计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(11):349−357 [doi: 10.15888/j.cnki.csa.008778]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于增量预训练和对抗训练的文本匹配模型①

司志博文[1,2], 李少博[2], 单丽莉[2], 孙承杰[2], 刘秉权[2]

[1](人民网 传播内容认知国家重点实验室, 北京 100733)
[2](哈尔滨工业大学 计算学部, 哈尔滨 150001)
通信作者: 刘秉权, E-mail: liubq@hit.edu.cn

摘　要: 文本匹配是自然语言理解的关键技术之一, 其任务是判断两段文本的相似程度. 近年来随着预训练模型的发展, 基于预训练语言模型的文本匹配技术得到了广泛的应用. 然而, 这类文本匹配模型仍然面临着在某一特定领域泛化能力不佳、语义匹配时鲁棒性较弱这两个挑战. 为此, 本文提出了基于低频词的增量预训练及对抗训练方法来提高文本匹配模型的效果. 本文通过针对领域内低频词的增量预训练, 帮助模型向目标领域迁移, 增强模型的泛化能力; 同时本文尝试多种针对低频词的对抗训练方法, 提升模型对词级别扰动的适应能力, 提高模型的鲁棒性. 本文在 LCQMC 数据集和房产领域文本匹配数据集上的实验结果表明, 增量预训练、对抗训练以及这两种方式的结合使用均可明显改善文本匹配结果.

关键词: 文本匹配; 预训练模型; 增量预训练; 对抗训练; 低频词; 深度学习; 自然语言处理

引用格式: 司志博文,李少博,单丽莉,孙承杰,刘秉权.基于增量预训练和对抗训练的文本匹配模型.计算机系统应用,2022,31(11):349–357. http://www.c-s-a.org.cn/1003-3254/8778.html

Text Matching Model Based on Incremental Pre-training and Adversarial Training

SI Zhi-Bo-Wen[1,2], LI Shao-Bo[2], SHAN Li-Li[2], SUN Cheng-Jie[2], LIU Bing-Quan[2]

[1](State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing 100733, China)
[2](Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Text matching is one of the key techniques in natural language understanding, and its task is to determine the similarity of two texts. In recent years, with the development of pre-trained models, text-matching techniques based on pre-trained language models have been widely used. However, these text matching models still face the challenges of poor generalization ability in a particular domain and weak robustness in semantic matching. Therefore, this study proposes an incremental pre-training and adversarial training method for low-frequency words to improve the effect of the text matching model. The incremental pre-training of low-frequency words in the source domain helps the model migrate to the target domain and enhances the generalization ability of the model. Additionally, various adversarial training methods for low-frequency words are tried to improve the model's adaptability to word-level perturbations and the robustness of the model. The experimental results on the LCQMC dataset and the text-matching dataset in the real estate domain indicate that incremental pre-training, adversarial training, and the combination of the two approaches can significantly improve the text matching results.

Key words: text matching; pre-trained model; incremental pre-training; adversarial training; low-frequency word; deep learning; natural language processing (NLP)

① 基金项目: 国家自然科学基金 (62176074)
收稿时间: 2022-01-29; 修改时间: 2022-02-24, 2022-03-11; 采用时间: 2022-04-02; csa 在线出版时间: 2022-07-29

文本匹配是自然语言理解的关键技术之一, 通常可以将研究两段文本之间关系的问题看作文本匹配问题, 本文主要研究两个文本的语义相似问题. 传统的文本匹配方法如 TFIDF[1]、BM25[2] 等, 它们基于词频统计特征计算文本与文本之间的相似度, 但是这些方法并没有利用到文本的上下文信息和语义信息. 近些年来, 深度学习在文本语义匹配中起到重要作用. 基于深度学习的文本匹配方法主要分为两种, 基于表示的文本匹配方法[3] 和基于交互的文本匹配方法[4]. 基于表示的文本匹配模型以经典双塔模型为代表, 通过将文本表示成向量计算文本之间的相似度; 而基于交互的文本匹配模型通过不同策略构建匹配特征, 尽可能保留重要的相似句子信息. 此外, 随着预训练模型 BERT[5] 的广泛应用, 文本匹配的效果有了明显改善[6]. 但是, 目前的预训练模型通常采用通用领域的语料进行训练, 特定数据进行微调, 而特定数据的领域与预训练领域若存在较大差异, 会导致模型对目标领域样本的匹配效果不理想; 此外, 在现有的基于预训练模型的文本匹配模型中, 文本格式的高度相似导致模型的泛化能力变弱.

