第 12 卷 第 2 期
2023 年 3 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 12 No. 2
Mar. 2023

引文格式：
卢美情, 申妍燕. 一种基于孪生网络预训练语言模型的文本匹配方法研究 [J]. 集成技术, 2023, 12(2): 53-63.
Lu MQ, Shen YY. A text matching method based on a pretraining language model: sentence embeddings using Siamese BERT-Networks [J]. Journal of Integration Technology, 2023, 12(2): 53-63.

一种基于孪生网络预训练语言模型的文本匹配方法研究

卢美情[1,2] 申妍燕[2*]

[1](五邑大学智能制造学部 江门 529020)
[2](中国科学院深圳先进技术研究院先进计算与数字工程研究所 深圳 518055)

摘 要 孪生网络预训练语言模型(Sentence Embeddings using Siamese BERT-Networks，SBERT)在文本匹配的表示层面上存在两个缺点：(1)两个文本查询经 BERT Encoder 得到向量表示后，直接进行简单计算；(2)该计算不能考虑到文本查询之间更细粒度表示的问题，易产生语义上的偏离，难以衡量单个词在上下文中的重要性。该文结合交互方法，提出一种结合多头注意力对齐机制的 SBERT 改进模型。该模型首先获取经 SBERT 预训练的两个文本查询的隐藏层向量；然后，计算两文本之间的相似度矩阵，并利用注意力机制分别对两个文本中的 token 再次编码，从而获得交互特征；最后进行池化，并整合全连接层进行预测。该方法引入了多头注意力对齐机制，完善了交互型文本匹配算法，加强了相似文本之间的关联度，提高了文本匹配效果。在 ATEC 2018 NLP 数据集及 CCKS 2018 微众银行客户问句匹配数据集上，对该方法进行验证，实验结果表明，与当前流行的 5 种文本相似度匹配模型 ESIM、ConSERT、BERT-whitening、SimCSE 以及 baseline 模型 SBERT 相比，本文模型在 F_1 评价指标上分别达到了 84.7% 和 90.4%，比 Baseline 分别提高了 18.6% 和 8.7%，在准确率以及召回率方面也表现出了较好的效果，且具备一定的鲁棒性。

关键词 文本匹配；Sentence-BERT；多头注意力对齐机制
中图分类号 TP 399 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20220817001

A Text Matching Method Based on a Pretraining Language Model: Sentence Embeddings Using Siamese BERT-Networks

LU Meiqing[1,2] SHEN Yanyan[2*]

[1](*Faculty of Intelligent Manufacturing*, *Wuyi University*, *Jiangmen* 529020, *China*)
[2](*Institute of Advanced Computing and Digital Engineering*, *Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[*]*Corresponding Author*: *yy.shen@siat.ac.cn*

收稿日期：2022-08-17 **修回日期**：2022-10-15
基金项目：国家重点研发计划项目(2019YFB1405200)；广东省2019年省拨高建“冲补强”专项项目(5041700175)；教育部第二批新工科研究与实践项目(E-RGZN20201036)
作者简介：卢美情，硕士研究生，研究方向为自然语言处理-文本匹配；申妍燕(通讯作者)，副研究员，研究方向包括服务机器人流程自动化、深度强化学习算法在无线通信中的应用，E-mail：yy.shen@siat.ac.cn。

Abstract The sentence embeddings using Siamese BERT-Networks pre-trained language model has two shortcomings in its presentation layer for text matching, that is, (1) two queried texts are directly computed after they are represented in vectors by the BERT Encoder, (2) such computation does not consider the needs to refine the granular representation of the two queried texts. As such presented semantics could be deviated and it is also difficult to assess the importance of single words in text matching. This paper proposes an improved text similarity matching model SBMAA based on SBERT pre-trained language model. Firstly, the hidden layer vectors of the two queries passing through the SBERT model are obtained, and then the similarity matrix between the two is calculated. The attention mechanism is used to encode the tokens in the two sentences again to obtain interactive features and pool them. Finally, the fully connected layer is connected for prediction. This method introduces the multi-head attention alignment mechanism, which is a common way of interactive text matching algorithm, and strengthens the correlation degree between similar texts, so that the model can achieve more accurate matching effect. The experimental results on ATEC 2018 NLP data set and CCKS 2018 Webank Customer Question Matching dataset show that compared with the five popular text similarity matching models ESIM, ConSERT, BERT-whitening, SimCSE and Baseline model SBERT, The proposed SBMAA model achieves 84.7% and 90.4% in F_1 evaluation index, 18.6% and 8.7% higher than Baseline, respectively. It also shows good effect in accuracy and recall rate, and has certain robustness.

Keywords text matching; Sentence-BERT; Multiple attention alignment mechanisms

Funding This work is supported by National Key Research and Development Program of China (2019YFB1405200), High-Level University Construction Special Project of Guangdong Province, China in 2019 (5041700175) and Second Batch of New Engineering Research and Practice Projects of Ministry of Education (E-RGZN20201036)

1 引 言

在计算力不断增强的助推下，深度学习技术迅猛发展，利用神经网络模型，可隐性地学习自然语言序列的语义及其内在表示的特性。2018年，谷歌公司提出的 Transformer 预训练语言模型成功引起了研究者们对“预训练＋微调”（端到端的学习方式）的强烈兴趣。该预训练语言模型具备超多参数的处理功能和强大的拟合能力，在各种下游任务中表现出色。预训练语言模型下游的一个应用——文本匹配，旨在判定两个文本是否表达相同的语义信息，其在信息检索、问题回答、语义鉴别等任务中扮演着重要角色[1]。利用文本匹配，可将用户的检索查询内容与数据库中大量的文档集合根据特定条件进行相关性建模，计算检索语句与文档间的相关程度，召回相关文档，并对召回的文档根据用户需求重新排序后，返回给用户[2]。文本匹配的主要任务是利用预训练语言模型理解检索语句与文档间的语义信息。

