性和日常性,专门用于测试模型的鲁棒性;PKU-Paraphrase-Bank数据集包含了8万余条问题匹配数据,且涵盖了多种语言的句子对,支持多语言任务,每条数据的序列长度都超过了20个字符,适合用于垂直领域的语义匹配任务;Chinese-MNLI数据集包含60万个假设对集合,数据语义关系标签标注为释义或非释义,且包含大量复述句子对,适合用于评估模型对复述句子的识别能力;OCNLI数据集中文本对重合度较高,是在不同领域的用户问题中抽取的实际问题构建的,包含句子对及其逻辑关系标注;CAIL2019-Small语义匹配数据集聚焦于法律领域,包含大量法律相关的文本对,如法律条款、案例描述等,可用于评估模型在法律专业文本上的语义匹配能力;CMeIE语义匹配数据集专注于医学领域,涵盖医学术语、疾病描述、诊断建议等医学文本对,有助于验证模型在医学专业场景下的性能。以上6个数据集的统计信息如表1所示。

数据预处理流程如下。

1)去重处理。对所有数据集,使用哈希算法计算每条文本对的唯一标识,通过比较标识来去除重复的文本对,确保数据集中不存在完全相同的样本,避免对模型训练和评估产生干扰。

2)去除特殊字符。使用正则表达式匹配并去除文本中的特殊字符,如标点符号(除中文常用的句号、逗号、问号等外)、特殊符号(@、#、$等)、控制字符等,只保留有效的文本信息。

3)统一文本格式。将所有文本转换为统一的编码格式,并将文本中的全角字符转换为半角字符,避免因字符格式不一致导致模型处理错误。

4)处理空白字符。去除文本开头和结尾的空白字符,并将文本中间的多个连续空白字符替换为单个空格,使文本格式更加规范。

5)分词处理(针对中文数据集)。使用中文分词工具对中文文本进行分词,将连续的中文句子分

表 1　数据集统计信息

数据集	类型	数据规模
CCKS2018	训练集	152 364
	开发集	10 000
	测试集	5 000
PKU-Paraphrase-Bank	训练集	254 689
	开发集	1 569
	测试集	2 236
Chinese-MNLI	训练集	361 245
	开发集	1 000
	测试集	5 589
OCNLI	训练集	330 266
	开发集	1 147
	测试集	2 598
CAIL2019 - Small	训练集	180 000
	开发集	6 000
	测试集	4 000
CMeIE	训练集	150 000
	开发集	5 000
	测试集	3 000

割成单个的词语序列,以便模型更好地理解文本语义。

实验在上述6个数据集上分别进行测试,由此验证本文所提方法在文本语义匹配中的有效性。

2.2 实验准备

为了更直观地了解各个数据集的样本长度分布情况,便于判断是否需要对数据进行预处理,本文对所有数据集的样本序列长度进行统计,结果如图3所示。

由图3可知,6个数据集的样本序列长度呈现出较为集中的态势,基本分布在12~25这一范围内。进一步深入分析可以发现,这6个数据集在样本序列长度的分布形态上具有一定的相似性,这种相似性反映出它们在数据特性方面存在共通之处。而改进型BERT预训练模型对于语料输入长度有着明确要求,即输入长度需为30。鉴于当前6个数据集

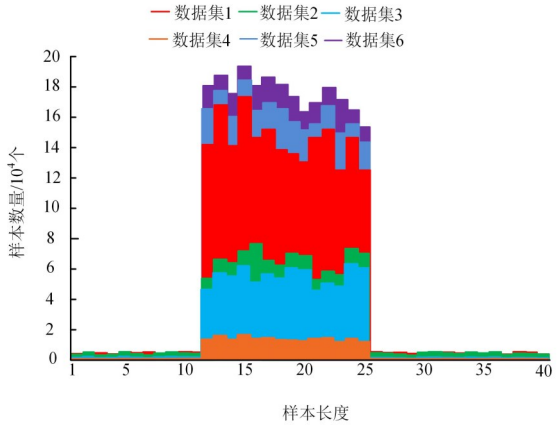


图3　数据集样本长度分布

的样本序列长度均远小于该模型的要求长度,且分布范围相对紧凑合理,不存在过长或过短的极端情况,这表明现有数据集的样本序列长度能够很好地适配改进型BERT预训练模型,无需额外对数据集样本序列进行诸如截断、填充等预处理操作,可直接将其输入模型进行语义匹配计算,以确保后续实验能够基于原始且适配的数据高效开展。