针对上述问题, 本文在一个基于文心 (ERNIE[7]) 的基线文本匹配模型, 利用逆文档频率 (IDF) 提取文本匹配的低频词信息, 然后对领域样本中低频但信息量大的词进行增量预训练和对抗训练以提高文本匹配的效果. 一方面, 通过对低频词进行全词掩码 (whole word MASK) 的形式, 对特定领域的语料进行增量预训练, 鼓励模型更多的重视文本中低频词带来的信息, 并改善模型对特定领域文本的编码表示; 另一方面, 通过使用对抗训练, 例如 FGM (fast gradient method)[8] 等, 对低频词的词向量梯度施加扰动, 提高模型对词级别扰动的适应能力, 相当于变向地对低频词进行数据增强. 在两个不同领域数据集上的实验验证表明, 上述两个策略可以提升文本匹配性能, 增强模型的泛化能力和鲁棒性.

1 相关工作

1.1 文本匹配

文本匹配是自然语言理解的关键技术之一, 它研究两个文本语义相似程度的问题. 传统的文本匹配方法利用句法信息、统计信息等浅层特征进行相似度的计算, 例如 Song 等人[9] 利用生成词语的概率来计算文本之间的相似度, 然而这些基于浅层语义特征的匹配方式难以处理复杂文本的匹配. 近些年随着深度学习的发展, 各种神经网络模型被用于文本匹配任务中. 例如, 微软提出了 DSSM[3], 将两个文本映射成低维稠密向量, 然后通过多层感知机进行编码, 最终使用余弦相似度来计算文本向量的语义匹配程度; Chen 等人[10] 综合 Bi-LSTM 和注意力机制, 有效提取了文本之间蕴含的全局信息和局部信息, 提出了 ESIM 模型; Wang 等人[11] 提出了 BiMPM, 其最大的创新点是双向多角度匹配, 解决了文本匹配模型交互不充分的问题. 这些深度学习模型虽然能够充分利用语义交互信息进行相似度计算, 但仍使用词向量 Word2Vec[12] 等方式进行句子的编码, 语义信息表达不够充分, 使得一词多义的情况没有解决. 近些年, 基于预训练模型的文本表示方法逐渐成为主流, 谷歌提出 BERT 模型[5], 刷新了各项自然语言处理任务的榜单; Reimers 等人[13] 提出了 sentence-BERT 模型, 利用孪生网络结构在保证准确性的同时大大提高了速度. 上述基于预训练的文本匹配模型大多采用了“预训练+微调”的形式, 即首先在大量无监督语料进行训练, 然后针对自己的数据集进行微调, 这种模式在针对特定领域时, 例如房产领域, 微调效果提升有限. 对此, Gururangan 等人[14] 设计了多个领域的分类任务, 进行领域自适应训练 (DAPT). 实验表明, 对于特定任务可以用任务相关数据再对语言模型做二次预训练, 能有效提高模型性能; Gu 等人[15] 将这种训练方式总结出一个通用的范式, 即“预训练-领域增量训练-微调”.

本文在增量预训练时, 采用的掩码策略并非随机选取词语, 而是遮盖文本匹配中的低频词, 使得模型根据上下文去预测这个词, 有效提高了模型特定领域的泛化性.

1.2 对抗训练

对抗训练是一种引入噪声的训练方式, 在训练过程中人为地在样本中混入一些微小的扰动, 它对于模型参数改变很小, 但是却可以造成误分类. 因此, 对扰动后的对抗样本进行训练, 能够增强模型的泛化能力[16]. Miyato 等人[8] 提出了快速梯度对抗方法 (FGM) 的对抗训练方式, 对词向量层的参数施加扰动, 取得了很好的效果; 但由于 FGM 是一次性施加扰动, 不能找到最优的扰动范围, 所以 Madry 等人[17] 提出了投影梯度下降 (projected gradient descent, PGD) 的对抗训练方式. 该方法采取迭代式的 k 步扰动策略, 虽然对抗扰动

效果更好, 但是对抗训练 k 次消耗了更多计算资源, 训练速度大大减缓; Shafahi 等人[18] 综合两者特点提出了自由对抗训练 (free adversarial training, FreeAT), 在 PGD 的基础上进行训练速度的优化, 找到了最优的对抗训练的参数. 然而这些对抗训练的方式都对整个词表扰动, 针对性不强, 本文在微调时针对低频词有目的性地进行对抗训练以增强模型的鲁棒性, 使得模型更好表达语义匹配中低频词的信息.

2 基于增量预训练和对抗训练的文本匹配模型

在本节中, 我们首先介绍了基于 ERNIE[7] 的文本匹配基线模型, 随后给出了基于低频词掩码的增量预训练方法和基于低频词对抗训练的方法, 并说明两种方法怎样融合来实现文本匹配.