孪生网络预训练语言模型（Sentence Embeddings using Siamese BERT-Networks，SBERT）[3]是一种 BERT 预训练语言模型，旨在获取检索语句和文档语句的语义表示，并进行对比。在对比过程中，SBERT 负责提取两个语句的主语义，然后将两个语句的主语义在同一向量空间中进行编码、建模，最后计算其相似度。典型的匹配方法有 DSSM[4]、CDSSM[5]和

ABCNN[6]等。这些匹配方法通过对句子进行预处理和索引提高效率，但由于只提取文本语句的表示向量，没有考虑文本对在词级别的交互信息[7]，失去了语义焦点，易产生语义上的偏差，且很难在上下文中度量单个单词的含义。本文提出了一种结合多头注意力对齐机制的 SBERT 改进模型(Improved SBERT algorithm integrating Multiple Attention Alignment mechanism，SBMAA)。该模型通过将孪生网络 SBERT 中的两个 BERT 模块输出的最后一层隐藏层向量进行对齐，从而获取交互特征，并对其进行融合，可有效加强文本间的交互，提升文本信息匹配的效果。

本文第 2 节简要介绍了基于深度学习的文本匹配算法的相关工作；第 3 节详细描述了本文提出的 SBMAA 模型的体系结构；第 4 节介绍 ATEC 2018 NLP 实验数据集，并对实验结果进行了比较和分析；第 5 节总结全文。

2　基于深度学习的文本匹配算法相关工作

本节将详细介绍深度学习现有的 3 种主流文本匹配算法模型，并对其结构及基本原理进行阐述。

2.1　基于表示型的文本匹配模型

基于表示型的文本匹配模型侧重于对文本表示层的构建[8]，使用神经网络将两个文本转换为相应向量，然后结合表示向量提取内部交互信息[2]。此类方法的主要研究内容涵盖两个方面：(1)针对表示层，研究如何更好地表示两段需要判断相似性的文本，从而获取文本中包含的重要语义信息，常见的对于表示层进行编码的模型包括全连接网络、卷积神经网络、循环神经网络和基于注意力机制的模型等[8]；(2)针对匹配层，研究如何利用得到的两个向量计算相似度，常用的方法包括点积、余弦距离、高斯距离、多层感知机和关联矩阵方法等。图 1 为基于表示型匹配模型的基本结构。

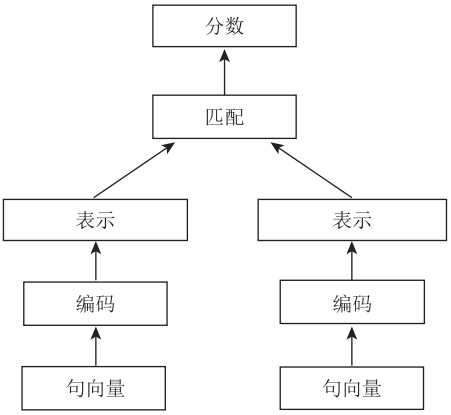

图 1　表示型匹配模型原理示意图

Fig. 1　Schematic diagram of presentation-based text matching models

He 等提出的 DSSM 模型是最早基于表示型的文本匹配模型，该模型是利用深度神经网络和词袋生成模型来提取检索语句(Query)和文档语句(Document)中的内在词袋，构建词袋向量，并将其表示为低维度的语义向量，再通过余弦公式计算两个向量的相似度[2]。在 DSSM 模型的基础上，研究者们又开发了一系列模型，如 CDSSM、MV-DSSM[9]、SiamLSTM[10]、ARC-I[11]和 InferSent[12]等。这些模型的基本结构均如图 1 所示，仅在表示层或匹配层有所不同。事实表明，上述模型在文本匹配、信息检索领域等均具有良好的效果。但基于表示型的文本匹配模型存在两大问题：(1)每个文本只提取最后的语义向量，信息损失难以度量；(2)文本对之间缺乏词汇和句子信息的比较。

2.2　基于交互型的文本匹配模型

基于交互的模型弃用了先编码表示再匹配的方式，在表示层之前就进行文本间的匹配与信息交互，更加关注于字词之间的匹配信息，很好地克服了表示型文本匹配存在的问题。因此，如何提取两段文本的交互信息成为此类模型的关键技术。该类模型通常采用注意力机制使一个文本与

另外一个文本对齐，以获取不同粒度的交互，然后融合各粒度的匹配信息，最终得到一个独立的特征矩阵，用以表示文本对的关系。图 2 为基于交互型匹配模型的基本原理示意图。

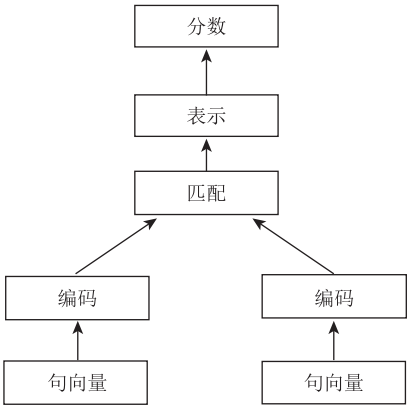

图 2 交互型匹配模型的原理示意图

Fig. 2 Schematic diagram of interactive matching models

典型的基于交互型的模型有：ARC-II[11]、MatchPyramid[13]、ABCNN 和 ESIM[14]等，以及大量基于上述模型的新模型。虽然此类模型可捕捉更深层次的语义信息，更好地把握语义焦点，但随着层次结构更加复杂，交互关系也变得更复杂，耗费的时间呈指数增长。此外，此类模型缺乏对句法、句间的对比，对全局信息的刻画，难以完美地描绘出全局信息。

