文章编号:1671-6833(2024)06-0056-09

面向中文科学数据集的句子级语义匹配模型

刘建平[1,2], 初新涛[1], 王 健[3], 顾勋勋[1], 王 萌[1], 王影菲[1]

(1.北方民族大学 计算机科学与工程学院,宁夏 银川 750021;2.北方民族大学 图像图形智能处理国家民委重点实验室,宁夏 银川 750021;3.中国农业科学院 农业信息研究所,北京 100081)

摘 要:针对现有以词为粒度的语义匹配模型难以理解句子级科学数据集元数据的问题,提出了一个面向中文科学数据集的句子级语义匹配(CSDSM)模型。该模型使用 CSL 数据集对 SimCSE 和 CoSENT 进行训练生成 CoSENT 预训练模型。基于 CoSENT 模型,引入多头自注意力机制进行特征提取,通过余弦相似度与 KNN 分类结果加权求和得到最终输出。以国家地球系统科学数据中心开放的语义元数据信息作为自建科学数据集进行实验,实验结果表明:与中文 BERT 模型相比,所提模型在公共数据集 AFQMC、LCQMC、Chinese-STS-B 和 PAWS-X 上的 Spearman 指标 ρ 分别提升了 0.044 8,0.029 0,0.177 7 和 0.050 9;在自建科学数据集上的 $F1$ 和 Acc 分别提升了 0.078 8 和 0.063 4,所提模型能够有效地解决科学数据集句子级语义匹配问题。

关键词:文本匹配; 语义匹配; 预训练模型; 科学数据集; 自然语言处理

中图分类号: TP3-05;TP391.1 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2024.03.008

在数据驱动的科学研究中,科学数据作为重要的证据性材料,其基础作用日益凸显。开放共享环境下,科学数据集主要存储于各大数据仓储、数据集共享平台和政府开放数据网站中[1]。在开放科学运动驱动下,传统的关键词匹配方法已无法满足科研人员快速、准确获得相关数据集的需求[2]。语义匹配任务可以简单地描述为确定句子对是否语义等价,广泛应用于信息检索、机器翻译、问答、推荐系统等领域,能够帮助用户快速捕捉所需信息[3]。语义匹配技术的发展与自然语言处理(natural language processing, NLP)的发展一脉相承,自 Transformer 框架提出后,预训练模型开始迅速发展[4]。采用预训练模型对下游任务进行微调是目前语义匹配的主流方法之一。

科学数据集的语义信息主要由其元数据进行表达,如标题、关键词和描述等。因此,从任务类型角度,科学数据集元数据语义匹配任务属于一类经典的 NLP 匹配任务,但是,相比于传统的文本匹配,科学数据集具有其独特的任务特点,主要问题包括:①现有的预训练模型在对文本进行表示时主要基于词级别,对科学数据集句子级匹配模型鲜有研究;②中英文混合文本中具有高度重叠但语义不相似的文本较多。

针对以上问题,本文提出了一个面向中文科学数据集的句子级语义匹配模型(chinese scientific dataset sentence-level match,CSDSM)模型。首先,使用 CSL(chinese scientific literature dataset)[5]数据集对 SimCSE[6] 和 CoSENT(cosine sentence)进行训练,生成 CoSENT 预训练模型。基于 CoSENT 模型,引入多头自注意力机制[4]进行特征提取,通过余弦相似度与 KNN 分类结果加权求和得到最终输出。其次,使用网络采集技术收集了国家地球系统科学数据中心的标题信息,通过多人标注和投票机制,构建了一个超过 20 000 条数据的中文科学数据集。在公共数据集和自建数据集上的实验结果表明,在语义匹配任务中 CSDSM 模型比现有模型性能更好。

1 相关工作

1.1 传统语义匹配技术

计算语义相似度是语义匹配模型中的一项关键任务。早期方法是从文本对象中提取关键词,构建

收稿日期:2023-10-02;**修订日期**:2023-11-13

基金项目:宁夏回族自治区重点研发计划(2022BSB03044);宁夏回族自治区自然科学基金资助项目(2021AAC03205);北方民族大学学科研启动金项目(2020KYQD37)

作者简介:刘建平(1989—),男,宁夏固原人,北方民族大学讲师,博士,主要从事智能信息检索与推荐的研究,E-mail:liujianping01@nmu.edu.cn。

引用本文:刘建平,初新涛,王健,等. 面向中文科学数据集的句子级语义匹配模型[J]. 郑州大学学报(工学版),2024,45(6):56-64.(LIU J P, CHU X T, WANG J, et al. Semantic matching model for chinese scientific datasets[J]. Journal of Zhengzhou University (Engineering Science),2024,45(6):56-64.)

向量空间模型,然后通过余弦相似度等方法计算相似性。传统基于词的特征关键词匹配模型无法解决语义问题[7]。随着深度学习的发展,通过构建大型语料库的方式,为挖掘上下文语境信息提供了材料和方法支持。从最早的深度神经网络(DNN)使用前馈神经网络映射文本序列,再到卷积神经网络(CNN)在固定大小的滑动窗口中共享参数[8],为了进一步获取文本序列的长期依赖性,出现了循环神经网络(RNN)。

早期工作受到孪生网络架构的启发,分配各自的神经网络将 2 个输入序列编码为高级表示。如 Huang 等[9]使用 DNN 将查询和标题表示为低维语义向量。Shen 等[10]通过使用 CNN 从文本中提取局部上下文文本表示。Mohan 等[11]利用 RNN 及其变体来捕获上下文依赖。这些传统神经网络在面对新的数据时,其泛化能力有限。这意味着它们可能会在未知数据上表现不佳,需要重新训练或进行调整以适应新的数据分布。近来,学者们大量的研究工作表明,在大型语料库上预训练模型可以学习通用语言表示,这有利于 NLP 的下游任务,并且可以避免从头开始训练新模型[12]。

