文章编号：1003-0077(2021)11-0043-08

基于动态词遮掩的句子匹配预训练模型

宋 挺[1,2]，郭展成[1,2]，何世柱[1,2]，刘 康[1,2]，赵 军[1,2]，刘升平[3]

（1. 中国科学院 自动化研究所，北京 100190；
2. 中国科学院大学 人工智能学院，北京 100049；
3. 云知声智能科技股份有限公司，北京 100083）

摘 要：BERT 通过遮掩语言模型、下一句预测等自监督学习任务学习通用语言规律，在自然语言理解任务中取得了良好效果。但 BERT 的下一句预测任务不能直接建模句子的语义匹配关系，且随机遮掩策略也不能高效处理句子的关键内容。针对上述问题，该文提出基于动态词遮掩的预训练模型：基于预训练模型获得句子的向量表示，并通过近似语义计算获取大规模“句子对”预训练数据，最后遮掩重要字词训练遮掩语言模型。在 4 个句子匹配数据集上的实验表明，使用该文提出的预训练方法，RBT3 和 BERT base 的效果都有一定提升，平均准确率分别提升 1.03%和 0.61%。

关键词：句子匹配；预训练模型；自然语言理解

中图分类号：TP391 **文献标识码**：A

Dynamic Word Masking Based Pre-trained Model for Sentence Matching

SONG Ting[1,2], GUO Zhancheng[1,2], HE Shizhu[1,2], LIU Kang[1,2], ZHAO Jun[1,2], LIU Shengping[3]

(1. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;
2. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China;
3. Beijing Unisound Information Technology Co. Ltd., Beijing 100083, China)

Abstract: Pre-trained model such as BERT achieves good results in natural language understanding tasks via random masking strategy and next sentence prediction task. To capture the semantic matching relationship between sentences, this paper proposes a pre-trained model based on dynamic word masking. A large-scale sentence-pairs are obtained through sentence embeddings, and then the important words are masked to train a new kind of masked language model. Experimental on four datasets show that, the performance of RBT3 and BERT base are improved by the proposed method by 1.03% and 0.61%, respectively, according to average accuracy.

Keywords: sentence matching; pre-trained model; natural language understanding

0 引言

句子匹配也被称为复述识别，既是自然语言处理的重要任务之一，也是智能问答、对话系统、信息检索、机器翻译等应用的重要基础和关键模块[1]。作为一个典型的语义匹配任务，其目标是：对于输入的两个句子，模型需要判断它们的语义是否一致、是否表达了相同意图。图 1 是一个句子匹配任务的示例，对于句子 A“听歌用什么软件比较好”和句子 B“现在好用的听音乐软件是什么”，模型需要识别出这两个句子的语义是一致的。

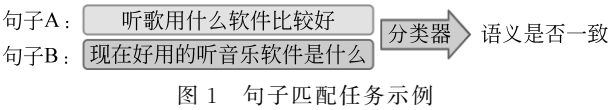

图 1 句子匹配任务示例

基于预训练语言模型的句子匹配方法是目前常用的方法，在众多自然语言理解任务中也取得了不错的效果。以 BERT[2-3] 为代表的预训练语言模型刷新了多项自然语言理解任务的最高水平，开创了

收稿日期：2021-0-12 **定稿日期**：2021-06-06

基金项目：国家重点研发计划(2017YFB1002101)；国家自然科学基金(61533018，U1936207，61976211，61702512)

自然语言处理研究的新范式：先基于大量无监督数据进行自监督预学习完成通用语言建模，再使用少量标注数据微调模型来完成文本分类、序列标注、句子匹配和机器阅读理解等下游任务。具体来说，BERT 基于 Transformer 架构[4]，并通过遮掩语言模型(Masked Language Model，MLM)和下一句预测(Next Sentence Predication，NSP)两项自监督任务在大规模语料上进行预训练。遮掩语言模型根据周围的上下文预测被遮掩的词，下一句预测任务预测输入的两个“句子”在预训练语料中是否按顺序排列。