2.3 基于预训练语言模型的文本匹配方法

基于预训练模型的匹配方法利用预训练加微调的方式完成文本匹配任务。该文本匹配训练方法可分为有监督学习、无监督学习与自监督学习等方法。在文本匹配任务中，之所以可以使用预训练模型技术，是因为预训练可以学习语言间的大量隐式联系与全局的一些知识，能够很好地利用语言模型的统计特征[15]。此类预训练语言模型一般具有千万级学习参数，因此，其匹配效果优于其他模型。

2018 年，Peters 等[16]提出了 ELMo 模型，该模型通过构造大规模的无监督预训练，得到性能出色且与特定任务无关，在各种文本任务上均通用的语义表征向量。ELMo 是一个双向长短期记忆网络模型，前向 LSTM 可通过预先给定的嵌入标记预测后一个嵌入标记，后向 LSTM 也可通过预先给定的嵌入标记预测前一个嵌入标记。此外，研究者们还研发了 BERT[17]和 GPT-2[18]模型。这些模型首先进行无监督的大规模预训练，然后针对自然语言处理具体下游任务进行有监督/无监督微调，模型匹配效果较好[15]，大幅度地刷新了评测的各项指标。2019 年，Yang 等[19]运用 BERT 实现了长文本匹配，其基本思想是将长文本分解成多个长度较短的句子，将每个句子在 BERT 上独立计算匹配分数，最后聚合这些句子的分数。同年，Reimers 等[3]在 BERT 模型的基础上提出了有监督模型 SBERT，将文本匹配任务带到了新的高度。相较于 Word2vec 和 BERT 模型，SBERT 更适用于文本相似度度量、语义搜索和语义聚类等任务[20]。但 SBERT 作为一个有监督模型，存在大多数有监督模型具有的缺点，其也需要大量的标注信息，在类似文本检索的现实场景中，获取大规模句子对标签的代价非常高。因此，2020 年，围绕无监督条件下如何更有效地处理 BERT 句向量，计算语义相似度的问题，Li 等[21]提出了 BERT-flow 模型。BERT-flow 模型利用标准化流将向量的分布变换成规整的高斯分布，该模型在多项评测中均表现良好，验证了模型的有效性。2021 年，Su 等[22]提出的 BERT-whitening 模型采用无监督的训练方式，通过传统机器学习中的白化变换操作，取得了和 BERT-flow 相近的效果。此外，BERT-whitening 还可降低向量维度，提高匹配速度。BERT-flow 和 BERT-whitening 均通过处理 BERT 输出的向量，进而解决文本相似度的计算问题。2021 年，Gao 等[23]提出的无监督模型 SimCSE 通过构建正负样本对解决文本相似度的计算问题。具体做法是将同一个文本输入两次，利用不同的 Dropout 机制，得到不同但相近的向量作为正

例，将同一批次内其他文本输入得到的向量作为负例。Dropout 机制的作用是通过参数的设定使整个连接层的神经元随机失活。与 BERT-flow 和 BERT-whitening 两种模型相比，采用 Dropout 机制的 SimCSE 文本匹配效果更优。目前，新提出的一些模型大多集中在研究无监督的模型上，如 Liu 等[24]提出的 Trans-Encoder 模型结合了表示型与交互型的优势，并以无监督的方式引导一个准确的句子对模型。将 SimCSE 作为 Baseline 的 7 个句子文本相似性基准进行对比实验，实验结果表明，在所有数据集上，Trans-Encoder 比以前的无监督句子对模型效果均有显著提升。由此可知，无监督/弱监督学习将是以后的研究趋势。

3　融合多头注意力对齐机制的孪生网络改进算法

本节将详细介绍融合多头注意力对齐机制的 SBERT 算法改进模型(SBMAA)。该模型的结构如图 3 所示。

为克服 BERT 模型结构不适用于聚类及句子回归等无监督训练任务的问题，Reimers 等[3]于 2019 年提出了预训练语言模型 SBERT，基于该模型，又提出了有监督模型 SBMAA。Reimers 等[3]指出“直接用 BERT 最后一层的结果作为句子向量，甚至比词向量的效果要差，而直接使用[CLS]的效果最差”。SBERT 利用孪生和三元网络结构对 BERT 预训练模型进行微调，从而得出具备语义信息的句子嵌入，并计算相似度[25]。该结构通过对每句话单独编码，极大地提高了计算效率。如在 10 000 条文本中检测出最相似的两条文本，单纯使用 BERT 将造成巨大的计算开销，模型将运算 $n\times(n-1)/2=49\ 995\ 000$ 次(约 65 h)，耗时较长；在同等条件下，SBERT 仅需运算 10 000 次(约 5 s)即可获取句子向量表示，极大地提高了效率[3]。

