计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700008

基于细粒度差异特征的文本匹配方法

王 胜 张仰森 陈若愚 向 尕
北京信息科技大学智能信息处理研究所 北京 100101
(1028742881@qq.com)

摘 要 文本匹配是检索系统中的关键技术之一。针对现有文本匹配模型对文本语义差异捕获不准确的问题,文中提出了一种基于细粒度差异特征的文本匹配方法。首先,使用预训练模型作为基础模型对匹配文本进行语义的抽取与初步匹配;然后,引入对抗学习的思想,在模型的编码阶段人为构造虚拟对抗样本进行训练,以提升模型的学习能力与泛化能力;最后,通过引入文本的细粒度差异特征,纠正文本匹配的初步预测结果,有效提升了模型对细粒度差异特征的捕获能力,进而提升了文本匹配模型的性能。在两个数据集上进行了实验验证,其中在LCQMC数据集上的实验结果显示,所提方法在ACC性能指标上达到了88.96%,优于已知的最好模型。

关键词:文本匹配;预训练模型;语义相似度;对抗学习;差异特征
中图法分类号 TP391.1

Text Matching Method Based on Fine-grained Difference Features

WANG Sheng,ZHANG Yang-sen,CHEN Ruo-yu and XIANG Ga
Institute of Intelligent Information Processing,Beijing Information Science and Technology University,Beijing 100101,China

Abstract Text matching is one of the key technologies in the retrieval system. Aiming at the problem that the existing text matching models can't capture the semantic differences of texts accurately,this paper proposes a text matching method based on fine-grained difference features. Firstly,the pre-trained model is used as the basic model to extract the matching text semantics and preliminarily match them. Then,the idea of adversarial learning is introduced in the embedding layer,and by constructing the virtual confrontation samples artificially for training,the learning ability and generalization ability of the model are improved. Finally,by introducing the fine-grained difference feature of the text to correct the preliminary prediction results of the text matching,the capture ability of the model for fine-grained difference features is effectively improved,and then the performance of the text matching model is improved. In this paper,two datasets are tested,and the experiment on LCQMC dataset shows that the performance index of ACC is 88.96%,which is better than the best known model.

Keywords Text match,Pre-trained model,Semantic similarity,Adversarial learning,Difference feature

1 引言

随着互联网的迅速发展,网络上的信息呈指数级增长。面对海量的信息数据,如何从其中检索出有用的信息显得尤为重要[1]。若采用人工方式来检索这些信息,则工作量太大,不符合高精度、高时效性的需求,因此各种各样的信息检索系统(如搜索引擎、问答系统等)应运而生。相比人工检索信息,这些系统不仅极大地提高了检索效率与精度,而且更能从语义角度分析用户提出的问题,深层次理解用户的检索意图,最终将尽可能准确的结果返回给用户。检索系统的一般工作流程为:获取用户向检索系统提出的问题,利用文本匹配技术在候选文本标题数据中进行匹配,将匹配到的文本标题及其对应的详细信息返回给用户。因此,文本匹配的效果将直接影响到检索系统的各项性能指标。传统方法在完成文本匹配的任务时,通常只使用文本之间的词频、TFIDF等浅层特征,这类方法难以挖掘文本之间的深层语义关系。近年来,随着深度学习等自然语言处理技术的发展,研究者们开始利用深层语义网络、预训练模型等技术挖掘文本之间的语义关系,衡量文本之间的匹配程度。这类基于深度语义网络的模型虽然能够在全局语义上衡量文本之间的匹配程度,但在一定程度上忽略了两者之间的局部差异。因为对于格式高度相似的匹配文本,其关注的重点往往体现在两者之间的局部差异上,例如在对陈述句对进行匹配时,模型更需要关注对陈述文本具有概括意义的中心词之间的差异。另外,现有的基于预训练模型的文本匹配模型往往直接利用训练数据进行微调,这对于文本格式高度相似的文本匹配任务来说,模型的泛化能力较

到稿日期:2020-07-01 返修日期:2020-08-20 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61772081);国家重点研发计划(2018YFB1403104);北京信息科技大学学科研基金(2035008)
This work was supported by the National Natural Science Foundation of China(61772081),National Key Research and Development Plan(2018YFB1403104) and Research Fund of Beijing Information Science and Technology University(2035008).
通信作者:张仰森(zhangyangsen@163.com)