然而，在预训练阶段，BERT 通过下一句预测任务建模句子间的关系，但从原理上没有直接建模两个句子的语义匹配关系，因此该预训练任务得到的表示不适合应用于句子匹配任务。此外，BERT 中词遮掩策略是随机遮掩输入序列中 15% 的 token，随机遮掩策略不能很好地建模句子的关键信息，而关键信息是判断两个句子语义关系的重要特征，如图 1 中两个句子的关键词“听歌”和“听音乐”是判断语义是否一致的重要特征。

为解决 BERT 等预训练模型在句子匹配任务中的上述两个问题，本文从预训练阶段与下游的句子匹配任务之间数据分布的差异以及如何有效建模关键信息的角度出发，探索面向句子匹配任务的预训练方法。具体地，针对 BERT 在预训练阶段不能建模句子的语义匹配信息及随机遮掩策略的问题，本文提出了一个基于动态词遮掩的句子匹配预训练方法：构造与句子匹配任务数据分布接近的预训练数据，以及能建模关键信息的动态词遮掩策略。为了验证上述方法，本文在 4 个公开中文句子匹配数据集上进行了实验，实验结果表明，使用本文提出的“二次”预训练方法，三层版的 RoBERTa base (RBT3)和 BERT base 的效果都有一定的提升，取得了当前最好的效果，在 AFQMC、LCQMC、BQ、cMedQQ 等四个数据集上 F_1 值分别提升了 1%、0.7%、0.6%、1.9%和 0.4%、1.1%、0.01%、1%。此外，对比实验也证明了“句子对”预训练数据构造方法与动态词遮掩策略的有效性。本文的主要贡献如下：

(1) 针对 BERT 等原始预训练模型无法有效建模句子对匹配任务的问题，提出了一种面向句子匹配任务的预训练数据构造方法。

(2) 针对当前预训练模型 BERT 随机掩码策略无法建模句子匹配等任务的关键信息问题，提出了一种能高效建模关键信息的动态词遮掩策略。

(3) 在 4 个中文句子匹配数据集上的实验结果表明：使用本文提出的预训练方法，RBT3 和 BERT base 的性能都有一定的提升，取得了当前最好的效果。

1 相关工作

1.1 句子匹配

随着深度学习在自然语言处理领域的成功，深度神经网络模型在句子匹配任务中占据了主导地位。近年来，句子匹配任务的模型可以分为两种框架：①基于表示的孪生神经网络模型，②基于交互的句子匹配模型。

在基于表示的模型中，两个句子分别输入到相同的编码器中，编码器将两个句子映射到同一向量空间中，然后利用句子的向量直接计算相似度或构建一个神经网络分类器。Huang 等提出 DSSM 模型[5]，该模型使用词袋模型作为模型的输入表示，使用前馈神经网络作为句子编码器，最后对句子向量使用余弦距离来度量相似性。Mueller 等提出了 Siamese-LSTM 模型[6]，该模型使用循环神经网络作为句子编码器，对句子向量使用曼哈顿距离评价句子相似度。基于表示的模型的优势在于共享参数，使模型更小且更易于训练，并且句子向量可用于可视化、聚类等。但该框架的缺点也很明显：在编码过程中句子之间没有显式的交互，这可能会丢失一些重要信息。

为解决基于表示模型的问题，研究者提出了基于交互的句子匹配模型：得到句子表示后，先对两个句子的子单元进行对齐，对齐后得到两个句子的交互表示，然后对交互表示进行聚合，最后通过多层全连接层与非线性激活函数得到句子的匹配得分。例如，Parikh 等提出了 DecAtt 模型[7]，该模型使用注意力机制获取句子对齐后的交互表示，然后使用全连接神经网络来聚合对齐的表示。Chen 等在 DecAtt 的基础上提出了 ESIM 模型[8]，该模型使用 Bi-LSTM 编码输入句子和聚合句子的对齐表示。虽然基于交互的语义匹配模型在句子匹配任务上表现很好，但因其模型结构复杂、参数量大，训练需要大规模的标注数据及繁琐的超参数选择。