1.2　基于预训练模型的语义匹配技术

由于注释成本昂贵,为 NLP 任务构建大规模标签数据集是一个巨大的挑战,而大规模无标签语料库相对容易构建。利用这些语料库,模型可以从中学习良好的表示,将这些表示应用于其他任务。早期方法是将单词表示为密集向量进行单词嵌入。Word2vec[13]是典型代表,它通过预训练单词嵌入用于不同的 NLP 任务。由于多数 NLP 任务超出了单词级别,因此出现了句子等更高级别的预训练神经编码器[14]。预训练模型在学习通用语言表示方面具有强大能力,如 BERT、GPT、XLNet、T5 等[15]。

以 BERT 为代表的预训练语言模型广泛应用于语义匹配任务中。这些模型先在大型语料库上进行预训练,然后对特定的任务进行微调。Choudhary 等[16]通过 BERT 获取丰富的语义信息。Esteva 等[17]通过 BERT 模型创建大量的元组(引文、标题和段落)用于训练匹配模块。虽然 BERT 等通用预训练模型已广泛应用于各个领域,但面对科学数据集领域时,通用预训练模型缺乏对特定科学领域的深入学习,无法充分利用该领域的专业知识,尤其是在需要深入理解领域特定概念和关系的科学任务中,这限制了模型的使用范围。

1.3　科学数据领域的预训练模型

随着预训练模型的发展,已出现一些针对科学领域的预训练语言模型。其中,SciBERT[18]是一种专门为科学领域设计的模型。SciBERT 在大量科学文献数据上进行预训练,相较于通用的 BERT 模型,SciBERT 在预训练过程中采用了一些特殊的处理方式。首先,它使用了与 BERT 相同的基础架构——Transformer 编码器。其次,SciBERT 使用了领域特定的词汇表,以便覆盖科学领域的术语和专有名词。除此之外,还有一些其他领域的模型,如 BioBERT、ClinicalBERT、BEHRT 以及 MT-BERT 等[19],它们在不同的科学领域中具有特定的应用。

以上模型均采用基于 BERT 结构的方法,主要应用于英文科学数据集任务,对中文数据的适应性较差。此外,由于原始 BERT 模型主要关注词级上下文的表示,对句子级语义理解的能力有限,可能导致部分信息的丢失。同时,针对科学数据集的匹配任务,实现句子级别的交互较为困难。因此,针对中文科学数据集的句子级语义匹配问题,本文使用 CSL 数据集对 SimCSE 和 CoSENT 进行训练生成预训练模型;在预训练模型的基础上,利用 KNN 算法以及多头自注意力机制训练获得了 CSDSM 模型。

2　CSDSM 模型

本文 CSDSM 模型的具体结构如图 1 所示。

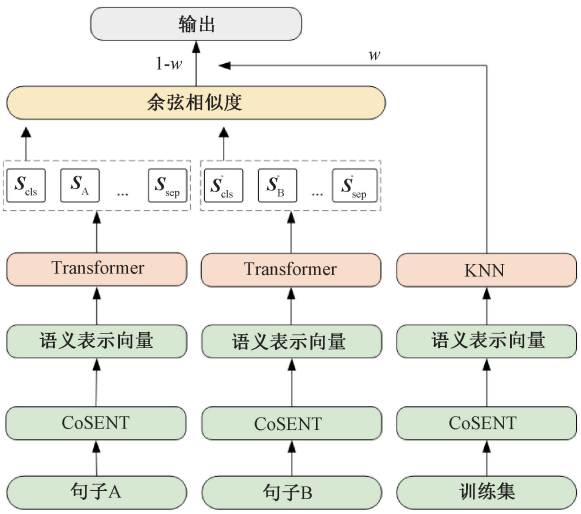

图 1　CSDSM 模型

Figure 1　CSDSM model

模型采用 Sentence-BERT 模型[20]作为基础方法,预训练 CoSENT 模型嵌入科学数据集句子对,获得句子 A 和句子 B 的向量,分别为 $\boldsymbol{S}_{\mathrm{A}}=\{\boldsymbol{s}_{\mathrm{A}}^{1},\boldsymbol{s}_{\mathrm{A}}^{2},\cdots,\boldsymbol{s}_{\mathrm{A}}^{m}\}$ 和 $\boldsymbol{S}_{\mathrm{B}}=\{\boldsymbol{s}_{\mathrm{B}}^{1},\boldsymbol{s}_{\mathrm{B}}^{2},\cdots,\boldsymbol{s}_{\mathrm{B}}^{n}\}$,其中 $\boldsymbol{s}_{\mathrm{A}}^{m}$ 和 $\boldsymbol{s}_{\mathrm{B}}^{n}$ 分别表示第 m 个和第 n 个句子;送入多头自注意力机制模块,提取句子的 CLS token 向量,CLS token 通常用于表示整个输入序列的句子级别的语义信息,最后使

用余弦相似度和 KNN 分类结果加权求和得到最终输出。

2.1 CoSENT 函数

Sentence-BERT 模型如图 2 所示,将 2 个文本向量表示进行拼接,并乘上一个可训练的权重和偏置,用于优化分类任务,具体公式为

$$\boldsymbol{S}_{\text{sent}} = \text{ReLU}(\boldsymbol{a}_{\text{out}} \cdot [\boldsymbol{u};\boldsymbol{v}] + \boldsymbol{b}_{\text{out}})。 \tag{1}$$