低，以致缺乏对新样本之间的细粒度差异的捕获能力。

针对上述现有文本匹配模型存在的问题，本文在预训练模型的基础上，提出了一种基于细粒度差异特征的文本匹配模型（RoBERTa＋FGM＋DF）。一方面，在微调过程中引入对抗学习对模型训练过程中的编码层进行扰动，以提升模型的泛化能力；另一方面，通过对文本差异特征的提取，来提升模型对文本之间细粒度差异的捕捉能力，进而提升文本匹配任务的效果。

2　相关工作

文本匹配作为信息检索任务中的核心技术，通常以文本相似度计算和文本相关度计算的形式进行展现，主要分为基于传统特征的文本匹配方法和基于深度学习的文本匹配方法。

基于传统特征的无监督文本匹配方法主要结合各种词法、句法等知识对文本进行特征抽取，并利用浅层特征进行文本相似度计算，进而实现文本匹配。例如 Salton 等[2]提出利用词语频次加权特征等统计学特征进行文本表示，并以此计算文本之间的相似度。Song 等[3]选择使用基于 LDA 的主题模型对文档进行建模，利用词语生成概率对文本进行相似度计算。虽然这些无监督的文本匹配方法能够衡量文本之间的匹配程度，但衡量质量的好坏取决于待匹配文本的质量，面对复杂文本时的处理能力较弱。

随着深度学习的发展，各种神经网络模型开始应用于文本匹配的任务中，主要分为文本表示模型和文本交互模型。文本表示模型指通过各种方法获得句子的语义表示、通过句子的语义表示进行匹配程度的计算。例如 Le 等[4]受 word2vec 等技术的启发，利用待匹配的文本数据训练句向量，通过拼接句向量与字向量的表示进行文本相似程度的计算。Logeswaran 等[5]通过将解码器修改为分类器，来预测候选文本与当前文本是否为上下文关系。该方法相比传统的编码器-解码器模型，在保证精度的前提下，训练速度有了明显的提升。Cer 等[6]为了提升不同语境下各个文本的编码效果，设计了基于迁移学习的编码模型，使模型动态适应各种自然语言。文本交互模型指通过强化文本之间的交互程度来提升文本匹配的效果，通常使用各种注意力机制来实现待匹配文本的对齐。例如 Yin 等[7]在卷积神经网络（CNN）的基础上引入注意力机制，利用注意力权重对卷积层的输出进行加权融合，最终得到可以对文本之间的关系进行建模的 ABCNN 模型。Chen 等[8]提出了一种简单高效的语义匹配网络 ESIM，首先使用双向长短期记忆神经网络（BiLSTM）对文本编码进行上下文表示，同时引入注意力机制强化文本之间的交互，然后将文本表示向量与注意力向量进行拼接，使拼接后的向量最大程度地蕴含文本的局部信息与全局信息，最后再次使用双向循环神经网络抽象各个文本的特征表示，并得到最终的特征向量。Wang 等[9]为了解决文本之间交互程度不充分的问题，提出了 BiMPM 模型，该模型包括语义向量获取层、文本表示层、信息交互层、信息聚合层和预测层五大部分。他们在信息交互层中提出了多种信息交互的方法来提升文本之间的交互程度。上述模型虽然能够利用文本之间的语义信息与交互信息进行匹配程度计算，但在语义编码阶段仍使用传统的静态文本语义获取方法（word2vec，glove 等）。这些方法对复杂语境下的文本表示能力不足，因此基于预训练模型的研究方法成为了近年来的研究热点。Radford 等[10]利用单个 Transformer 来实现文本语义获取。Devlin 等[11]在超大规模数据集的基础上，提出了基于自注意力机制的双向 Transformer 框架（Bidirectional Encoder Representations from Transformers，BERT），刷新了多项自然语言处理任务的最优成绩。随后，各种各样的预训练模型[12-14]被提出，不断刷新着各领域任务的成绩。在文本匹配领域，Hu 等[15]尝试将多种预训练模型引入文本匹配任务中，并分析了各种模型的特点。Sakata 等[16]利用 BERT 计算问题与答案之间的相似性，并将其应用于问答系统中。Wu 等[17]利用 BERT 获取文本的向量表示后，利用注意力机制进行文本匹配程度的计算，与传统模型相比，该方法取得了明显的效果提升。Wang[18]引入独立双向多头注意力机制以及层级密集神经网络，有效地提升了 BERT 模型的预训练收敛速度和最终精度。与传统方法相比，基于预训练模型的文本匹配方法虽已取得不错的成绩，但对文本之间细粒度差异的捕获能力仍然存在不足。因此，本文在预训练模型的基础上，融入对抗学习与差异特征的提取，强化模型对细粒度差异特征的识别与计算能力，从而提升模型的性能。