1.2 预训练模型

自 BERT 发布以来，学术界在改进遮掩策略、

修改预训练任务及面向特定任务的预训练方面做了很多努力。谷歌在原生 BERT 的基础上，发布了全词遮掩版 BERT WWM[3]，与原生 BERT 的遮掩 sub-word 不同，BERT WWM 以词的粒度进行遮掩。Facebook 提出的 SpanBERT[9] 不需要先验的词、实体、短语等边界信息进行遮掩，而是采取随机遮掩一段连续的 token，这种遮掩策略增加了模型的预测难度。Facebook 提出的 RoBERTa[10]，其实验证明 BERT 的下一句预测任务意义不大，在只训练遮掩语言模型的情况下，BERT 在下游任务的性能上有所提升。谷歌提出的 ALBERT[11] 引入了新的自监督任务——句子顺序预测（Sentence Order Predication，SOP），与 BERT 的下一句预测任务不同，句子顺序预测任务中同一文档的两个连续句子作为正样本，其顺序互换为负样本，该任务旨在建模两个输入句子的连贯性，以解决下一句预测任务低效的问题。

本文主要关注面向句子匹配任务的预训练方法，而最近有一些工作是面向特定下游任务的预训练模型。针对文本分类任务，AI2 提出了在目标领域和下游任务领域增量训练的方法[12]。实验表明，在目标领域和下游任务领域上继续进行预训练，下游任务的效果会有明显提升。针对机器阅读理解任务，IBM 提出了在维基百科上进行自监督学习的预训练方法：跨度选择预训练（Span Selection）[13]，实验表明，该预训练方法在 4 个阅读理解数据集上都得到了有效的提升。

2　背景

本节将简要介绍 BERT 的模型结构及其预训练阶段的自监督学习任务。

2.1　BERT 模型结构

BERT 由 L 层 Transformer[4] 堆叠而成，Transformer 层的隐层维度为 H，其中多头自注意力的头数为 A。BERT 的输入是两个“句子”拼接而成：[CLS]x_1，…，x_m，[SEP]y_1，…，y_m，[SEP]，输入序列的第一个 token 是特殊标记[CLS]，每个“句子”的末尾用特殊的分隔符[SEP]标记，其中第一个句子 x 的长度为 m，第二个句子 y 的长度为 n，$m+n\leqslant S$，其中 S 是模型允许的最大输入序列长度。

2.2　BERT 预训练任务

在预训练阶段，BERT 使用了两个预训练任务进行训练：遮掩语言模型和下一句预测。

（1）遮掩语言模型：BERT 随机选择输入序列中 15％的 token 进行“替换”，被选择的 token 中有 80％的概率被替换为特殊标记[MASK]，而 10％的概率保持不变，10％的概率被替换为词汇表中一个随机的 token。

（2）下一句预测：该任务的目标是预测两个“句子”在预训练语料中的顺序是否正确，是一个二分类任务。正负样本以等概率采样，其中正样本是两个来自同一文档的连续的句子，负样本通过采样两个不同文档中的句子得到。

3　基于动态词遮掩的句子匹配预训练

本节将详细介绍本文提出的基于动态词遮掩的句子匹配预训练方法，首先介绍如何获得句子匹配预训练模型所需的数据（“句子对”语料），然后介绍针对句子匹配预训练所采取的词遮掩策略，最后介绍模型的预训练过程。

3.1　预训练数据构造

在预训练阶段，BERT 通过下一句预测任务建模句子间关系，但从原理上没有直接建模两个句子的语义匹配关系。RoBERTa[10] 的实验表明，下一句预测任务的效果有限，去除该任务反而可以提高 BERT 在下游任务的性能。ALBERT[11] 的实验表明，下一句预测任务效果不佳的原因是其设置不合理且难度较小，该任务将连贯性预测和主题预测结合在了一起，但主题预测比连贯性预测简单得多，这使得模型在进行预测时仅依赖于主题建模。