2.1 基于 ERNIE 的文本匹配基线模型

本节介绍基线文本匹配模型, 采用了百度发布的文心 (ERNIE[7]) 预训练模型作为基础模型, 它相较于 BERT[5]、RoBERTa[19] 等模型, 采用了知识掩码的训练策略, 一次掩码 (MASK) 一个中文的实体和短语, 能够取得更好的性能, 具体的匹配模型如图 1 所示.

在进行文本匹配任务时, 首先将待匹配的两段文本拼接成一个句子, 在句子头部添加 [CLS] 符号, 在文本之间添加 [SEP] 分隔符, 作为模型的输入.模型使用 ERNIE[7] 对文本的语义特征进行编码, 最终的输出T_i是一个 768 维的向量. 在得到输出后, 本文使用句首标识符 [CLS] 对应的向量作为整个句子的向量表征; Reimers 等人[13] 探究以平均池化和最大池化后的向量代表整个句子的向量. 因为最大池化可以表征一个句子中关键的语义特征信息, 而平均池化则可以最大程度保留文字的背景信息. 所以在基线文本匹配模型的构建中, 我们将 [CLS] 标签代表的句向量$T_{[CLS]}$、平均池化向量T_{avg}和最大池化向量T_{max}拼接得到向量T_{all}, 如式 (1) 所示:

$$T_{all} = Concat\left(T_{[CLS]}, T_{avg}, T_{max}\right) \tag{1}$$

将T_{all}通过多层感知机建模得到包含各个特征维度的语义向量T_m, 如式 (2) 所示, 利用T_m进行文本匹配:

$$T_m = Linear\left(T_{all}\right) \tag{2}$$

在获得文本匹配的基线模型后, 我们希望将特定任务的数据集进行增量的预训练, 将特定领域的语义知识嵌入到通用领域的预训练模型中.

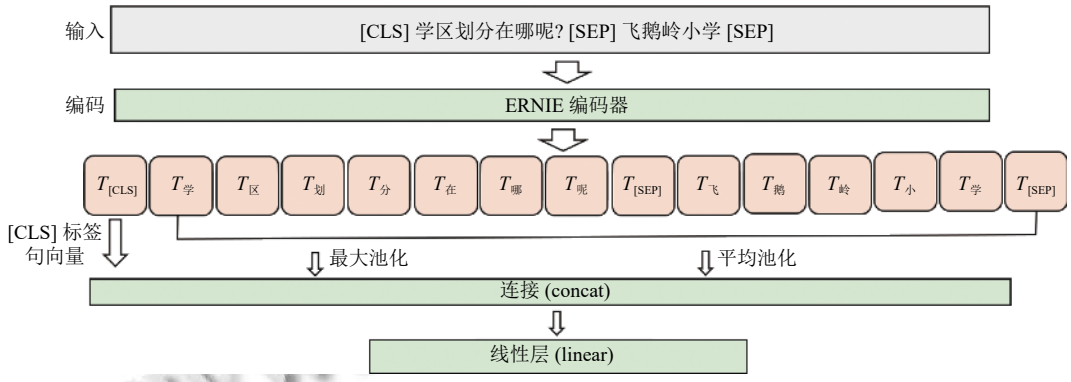

图 1 文本匹配基线模型

2.2 低频词增量预训练

原始的预训练模型 BERT[5] 是在通用领域语料上无监督训练, 其训练任务有两个, 第一个是掩码语言模型 (MLM), 第二是下一句预测 (NSP). 因为 Liu 等人[19] 实验发现 NSP 任务对于提升下游任务的性能并没有太大作用, 我们只选取了 MLM 任务, 如图 2 所示.

由于 ERNIE 模型[7] 是在通用领域上训练的, 可以表达中文基本语法语义, 但是对于特定领域例如房产、娱乐、体育等领域的数据集进行直接微调 (fine-tune) 效果有时候不理想. 因此, 为了更好地编码特定领域的数据集, 本文提出了针对文本匹配中低频词的增量预训练方法.

首先对特定领域数据集进行预处理, 按照 BERT 原文[5] 的方式, 将数据集复制 10 份, 对每条训练语句进行全词掩码. 因为随机掩码的字词可能没有意义, 例如 “的”. 所以, 我们使用 IDF 对分词后的语料进行打分, 掩码 IDF 值高的词语, 以强化模型对语料中低频但信息量大词的语义表示. IDF 计算如式 (3) 所示:

$$IDF_w = \log \frac{N}{1+\sum_{i=1}^{N} I(w, D_i)} \tag{3}$$

其中, w是词语, N是语料库中的文档总数, $I(w, D_i)$表示文档D_i是否包含词语w, 为了防止分母为 0, 选取了加一平滑.