本文所提方法采用的预训练模型是谷歌发布的中文改进型BERT预训练模型,模型参数配置如表2所示。

为体现本文所提方法的适用性,在基于上述6个数据集构建新数据集A和B并开展实验时,训练/测试集的划分遵循以下标准。对于从每个数据集中随机抽取100条句对构成的新数据集,在划分训练集和测试集时,首先不采用分层抽样,直接进行随机划分,以保持数据的随机性和自然分布状态。同时,为保证类别平衡,在划分前先统计抽取句对中“语义相同”和“语义不同”2类样本的数量,在划分训练集和测试集时,按照2类样本在原抽取数据中的比例,分别从2类样本中随机选取相应数量构成训练集和测试集,确保训练集和测试集中“语义相同”和“语义不同”2类样本的比例与原抽取数据基本一致,从而避免因类别不平衡对模型训练和评估结果产生偏差。

在实验的软件和硬件平台方面,硬件上采用配

表2　改进型BERT训练模型参数配置

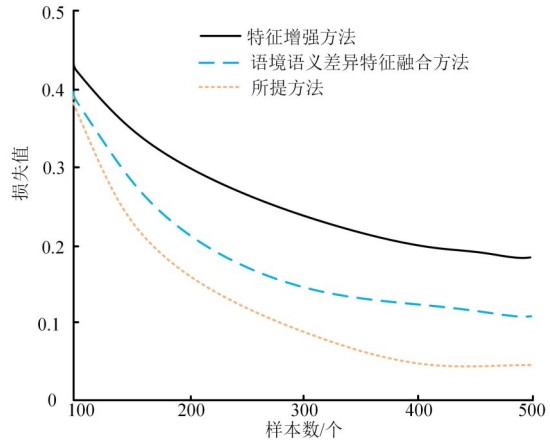
参数	数值	调参过程
批处理大小	32	通过在不同批处理大小(16、32、64)下进行初步实验,发现批处理大小为32时,模型在训练效率和稳定性之间能达到较好平衡。较小的批处理大小会导致训练时间过长且梯度估计不够稳定,较大的批处理大小则可能使模型陷入局部最优解,且对硬件资源要求较高。
学习率	0.03	采用学习率预热策略结合手动调整。初始学习率在0.01、0.03、0.05等值中进行尝试,发现学习率为0.03时,模型在训练初期能较快收敛,同时在后续训练中也能保持较好的性能提升。配合预热步数为10的预热策略,使模型在训练初期以较小的有效学习率开始,逐步增加到设定值,有助于模型稳定训练。
总层数	12	遵循改进型BERT的原始结构设计,该层数在多个自然语言处理任务中已被验证具有较好的特征提取能力,能够捕捉到文本中不同层次的语义信息。
训练轮数	10	最初经过在验证集上的多次实验,当训练轮数为3时,模型在验证集上的部分性能指标有趋于稳定迹象,但整体性能提升仍有空间。进一步将训练轮数增加至10,此时模型在验证集上的各项性能指标(准确率、F1值等)基本达到稳定且较优状态,继续增加训练轮数对模型性能提升不明显,且会增加训练时间和计算资源消耗。
最大序列长度	128	综合考虑文本数据的实际长度分布和计算资源限制。对数据集中的文本长度进行统计分析,发现大部分文本长度在128以内,设置最大序列长度为128可以覆盖大部分文本,同时避免因序列过长导致计算量过大。
隐藏层维度	32	通过在不同隐藏层维度(16、32、64)下进行实验对比,发现隐藏层维度为32时,模型在参数数量和性能表现上达到较好的折中。较小的隐藏层维度可能无法充分提取文本特征,较大的隐藏层维度则会导致模型参数过多,容易过拟合。
注意力头数	12	注意力头数的选择参考了改进型BERT的原始设计及相关研究经验。12个注意力头可以从不同角度关注文本的不同部分,有助于模型捕捉到更丰富的语义信息。
中间层维度	16	在模型设计过程中,通过实验调整中间层维度,发现中间层维度为16时,模型在保证性能的同时,能有效控制参数数量和计算复杂度。
Dropout概率	0.1	通过在不同Dropout概率(0.05、0.1、0.15)下进行实验,发现Dropout概率为0.1时,模型在防止过拟合方面表现较好,能够在训练集和验证集上保持较好的性能一致性。
权重衰减	0.01	权重衰减是一种常用的正则化方法,通过在不同权重衰减值(0.005、0.01、0.02)下进行实验,发现权重衰减值为0.01时,模型在防止过拟合的同时,不会对模型的表达能力产生过大影响。
预热步数	10	结合学习率预热策略,通过实验尝试不同的预热步数(5、10、15),发现预热步数为10时,模型在训练初期能够更平稳地过渡到正常训练阶段,有助于提高模型的收敛性和性能。
分类头层数	2	通过实验对比不同分类头层数(1、2、3)对模型性能的影响,发现分类头层数为2时,模型能够更好地对文本进行分类,在准确率和召回率等指标上表现较优。
词汇表大小	3 000	根据数据集的文本特点和词汇分布,统计出现频率较高的词汇,构建大小为3 000的词汇表。该词汇表能够覆盖数据集中大部分常见词汇,同时避免词汇表过大导致计算资源浪费和模型复杂度增加。