SBERT 作为一种基于表示型的模型，只能提取文本句子级别的表示向量，未能考虑文本对之间在词级别的交互信息[4]。因此，本文提出引入多头注意力对齐机制，将表示型方法与交互型方法融合，让模型更好地捕捉到原始文本中不同层次的信息，提高文本匹配的准确性。融合多头注意力对齐机制的 SBERT 改进模型 SBMAA 主要由输入层、BERT 编码层、交互式句子表示层、融合层以及输出层组成。该模型将交互注意力与表示注意力融合，使模型可以充分利用和捕捉文本中存在的多个粒度信息，从而提高文本匹配精度，下面将对模型的每个部分进行详细介绍。

3.1　输入层

输入层沿用 BERT 模型的输入设置，包括 A 和 B 两个句子，每个句子分别包含位置向量、段向量和词向量 3 个部分，将这 3 部分向量相

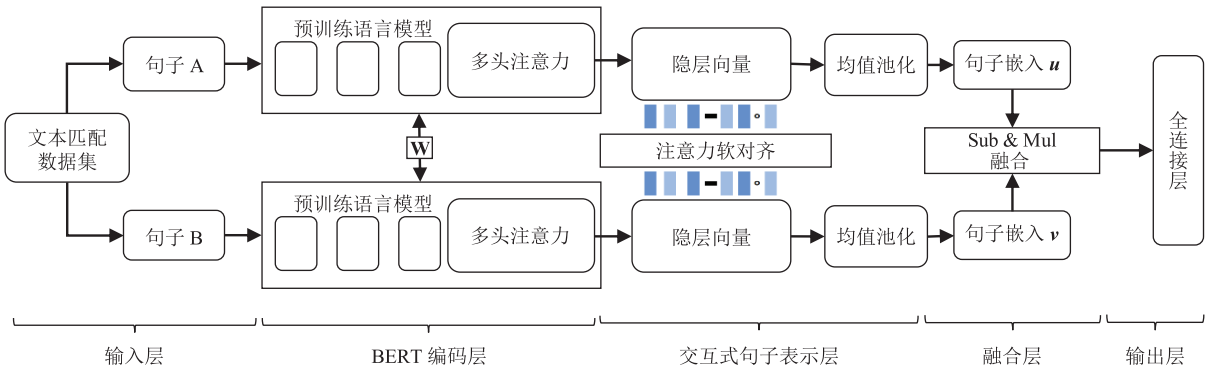

图 3　SBMAA 结构图

Fig. 3　Structure diagram of SBMAA

加后输入 BERT 编码层。其中，位置向量包括输入句子中每一个 token 的时序信息；段向量负责将文字逐句定位，并用标记[CLS]和[SEP]区别不同的句子，[CLS]代表分辨输出的特定符号，[SEP]代表分隔非连续 token 序列的特定符号，同时保持各句子的位置信息；词向量为输入句子中每一个 token 对应的词向量。BERT 模型如图 4 所示。

累加 3 层向量后，将结果引入 Transformer 编码器[26]中，用双向编码的结果表示各个 token。Transformer 编码器包括自注意层、残差层、归一化层、前馈神经网络层。编码器将叠加后的字符级向量作为输入，最终得到具有语义信息的隐层向量[27]，即 BERT 模型的最后一层输出，其包含[CLS]和[SEP]。

3.2 交互式句子表示层

获取两个隐藏层向量 $\boldsymbol{p}_{a_i}$ 和 $\boldsymbol{p}_{b_j}$ 后，需进行句向量间的注意力对齐。首先，计算两个经 BERT 后隐藏层向量间的相似度矩阵 $\boldsymbol{e}_{ij}$，得到两者之间的相似度；然后，使用注意力机制分别对两个句子中的 token 再次编码得到 $\boldsymbol{s}_{a_i}$ 和 $\boldsymbol{s}_{b_j}$。该过程可表示如下：

$$\boldsymbol{e}_{ij}=\boldsymbol{p}_{a_i}^{\mathrm{T}}\boldsymbol{p}_{b_j} \tag{1}$$

$$\boldsymbol{s}_{a_i}=\sum_{j=1}^{l_b}\frac{\exp\left(\boldsymbol{e}_{ij}\right)}{\sum_{k=1}^{l_b}\exp\left(\boldsymbol{e}_{ik}\right)}\boldsymbol{p}_{b_j},\forall j\in[1,\cdots,l_a] \tag{2}$$

$$\boldsymbol{s}_{b_j}=\sum_{i=1}^{l_a}\frac{\exp\left(\boldsymbol{e}_{ij}\right)}{\sum_{k=1}^{l_a}\exp\left(\boldsymbol{e}_{kj}\right)}\boldsymbol{p}_{a_i},\forall i\in[1,\cdots,l_b] \tag{3}$$

其中，$\boldsymbol{p}_{a_i}^{\mathrm{T}}$ 为句子 a 经过 BERT 后的隐藏层语义向量的转置；$\boldsymbol{p}_{b_j}$ 为句子 b 经过 BERT 后的隐藏层语义向量；$\boldsymbol{e}_{ij}$ 为两个隐藏层向量的相似度矩阵，即句子 a 中第 i 个词和句子 b 中第 j 个词的相似度；$\boldsymbol{s}_{a_i}$ 为经过注意力对齐后提取出的句子 a 与句子 b 的相似性信息；$\boldsymbol{s}_{b_j}$ 为经过注意力对齐后提取出的句子 b 与句子 a 的相似性信息。获取到用于预测的重要特征 $\boldsymbol{s}_{a_i}$ 和 $\boldsymbol{s}_{b_j}$ 后，分别进行 mean pooling 得到 $\boldsymbol{u}$ 和 $\boldsymbol{v}$。