BERT 的下一句预测任务得到的表示不适合应用于句子匹配任务。针对该问题，本文从缩小预训练阶段与下游任务之间数据分布的差异出发，提出了面向语义匹配的“句子对”预训练数据构造方法。面向句子匹配任务做预训练的关键点在于：如何构造“句子对”预训练数据，并使得预训练数据与下游任务数据的分布尽量接近。因此，针对句子匹配预训练，其预训练语料应该是语义相似的“句子对”语料，且领域与下游任务相似或接近。构造句子对预训练数据的流程如图 2 左侧所示，构造预训练数据的方法分为如下三步：

（1）将所有句子用预训练模型 BERT 进行编码，得到句子的向量表示。给定一个长度为 T 的句

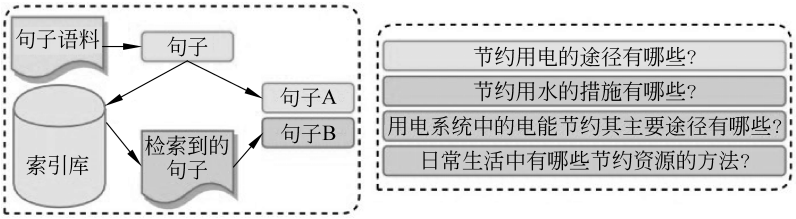

图2 “句子对”预训练数据构造

子序列 $x=[x_1,\cdots,x_T]$，经过BERT的多层Transformer编码后得到隐向量序列 $\boldsymbol{h}(x)=[h_1,\cdots,h_T]$，使用平均池化得到 H 维的向量 $\boldsymbol{h}=\text{MeanPooling}(h(x))$。

(2) 利用向量检索工具对所有句子的向量表示建立索引，以实现向量的快速最近邻搜索。

(3) 选取部分句子作为种子句子，根据种子句子的向量表示去检索相似的句子，从而构成语义相似的“句子对”。

为应对大规模向量检索的速度问题，本文使用快速近似最近邻搜索库Hnswlib作为向量检索工具，使用欧氏距离的平方作为向量的距离函数 $d(x,y)=\sum_{i=1}^{n}(x_i-y_i)^2$，其中 x，y 分别为两个句子的向量表示，n 为向量的维度。通过检索得到的问句对预训练数据示例如图2右侧所示，以句子A“节约用电的途径有哪些?”为例，我们检索到了3个语义相似的句子B，最后构造出3个“句子对”样本。我们爬取了近800万个来自百度知道的问句数据并构建索引库，随机取其中100万的问句作为种子句子，并利用这些种子问句去检索语义接近的相似问句，其中每个句子只检索出3至4个相似的句子。通过这种方式，最终构造了约400万个“问句对”样本作为预训练数据。

3.2 动态词遮掩策略

BERT的词遮掩策略随机遮掩输入序列中15%的token，所有的token都同等对待，随机进行遮掩。对于句子匹配任务来说，随机遮掩策略不能很好地建模句子关键信息，而关键信息是判断两个句子语义关系的重要特征。因此，建模句子的关键信息是句子匹配预训练模型的重点。

针对BERT随机遮掩策略的问题，本文提出了一种动态词遮掩策略：动态选择句子中的关键信息，优先遮掩句子中掩码恢复损失函数值高的关键词，在增加遮掩语言模型学习难度的同时，也能增强预训练模型对句子关键信息的建模能力。动态词遮掩策略细节如下：

(1) 利用Jieba(结巴)的TF-IDF算法抽取句子的关键词，每个句子最多抽取5个关键词。

(2) 对于句子中识别的所有关键词，利用预训练遮掩语言模型计算它们的掩码恢复损失值。细节如下：分别遮掩每一个关键词，然后将遮掩后的句子输入到遮掩语言模型，并预测这个被遮掩的关键词，最后通过交叉熵损失函数计算该关键词在掩码恢复任务上的损失值。给定一个长度为 T 的句子序列 $x=[x_1,\cdots,x_T]$，假设 $x_{i:j}$ 为句子中某个关键词序列，遮掩该关键词后的句子为 x'，通过BERT多层Transformer编码后最后一层的隐向量序列为 $h(x')=[h_1,\cdots,h_T]$，遮掩语言模型的交叉熵损失函数如式(1)所示，其中 $e(x)$ 为词 x 查表得到词嵌入的操作，V 为模型的词表。

$$L_{\text{MLM}}=-\log p(x_{i:j}\mid x')=-\sum_{t=i}^{j}\log p(x_t\mid x')$$
$$=-\sum_{t=i}^{j}\log\frac{\exp(h(x')_t^{\mathrm{T}}e(x_t))}{\sum_{x_k\in V}\exp(h(x')_t^{\mathrm{T}}e(x_k))} \quad (1)$$

