基于 MacBERT 与对抗训练的机器阅读理解模型

周昭辰[1],方清茂[2],吴晓红[1*],胡平[2],何小海[1]
(1. 四川大学电子信息学院,四川 成都 610065;2. 四川省中医药科学院,四川 成都 610041)

摘 要:机器阅读理解旨在让机器像人类一样理解自然语言文本,并据此进行问答任务。近年来,随着深度学习和大规模数据集的发展,机器阅读理解引起了广泛关注,但是在实际应用中输入的问题通常包含各种噪声和干扰,这些噪声和干扰会影响模型的预测结果。为了提高模型的泛化能力和鲁棒性,提出一种基于掩码校正的来自 Transformer 的双向编码器表示(MacBERT)与对抗训练(AT)的机器阅读理解模型。首先利用 MacBERT 对输入的问题和文本进行词嵌入转化为向量表示;然后根据原始样本反向传播的梯度变化在原始词向量上添加微小扰动生成对抗样本;最后将原始样本和对抗样本输入双向长短期记忆(BiLSTM)网络进一步提取文本的上下文特征,输出预测答案。实验结果表明,该模型在简体中文数据集 CMRC2018 上的 F1 值和精准匹配(EM)值分别较基线模型提高了 1.39 和 3.85 个百分点,在繁体中文数据集 DRCD 上的 F1 值和 EM 值分别较基线模型提高了 1.22 和 1.71 个百分点,在英文数据集 SQuADv1.1 上的 F1 值和 EM 值分别较基线模型提高了 2.86 和 1.85 个百分点,优于已有的大部分机器阅读理解模型,并且在真实问答结果上与基线模型进行对比,结果验证了该模型具有更强的鲁棒性和泛化能力,在输入的问题存在噪声的情况下性能更好。

关键词:机器阅读理解;对抗训练;预训练模型;掩码校正的来自 Transformer 的双向编码器表示;双向长短期记忆网络
源代码链接:https://github.com/qinghuanguayu/CMRC2018_AT
中图分类号:TP18 **文献标志码**:A **DOI**:10.19678/j.issn.1000-3428.0068121

Machine Reading Comprehension Model Based on MacBERT and Adversarial Training

ZHOU Zhaochen[1], FANG Qingmao[2], WU Xiaohong[1*], HU Ping[2], HE Xiaohai[1]
(1. School of Electronic Information, Sichuan University, Chengdu 610065, Sichuan, China;
2. Sichuan Academy of Chinese Medicine Sciences, Chengdu 610041, Sichuan, China)

【Abstract】 Machine reading comprehension is designed to allow machines to understand natural language texts, resembling humans, and perform question-answering tasks accordingly. In recent years, owing to the development of deep learning and large-scale datasets, machine reading comprehension has received widespread attention. However, input problems in practical applications typically involve various noises and interferences, which affect the prediction results of a model. To improve the generalizability and robustness of a model, a machine reading comprehension model based on Masked language modeling as correction Bidirectional Encoder Representations from Transformers (MacBERT) and Adversarial Training (AT) is proposed. First, MacBERT is used to convert input questions and texts into word embeddings and vector representations. Subsequently, a small perturbation is added to the original word vector based on the gradient change of the original sample backpropagation to generate an adversarial sample. Finally, the original and adversarial samples are input into a Bidirectional Long Short-Term Memory (BiLSTM) network to further extract the contextual features of the text and output the predicted answer. Experimental results show that the F1 and Exact Matching (EM) values of this model on the simplified Chinese dataset CMRC2018 improve by 1.39 and 3.85 percentage points, respectively, compared with those of the baseline model. Meanwhile, the F1 and EM values on the traditional Chinese dataset DRCD improve by 1.22 and 1.71 percentage points, respectively, compared with those of the baseline model. Moreover, the F1 and EM values on the English dataset SQuADv1.1 improve by 2.86 and 1.85 percentage points, respectively, compared with those of the baseline model. The experimental results are better than those of most existing machine reading comprehension models. Based on actual question-answering results, the proposed model outperforms the baseline model in terms of robustness and generalizability; additionally, it performs better when the input problems contain noise.