3　基于细粒度差异特征的文本匹配方法

本文方法主要由 3 部分组成，分别是基于预训练模型的文本匹配基础模型、基于对抗学习的文本匹配模型以及基于细粒度差异特征的文本匹配模型。

3.1　基于预训练模型的文本匹配基础模型

本文中的文本匹配基础模型采用 Googleresearch 公布的句对分类框架体系，如图 1 所示。

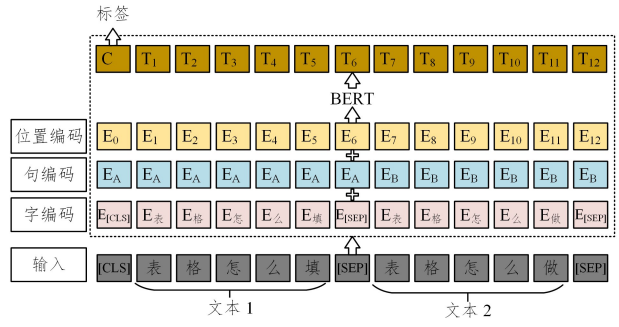

图 1　文本匹配的基础模型

Fig. 1　Basic model of text matching

当进行文本匹配任务时，待匹配的两段文本首先会拼接成一个句子，并在句首加入[CLS]标识符，同时使用标识符[SEP]分割两段文本，并将此句子作为模型的输入。模型的编码部分主要包括 3 部分。

（1）词嵌入层：将单词转化为固定维度的词向量。

（2）段嵌入层：获取文本对中两个文本的句向量，用于判断两个文本在语义上是否相似。

（3）位置嵌入层：为 Transformer 层提供文本的顺序特征，使模型更能理解输入的文本信息。

经过模型计算得到语句编码之后，仅使用句首标识符[CLS]对应的向量作为整个句子的向量表征，该向量对应的标签即为文本匹配任务的匹配结果。

在获得文本匹配的基础模型后，还需进一步提升模型的泛化能力，从而提升对文本的语义编码效果。具体来说，在文本匹配任务中，更希望模型具有较高的泛化能力，并能够识别文本之间的细微差异。

3.2 基于对抗学习的文本匹配模型

对抗学习可以使模型在训练过程中尽可能地犯错，进而提升模型的学习能力，例如在模型的训练阶段，人为添加一些对抗样本去提升模型的抗扰动能力，进而提升模型在下游任务中的泛化能力，增强模型在微调过程中的鲁棒性[19]。具体流程如图 2 所示。

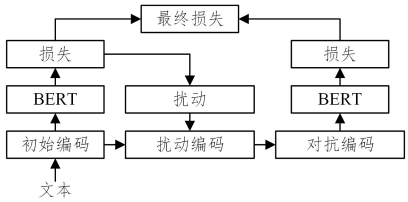

图 2 基于对抗学习的文本匹配模型

Fig. 2 Text matching model based on confrontation learning

对抗样本生成方法主要分为黑盒方法与白盒方法，前者主要通过对模型的输入输出进行扰动，后者则可以直接对模型的参数进行扰动。本文使用基于白盒的扰动方法，直接对模型的编码层进行扰动。经典的扰动样本的生成方法[19]如式(1)、式(2)所示：

$$\boldsymbol{g}=\nabla_x L(\theta,x,y) \tag{1}$$

$$\boldsymbol{r}_{adv}=\varepsilon\cdot \mathrm{sgn}(g) \tag{2}$$

其中，$\boldsymbol{g}$ 为梯度向量，L 为损失函数，∇_x 为损失函数的梯度，θ 为模型的参数，x 为原始输入样本，y 为目标标签，$\boldsymbol{r}_{adv}$ 为需要计算的扰动向量，ε 为扰动步长，sgn 为符号函数。由于模型提升了输入样本在损失方向上的权重，进而提升了样本经过模型计算得到的损失，因此仅在细微变化输入样本的前提下，能够使模型尽可能地犯错，从而提升模型的泛化能力。上述方法已被广泛用于图像领域，然而当将其应用到自然语言处理领域时，由于文本数据是以离散的 Token 序列存储的，不同于图像数据以连续的 RGB 数值存储，因此无法同图像数据一样，直接在原始文本数据中添加扰动，例如直接替换文本中的单词、词语等。为解决上述问题，可以采用创建“虚拟”对抗文本[20]的方式，即直接对编码过程中的字嵌入进行扰动，同时使用合适的方法修改扰动方向，扰动过程如式(3)所示：