置为 Intel Xeon Platinum 8280L 的 CPU,该 CPU 具备多核心、高主频的特点,能够高效处理多线程任务,为数据预处理和模型推理过程中的计算任务提供强大的支持;同时配备 NVIDIA Tesla V100 SXM2 32GB 的 GPU,其拥有大量的 CUDA 核心和高速显存,可显著加速深度学习模型的训练过程,大幅缩短训练时间。软件上,选用 PyTorch 作为深度学习框架,PyTorch 凭借其动态计算图特性,在模型开发和调试过程中具有更高的灵活性和便捷性,能够更高效地实现改进型 BERT 预训练模型以及文本语义匹配算法的搭建与训练。同时,结合 Python 科学计算库 NumPy、数据处理库 Pandas 及可视化库 Mat-

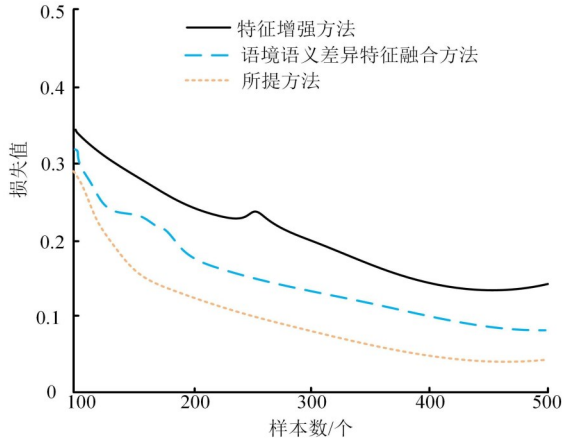
plotlib 等工具,完成数据的预处理、结果分析以及可视化展示等工作。

2.3 实验结果

引入文献[3]基于特征增强方法和文献[4]融合语境语义差异特征方法作为本文所提方法的对比方法,分别采用3种方法对数据集A和数据集B进行文本语义匹配,采用损失值作为评估指标,损失值越低,表明预测的文本语义一致性越高,匹配精度越高,实验结果如图4所示。



(a)数据集A



(b)数据集B

图4　不同数据集上文本语义匹配损失值对比

由图4可知,不论是在数据集A还是数据集B上,本文所提方法匹配结果的损失值均显著低于2种对比方法,说明本文所提方法通过大规模无监督学习,能够准确捕捉文本的上下文语义,可以处理隐含语义和多模态匹配任务,使得快速准确地找到语义一致性的2个文本进行匹配,保证了匹配精度。而文献[3]的特征增强方法可能仅从局部特征进行强化,未能全面考虑文本整体的语义关联,在处理复杂语义结构时难以有效整合信息。文献[4]融合语境语义差异特征方法虽考虑了语境差异,但对语序的敏感性不足,且在面对多模态语义时,无法充分挖掘不同模态间的语义关联,对复杂语义特征提取能力不足,导致匹配结果不理想。通过实验结果可以证明本文所提方法的实际应用可靠性。

采用平均倒数排名(mean reciprocal rank,MRR)指标对不同方法的匹配精度进一步测试,MRR值越高,表明方法能够正确匹配大部分文本,匹配效果越好,实验结果如图5所示。