式中:$\boldsymbol{u}$、$\boldsymbol{v}$ 分别表示 2 个句子的 BERT 嵌入;$[\boldsymbol{u};\boldsymbol{v}]$ 表示 2 个向量拼接;$\boldsymbol{a}_{\text{out}}$ 和 $\boldsymbol{b}_{\text{out}}$ 分别为模型可训练的权重向量和偏置向量;ReLU 表示激活函数,当输入小于 0 时,ReLU 函数的输出为 0,这意味着一些神经元的激活值会被抑制为 0。这种稀疏性有助于减少特征空间的维度,提取更加重要和有区别性的特征,并降低模型的复杂度,其表达式为 $\text{ReLU}(\boldsymbol{x}) = \max(0,\boldsymbol{x})$,其中 $\boldsymbol{x}$ 表示输入。

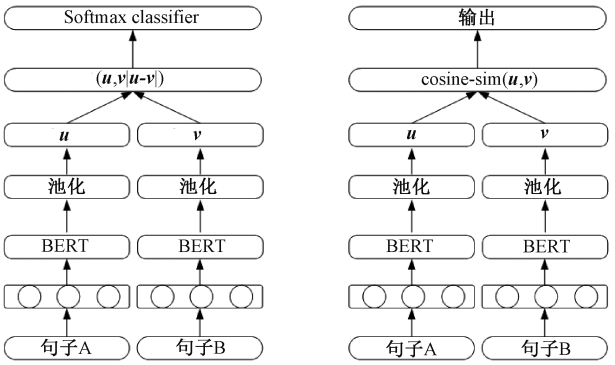

图 2 Sentence-BERT 模型

Figure 2 Sentence-BERT model

然而,模型的预测阶段使用余弦相似度进行预测,导致模型的训练阶段和预测阶段不一致。此外,对于标注样本,损失函数的目标对于正样本尽可能大,对于负样本则尽可能小,导致模型失去泛化能力或难以优化,模型的损失函数为

$$t \cdot (1 - \cos(\boldsymbol{u},\boldsymbol{v})) + (1 - t) \cdot (1 + \cos(\boldsymbol{u},\boldsymbol{v})),\ t \in \{0,1\}。 \tag{2}$$

式中:$t=1$ 表示相似,$t=0$ 表示不相似。

为解决上述问题,本文将损失函数改为 CoSENT 损失函数:

$$\log\left(1 + \sum_{(i,j)\in\Omega_{\text{pos}},(k,l)\in\Omega_{\text{neg}}} e^{\lambda(\cos(\boldsymbol{u}_k,\boldsymbol{u}_l)-\cos(\boldsymbol{u}_i,\boldsymbol{u}_j))}\right)。 \tag{3}$$

式中:Ω_{pos} 为正样本对集合;Ω_{neg} 为负样本对集合;$\boldsymbol{u}_i$ 和 $\boldsymbol{u}_j$ 均为正样本对的句向量;$\boldsymbol{u}_k$ 和 $\boldsymbol{u}_l$ 均为负样本对的句向量;λ 表示缩放因子,为了避免某些特征对模型的影响过大,通过缩放因子将其缩放到相似的范围,本文 λ 取值为 20,该值参考 CoSENT 源码及其实验。

GoSENT 函数是一种有监督损失函数,它直接优化余弦相似度和标签的差异,而不是使用传统的交叉熵或均方误差等损失函数,如图 3 所示。CoSENT 损失函数考虑了所有正样本对和负样本对的余弦值的差异,并确保正样本对的相似度大于负样本对的相似度。以使模型的训练阶段和推理阶段的任务保持一致,提高模型语义匹配的能力,具体公式为

$$\cos(\boldsymbol{u}_i,\boldsymbol{u}_j) > \cos(\boldsymbol{u}_k,\boldsymbol{u}_l)。 \tag{4}$$

式中:$\cos(\cdot)$ 表示相似度。

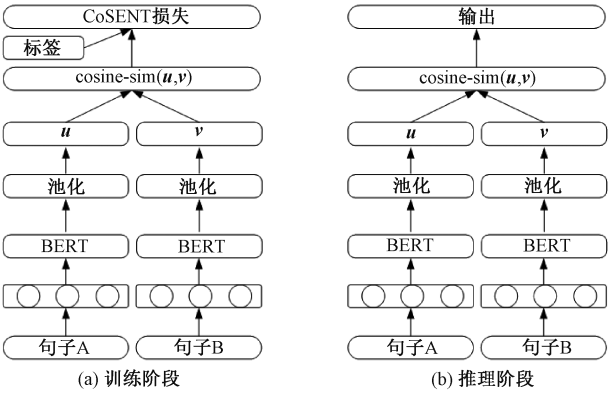

图 3 CoSENT 模型

Figure 3 CoSENT model

2.2 SimCSE 对比学习框架

对比学习在自监督学习中被广泛应用,它能够利用未标记的数据集来增强潜在的语义表示,通过学习样本之间的相似性来增强语义表示,具体公式为

$$D = \{(\boldsymbol{x}_i,\boldsymbol{x}_i^+)\}_{i=1}^m。 \tag{5}$$

式中:$\boldsymbol{x}_i$ 和 $\boldsymbol{x}_i^+$ 为 2 个语义相关的句子向量。

具体来说,首先将样本通过一个编码器映射到潜在的语义空间中,然后通过度量样本之间的距离来衡量它们在语义空间中的相似度。在对比学习中,通过拉近正样本的距离并拉远负样本的距离来增强语义表示的区分度。