$$\boldsymbol{r}_{adv}=\varepsilon\cdot \boldsymbol{g}/\|g\|_2 \tag{3}$$

其中，$\boldsymbol{r}_{adv}$ 为扰动向量，$\boldsymbol{g}$ 为梯度向量，$\|g\|_2$ 为向量 $\boldsymbol{g}$ 的二范式，ε 为扰动步长。此时的原始输入 x 为经过模型编码的中间结果，通过在嵌入层中加入扰动向量，来在嵌入空间中生成新的对抗样本，将对抗样本经过模型计算产生的损失函数与原始损失函数相加，即可得到模型最终的损失函数。对于扰动步长 ε 的取值，我们将在实验部分进行讨论。

3.3 基于细粒度差异特征的文本匹配模型

当待匹配的文本数据在形式上高度一致且仅存在个别差异时，仅采用上述模型较难发现文本数据之间的差异，具体来说，体现在模型对文本匹配的结果概率值的差异较小。当模型对文本进行匹配程度预测时，经过模型计算得到的概率结果为$[p_0,p_1]$，p_0 表示文本匹配结果为错误的概率值，p_1 表示文本匹配结果为正确的概率值，p_0 和 p_1 满足 $p_0+p_1=1$。例如，当匹配文本为“表格怎么填”和“表格怎么做”时，此时 p_0 和 p_1 均十分接近 1/2，因此认为模型对该文本对的匹配结果可信度较低，经分析可得模型难以识别其细粒度的差异。针对上述问题，本文引入置信区间与细粒度差异特征的概念，当匹配概率值位于置信区间内时，对其细粒度差异特征进行提取以修正模型的初步匹配结果。其中，置信区间 P 由置信度 P_c 确定，置信区间 P 满足式(4)。

$$p_0,p_1\in[1-P_c,1+P_c] \tag{4}$$

本文将文本中具有代表性的主要动词当作细粒度特征[21]，具体分为普通动词、动副词以及名动词 3 种。例如，文本对“表格怎么填”和“表格怎么做”的匹配重点应体现在“填”和“做”的差异上；文本对“穿什么衣服好看”和“有什么好看的衣服”的匹配重点应体现在“有”跟“穿”的差异上。

对于细粒度差异特征的提取，本文使用词法分析工具 LAC 进行提取，例如待匹配文本对 A：“穿什么衣服好看”，B：“有什么好看的衣服”，经过词法分析可以得到如图 3 所示的结果。

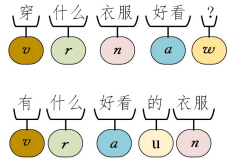

图 3 词法分析结果

Fig. 3 Lexical analysis results

图 3 中，v 为普通动词，r 为代词，a 为形容词，u 为助词，n 为名词，w 为标点符号。提取“穿”为文本 A 的差异特征 F_{va}，提取“有”为文本 B 的差异特征 F_{vb}，若文本中有多个动词，则将其全部提取到差异特征集合中；若文本中无主要动词，则特征集合为空。在得到文本的差异特征集合后，利用 F_{va} 与 F_{vb} 的交集对特征集合进行清洗，清洗方法如式(5)－式(7)所示：

$$F_v=F_{va}\cap F_{vb} \tag{5}$$

$$F'_{va}=F_{va}-F_v \tag{6}$$

$$F'_{vb}=F_{vb}-F_v \tag{7}$$

在得到经过清洗的差异特征集合后，利用特征词对应的预训练词向量[22]作为差异特征向量 $\boldsymbol{F}_{Va}$ 与 $\boldsymbol{F}_{Vb}$。利用余弦相似度计算特征向量 $\boldsymbol{F}_{Va}$ 与 $\boldsymbol{F}_{Vb}$ 之间的相似值，计算式如式(8)所示：

$$sim=\frac{\boldsymbol{F}_{Va}\cdot\boldsymbol{F}_{Vb}}{\|\boldsymbol{F}_{Va}\|\|\boldsymbol{F}_{Vb}\|} \tag{8}$$