其次, 在进行增量预训练时, 将每个句子中 15% 的低频词会被掩码, 这其中 80% 的低频词被替换为 [MASK], 10% 用词表中其他词替换, 10% 的词语保持不变, 保证增量模型能够充分学习到中文短句级别的语义上下文关系; 最后, 训练时使用多分类交叉熵损失函数, 只对被掩码的词语计算损失值并进行反向传播, 具体如式 (4) 所示:

$$L = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic}) \tag{4}$$

其中, M 是类别的数量, p_{ic}是观测样本i属于类别c的预测概率, y_{ic}是指示函数, 其形式如式 (5):

$$y_{ic} = \begin{cases} 1, & i = c \\ 0, & i \neq c \end{cases} \tag{5}$$

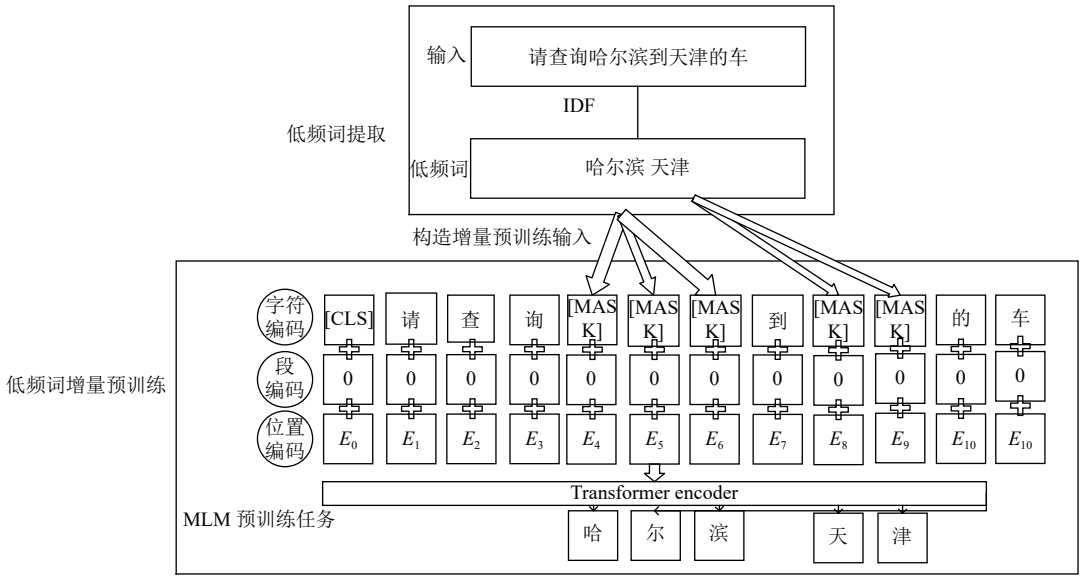

图 2 低频词增量预训练

2.3 低频词对抗训练

常见的文本对抗训练的方式都是在模型的词向量层对梯度参数的扰动, 输入是模型反向传播的梯度g, 目的是在g的基础上添加扰动向量r_{adv}. 本文中我们实验了 3 种对抗训练的方式, 分别是 FGM、PGD 和 FreeAT, 它们的优缺点如表 1 所示. 具体的梯度扰动如式 (6) 和式 (7) 所示:

$$g = \nabla_x L(\theta, x, y) \tag{6}$$

$$r_{\text{adv}} = \epsilon \times \frac{g}{\|g\|_2} \tag{7}$$

其中, g为梯度向量, ∇_x求损失函数的梯度, L是损失函数, θ是模型参数, x是输入, y是目标标签, 是扰动步长, $\|g\|_2$是向量 g 的二范数.

表 1 对抗训练方法对比

优/缺点	FGM	PGD	FreeAT
优点	梯度扰动方式直接且有效	k步迭代, 找到最优的扰动参数扰动范围过大则映射到规定范围	保证梯度的最优扰动参数同时加快训练速度
缺点	需要损失函数局部线性无法得到最优的扰动步长	训练速度慢消耗大量计算资源	扰动参数相较于上一步最优, 相较于全局次优

于是, 本文在下游微调阶段对 ERNIE 模型[7] 的词向量层采取了有针对性的对抗训练的方式. 本文采用 IDF 分值较大的词作为需要对抗训练的词, 因为这些词大部分是一些低频词且与领域强相关, 对这些词的扰动相当于做数据增强以缓解词语过少带来的欠拟合, 从而提升了文本匹配任务的鲁棒性. 具体的流程如算

法 1 所示.