在传统的 Transformer 中,全连接层放置了 dropout 掩码(默认概率值 P_{d} 为 0.1),把它表示为 $F_i^z = f_\theta(\boldsymbol{x}_i,z)$,其中 dropout 代表以一定概率 P_{d} 失活部分神经元,z 代表随机 dropout 掩码。在无监督 SimCSE 方法中,将相同的输入送入编码器 2 次,可以得到 2 个具有不同 dropout 掩码的嵌入,由此得到无监督 SimCSE 的训练目标,如式(6)所示。通过预训练模型编码,将句子的余弦相似度作为相似性得分,根据式(6),通过对数函数的运算,使得正样本对的相似性得分尽可能大,负样本对的相似性得分尽可能小。在训练中,通过最小化损失函数值,模型

学习到更好的句子表示,使得相似的句子在嵌入空间中更加接近,而不相似的句子则更加远离,从而达到无监督对比学习的目标。

$$l_i = -\log \frac{e^{\frac{sim(\boldsymbol{h}_i^{z_i},\boldsymbol{h}_i^{z_i^+})}{t}}}{\sum_{j=1}^{N} e^{\frac{sim(\boldsymbol{h}_i^{z_i},\boldsymbol{h}_j^{z_j^+})}{t}}}。\tag{6}$$

式中:t 为温度系数控制 softmax 分布的一个超参数;$sim(\cdot)$ 代表相似度;N 代表 batch size 的大小;$\boldsymbol{h}_i^{z_i^+}$ 代表 $\boldsymbol{h}_i^{z_i}$ 在 dropout 后的样本;l_i 代表对比损失函数的第 i 个样本的损失值。

为了在科学数据集领域应用性能更好,本文使用 CSL 科学数据集来训练 SimCSE 和 CoSENT 模型。首先,采用预训练模型 Chinese-Roberta-wwm-ext 对 SimCSE 进行了无监督训练,以增强领域知识;其次,将训练好的 SimCSE 无监督模型输入到 CoSENT 中进行有监督训练,以进一步优化模型性能;最后,得到针对科学数据集领域的 CoSENT 模型,如图 4 所示。

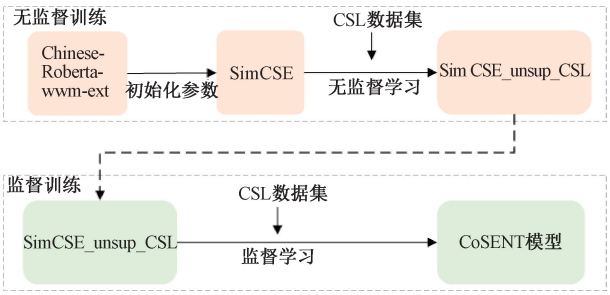

图 4　CoSENT 模型训练过程

Figure 4　CoSENT model training process

2.3　多头自注意力机制

本文使用 Transformer 中的多头自注意力机制对输入的句子对进行建模。这种机制可以在一个句子中捕捉到不同词语之间的关联,并且能够在多个句子之间进行信息交互。通过多头自注意力机制,能够有效地捕捉到句子之间的语义关系,从而提高模型的准确度和表现力。具体地,对上述 2 个句子向量 $\boldsymbol{S}_A \in \mathbf{R}^d$ 和 $\boldsymbol{S}_B \in \mathbf{R}^d$ 进行拼接得到 $\boldsymbol{x} = [\boldsymbol{S}_A : \boldsymbol{S}_B] \in \mathbf{R}^{2d}$,通过多头自注意力机制计算向量中每个词和其他词之间的相似度得分,得到一个向量 $\boldsymbol{s} \in \mathbf{R}^d$,其中向量 $\boldsymbol{s}_i$ 表示向量 $\boldsymbol{x}$ 中第 i 个词和其他词之间的相似度得分,具体公式如下:

$$\boldsymbol{s}_i = \text{softmax}\left(\frac{\boldsymbol{x}\boldsymbol{W}_1(\boldsymbol{x}\boldsymbol{W}_2)^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{x}\boldsymbol{W}_3;\tag{7}$$

$$head_i = \boldsymbol{s}(\boldsymbol{x}\boldsymbol{W}_1, \boldsymbol{x}\boldsymbol{W}_2, \boldsymbol{x}\boldsymbol{W}_3);\tag{8}$$

$$O = \text{concat}(head_1, head_2, \cdots, head_h)\boldsymbol{W}^o。\tag{9}$$

式中:$\boldsymbol{W}_1 \in \mathbf{R}^{d\times h}$、$\boldsymbol{W}_2 \in \mathbf{R}^{d\times h}$ 和 $\boldsymbol{W}_3 \in \mathbf{R}^{d\times h}$ 为线性变换矩阵;h 表示每个头部的向量维度,通过 softmax 函数将结果映射到 [0,1] 的概率分布空间中;$\boldsymbol{x}\boldsymbol{W}_1$ 和 $\boldsymbol{x}\boldsymbol{W}_2$ 分别为输入向量 $\boldsymbol{x}$ 经过线性变换得到的矩阵;$\sqrt{d}$ 为标量;$\boldsymbol{x}\boldsymbol{W}_3$ 表示对得到的相似度得分加权并进行线性变换得到最终的向量;$head_i$ 代表第 i 个自注意力层;$\boldsymbol{W}^o$ 为输出矩阵,用于将多个注意力头输出映射到向量空间中;concat(·) 表示将多个注意力头的输出结果按维度拼接;O 表示多头自注意力的输出结果。