得到差异特征之间的相似值后，通过阈值对匹配结果概率值 p_0，p_1 进行适当的放缩，放缩方法如式(9)、式(10)所示：

$$p_0=\begin{cases}p_0+sim\times 2P_c, & sim<thr\\ p_0-sim\times 2P_c, & sim>thr\end{cases} \tag{9}$$

$$p_1=\begin{cases}p_1-sim\times 2P_c, & sim<thr\\ p_1+sim\times 2P_c, & sim>thr\end{cases} \tag{10}$$

其中，p_0 表示文本匹配结果为错误的概率值，p_1 表示文本匹配

结果为正确的概率值，P_c 为置信度，$2P_c$ 则表示置信区间的长度，thr 为相似度阈值，$sim\times 2\ P_c$ 为放缩因子，使用该因子进行放缩，既满足边界值情况下的放缩要求，又不会产生过量放缩的情况。关于相似度阈值的取值会在实验参数分析部分进行讨论。对于相似度小于阈值的数据，强化 p_0 的权重，对于相似度大于阈值的数据，则增加 p_1 对应的权重。对匹配结果概率值进行适当的放缩，可有效提升模型对细粒度差异特征的识别能力。

4　实验与分析

为了验证模型的效果，本文在公共数据集 LCQMC 与数据集 SECP 上进行了验证。

4.1　数据集介绍

4.1.1　LCQMC 数据集

LCQMC 数据集是由 Liu 等[23]在 COLING2018 上提出的中文语义匹配数据集，该数据集在匹配重点上更偏重语义匹配而不是词语匹配，因此被广泛用于中文文本匹配任务的验证。该数据集一共包括训练集、验证集和测试集 3 部分，由 256 433 个句子和 716 298 条文本对组成，具体信息如表 1 所列。

表 1　LCQMC 数据集的数据分布

Table 1　Data distribution of LCQMC dataset

	train	dev	test	all
PAIRS	238 766	8 802	12 500	260 068
SENS	245 951	15 917	23 557	256 433
POS	138 574	4 402	6 250	149 226
NEG	100 192	4 400	6 250	110 842
LEN	6.12	7	5.65	6.26

表 1 中，PAIRS 表示文本对的个数，SENS 表示句子的数量，POS 表示正例的个数，NEG 表示负例的个数，LEN 表示匹配文本的平均长度。数据集的评价指标采用准确率 ACC 进行评测，准确率的计算式如式(11)所示：

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

其中，TP 为正例预测正确的个数，TN 为负例预测正确的个数，FP 为正例预测错误的个数，FN 为负例预测错误的个数。

4.1.2　SECP 数据集

为了验证模型在其他数据集上的效果，本文收集并整理了 SECP 数据集。该数据集以新浪微博为原始数据，对有关社会安全事件的相关微博标题及文本进行收集、清洗过滤并进行人工标注，最终得到由 2 498 个文本对、3 550 个句子组成的数据集，数据集的具体信息如表 2 所列。

表 2　SECP 数据集数据分布

Table 2　Data distribution of SECP dataset

	train	dev	test	all
PAIRS	1 724	250	524	2 498
SENS	2 601	395	954	3 550
POS	862	125	255	1 242
NEG	862	125	269	1 256
LEN	11.8	10.37	11.3	11.16

本数据集的评价指标采用准确率 ACC 进行测评。

4.2　实验对比模型

本文主要选取以下模型进行对比。

(1)BiLSTM[24]。该方法利用双向长短期记忆神经网络对文本进行语义表示，将文本的语义表示进行相似度计算得到文本的匹配结果。

(2)ESIM[8]。该方法在 BiLSTM 的基础上引入注意力机制来强化文本之间的交互，以提升模型对文本局部信息与全局信息的捕捉能力。

(3)BiMPM[9]。该方法在 ESIM 模型的基础上对文本交互方法进行了改进，提升了文本之间的交互程度。

(4)BERT[11]。该方法是直接利用 BERT 进行文本匹配的基础模型。

(5)MGF[24]。该方法提取文本中字粒度与词粒度的语义特征，通过对多粒度特征的融合来更好地优化下游句子的语义匹配。

(6)K-BERT[25]。该方法提出了一种基于知识图谱的语言模型，通过将知识图谱中的三元组作为领域知识添加到训练文本中，以强化语言模型中的领域知识的表示能力，同时引入软定位和可见矩阵来解决知识噪声的问题。