3.3 融合层及全连接输出层

交互式句子表示层提取了每个句子中的交互特征 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 后，融合层根据公式(4)将两部分交互特征进行融合。

$$\boldsymbol{f}=[\boldsymbol{u}-\boldsymbol{v};\boldsymbol{u}\cdot\boldsymbol{v}] \tag{4}$$

其中，$\boldsymbol{u}-\boldsymbol{v}$ 为向量间的减法操作，其目的是获取差异特征；$\boldsymbol{u}\cdot\boldsymbol{v}$ 为向量矩阵相乘操作，目的是获取交互特征。将减法和乘法操作的结果进行向量拼接，得到特征融合向量 $\boldsymbol{f}$。然后，将 $\boldsymbol{f}$ 输

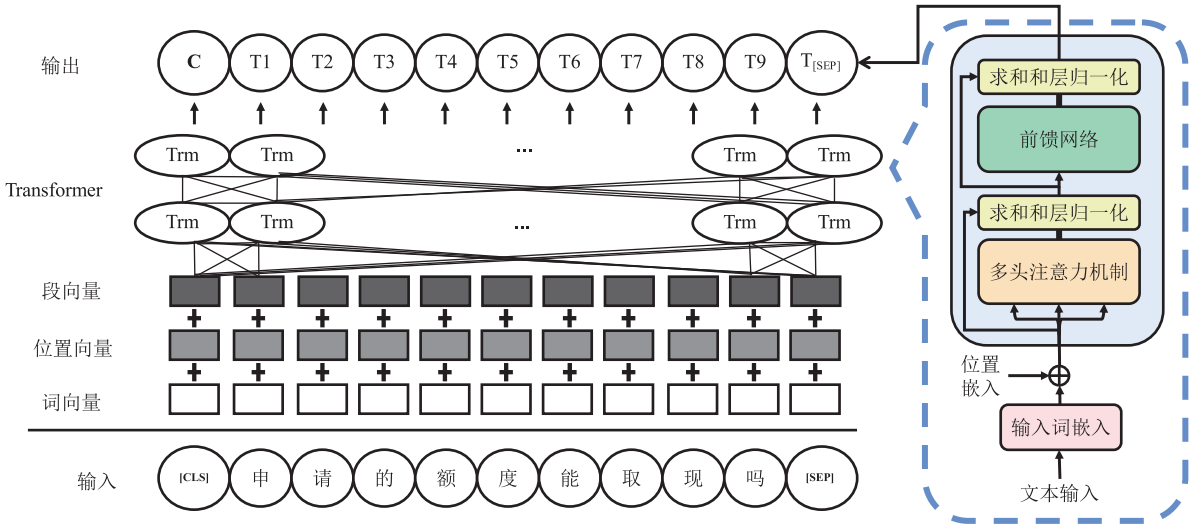

图 4 BERT 模型简图

Fig. 4 BERT model diagram

入到一个全连接网络，调整特征的权重，使用 softmax 函数预测分类的结果，损失函数使用交叉熵损失函数。全连接层的计算公式如下所示。

$$\hat{p}(y|u,v)=softmax(W^f f+b^f) \tag{5}$$

$$\hat{y}=argmax(\hat{p}(y|u,v)) \tag{6}$$

其中，y 为预测的值；W^f 为维度与 f 相同的矩阵；b^f 为维度与 f 相同的一维向量。当进行预测分类时，可得到 $\hat{p}(y|u,v)$范围内的最大值 $\hat{y}$。有研究表明[28]，与使用余弦相似度或欧氏距离作为输出相比，使用全连接层作为输出的效果更好。SBMAA 的详细结构如图 5 所示。

4　实验结果与分析

4.1　数据集

基于中文数据集 ATEC 2018 NLP 蚂蚁金服金融大脑赛题标注数据和 CCKS 2018 微众银行客户问句匹配数据，本文进行了相关实验。其中，ATEC 2018 NLP 数据是为了解决金融领域智能客服遇到的自然语言处理问题而构建的客服专用问答库。人工对数据集中两个句子的语义进行分析，若语义相似或相同，标注为 1，否则标注为 0。数据集共有 10 万条数据，其中 8 万条作为训练集，2 万条作为测试集。数据集示例如表 1 所示。

表 1　ATEC 2018 NLP 数据集示例

Table 1　Sample ATEC 2018 NLP dataset

问题 1	问题 2	标签
为什么我不能开通花呗	开通花呗提示安全不通过	0
怎么花呗不能支付	花呗付款不了怎么回事	1
花呗还款有利息没	花呗还钱涨利息吗	0

CCKS 2018 微众银行数据集是针对中文的真实客服语料，进行问句意图匹配的银行领域智能客服日志数据。该数据集由 10 万个中文问句对组成，包括 8 万个训练样本和 2 万个测试样本。每一对问句与一个二进制标签相关联，标签用于指示这两个句子是否具有相同的含义。数据集示