(3) 对于所有识别的关键词，选择掩码恢复损失函数值大的关键词进行遮掩，在生成预训练数据时，优先遮掩句子中损失函数值前三的关键词。

为验证动态词遮掩策略的效果，本文比较了3种遮掩策略：①随机遮掩策略(RAN)：BERT的随机遮掩策略；②关键词遮掩策略(KW)：利用TF-IDF算法抽取关键词，优先遮掩关键词；③动态词遮掩策略(MLM)：在关键词遮掩策略的基础上，利用预训练的遮掩语言模型，优先遮掩句子中掩码恢复损失函数值高的关键词。以“节约用电的途径有哪些”为例，使用TF-IDF算法识别的关键词为“节约”“用电”“途径”和“哪些”，使用遮掩语言模型分别计算关键词的损失函数后，损失值前三的关键词为“节约”“用电”和“途径”。三种遮掩策略的对比结果如表1所示。

表 1　三种遮掩策略的对比

遮掩策略	遮掩后的句子
RAN	节 约 用 电 **的** 途 径 **有** 哪 些
KW	节 约 用 电 的 途 径 有 **哪 些**
MLM	节 约 **用 电** 的 途 径 有 哪 些

3.3　预训练过程

本文沿用 BERT 的遮掩语言模型作为预训练任务：将输入序列中的某些 token 替换为特殊标记[MASK]，遮掩语言模型预测这些被遮掩的 token 并利用交叉熵损失函数计算损失值，通过最小化掩码恢复损失训练模型参数。遮掩语言模型及其预训练如图 3 左侧所示。

参照过去的工作[12-14]，本文不从头开始训练 BERT，而使用预训练的模型权重进行初始化，并在此基础上使用遮掩语言模型任务及构造的“句子对”预训练数据进行预训练。预训练实验设置：batch size 为 768，最大序列长度为 64，学习率为 3e-5，优化器为 Adam，预训练轮数为 5。我们使用哈工大讯飞联合实验室发布的两个模型权重[14]：RBT3($L=3, H=768, A=12$，总参数量 38M)与 BERT base ($L=12, H=768, A=12$，总参数量 110M)，其中 RBT3 是 3 层版的 RoBERT，BERT base 为中文版的全词遮掩 BERT 模型。

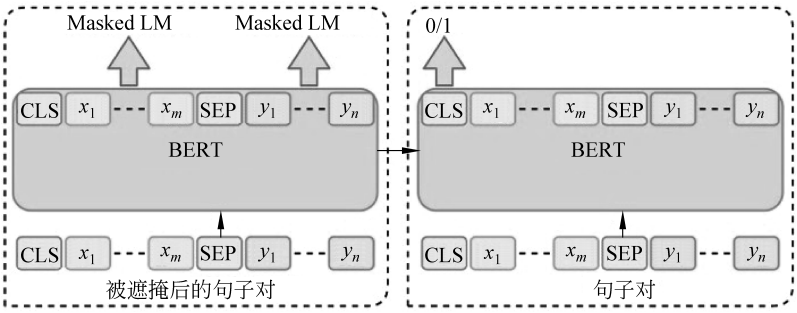

图 3　句子匹配预训练与下游任务微调示例

4　实验与分析

本节将详细描述句子匹配任务的实验数据、实验设置，并评估基于动态词遮掩的句子匹配预训练方法。

4.1　实验数据

我们在 4 个公开中文句子匹配数据集上进行实验，这 4 个数据集来源于 4 个不同领域，有着不同的特点，适合用来评估本文提出的预训练方法。句子匹配数据集的详细信息如表 2 所示。