(7)3SiBert[26]。该方法在预训练模型的下游任务中引入了定制的自监督任务，使得预训练模型能够获得更多的领域知识与语义信息。

(8)Glyce-BERT[27]。为了利用中文文本中丰富的符号信息，该方法提出了利用历史汉字(隶书、繁体等)作为训练语料，设计适用于汉字图像处理的田字格 CNN 网络，最终将图像处理结果作为多任务学习的辅助知识，以提升预训练模型的泛化能力。该模型是目前 LCQMC 数据集上的 SOTA(State of the Art)模型。

4.3　实验参数分析

本实验采用 Pytorch 框架进行相关模型的编码实现，在 Centos7 系统上采用 GPU(RTX 2080Ti)进行模型的训练和调试。使用 huggingface 公布的 chinese-roberta-wwm-ext[28]作为本文的基础匹配模型。模型的主要参数 Batch_size 设置为 64，学习率设置为 2×10^{-5}，置信度设置为 0.2。

针对对抗学习中扰动步长 ε 的选取，我们利用 RoBERTa-FGM 模型在 LCQMC 验证集上进行了实验，其性能随其取值变化而变化的情况如图 4 所示。

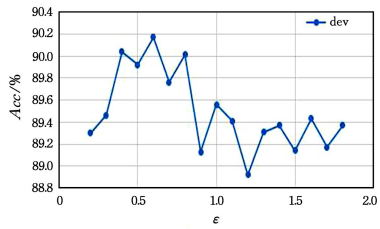

图 4　ε 值的变化趋势图

Fig. 4　Epsilon trend chart

由实验可知，扰动步长过大或过小均会对模型产生影响，进而影响最终匹配模型的性能。具体来说，随着扰动步长 ε 值的增加，模型的性能呈先上升后下降的趋势。最终，我们选择在验证集上性能最好的 ε 值作为模型最终的取值。

由于在细粒度差异特征模型中的相似度阈值的选取会直接影响模型匹配的性能，因此我们在 LCQMC 数据集的验证集上对模型进行了实验，其性能随阈值 thr 的变化趋势如图 5 所示。

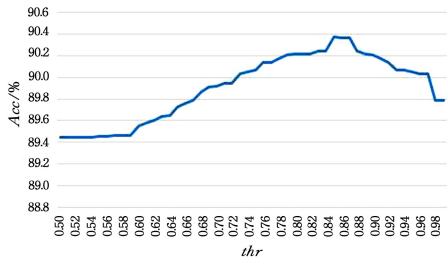

图 5 *thr* 值的变化趋势图

Fig. 5 *thr* trend chart

由实验可知，相似度阈值 *thr* 在 0.85 左右时获得最优的模型性能，选取该值作为模型的最终取值。

4.4 实验结果分析

按照相关数据集的实验流程和测评指标，本文对 LCQMC 数据集和 SECP 数据集进行了实验与分析，具体实验结果如表 3 所列，其中 EISM 模型、BIMPM 模型与 BERT 模型的性能指标为本文复现论文中的方法并进行实验得到的模型性能。另外，由于在 LCQMC 数据集的实验中数据划分与实验流程均按照规范进行，因此 MGF，K-BERT，3SiBERT 和 Glyce+BERT 的实验结果均来自对应论文中报告的结果。

表 3 实验对比结果

Table 3 Experimental comparison results

Model	LCQMC		SECP
	dev	test	test
BiLSTM[24]	—	76.1	—
EISM[8]	85.44	84.28	78.24
BIMPM[9]	84.56	82.98	73.85
MGF[24]	—	85.83	—
BERT[11]	88.70	87.08	86.07
K-BERT[25]	89.20	87.1	—
3SiBERT[26]	89.0	88.1	—
Glyce+BERT[27]	—	88.7	—
RoBERTa	89.16	87.47	87.02
RoBERTa+FGM	89.54*	88.61**	87.21*
RoBERTa+FGM+DF	**89.87****	**88.96****	**87.60****

其中，RoBERTa 模型表示第 3.1 节对应的基础文本匹配模型，RoBERTa+FGM 模型表示第 3.2 节对应的基于对抗学习的文本匹配模型，RoBERTa+FGM+DF 模型表示第 3.3 节对应的基于细粒度差异特征的文本匹配模型。