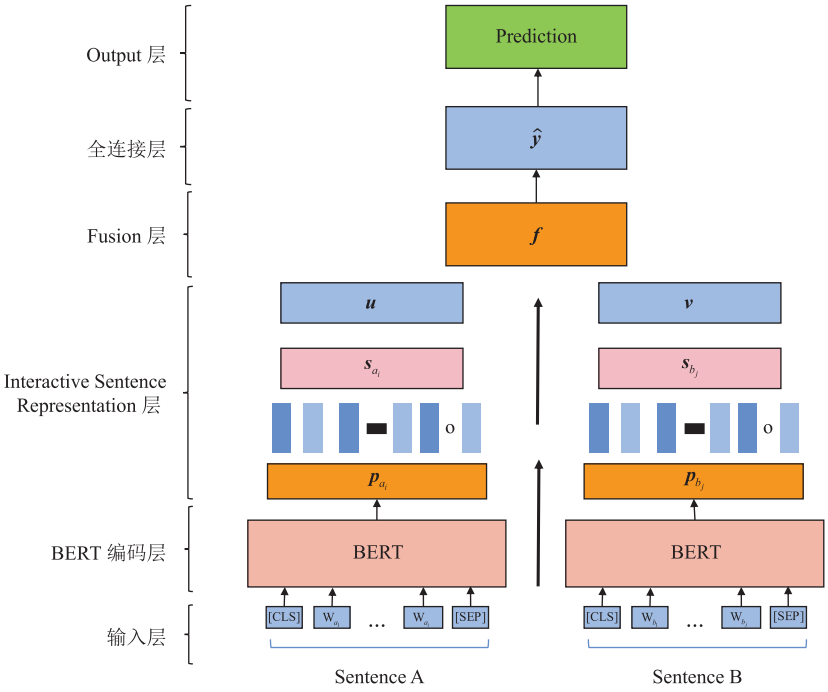

图 5　SBMAA 详细结构图

Fig. 5　Detailed structure diagram of SBMAA

例如表 2 所示。

表 2 CCKS 2018 微众银行数据集示例

Table 2 Sample CCKS 2018 WeBank dataset

问题 1	问题 2	标签
申请的额度能取现吗	取现一次性取完可以吗	0
如何获得微粒贷资格	为什么没微粒贷啊	1
微粒咨询电话号码是多少	你们的人工客服电话是多少	1

4.2 实验设置

实验参数设置如下：BATCH_SIZE＝64，输入句子最大长度 Text_SIZE＝128，学习率 LEARNING_RATE＝1×10^{-5}，BERT 模型隐藏单元数 HIDDEN_NUM＝768，Dropout＝0.3。为防止训练过程不稳定，优化方法采用 Adam 算法，设置 detect_imp＝3 500，在一定 batch 的条件下，若设定模型没有明显提升，则结束训练。实验涉及的 BERT 预训练模型均使用 12 层的 BERT-Base-Chinese 模型对文本进行向量化。

实验使用 TESLA V100 32 GB 显存 GPU 进行运算，开发框架使用 Pytorch 1.8.0。

4.3 评价指标

本实验采用 F_1 值、准确率 *Acc* 和召回率 *R* 作为评估模型效果的指标。其中，召回率 *R* 代表真实标签为正的样例中有多少被预测正确，精确率 *Pre* 代表预测为正的样例中有多少是预测正确的，准确率 *Acc* 代表在总样本中预测正确的比例。F_1 值与精确率 *Pre* 和召回率 *R* 成正相关，相关计算公式如下:

$$R=\frac{TP}{TP+FN} \tag{7}$$

$$Pre=\frac{TP}{TP+FP} \tag{8}$$

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$F_1=\frac{2\times Pre\times R}{Pre+R}=\frac{2TP}{2TP+FP+FN} \tag{10}$$

其中，*FP* 为真实标签是负预测为正的样本总数；*TP* 为真实标签是正预测为正的样本总数；*TN* 为真实标签是负预测为负的样本总数；*FN* 为真实标签是正预测为负的样本总数。

4.4 实验结果与分析

本研究设计了 3 组实验，第一组用于验证 SBMAA 模型的效果，选取 5 种经典的文本匹配模型：ESIM、SBERT、ConSERT[29]、BERT-whitening、SimCSE，并分别基于两个数据集，进行实验对比。其中，ESIM 是 Chen 等[14]在 2016 年基于双向长短期记忆网络和 tree-LSTM 的模型，提出的一种专为自然语言推理而设计的加强版 LSTM[30]；ConSERT 是 Yan 等[29]在 2021 年提出的采用无监督和数据增强方式微调 BERT，以进行对比学习的模型；BERT-whitening 是 Su 等[22]在 2021 年提出的文本匹配模型，该模型通过简单的白化操作，将嵌入向量转化为各向同性，文本匹配效果可媲美 BERT-flow；SimCSE 是 Gao 等[23]在 2021 年提出文本匹配模型，该模型利用对比学习优化目标函数进行模型微调，从而获得文本向量表示[31]。

表 3 和表 4 为 6 种不同的深度学习模型在两个数据集上的实验结果。由实验结果可知，在两个数据集上，SBMAA 模型的 F_1 值均高于其他 5 个模型，即其整体效果优于其他模型。与基线(baseline)模型 SBERT 相比，其 F_1 值提升超过 8%，可能是由于增加了句向量间的交互，与直接捕获句子之间的相似性信息相比，该模型可以捕获更细粒度上的语义信息。经典模型如 ESIM 由于综合应用了双向长短期记忆网络和注意力机制也取得了不错的实验效果，但因为缺乏全局语义表示和多维度的交互信息，实验效果略差于 SBMAA。

在两个数据集上，一些较新的模型如 ConSERT、BERT-whitening 和 SimCSE 的实验效果不佳，某些指标低于 50%，原因可能是不同的数据集间样本分布差异较大，且中英文数据集的处理方式也有所差别，导致较新的模型在原文

表 3　在 ATEC 2018 NLP 数据集上的实验结果统计

Table 3　Results of text matching model on ATEC 2018 NLP dataset

(%)