表 2　4 个中文句子匹配数据集

数据集	领域	＃训练集	＃开发集	＃测试集
AFQMC	金融	34 334	4 316	—
LCQMC	开放域	238 766	8 802	12 500
BQ	银行	100 000	10 000	10 000
cMedQQ	医疗	16 071	1 793	1 935

(1) AFQMC：蚂蚁金融问句匹配数据集，来源于蚂蚁技术探索大会开发者竞赛[15]。

(2) LCQMC：大规模中文问句匹配数据集，来源于百度知道的问题数据[16]。

(3) BQ：银行客服领域问句匹配数据集，语料来自银行领域智能客服日志，并经过了筛选和人工的意图匹配标注[17]。

(4) cMedQQ：医疗领域问句匹配数据集，包含来自 5 个不同病种的问句对[18]。

4.2　实验结果

在句子匹配任务中，BERT 将输入的两个句子用特殊标记符[SEP]拼接为一个序列。在微调阶段，BERT 使用[CLS]位置最后一层的隐向量 $\boldsymbol{h}$ 作为两个句子的整体表示，最后在 BERT 的顶部添加一个简单的 Softmax 分类器来预测语义是否一致，并使用交叉熵作为损失函数，BERT 在句子匹配任务上微调，如图 3 右侧所示。使用本文的句子匹配预训练方法后，RBT3 和 BERT base 在 4 个问句匹配数据集上的实验结果如表 3 所示。下游任务实验设置：batch size 为 64，最大序列长度为 40，学习率为 3e-5，优化器为 Adam，训练轮数为 3。

表3 模型在4个数据集上的准确率

（单位：%）

模型	AFQMC	LCQMC	BQ	cMedQQ	AVG
S-LSTM	65.11	73.50	73.51	72.11	71.06
DecAtt	66.11	80.04	77.78	72.96	74.22
ESIM	69.20	86.26	81.45	82.38	79.82
RBT3	70.57	85.97	81.70	84.39	80.66
RBT3-Ours	**71.52**	**86.70**	**82.30**	**86.25**	**81.69（+1.03）**
BERT	73.47	85.94	84.29	86.61	82.58
RoBERTa	73.22	**87.38**	83.63	86.51	82.68
BERT-Ours	**73.86**	87.00	**84.30**	**87.60**	**83.19（+0.61）**

注：表中BERT和RoBERTa均为base版本。

为与传统句子匹配模型比较，我们选取Siamese-LSTM(S-LSTM)、DecAtt、ESIM等传统的句子匹配模型作为基线模型。实验结果表明：在4个问句匹配数据集上，相对于传统的句子匹配模型，RBT3、BERT base等预训练模型的效果全面优于传统句子匹配模型。

为验证本文提出的句子匹配预训练方法，我们选取RBT3、BERT base和RoBERTa base作为基线模型。

相对于RBT3模型，基于动态词遮掩的预训练模型在4个数据集上均取得最好的效果，平均准确率达到81.69%，提升了1.03%；在AFQMC、LCQMC、BQ、cMedQQ等4个数据集上准确率分别提升了0.98%、0.77%、0.60%、1.86%。在LCQMC上的准确率达到86.70%，超过了大模型BERT base的85.94%。

相对于BERT base模型，本文提出的方法在四个数据集上的平均准确率达到83.19%，提升了0.61%；在AFQMC、LCQMC、BQ、cMedQQ等4个数据集上准确率分别提升了0.39%、1.06%、0.01%、0.99%；在LCQMC数据集上，我们的方法达到了87.00%的准确率。

与RoBERTa base模型同等大小的BERT base在使用本文的预训练方法后，平均准确率比RoBERTa base高出0.51%，在AFQMC、BQ、cMedQQ等3个数据集上准确率分别高出0.64%、0.67%、1.11%，但在LCQMC上的效果比RoBERTa base略低（87.38%）。相对来说，RoBERTa使用了更多的预训练数据，并且其训练步数更多(RoBERTa使用了11G文本语料，训练了1百万步)。