2.4　KNN 算法

为进一步提高语义匹配准确性,本文在模型中引入了 KNN 方法。该方法将有监督的科学数据集输入到 KNN 算法中,使用余弦相似度来度量最近邻数量 K,通过将 KNN 分类结果与模型预测的余弦相似度结果相结合来提高语义匹配的准确性,如图 5 所示,具体公式如下:

$$\cos\theta = \frac{\boldsymbol{S}_A \cdot \boldsymbol{S}_B}{\|\boldsymbol{S}_A\| \quad \|\boldsymbol{S}_B\|};\tag{10}$$

$$P = (1-w)\cos(\boldsymbol{S}_A, \boldsymbol{S}_B) + wK_{NN}(\boldsymbol{x})。\tag{11}$$

式中:P 为概率;w 为权重;$\boldsymbol{S}_A$ 和 $\boldsymbol{S}_B$ 为输入句子的向量表示;$K_{NN}(\boldsymbol{x})$ 为待分类句子 $\boldsymbol{x}$ 利用 KNN 算法输出分类概率值。

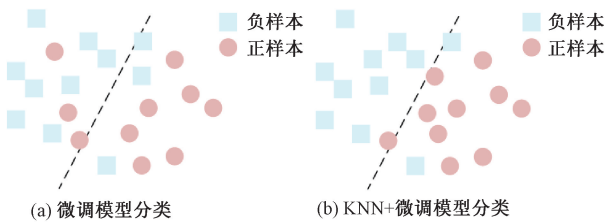

图 5　KNN 算法辅助模型分类结果

Figure 5　Classification results of the model aided by KNN algorithm

2.5　余弦相似度

多头自注意力机制输出 CLS token 向量空间来表示句子的语义信息。该向量空间模型将句子映射到一个多维向量空间中,并通过余弦相似度来计算 2 个句子之间的相似度。

具体地,本文使用预训练的语言模型来计算每个句子的嵌入向量并通过多头自注意力机制模块输出 CLS token 向量。向量表示句子在向量空间中的位置,其在空间中的距离可以用来表示它们之间的相似度。模型通过余弦相似度输出结果。余弦相似度是一种用于比较 2 个向量之间的相似度的度量方法,其取值在 -1 到 1 之间。2 个向量越相似,其余弦相似度越接近 1,反之亦然。

3 实验

3.1 实验环境及参数

实验中使用操作系统为 Linux,GPU 为 NVIDIA Tesla V100,CUDA 为 10.1。深度学习框架为 Pytorch1.7.1,Adam optimizer 优化网络。在模型的初始训练中进行数据集长度×0.05 个步长的 warm-up 训练,文本向量输出 CLS。本文进行了大量实验,研究了 K 值(例如 3,5,7 等)以及相似度阈值(从 0.6 至 0.8)。通过平衡模型性能和时间消耗,实验将 K 设置为 3,预测相似度的阈值设定为 0.79。模型的主要超参数如表 1 所示。其中参数 w 的值参考文献[21],初始值设定为 0.5,详细实验结果如表 2 所示。实验结果表明,将权重值设为 0.29 可以在此实验中达到最佳的效果。为了确保实验结果的一致性,后续实验中权重值不变。

表 1 超参数设置

Table 1 Hyperparameter of CSDSM

参数	取值	参数	取值
学习率	0.000 02	*embedding size*	768
N	32	P_d	0.1
epoch	10	*head number*	8
Transformer 块	1	w	0.29

表 2 不同权重在科学数据集上的实验结果比较

Table 2 Comparison of experimental results on scientific datasets with different weights

w	$F1$	Acc	w	$F1$	Acc
0.25	0.919 1	0.889 6	0.29	0.924 9	0.899 5
0.26	0.919 7	0.892 1	0.30	0.924 1	0.896 3
0.27	0.922 6	0.894 2	0.50	0.896 3	0.871 3
0.28	0.925 8	0.896 1	0.70	0.845 5	0.835 1

3.2 数据集

目前公开的科学领域的数据集相对有限,为了丰富科学数据集,本文采用 Python 中的 Requests 库和 Scrapy 框架从国家地球系统科学数据中心采集数据集的标题等元数据超过 30 000 条。使用数据清洗、去重等手段对数据进行预处理。人工标注构建相似语义句子对并收集术语添加到模型词表中。相似度标签通过多人标注,利用投票机制取平均结果。标签分为 2 个等级:相似为 1,不相似为 0,并按照 8∶2 的比例构建训练集和测试集(简称科学数据集)。其中训练集 8 607 对,测试集 2 353 对,样例见表 3。

中文科学文献数据集(CSL)包含 396 209 篇中文核心期刊论文的元数据(标题、摘要、关键词、学科、门类)。数据源自国家科技资源共享服务工程技术研究中心。

表 3 部分科学数据集样例

Table 3 Some examples of scientific datasets

序号	句子 A	句子 B	标签
1	吉鄂湘粤基层调查联系点分地区流动人口就业收支与居住特征数据	吉鄂湘粤基层调查联系点分地区流动人口流动特征数据	0
2	基于“两叶模型”的贵州省 500 m 分辨率 8 天 LAI 数据集	基于“两叶模型”的贵州省 500 m 分辨率 8 天 APAR 数据集	0
3	新疆维吾尔自治区精河县地震区域公里格网人口密度数据	四川省九寨沟地震区域公里格网人口密度数据	1
4	全球未来 30 m 土地利用预测数据集 FROM-GLC-Simulation_6.0	中国未来 30 m 土地利用预测数据集 FROM-GLC-Simulation_6.0	1
5	全球未来土地利用预测数据集 FROM-GLC-Simulation2.6	全球 250 m 土地覆盖数据集 FROM-GLC-Hierarchy	0