模型	准确率	召回率	F_1 值
ESIM	86.9	51.2	60.8
SBERT	77.4	100	66.1
ConSERT	48.1	57.2	66.0
BERT-whitening	36.8	73.7	67.1
SimCSE	86.9	51.2	60.8
SBMAA	78.0	92.6	84.7

表 4　在 CCKS 2018 数据集上的实验结果统计

Table 4　Statistics of experimental results on CCKS 2018 dataset

(%)

模型	准确率	召回率	F_1 值
ESIM	85.4	87.0	85.7
SBERT	77.8	99.7	81.9
ConSERT	78.9	96.3	66.0
BERT-whitening	64.7	93.5	72.5
SimCSE	86.9	51.2	60.8
SMBAA	90.2	91.1	90.6

数据集上效果虽好，但不适用于本文数据集。此外，这些模型在原文中使用的指标不同，当主要考量指标为 F_1 值时，实验效果较差，但有些指标正常，甚至高于本文模型，因此实验结果也与模型侧重效果的不同相关联。

第 2 组实验基于 ATEC 2018 NLP 数据集，研究批处理规模对 SBMAA 模型的性能及收敛速度的影响，实验结果如表 5 所示。由表 5 可知，批处理规模越大，模型的性能越好，但提升并不明显。此外，更大的批处理加快了训练过程，但也意味着需要更多的 GPU 显存。

表 5　不同的批处理规模下 SBMAA 模型在 ATEC 2018 NLP 数据集上的实验结果统计

Table 5　Experimental results statistics of SBMAA model on ATEC 2018 NLP dataset under different batch sizes

批处理规模	F_1 值	准确率	召回率	Number of Steps
8	84.61%	77.92%	92.56%	31 479
16	85.04%	78.38%	92.93%	15 739
32	84.99%	78.33%	92.88%	7 869
64	84.70%	78.02%	92.65%	3 934

由于实验中使用的两个数据集属于金融领域，其中含有大量金融领域的高频词，对模型效果可能产生影响。因此，在 ATEC 2018 NLP 数据集上，第 3 组实验主要研究高频词对 SBMAA 模型性能的影响。本文测试了隐藏出现频次排名最高的 k 个高频词，Top-k 分别取 5，10，15，20，25，实验结果如图 6 所示。

由图 6 可知，当移除 15 个最频繁的 token 时，各项性能指标均达到最好。虽然删除一些最常见的标记，各项指标有所提升，但幅度小于

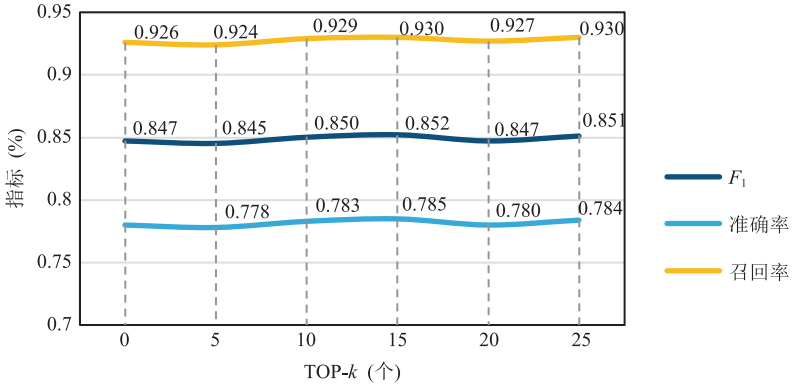

图 6　在 ATEC 2018 NLP 数据集中移除不同频繁词 Top-k 数量后的实验结果统计

Fig. 6　Experimental results statistics after removing frequent words Top-k in ATEC 2018 NLP dataset

0.1。实验结果表明，该方法具有鲁棒性，常用频繁词的存在对句子表征的影响不大。

综上所述，本文提出的模型 SBMAA 提升效果较为明显，鲁棒性较好。

5 总 结

针对文本相似度匹配的问题，本文提出一种基于 SBERT 的文本匹配改进模型 SBMAA。该模型首先利用 SBERT 实现文本的向量化表示，在孪生网络架构的基础上，引入多头注意力的对齐，增加了句向量的交互，并通过拼接融合层，使得模型自身拥有获取交互信息的能力。由实验结果可知，本文提出的 SBMAA 模型能够有效提升文本匹配的效果，且具有一定的鲁棒性。

参 考 文 献

[1] 吕乐宾, 刘群, 彭露, 等. 结合多粒度信息的文本匹配融合模型 [J]. 计算机科学, 2021, 48(6): 196-201.
Lv LB, Liu Q, Peng L, et al. Text matching and fusion model with multi-granularity information [J]. Computer Science, 2021, 48(6): 196-201.

[2] 周献杭, 申妍燕. 基于多粒度语义交互的无监督法律裁判文书检索 [J]. 集成技术, 2022, 11(2): 55-66.
Zhou XH, Shen YY. Unsupervised retrieval of court documents based on multi-granularity semantic interaction [J]. Journal of Integration Technology, 2022, 11(2): 55-66.

[3] Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks [Z/OL]. arXiv Preprint, arXiv: 1908.10084, 2019.

[4] Huang PS, He XD, Gao JF, 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, 2013: 2333-2338.

[5] Shen YL, He XD, Gao JF, et al. A latent semantic model with convolutional-pooling structure for information retrieval [C] // Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, 2014: 101-110.