4.3 实验分析

4.3.1 对比实验

为验证“句子对”预训练数据构造方法与动态词遮掩策略的有效性，我们进行了对比实验。RBT3和BERT base在4个问句匹配数据集上的遮掩策略对比实验结果如表4所示。为公平比较，三种遮掩策略的预训练数据与规模都一致，只有遮掩策略上的差异。三种待比较遮掩策略的详细说明如下：

表4 三种遮掩策略的准确率

（单位：%）

模型	AFQMC	LCQMC	BQ	cMedQQ	AVG
RBT3	70.57	85.97	81.70	84.39	80.66
+RAN	70.62	86.14	82.13	84.91	80.95（+0.29）
+KW	70.67	86.32	82.22	85.37	81.15（+0.51）
RBT3-MLM	**71.52**	**86.70**	**82.30**	**86.25**	**81.69（+1.03）**
BERT	73.47	85.94	84.29	86.61	82.58
+RAN	72.94	86.94	84.10	86.22	82.55（+0.03）
+KW	73.24	87.14	83.79	86.51	82.67（+0.09）
BERT-MLM	**73.86**	87.00	**84.30**	**87.60**	**83.19（+0.61）**

（1）+RAN(随机遮掩策略)：使用构造的“句子对”预训练数据，在预训练数据上使用BERT的随机遮掩策略。

（2）+KW(关键词遮掩策略)：在预训练数据上使用TF-IDF算法抽取句子的关键词，优先遮掩句子的关键词。

（3）+MLM(动态词遮掩策略)：在关键词遮掩策略基础上，利用预训练遮掩语言模型，优先遮掩句子中掩码恢复损失函数值高的关键词。

相对于RBT3模型，使用动态词遮掩策略后在4个数据集上均取得最好的效果，平均准确率提升了1.03%；在cMedQQ数据集上，本文提出的方法提升了1.86%。使用随机遮掩策略时，在4个数据集上都有提升，平均准确率提升0.29%，这说明面向

句子匹配的预训练数据构造方法是有效的。使用基于关键词的遮掩策略时，平均准确率比 RBT3 提升了 0.51%，在 cMedQQ 数据集上平均准确率提升了 0.98%，在 BQ 上提升了 0.52%，这说明在预训练阶段增强对关键词信息的建模能力有助于句子匹配任务。

相对于 BERT base 模型，使用动态词遮掩策略后在 4 个数据集上的平均准确率提升了 0.61%；在 LCQMC 与 cMedQQ 数据集上，平均准确率比 BERT base 分别提升 1.06% 和 0.99%。使用随机遮掩策略与基于关键词的遮掩策略时，虽然二者的平均准确率比 BERT base 提升有限，但在 LCQMC 数据集上分别提升了 1.00%、1.20%。相比小模型 RBT3，训练 BERT base 需要更多的训练数据与训练轮数，在数据量和训练轮数与 RBT3 一致的情况下，在 BERT base 上的提升不大。

4.3.2　实例分析

表 5 显示了在 LCQMC 测试集上的实例分析结果。实例 1 中句子 A 的关键信息为“百度”，句子 B 的关键信息为“百度云”，RBT3 将其错分为正样本，而本文的方法捕捉到了两个句子关键信息的不匹配并预测正确。实例 2 中句子 A 的关键信息为“种什么”，句子 B 为“办什么厂”，RBT3 将其错分为正样本，而本文的方法捕捉到了关键信息的不匹配并预测正确。实例 3 中句子 A 的关键信息为“红酒”，而句子 B 为“葡萄酒”，RBT3 将其错分为负样本，而本文的方法捕捉到了关键信息的多种表达方式并预测正确。实例 4 中句子 A 的关键信息为“读”，而句子 B 为“看”，这两个词在这两个句子的上下文中应该表达相同语义，RBT3 将其错分为负样本，而本文的方法捕捉到了关键信息在上下文中的多种表达方式并预测正确。实例分析结果表明，在需要关键信息帮助判断两个句子的语义关系时，相比 RBT3，基于动态词遮掩的句子匹配预训练模型确实能捕捉到两个句子在关键信息上的异同。