算法 1. 低频词对抗训练伪代码

输入: 训练集合X, 训练轮数N_{ep}, 扰动步长ϵ, 学习率τ
1. 初始化模型参数θ
2. 扰动向量$\delta \leftarrow 0$
3. for $epoch=1,\cdots,N_{ep}$ do
4. for minibatch $B \in X$ do
5. 用随机梯度下降更新参数
6. $g_\theta \leftarrow \mathbb{E}_{(x,y)\in B}\left[\nabla_x l(\theta,x+\delta,y)\right]$
7. $g_{\text{adv}} \leftarrow \nabla_x l(\theta,x+\delta,y)$
8. $\theta \leftarrow \theta-\tau\cdot g_\theta$
9. 更新扰动向量
10. $\delta \leftarrow \delta+\epsilon\cdot g_{\text{adv}}\cdot M$
11. end for
12. end for

算法 1 中的 x 代表输入, y 代表目标标签, $\mathbb{E}_{(x,y)\in B}$ 代表在此 minibatch 下最优化模型梯度参数g_θ, ∇_x代表求损失函数l的梯度. 对于原始输入文本, 经过嵌入层和编码层后计算损失函数, 得到反向传播后的梯度. 在词向量的梯度上添加针对低频词的对抗扰动, 即将算法第 11 行的梯度g_{adv}与低频词掩码 (MASK) 矩阵M相乘, 生成新的扰动向量, 将叠加扰动向量后的输入通过模型编码层计算产生的损失函数重新进行反向传播, 更新模型参数.

2.4 融合增量预训练和对抗训练的文本匹配

本节主要介绍如何将第 2.2 节增量预训练和第 2.3 节对抗训练的内容融合在一起, 形成一个完整的文本匹配.

如图 3 上半部所示, 增量预训练是文本匹配模型训练的前置部分, 利用语料中低频词增量后的模型改善了预训练模型对于某一特定领域的语言表示, 增强了模型的泛化能力, 得到包含领域知识的模型 CONTINUAL-ERNIE. 在后续文本匹配时, 使用的也是这个模型.

对抗训练则是与文本匹配模型一起训练, 在训练过程中通过对低频词的词向量施加扰动, 学习这些领域低频词带来的重要信息, 从而增强模型的鲁棒性, 具体如图 3 字符编码处所示.

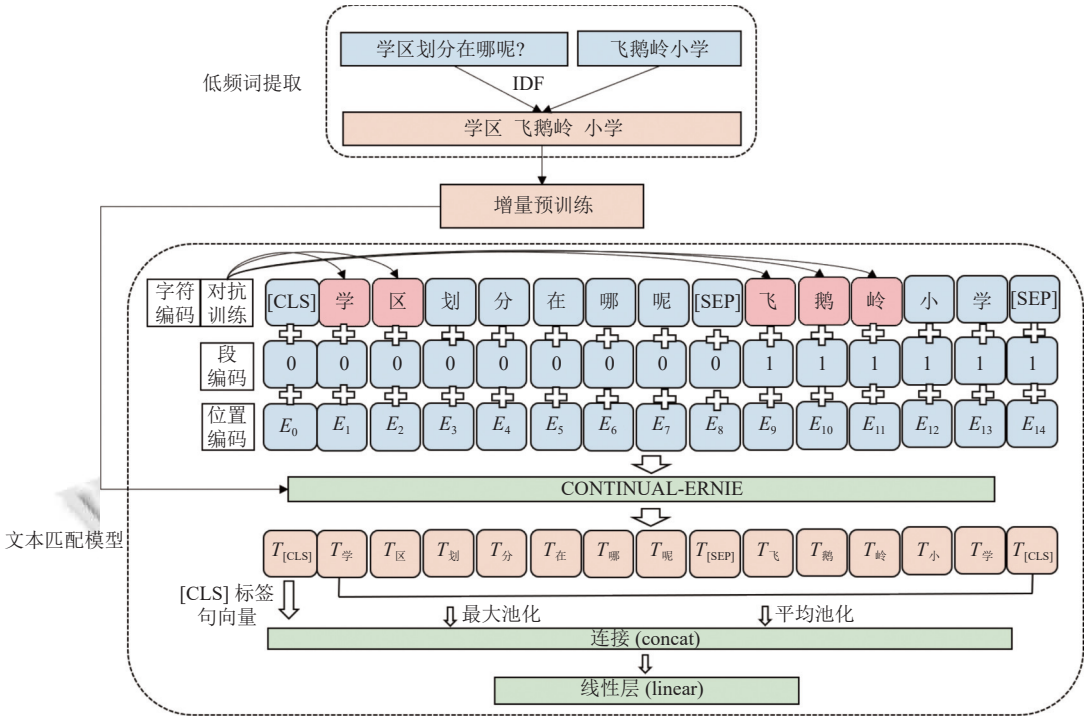

图 3 融合增量预训练和对抗训练的文本匹配

3 实验与分析

3.1 数据集介绍

为了验证方法的效果, 本文选取了 LCQMC[20] 的领域中意图匹配数据和房产问答匹配的数据, 皆为工业界文本匹配的数据集.