表 3 中，* 表示显著性水平 $p<0.05$，* * 表示显著性水平 $p<0.01$，本文中的显著性验证参考文献[29]中的方法，在数据集上采用 1 000 次有放回的抽样进行评估。具体来说，RoBERTa+FGM 模型和 RoBERTa+FGM+DF 模型都是对 RoBERTa 模型进行显著性检验。从表 3 可以看出，相比对比模型，本文模型的性能都有显著性提高。

由表 3 中的实验数据可知，从是否添加注意力来看，在神经网络中添加注意力机制的方法明显优于传统神经网络的方法，这是由于 BiLSTM 模型仅从文本本身对语义进行建模，忽略了文本之间的联系，而 EISM 等模型则通过注意力机制强化待匹配文本之间的交互，提升了模型的语义建模能力。从语义编码的角度来看，基于预训练语言模型的方法明显优于基于传统神经网络的方法，这是由于预训练模型经过大规模语料的训练后，获得了较好的语义编码效果。从预训练模型的效果来看，RoBERTa 模型优于传统的 BERT 模型，这是由于经过优化训练过程，例如使用动态掩盖、省去下一句预测任务以及增大训练语料等机制，可以使预训练模型的性能得到提升。从外部知识的引用来看，将知识图谱、文本符号信息等知识引入预训练模型(K-BERT，Glyce+BERT 等)中能有效提升模型的泛化能力。

在 LCQMC 数据集中，本文提出的 RoBERTa+FGM+DF 模型在验证集和测试集的性能指标上均明显优于其他模型，证明了本文模型的有效性。具体来说，在基础匹配模型上添加对抗学习之后，模型取得了较高的性能，并超过了除 SOTA 模型外的其他模型。在添加了细粒度差异特征之后，模型的性能超过了 Glyce+BERT 模型，取得了最优效果，进一步说明了本文中细粒度差异特征的引入是有效的。

在 SECP 数据集中，本文提出的 RoBERTa+FGM 模型在性能指标上高于传统模型及基于 BERT 的文本匹配模型，在添加了细粒度差异特征之后，模型性能得到了进一步的提升，这说明了本文方法的有效性。

另外，在 SECP 数据集的实验中，本文探究了迁移学习对模型性能的影响，具体实验结果如表 4 所列，具体来说，包括如下 3 种模型。

(1)SECP。训练集仅使用 SECP 数据集的训练集进行微调，并在测试集上进行验证。

(2)Fine-tune。使用 LCQMC 数据集中的训练集进行预微调，使模型预先学习到文本匹配领域的特征知识，然后在 SECP 训练集上进行再微调，最后在 SECP 测试集上进行验证。

(3)MultiSet。模型在 LCQMC 和 SECP 组成的联合训练集上进行微调，并在 SECP 测试集上进行验证。

表 4 迁移学习实验对比结果

Table 4 Comparison results of transfer learning experiment

Model	SECP		
	SECP	Fine-tune	MultiSet
RoBERTa	83.97	84.35	87.02
RoBERTa+FGM	85.50**	85.68**	87.21*
RoBERTa+FGM+DF	**85.88****	**85.88****	**87.60****

由表 4 中的实验数据可得，经过在多数据集上进行预微调，并在目标任务上进行再微调，能有效提升模型的性能，但性能的提升效果不如直接使用联合训练集进行微调，可见通过增加模型训练集的样本量可以有效地提升模型的泛化能力，进而有效提升小数据量任务的性能。

结束语 本文提出了一种基于细粒度差异特征的文本匹配模型，在预训练模型的基础上，提取文本之间的细粒度差异，将差异特征融入匹配模型中，有效地提升了模型的性能，强化了模型对细粒度差异特征的捕捉能力。同时，将对抗学习应用到模型训练的过程中，通过在模型的编码阶段构造虚拟的对抗样本对模型进行扰动，有效提升了模型的泛化能力，这进一步说明了本文方法的有效性。

但是，在已有的实验中，我们主要在中文数据集上进行实验验证。在以后的工作中，我们将尝试对英文语料进行处理，以验证模型是否具有普适性。同时，文本匹配任务在实际应用中与搜索任务相似，一般需要事实返回计算结果，因此对计算时间的要求较高，在以后的工作中，我们也将进一步优化模

型的执行效率。另外，本文模型在匹配短文本时的效果较好，而在面对匹配文本长度较长时的效果欠佳，这可能是因为长文本中的细粒度差异特征较复杂，难以全面地抽取。在未来的工作中，我们还需要进一步研究更多情形下的细粒度特征抽取方法，以寻求泛化能力更强的文本匹配模型。

参 考 文 献

[1] LI X. Research on paragraph retrieval technology for question answering system [D]. Chengdu: University of Science and Technology of China, 2010.