表 5　LCQMC 测试集上的实例分析

ID	句子 A	句子 B	RBT3	Ours
1	**百度**最近怎么了	最近**百度云**怎么了	1	**0**
2	现在农村**种什么**最赚钱	农村现在**办什么厂**最赚钱	1	**0**
3	哪些食物不能和**红酒**一起吃	哪些食物不能和**葡萄酒**一起吃	0	**1**
4	怎么**读**燃气表	燃气表怎么**看**	0	**1**

5　总结与展望

针对 BERT 在预训练阶段不能建模句子的语义匹配信息及随机遮掩策略无法高效建模句子关键信息的问题，本文提出了基于动态词遮掩的句子匹配预训练方法：通过构造与句子匹配任务数据分布接近的预训练数据，用动态词遮掩策略建模句子关键信息。在 4 个中文句子匹配数据集上的实验表明，使用本文提出的“句子对”预训练数据构造方法与动态词遮掩策略，RBT3 和 BERT base 模型的效果都有一定的提升（准确率分别提升 1.03% 和 0.61%）。同时，对比实验也证明了“句子对”预训练数据构造方法与动态词遮掩策略的有效性。

以后的工作中，更好的预训练数据构造方式与高效建模句子的关键信息是我们的重点，如以端到端的方式构造预训练数据、使用预训练模型直接提取句子的关键信息是以后的研究方向。

参考文献

[1] Lan W, Xu W. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering[C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018: 3890-3902.

[2] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training [EB/OL]. [2018-10-12]. https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[3] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 2019: 4171-4186.

[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, 2017, 30: 5998-6008.

[5] Huang P S, He X, Gao J, 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.

[6] Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity[C]//Proceedings

of the 13th AAAI Conference on Artificial Intelligence, 2016: 2786-2792.

[7] Parikh A, Täckström O, Das D, et al. A decomposable attention model for natural language inference [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016: 2249-2255.

[8] Chen Q, Zhu X, Ling Z H, et al. Enhanced LSTM for natural language inference [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1657-1668.

[9] Joshi M, Chen D, Liu Y, et al. SpanBERT: Improving pre-training by representing and predicting spans [J]. Transactions of the Association for Computational Linguistics, 2020, 8: 64-77.

[10] Liu Y, Ott M, Goyal N, et al. RoBERTA: A robustly optimized BERT pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[11] Lan Z, Chen M, Goodman S, et al. ALBERT: A Lite BERT for self-supervised learning of language representations[C]//Proceedings of the 8th International Conference on Learning Representations, 2020.

[12] Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: Adapt language models to domains and tasks [J]. arXiv preprint arXiv: 2004.10964, 2020.

[13] Glass M, Gliozzo A, Chakravarti R, et al. Span selection pre-training for question answering[J]. arXiv preprint arXiv:1909.04120, 2019.

[14] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for Chinese BERT[J]. arXiv preprint arXiv:1906.08101, 2019.

[15] Xu L, Zhang X, Li L, et al. CLUE: A Chinese language understanding evaluation benchmark[J]. arXiv preprint arXiv:2004.05986, 2020.

[16] 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.

[17] Chen J, Chen Q, Liu X, et al. The BQ corpus: A large-scale domain-specific Chinese corpus for sentence semantic equivalence identification [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 4946-4951.

[18] Zhang N, Jia Q, Yin K, et al. Conceptualized representation learning for Chinese biomedical text mining [J]. arXiv preprint arXiv:2008.10813, 2020.

宋挺(1996—),硕士研究生,主要研究领域为语义匹配、问答系统和自然语言处理。
E-mail: ting.song@nlpr.ia.ac.cn

何世柱(1987—),博士,副研究员,主要研究领域为知识图谱和自然语言处理。
E-mail: shizhu.he@nlpr.ia.ac.cn

郭展成(1996—),硕士研究生,主要研究领域为智能问答、知识图谱和自然语言处理。
E-mail: zhancheng.guo@nlpr.ia.ac.cn