扩展实验中使用公开的中文语义匹配数据集,包括 AFQMC(蚂蚁金融语义相似度数据集)、LCQMC(哈尔滨工业大学语义匹配数据集)、Chinese-STS-B(人工翻译修正的中文 STS-B 语义匹配数据集)以及 PAWS-X(谷歌同义句识别数据集)。

3.3 评价指标

语义匹配任务的评价指标是判断输入句子对的语义是否等价。本文采用该领域常用的 3 个评价指标:$F1$、Acc 和 Spearman 相关系数来评价模型性能。评价指标数值越大,模型性能越好。

精确率 Pre[22]:

$$Pre = \frac{TP}{TP + FP}。 \tag{12}$$

式中:TP 代表样本为真预测为真的数量;FN 代表样本为真预测为假的数量;FP 代表为样本为假预测为真的数量;TN 代表样本为假预测为假的数量。

召回率 Rec[22]:

$$Rec = \frac{TP}{TP + FN}。 \tag{13}$$

$F1$ 为综合考虑精确率和召回率的指标[22]:

$$F1 = \frac{2Pre \cdot Rec}{Pre + Rec}。\tag{14}$$

准确率 Acc[22]:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}。\tag{15}$$

Spearman 相关系数 ρ[23]:

$$\rho = 1 - \frac{6\sum d_i}{n(n^2 - 1)}。\tag{16}$$

式中:d_i 表示第 i 个数据对的等级差;n 为总的观测样本数。

3.4　基准模型

目前,预训练-微调方法已成为语义匹配领域的主流方法之一。为了验证本文所提出的方法在科学数据集语义匹配中的有效性,本文选择了一些代表性的预训练模型进行微调,并比较它们的性能表现。

BERT[24]:中文 BERT-base-Chinese 预训练模型。

RoBERTa[25]:基于全词掩码的 Chinese-Roberta-wwm-ext 预训练模型。PERT[26]:一种基于乱序语言模型的预训练模型。

LERT[27]:融合了多种语言学知识的语言学信息增强的预训练模型。

4　结果分析

为了验证不同模块在科学数据集语义匹配的作用,在科学数据集上进行了消融实验,结果如表 4 所示。

表 4　科学数据集上的消融实验结果比较

Table 4　Comparison of ablation experimental results on scientific datasets

方法	$F1$	Acc
BERT	0.846 1	0.836 1
SimCSE	0.869 1	0.847 9
SimCSE+CoSENT	0.880 7	0.858 4
SimCSE+CoSENT+muatt	0.884 1	0.859 0
SimCSE+CoSENT+muatt+KNN	0.924 9	0.899 5

注:SimCSE 表示预训练模型为 SimCSE_unsup_CSL 模型;muatt 代表多头自注意力机制。

由表 4 可知,在模型中使用 SimCSE 后 $F1$ 提升了 0.021 8,说明具有领域知识的预训练模型能够理解科学数据语义;使用 CoSENT 函数后 $F1$ 提升了 0.034 6,训练和预测阶段任务一致使得模型性能进一步提升;加入多头自注意力机制,模型 $F1$ 提升了 0.003 8,将所有方法加入后,$F1$ 提升了 0.078 8,此时 $F1$ 达到了 0.924 9。具有科学领域知识的预训练模型能够更准确地理解和捕捉科学领域的特定语义,从而更好地完成科学数据集上的语义匹配任务。

为了体现公平性,本文在科学数据集上微调了基线模型,对比实验结果如表 5 所示。

表 5　不同模型在科学数据集上语义匹配结果比较

Table 5　Comparison of semantic matching results of different models on scientific datasets

模型	$F1$	Acc	耗时/min
BERT	0.846 1	0.836 1	20
RoBERTa	0.867 9	0.846 1	12
PERT	0.887 3	0.889 0	22
LERT	0.887 5	0.888 9	15
CSDSM	**0.924 9**	**0.899 5**	24

由表 5 可知,在科学数据集上,相比于原始预训练微调的方法,本文模型 CSDSM 的 Acc 和 $F1$ 都有一定提升。由于在模型中增加了多头自注意力机制,并且首次训练需要将科学数据集嵌入到 KNN 算法中,导致了模型的训练时间增加。从测试结果看,相较于中文 BERT 模型,CSDSM 模型在科学数据集上 $F1$ 值和 Acc 分别提高了 0.078 8 和 0.063 4。图 6 显示了不同模型在 10 轮训练中损失值的大小。可以看出,CSDSM 模型使用 CoSENT 损失函数,在训练过程中损失值收敛速度更快。

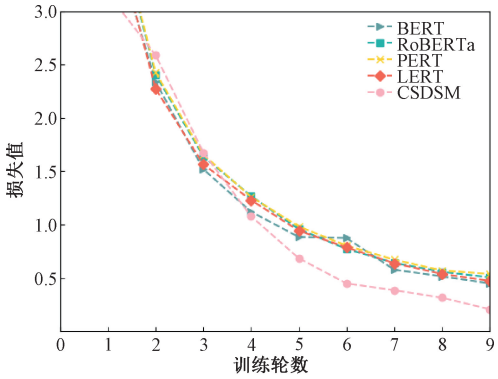

图 6　科学数据集上训练的损失值变化图

Figure 6　Loss curve of training on the scientific dataset

为了验证提出模型的可扩展性,本文在各个中文语义匹配数据集上进行了对比实验。由于中文语义匹配数据上多数作者使用 Spearman 相关性系数,所以本文使用此评价指标,实验结果如表 6 所示。