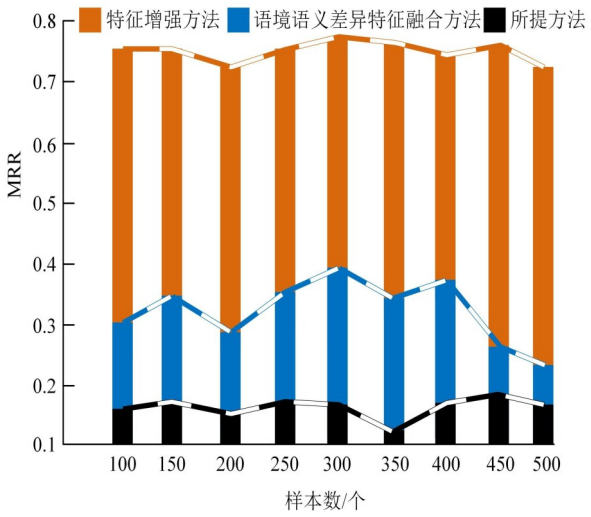


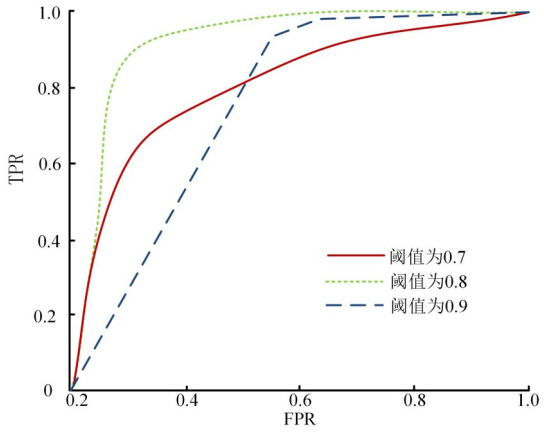
图5　MRR对比结果

由图5可知,在100~500个匹配样本中,本文所提方法的平均倒数排名MRR远高于2种对比方法,始终处于0.7以上,说明本文所提方法凭借其大规模无监督学习机制,能深度挖掘文本语义信息,在处理各类文本时,可精准把握语义关键特征,从而在匹配过程中展现出卓越性能。而文献[3]的特征增强方法,由于过度聚焦局部特征,在整体语义匹配的宏观把控上存在欠缺,难以在复杂多变的文本匹配场景中实现精准匹配。文献[4]融合语境语义差异特征方法,虽考虑到语境因素,但在应对大规模样本和复杂语义关系时,其语义特征提取和整合

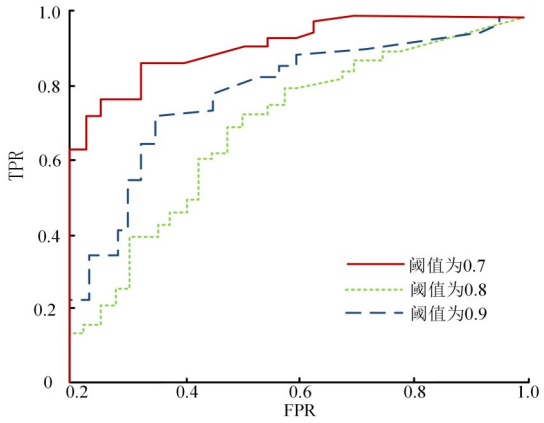
能力有限,无法有效提升匹配精度。这说明本文所提方法可以实现高精度的文本语义匹配任务,能够满足应用需求。

为了更全面地评估所提方法的性能,实验设定阈值分别为0.7、0.8、0.9,通过绘制受试者工作特征曲线(receiver operating characteristic curve,ROC)并计算曲线下面积(area under the curve,AUC)值,以分析阈值敏感性。ROC曲线展示了在不同阈值下,模型的真阳性率(true positive rate,TPR)与假阳性率(false positive rate,FPR)之间的关系。AUC值则反映了模型区分正负样本的能力,AUC值越接近1,模型的性能越好。实验结果如图6所示。

由图6可知,在不同数据集上,随着阈值从0.7增加到0.9时,ROC曲线的走势有所变化。在数据集A和B中,阈值为0.8时曲线更靠近左上角,表明在较低阈值下模型能获得相对较高的真阳性率和较低的假阳性率;随着阈值升高到0.9,曲线逐渐向右下方移动,意味着在提高阈值时,真阳性率有所下降或者假阳性率有所上升。通过计算AUC值可以进一步量化模型性能,阈值为0.8时AUC值更接近1,说明模型在阈值为0.8时区分正负样本的能力更强,但整体来看,不同阈值下的AUC值都应处于一定合理范围,反映出模型在不同阈值设定下具有一定的性能稳定性,但也存在阈值敏感性,阈值的选择会对模型的TPR值和FPR值产生明显影响。



(a)数据集A



(b)数据集B

图6 不同数据集上不同阈值的ROC曲线

3 结束语

利用改进型BERT模型强大的特征提取能力对文本进行深度编码并捕获语义信息,结合语义一致性预测完成文本语义匹配任务,该方法为大规模文本语义匹配提供了有效的解决方案,推动了自然语言处理技术的发展,并为后续研究提供了新思路。未来,将改进型BERT与其他前沿模型进行集成,以利用不同模型的优势,进一步提高语义匹配效果,并结合领域知识来增强模型在特定场景中的匹配能力。具体而言,需进一步探讨本文模型与GPT、Transformer-XL等前沿模型在模型结构融合、参数共享、训练策略协同等方面的集成可能性,挖掘不同模型在语义理解、长距离依赖捕捉等能力的互补优势,以实现更优的文本语义匹配性能。

参考文献:

[1] YANG L, FENG Y, ZHOU M L ,et al. .Multi-level network based on transformer encoder for fine-grained image-text matching [J].Multimedia Systems,2023,29:1981-1994.

[2] 赵云肖,李茹,李欣杰,等.基于汉字形音义多元知识和标签嵌入的文本语义匹配模型[J].中文信息学报,2024,38(3):42-55.

(下转第102页)