本文选取的 LCQMC 是中文问题匹配的语料库,

它更侧重两个问句的意图匹配, 是中文文本匹配经典的语料库之一. 该数据集包含了 260 068 个人工标注的问题对且已划分好训练集、验证集和测试集, 具体情况如表 2 所示.

表 2 LCQMC 数据统计

统计数据	训练集	验证集	测试集
问题对数量	238 766	8 802	12 500
正例	138 574	4 402	6 250
负例	100 192	4 400	6 250
平均长度	24.79	27.89	22.42

房产聊天问答匹配数据集是由 CCF 大数据与计算智能大赛在 2020 年发布的一个赛题的数据. 该数据集属于短文本匹配, 一共包含 6 000 组对话, 21 583 条问答对, 本文对问答对划分了训练集、验证集和测试集, 具体情况如表 3 所列.

表 3 房产问答数据统计

统计数据	训练集	验证集	测试集
问答对数量	15 108	2 158	4 317
正例	3 770	539	1 077
负例	11 338	1 619	3 240
平均长度	18.24	18.19	18.29

本数据集采用的评价指标为精确率、召回率和 $F1$ 值 3 项, 如式 (8)–式 (10) 所示:

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

$$R = \frac{TP}{TP+FN} \tag{9}$$

$$F1 = \frac{2\times P\times R}{P+R} \tag{10}$$

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

3.2 实验设置

3.2.1 实验参数

本文实验在 CentOS 7 系统上采用 RTX 3090 GPU 进行模型的训练和调试, 使用 PyTorch 框架[21] 对相关模型编码实现, 使用 Huggingface 公司公开的 ERNIE、chinese-RoBERTa-wwm 和 BERT-wwm 三个预训练模型进行实验. 模型的主要参数设置: batch size 设置为 32, 学习率调整为 2×10^{-5}, 并采用线性学习率预热的策略. 对于 ESIM 模型, 词向量为 300 维, 隐藏层为 256 维, dropout 设置为 0.2, 学习率设置为 0.001.

3.2.2 实验对比模型

本文对比了 ERNIE[7] 其他预训练模型以及 ESIM 模型的效果. 此外本文在 ERNIE 模型[7] 基础上进行了基于低频词的增量预训练及对抗训练, 具体的实验设置如下:

(1) ESIM 模型[10], 该方法利用 Bi-LSTM 和注意力机制进行文本之间的信息交互, 提升了模型对于文本全局信息和局部信息的捕获能力.

(2) BERT 系列模型, 包含 BERT、RoBERTa 和 ERNIE, 这些模型利用了 Transformer 结构强大的文本特征的表达能力, 是文本匹配的基础模型.

(3) ERNIE+增量预训练模型, 对 ERNIE 模型进行了低频词掩码的增量预训练, 对比了该种方式相对于随机全词掩码的效果.

(4) ERNIE+对抗训练模型, 通过选取 IDF 分值排名前 25%、50% 和 75% 的低频词进行对抗训练, 探究基于低频词对抗训练的文本匹配模型收益.

3.3 实验结果与分析

(1) ERNIE 模型的有效性: 本文在 LCQMC 和房产问答匹配数据集进行基线模型的对比实验, 具体结果如表 4 所列. LCQMC 问题匹配和房产领域问答匹配的实验结果显示, ERNIE 的 $F1$ 达到了 87.63% 和 79.95%, 高于其他预训练模型. 这印证了 ERNIE 模型在预训练时采用的实体和短句级别的掩码是有效的, 因此本文在 ERNIE 模型的基础上进行了其他方法的实验.

(2) 模型的鲁棒性: 在设置对抗训练的实验时, 本文选取了 IDF 分值高的前 25%、50% 和 75% 的词进行了对抗训练, 其结果如图 4 和图 5 所示.

表 4 预训练模型实验对比 (%)

预训练模型	LCQMC问题匹配			房产领域问答匹配		
	P	R	$F1$	P	R	$F1$
ESIM	77.26	93.44	84.58	70.23	74.42	72.26
BERT_wmm	79.47	95.63	86.80	74.18	81.75	77.78
RoBERTa_wmm	79.61	**95.91**	87.03	77.33	**82.36**	79.77
ERNIE	**80.90**	95.57	**87.63**	**77.85**	82.17	**79.95**