由表 6 可知,CSDSM 模型使用 CoSENT 函数优化模型的效果要好于预训练微调的结果,利用多头自注意力机制和 KNN 算法提升模型性能。实验结果表明,与中文 BERT 模型相比,所提模型在公共数据集 AFQMC、LCQMC、Chinese-STS-B 和 PAWS-X 上,Spearman 指标 ρ 分别提升了 0.044 8,0.029 0,0.177 7 和 0.050 9,证明了本文方法的可扩展性。

为了进一步说明 CSDSM 模型在科学数据集语

义匹配上的有效性,本文使用基线模型在科学数据集上进行微调,并对表 3 中的句子进行实例分析,其中样本序号分别对应表 3 中样本,具体结果如表 7 和表 8 所示。

表 6 不同模型在中文语义匹配数据上的 Spearman 相关性系数

Table 6 Spearman correlation coefficients of different models on Chinese semantic matching data

模型	ρ			
	AFQMC	LCQMC	Chinese-STS-B	PAWS-X
BERT	0.737 8	0.771 6	0.630 8	0.578 7
RoBERTa	0.433 7	0.773 4	0.707 9	0.589 0
PERT	0.367 1	0.785 3	0.689 8	0.599 0
LERT	0.710 9	0.785 7	0.682 8	0.598 1
CSDSM	0.782 6	0.800 6	0.808 5	0.629 6

表 7 不同模型在科学数据集上的实例分析

Table 7 Different models′ instance analysis on scientific datasets

模型	余弦相似度				
	样本 1	样本 2	样本 3	样本 4	样本 5
BERT	0.745 4	0.873 4	0.986 4	0.992 0	0.474 3
RoBERTa	0.734 3	0.960 3	0.982 4	0.993 3	0.363 0
PERT	0.636 7	0.942 8	0.974 3	0.993 2	0.141 9
LERT	0.748 0	0.897 5	0.979 1	0.994 6	0.558 3
CSDSM	0.289 3	0.779 6	0.968 4	0.993 6	0.461 0

表 8 不同模型在科学数据集上的实例预测结果

Table 8 Prediction results of different models on instances in the scientific datasets

模型	标签等级				
	样本 1	样本 2	样本 3	样本 4	样本 5
BERT	0	1	1	1	0
RoBERTa	0	1	1	1	0
PERT	0	1	1	1	0
LERT	0	1	1	1	0
CSDSM	0	0	1	1	0

从表 7 和表 8 中可以看出,基线模型在大部分样本中分类正确,而样本 2“基于‘两叶模型’的贵州省 500 m 分辨率 8 天 LAI 数据集”和“基于‘两叶模型’的贵州省 500 m 分辨率 8 天 APAR 数据集”,具有中英文混合、文本重叠但语义不相似的样本,基线模型对于该样本都没有正确分类。CSDSM 模型对所有实例中的样本都做出了正确的判断,这说明本文提出的改进策略在科学数据集语义匹配上的效果更好。

5 结论

本文提出了一个面向中文科学数据集的句子级语义匹配模型 CSDSM。采用爬虫技术对科学数据集进行了收集和标注。通过 CSL 数据集对 SimCSE 和 CoSENT 模型进行预训练以学习领域知识;加入多头自注意力机制模块捕获语义信息;使用 KNN 算法来增强语义判断的能力。模型在自建数据集和公共数据集上均取得了最优的结果。

下一步,基于本研究将开展以下工作:①模型的持续优化,进一步探索利用模型压缩技术,以提升模型的线上部署能力;②提升模型的泛化能力,面向不同领域的国家科技资源共享服务平台,收集和组织更为丰富的科学数据语义元数据集开展实验。

参考文献:

[1] 罗鹏程, 王继民, 王世奇, 等. 基于深度学习的科学数据集检索方法研究[J]. 情报理论与实践, 2022, 45(7): 49-56.
LUO P C, WANG J M, WANG S Q, et al. Research on deep learning based scientific dataset retrieval method [J]. Information Studies: Theory & Application, 2022, 45(7): 49-56.

[2] CHEN S H, XU T J. Long text QA matching model based on BiGRU-DAttention-DSSM[J]. Mathematics, 2021, 9(10): 1129.

[3] 冯皓楠, 何智勇, 马良荔. 基于图文注意力融合的主题标签推荐[J]. 郑州大学学报(工学版), 2022, 43(6): 30-35.
FENG H N, HE Z Y, MA L L. Multimodal hashtag recommendation based on image and text attention fusion [J]. Journal of Zhengzhou University (Engineering Science), 2022, 43(6): 30-35.

[4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[5] LI Y D, ZHANG Y Q, ZHAO Z, et al. CSL: a large-scale Chinese scientific literature dataset [EB/OL]. (2022-09-12)[2023-06-11]. https://arxiv.org/abs/2209.05034.

[6] GAO T Y, YAO X C, CHEN D Q. SimCSE: simple contrastive learning of sentence embeddings[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 6894-6910.

[7] LI S H, GONG B. Word embedding and text classification based on deep learning methods[J]. MATEC Web of Conferences, 2021, 336: 06022.

[8] LIU J P, CHU X T, WANG Y F, et al. Deep text retrieval models based on DNN, CNN, RNN and Trans-

former: a review[C]//2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS). Piscataway: IEEE, 2022: 391-400.

[9] HUANG P S, HE X D, GAO J F, et al. Learning deep structured semantic models for web search using clickthrough data[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. New York: ACM, 2013: 2333-2338.

[10] SHEN Y L, HE X D, GAO J F, et al. A latent semantic model with convolutional-pooling structure for information retrieval[C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York: ACM, 2014: 101-110.

[11] MOHAN S, FIORINI N, KIM S, et al. A fast deep learning model for textual relevance in biomedical information retrieval[C]//Proceedings of the 2018 World Wide Web Conference. New York: ACM, 2018: 77-86.