[2] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523.

[3] SONG F, CROFT W B. A general language model for information retrieval[C]// Proceedings of the Eighth International Conference on Information and Knowledge Management. 1999: 316-321.

[4] LE Q, MIKOLOV T. Distributed representations of sentences and documents [C] // International Conference on Machine Learning. 2014: 1188-1196.

[5] LOGESWARAN L, LEE H. An efficient framework for learning sentence representations[J]. arXiv: 1803. 02893, 2018.

[6] CER D, YANG Y, KONG S, et al. Universal sentence encoder [J]. arXiv: 1803. 11175, 2018.

[7] YIN W, 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.

[8] CHEN Q, ZHU X, LING Z, et al. Enhanced lstm for natural language inference[J]. arXiv: 1609. 06038, 2016.

[9] WANG Z, HAMZA W, FLORIAN R. Bilateral multi-perspective matching for natural language sentences [J]. arXiv: 1702. 03814, 2017.

[10] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding with unsupervised learning [R/OL]. Technical Report, OpenAI, 2018. https://openai. com/blog/language-unsupervised/.

[11] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810. 04805, 2018.

[12] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9.

[13] SUN Y, WANG S, LI Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv: 1904. 09223, 2019.

[14] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. arXiv: 1910. 10683, 2019.

[15] HU W, DANG A, TAN Y. A Survey of State-of-the-Art Short Text Matching Algorithms [C] // International Conference on Data Mining and Big Data. Singapore: Springer, 2019: 211-219.

[16] SAKATA W, SHIBATA T, TANAKA R, et al. FAQ retrieval using query-question similarity and BERT-based query-answer relevance[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 1113-1116.

[17] WU Y, WANG R J. Application of BERT-Based Semantic Matching Algorithm in Question Answering System[J]. Instrumentation Technology, 2020(6): 19-22, 30.

[18] WANG N Z. Research on improved text representation model based on BERT[D]. Southwest University. 2019.

[19] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: 1412. 6572, 2014.

[20] ZHU C, CHENG Y, GAN Z, et al. Freelb: Enhanced adversarial training for natural language understanding[C]// International Conference on Learning Representations. 2019.

[21] YAN J, BRACEWELL D B, REN F, et al. A semantic analyzer for aiding emotion recognition in Chinese[C]// International Conference on Intelligent Computing. Berlin, Heidelberg: Springer, 2006: 893-901.

[22] SONG Y, SHI S, LI J, et al. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 175-180.

[23] LIU X, CHEN Q, DENG C, et al. Lcqmc: A large-scale chinese question matching corpus[C]// Proceedings of the 27th International Conference on Computational Linguistics. 2018: 1952-1962.

[24] ZHANG X, LU W, ZHANG G, et al. Chinese Sentence Semantic Matching Based on Multi-Granularity Fusion Model[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer, 2020: 246-257.

[25] LIU W, ZHOU P, ZHAO Z, et al. K-BERT: Enabling Language Representation with Knowledge Graph[C]// AAAI. 2020: 2901-2908.

[26] CHEN J, CAO C, JIANG X. SiBert: Enhanced Chinese Pre-trained Language Model with Sentence Insertion[C]// Proceedings of The 12th Language Resources and Evaluation Conference. 2020: 2405-2412.

[27] MENG Y, WU W, WANG F, et al. Glyce: Glyph-vectors for Chinese character representations[C]// Advances in Neural Information Processing Systems. 2019: 2746-2757.

[28] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for chinese bert[J]. arXiv: 1906. 08101, 2019.

[29] KOEHN P. Statistical significance tests for machine translation evaluation[C]// Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004: 388-395.

WANG Sheng, born in 1996, postgraduate. His main research interests include natural language processing and machine learning.

ZHANG Yang-sen, born in 1962, postdoctoral, professor, is a distinguished member of China Computer Federation. His main research interests include natural language processing and artificial intelligence.