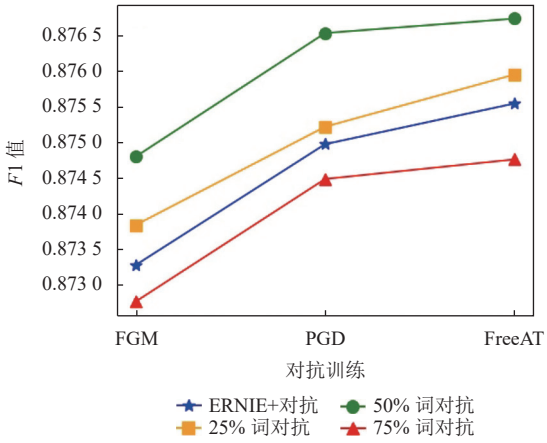

图 4 LCQMC 数据集的低频词对抗训练

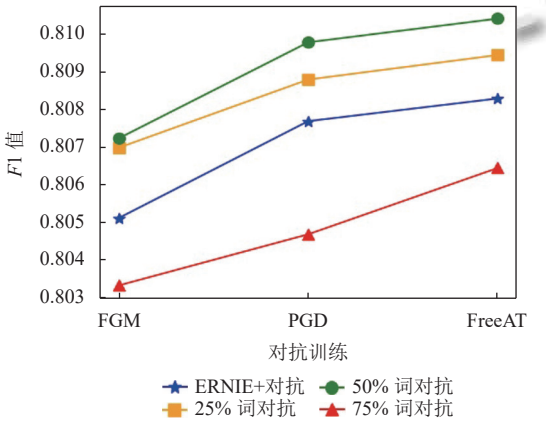

图 5 房产问答匹配的低频词对抗训练

图 4 和图 5 横轴是对抗训练的方法, 纵轴是 $F1$ 值. 分析可知, 选取 25% 和 50% 的低频词进行对抗训练, 可使模型在对抗训练的基础上再次提高, 这是因为低频词虽然频率低, 却是领域内“代表词”, 针对这类词语的对抗扰动大大增强了模型的鲁棒性. 例如基线模型对于问答对“学区划分在哪呢?”“飞鹅岭小学”识别困难, 但经过低频词“飞鹅岭”的对抗训练后, 模型成功将问答匹配.

(3) 模型特定领域的泛化能力: 本文还实验了在领域增量预训练时对低频词进行掩码的策略, 并将对抗训练和增量预训练两种方式结合在一起进行实验. 其结果如表 5 和表 6 所示. 由表 5 可知针对低频词的增量预训练 $F1$ 值相比 Baseline 分别提升了 0.65% 和 0.40%, 相比于普通增量预训练有 0.13% 和 0.09% 的提升. 这些都印证了本文的假设: 使用 IDF 提取出的词为低频但信息量大的词语, 更容易使得模型出错, 所以对低频词的增量预训练增强了模型泛化能力.

(4) 模型泛化能力和鲁棒性的结合: 如表 6 所示, 本文选取效果最好的 50% 词对抗, 将上述两种方法结合进行了实验, 低频词增量+FreeAT 的训练取得了最好的效果, 其 $F1$ 值在两个数据集分别提升了 0.23% 和 0.28%. 此外, 对于包含对抗训练的每组实验, 我们设置了 4 个随机种子数进行多次实验, 以多次实验的平均结果作为该组实验的最终结果, 这样做的目的排除了因参数初始化不同带来的偶然性.

对比两个数据集的实验结果, 可以发现相同实验设置条件下实验结果差别较大. 这是因为 LCQMC 数据集更侧重于意图的匹配, 两个匹配的句子含义相近, 例如“上班族, 做什么兼职好呢?”“上班族兼职做什么好?”. 房产数据则是问答对的匹配, 例如“房子几年了?”“两年了”. 可以明显感觉到, LCQMC 数据更加规整, 而房产数据则侧重于问答语义的流畅性, 任务难度相对大, 造成了整体结果相较于 LCQMC 偏低.

表 5 低频词增量预训练结果 (%)

增量预训练	LCQMC问题匹配			房产领域问答匹配		
	P	*R*	*F*1	*P*	*R*	*F*1
ERNIE-baseline	80.90	95.57	87.63	77.85	82.17	79.95
ERNIE 增量	81.54	95.92	88.15	**79.91 (+2.06)**	82.63	81.24
ERNIE 低频词增量	**81.63 (+0.73)**	**96.11 (+0.54)**	**88.28 (+0.65)**	79.87	**82.89 (+0.72)**	**81.35 (+0.40)**

表 6 ERNIE 增量+50% 词对抗 (4 次取平均) (%)

对抗训练	LCQMC问题匹配			房产领域问答匹配		
	P	*R*	*F*1	*P*	*R*	*F*1
ERNIE低频词增量	81.63	96.11	88.28	79.87	82.89	81.35
ERNIE低频词增量+FGM	81.61	96.41	88.39	**79.92 (+0.05)**	82.85	81.36
ERNIE低频词增量+PGD	81.64	96.48	88.43	79.46	83.56	81.46
ERNIE低频词增量+FAT	**81.74 (+0.11)**	**96.50 (+0.39)**	**88.51 (+0.23)**	79.33	**84.07 (+1.18)**	**81.63 (+0.28)**

同样使用 LCQMC 数据集, 本文模型 88.51% 的 *F*1 值高于 ISTM 模型的 86.30%[22] 和语义正交化匹配模型的 88.43%[23].