[12] KHURANA D, KOLI A, KHATTER K, et al. Natural language processing: state of the art, current trends and challenges[J]. Multimedia tools and applications, 2023, 82(3): 3713-3744.

[13] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems: Volume 2. New York: ACM, 2013: 3111-3119.

[14] 汪烨, 周思源, 翁知远, 等. 一种面向用户反馈的智能分析与服务设计方法[J]. 郑州大学学报(工学版), 2023, 44(3): 56-61.
WANG Y, ZHOU S Y, WENG Z Y, et al. An intelligent analysis and service design method for user feedback[J]. Journal of Zhengzhou University (Engineering Science), 2023, 44(3): 56-61.

[15] LIU P F, YUAN W Z, FU J L, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2021,55(9): 195.

[16] CHOUDHARY S, GUTTIKONDA H, CHOWDHURY D R, et al. Document retrieval using deep learning[C]// 2020 Systems and Information Engineering Design Symposium (SIEDS). Piscataway: IEEE, 2020: 1-6.

[17] ESTEVA A, KALE A, PAULUS R, et al. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization[J]. NPJ Digital Medicine, 2021, 4: 68.

[18] BELTAGY I, LO K, COHAN A. SciBERT: a pretrained language model for scientific text[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: Association for Computational Linguistics, 2019: 3615-3620.

[19] CHOWDHURY A, ROSENTHAL J, WARING J, et al. Applying self-supervised learning to medicine: review of the state of the art and medical implementations[J]. Informatics, 2021, 8(3): 59.

[20] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: Association for Computational Linguistics, 2019: 3982-3992.

[21] LI L Y, SONG D M, MA R T, et al. KNN-BERT: fine-tuning pre-trained models with KNN classifier[EB/OL]. (2021-10-06)[2023-06-11]. https://arxiv.org/abs/2110.02523.

[22] PALANIVINAYAGAM A, EL-BAYEH C Z, DAMAŠEVIČIUS R. Twenty years of machine-learning-based text classification: a systematic review[J]. Algorithms, 2023, 16(5): 236.

[23] CHICCO D. Siamese neural networks: an overview[J]. Methods in Molecular Biology, 2021, 2190: 73-94.

[24] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24)[2023-06-11]. https://arxiv.org/abs/1810.04805.

[25] CUI Y M, CHE W X, 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.

[26] CUI Y M, YANG Z Q, LIU T. PERT: pre-training BERT with permuted language model[EB/OL]. (2022-03-14)[2023-06-11]. https://arxiv.org/abs/2203.06906.

[27] CUI Y M, CHE W X, WANG S J, et al. LERT: a linguistically-motivated pre-trained language model[EB/OL]. (2022-11-10)[2023-06-11]. https://arxiv.org/abs/2211.05344.

Semantic Matching Model for Chinese Scientific Datasets

LIU Jianping[1,2], CHU Xintao[1], WANG Jian[3], GU Xunxun[1], WANG Meng[1], WANG Yingfei[1]

(1. College of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China; 2. The Key Laboratory of Images and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan 750021, China; 3. Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China)

Abstract: In order to address the difficulty of existing word-level semantic matching models in understanding sentence-level scientific dataset metadata, a sentence-level semantic matching (CSDSM) model for Chinese scientific datasets was proposed. The model used the CSL dataset to train and generate the CoSENT pre-training model based on SimCSE and CoSENT. Building upon the CoSENT model, a multi-head self-attention mechanism was introduced for feature extraction, and the final output was obtained by weighting the cosine similarity and KNN classification results. Experimental data from the National Earth System Science Data Center's open semantic metadata information was used as a self-built scientific dataset. The experimental results showed that compared to the Chinese BERT model, the proposed model improved the Spearman's ρ index by 0.044 8, 0.029 0, 0.177 7 and 0.050 9 on the public datasets AFQMC, LCQMC, Chinese-STS-B, and PAWS-X, respectively. Additionally, $F1$ and Acc on the self-built scientific dataset were improved by 0.078 8 and 0.063 4 respectively. The proposed model effectively addresses the problem of sentence-level semantic matching in scientific datasets.

Keywords: text matching; semantic matching; pre-training model; scientific datasets; natural language processing

(上接第55页)

Image Super-resolution Reconstruction Network Based on Double Feature Extraction and Attention Mechanism

BO Yangyu, WU Yongliang, WANG Xuejun

(College of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China)

Abstract: In the process of image super-resolution reconstruction, high frequency features might be ignored, which would lead to insufficient extraction features and fuzzy texture details in the reconstructed image. To solve this problem, an image super-resolution reconstruction network based on double feature extraction and attention mechanism was proposed. In particular, in this study, a two-branch network for feature extraction was proposed to solve the problem that high frequency features and multi-scale features could not be effectively extracted and uniformly fused during image reconstruction. In addition, in order to make the network obtain more accurate high-frequency features, a local spatial attention module was proposed, and combined with channel attention. A residual fusion attention module was constructed to improve the network's ability to locate high-frequency features. Finally, the atrous pyramid module was designed to enlarge the receptive field of the network and enable the multi-scale feature extraction. Experiments were carried out on four benchmark datasets, and the results were better than the current advanced methods. Especially when the super-resolution multiple was 4, the proposed method improved the optimal $PSNR$ by 0.16, 0.08, 0.03 and 0.20 dB, respectively, compared with the current mainstream models. The experimental results shown that the proposed method achieved better improvement in visual effect and quantitative analysis.

Keywords: image super-resolution reconstruction; local spatial attention; residual fusion attention; atrous pyramid; double feature extraction