[6] Yin WP, Schütze H, Xiang B, et al. ABCNN: attention-based convolutional neural network for modeling sentence pairs [J]. Transactions of the Association for Computational Linguistics, 2016, 4: 259-272.

[7] 余传明, 薛浩东, 江一帆. 基于深度交互的文本匹配模型研究 [J]. 情报学报, 2021, 40(10): 1015-1026.
Yu CM, Xue HD, Jiang YF. Research on text matching model based on deep interaction [J]. Journal of Information Science, 2021, 40(10): 1015-1026.

[8] 曹帅. 基于深度学习的文本匹配研究综述 [J]. 现代计算机, 2021, (16): 74-78.
Cao S. Survey of text matching based on deep learning [J]. Modern Computer, 2021, (16): 74-78.

[9] Elkahky AM, Song Y, He XD. A multi-view deep learning approach for cross domain user modeling in recommendation systems [C] // Proceedings of the 24th International Conference on World Wide Web, 2015: 278-288.

[10] Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2016: 2786-2792.

[11] Hu BT, Lu ZD, Li H, et al. Convolutional neural network architectures for matching natural language sentences [J]. Advances in Neural Information Processing Systems, 2014, 27: 2042-2050.

[12] Conneaua A, Kjelad D, Schwenk H, et al. Supervised learing of universal sentence representations from natural language inference data [Z/OL]. arXiv Preprint, arXiv: 1705.02364, 2017.

[13] Pang L, Lan Y, Guo J, et al. Text matching as image recognition [C] // Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016: 2793-2799.

[14] Chen Q, Zhu XD, Ling ZH, et al. Enhanced LSTM for Natural Language Inference [Z/OL]. arXiv Preprint, arXiv: 1609.06038, 2016.

[15] 周烨恒, 石嘉晗, 徐睿峰. 结合预训练模型和语言知识库的文本匹配方法 [J]. 中文信息学报, 2020, 34(2): 63-72.
Zhou YH, Shi JH, Xu RF. A text matching method combining pre-trained model and language knowledge base [J]. Journal of Chinese Information Science, 2020, 34(2): 63-72.

[16] Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations [Z/OL]. arXiv Preprint arXiv: 1802.05365, 2018.

[17] Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [Z/OL]. arXiv Preprint, arXiv: 1810.04805, 2018.

[18] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019, 1(8): 9.

[19] Yang W, Zhang HT, Lin J. Simple applications of BERT for ad hoc document retrieval [Z/OL]. arXiv Preprint, arXiv: 1903.10972, 2019.

[20] 唐晓波, 刘亚岚. 基于 Sentence-BERT 语义表示的咨询问题提示列表自动构建方法研究——以糖尿病咨询为例 [J]. 现代情报, 2021, 41(8): 3-15.
Tang XB, Liu YL. Research on automatic construction of counseling question prompt list based on Sentence-BERT semantic representation—taking diabetes counseling as an example [J]. Journal of Modern Information, 2021, 41(8): 3-15.

[21] Li BH, Zhou H, He JX, et al. On the sentence embeddings from pre-trained language models [Z/OL]. arXiv Preprint, arXiv: 2011.05864, 2020.

[22] Su JL, Cao JR, Liu WJ, et al. Whitening sentence representations for better semantics and faster retrieva [Z/OL]. arXiv Preprint, arXiv: 2103.15316, 2021.

[23] Gao TY, Yao XC, Chen DQ. SimCSE: simple contrastive learning of sentence embeddings [Z/OL]. arXiv Preprint, arXiv: 2104.08821, 2021.

[24] Liu FY, Jiao YL, Massiah J, et al. Trans-Encoder: unsupervised sentence-pair modelling through self- and mutual-distillations [Z/OL]. arXiv Preprint, arXiv: 2109.13059, 2021.

[25] 孙维远. 基于预训练模型的海关商品异常申报检测研究 [D]. 大连: 大连理工大学, 2021.
Sun WY. Research on abnormal declaration detection of customs goods based on pre-trained model [D]. Dalian: Dalian University of Technology, 2021.

[26] 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, 2017: 6000–6010.

[27] 张小川, 戴旭尧, 刘璐, 等. 融合多头自注意力机制的中文短文本分类模型 [J]. 计算机应用, 2020, 40(12): 3485-3489.
Zhang XC, Dai XY, Liu L, et al. Chinese short text classification model fused with multi-head self-attention mechanism [J]. Journal of Computer Applications, 2020, 40(12): 3485-3489.

[28] 魏垂沛, 李海华, 朱红杰. 基于语义的招标文件示范文本智能更新研究 [J]. 招标采购管理, 2021, (2): 23-27.
Wei CP, Li HH, Zhu HJ. Research on intelligent updating of bidding document model text based on semantics [J]. Bidding and Procurement Management, 2021, (2): 23-27.

[29] Yan Y, Li R, Wang S, et al. ConSERT: a contrastive framework for self-supervised sentence representation transfer [Z/OL]. arXiv Preprint, arXiv: 2105.11741, 2021.

[30] 黄静, 陈新府豪. 一种简化门控结构的增强序列文本语义匹配模型研究 [J]. 软件工程, 2022, 25(1): 50-55.
Huang J, Chen XFH. Research on a text semantic matching model of ESIM with simplified gating structure [J]. Software Engineering, 2022, 25(1): 50-55.

[31] 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/OL]. 情报理论与实践, 2022, 45(7): 49-56.
Luo PC, Wang JM, Wang SQ, et al. Research on deep learning based scientific dataset retrieval method [J/OL]. Intelligence Theory and Practice, 2022, 45(7): 49-56.