基于上述的实验结果及分析, 我们的模型通过利用领域内的低频词, 为下游文本匹配任务带来了巨大性能上的提升.

4 结论与展望

本文构建了基于低频词的增量预训练和对抗训练的文本匹配模型. 通过 IDF 提取文本中的低频词, 利用低频词掩码的形式进行增量预训练, 强化了模型对于特定领域数据的适配性, 提升了模型的泛化能力; 同时, 本文还将针对低频词的对抗训练融入到模型的训练过程中, 在模型的词向量编码阶段施加扰动, 探索了最优的扰动对抗方式, 有效提升了模型的鲁棒性. 增量预训练使整个模型向特定领域迁移, 而对抗训练更好地表征了低频词的词嵌入向量, 通过全局增量预训练和局部对抗训练的结合使模型效果有了更进一步提升, 在两个不同领域数据集的实验证明了本文方法的有效性.

对于文本匹配模型, 针对低频领域相关词进行增量预训练及对抗训练有效提升了文本匹配的效果. 除了选取低频领域相关词语外, 还可探究其他词语选取的方式进行增量预训练, 例如提取易使模型分类错误的词语; 此外, 如何在对抗训练中找到最优扰动的同时提高训练效率, 这也是一个值得研究的问题.

参考文献

1 唐贤伦, 李佳歆, 万亚利, 等. 基于条件随机场和 TF-IDF 的文本语义匹配及推荐. 第 28 届中国过程控制会议 (CPCC 2017) 暨纪念中国过程控制会议 30 周年摘要集. 重庆: 中国自动化学会过程控制专业委员会, 2017. 192.

2 Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 2009, 3(4): 333–389. [doi: 10.1561/1500000019]

3 Huang PS, He XD, Gao JF, *et al*. Learning deep structured semantic models for Web search using clickthrough data. Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco: ACM, 2013. 2333–2338.

4 Pang L, Lan YY, Guo JF, *et al*. Text matching as image recognition. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix: AAAI Press, 2016. 2793–2799.

5 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186.

6 Yang W, Zhang HT, Lin J. Simple applications of BERT for ad hoc document retrieval. arXiv: 1903.10972, 2019.

7 Zhang ZY, Han X, Liu ZY, *et al*. ERNIE: Enhanced language representation with informative entities. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 1441–1451.

8 Miyato T, Dai AM, Goodfellow IJ. Adversarial training methods for semi-supervised text classification. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

9 Song F, Croft WB. A general language model for information retrieval. Proceedings of the 8th International Conference on Information and Knowledge Management. Kansas City: ACM, 1999. 316–321.

10 Chen Q, Zhu XD, Ling ZH, *et al*. Enhanced LSTM for natural language inference. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017. 1657–1668.

11 Wang ZG, Hamza W, Florian R. Bilateral multi-perspective matching for natural language sentences. Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne: IJCAI.org, 2017. 4144–4150.

12 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 3111–3119.

13 Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019. 3980–3990.

14 Gururangan S, Marasović A, Swayamdipta S, *et al*. Don't stop pretraining: Adapt language models to domains and tasks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 8342–8360.

15 Gu YX, Zhang ZY, Wang XZ, *et al*. Train no evil: Selective masking for task-guided pre-training. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. ACL, 2020. 6966–6974.

16 Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. Proceedings of the 3rd International Conference on Learning Representations. San Diego, 2015.

17 Madry A, Makelov A, Schmidt L, *et al*. Towards deep learning models resistant to adversarial attacks. arXiv: 1706.06083, 2017.

18 Shafahi A, Najibi M, Ghiasi A, *et al*. Adversarial training for free! Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, 2019. 302.

19 Liu YH, Ott M, Goyal N, *et al*. RoBERTa: A robustly optimized BERT pretraining approach. arXiv: 1907.11692, 2019.

20 Liu X, Chen QC, Deng C, *et al*. LCQMC: A large-scale Chinese question matching corpus. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: ACL, 2018. 1952–1962.

21 Paszke A, Gross S, Massa F, *et al*. PyTorch: An imperative style, high-performance deep learning library. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, 2019. 721.

22 蔡林杰, 刘新, 刘龙, 等. 基于 Transformer 的改进短文本匹配模型. 计算机系统应用, 2021, 30(12): 268–272. [doi: 10.15888/j.cnki.csa.008196]

23 朱朦朦, 武恺莉, 洪宇, 等. 面向问句复述识别的语义正交化匹配方法研究. 中文信息学报, 2021, 35(11): 34–42.

(校对责编: 牛欣悦)