【Key words】 machine reading comprehension; Adversarial Training(AT); pre-trained model; Masked language

收稿日期:2023-07-20 **修回日期**:2023-11-01
基金项目:成都市重大科技应用示范项目(2019-YF09-00120-SN)。

通信作者 E-mail:* wxh@scu.edu.cn

modeling as correction Bidirectional Encoder Representations from Transformers (MacBERT); Bidirectional Long Short-Term Memory (BiLSTM) network

0 引言

在自然语言处理(NLP)的发展历程中,教会机器阅读文本并理解文本含义是一项尚未完全实现的研究目标。近些年大规模数据集、高算力计算机和深度学习模型的出现促进了机器阅读理解的发展,并使其广泛应用于信息检索、智能问答、机器翻译等领域。

早期机器阅读理解系统由于计算机水平和数据集规模的限制,主要是依赖于人工制定的规则建立的,后来深度学习的出现使得机器阅读理解发展到新的阶段。HERMANN 等[1]在传统的深度学习模型的基础上,提出了 The Attentive Reader 和 The Impatient Reader 模型,将注意力机制加入到深度学习的长短期记忆(LSTM)网络。SEO 等[2]提出了双向注意力流(BiDAF)网络,加入双向注意力层来整合问题和上下文表示。VASWANI 等[3]提出了 Transformer 模型,仅采用多头注意力机制就在多个任务中取得了优异成绩,在此基础上 DEVLIN 等[4]提出了来自 Transformer 的双向编码器表示(BERT)预训练模型,极大地提高了机器阅读理解模型的性能。BERT 预训练模型在 NLP 领域的优异表现使得越来越多学者对其进行了改进创新。LIU 等[5]提出一种鲁棒优化的 BERT 预训练方法(RoBERTa),在预训练时使用了更多的训练数据、更大的 Batch Size 和更长的训练时间。CUI 等[6]同样对预训练方式进行改进后提出了掩码校正的来自 Transformer 的双向编码器表示(MacBERT)模型。YANG 等[7]对 BERT 的模型结构进行改进并提出了名为 XLNet 的用于语言理解的广义自回归预训练模型,其中加入了双流注意力机制以及 Transformer-XL 的相关技术。之后,各类预训练模型[8-9]被相继提出,模型性能得到优化提升。在深度学习方法下,大规模机器阅读理解数据集的提出同样促进了相关领域的研究发展。HERMANN 等[1]提出大规模机器阅读理解数据集 CNN&Daily Mail,该数据集属于完形填空类数据集,与实际生活所使用的自然语言形式不符。在此基础上,RAJPURKAR 等[10]提出了抽取类机器阅读理解数据集 SQuADv1.1,TRISCHLER 等[11]提出问题较为复杂的数据集 NewsQA。随后,自由生成类数据集 TriviaQA[12]、多项选择类数据集 RACE[13]、抽取类数据集 SQuAD2.0[14]等一系列大规模机器阅读理解数据集被相继提出。中文机器阅读理解数据集虽然起步较晚,但是依旧有许多大规模的优质数据集被提出。CUI 等[15-16]先后提出了完形填空类数据集 CMRC2017 和以 SQuAD 为模板的抽取类数据集 CMRC2018。SHAO 等[17]提出了抽取类数据集 DRCD,该数据集成了中文机器阅读理解研究常用数据集。

现阶段机器阅读理解模型大多基于 BERT 或 BERT 的变体预训练模型进行优化改进,但是在实际的机器阅读理解问答应用中仍存在一些问题。由于不同人在进行提问时,对同一个问题的表述不尽相同,可能在问题中会出现简写、语法错误等情况,因此机器阅读理解任务中单一的文本和问题表述会造成模型的泛化能力较差、鲁棒性较低。针对此问题,本文提出了一种基于 MacBERT 与对抗训练(AT)的机器阅读理解模型。首先,利用 MacBERT 对文本进行词嵌入,将自然语言转化为向量表示,之后引入对抗训练,在每个词向量上根据梯度加入不同的微小扰动生成对抗样本,对抗样本会在训练的过程中不断攻击模型,从而增加了模型的鲁棒性和泛化能力。然后,引入双向长短期记忆(BiLSTM)网络,使模型可以更好地提取文本和问题中的上下文的特征。最后,在简体中文数据集 CMRC2018 和繁体中文数据集 DRCD 以及英文数据集 SQuADv1.1 上进行实验来验证模型的有效性。

1 基于对抗训练的机器阅读理解模型

为了解决现实问答场景中可能出现的简写、语法错误等情况,提高模型的泛化能力和鲁棒性,提出一种基于 MacBERT、BiLSTM 和对抗训练的机器阅读理解问答模型,其中利用对抗训练来提高模型的泛化能力和鲁棒性,利用 BiLSTM 来更好地融合上下文特征。首先,用 MacBERT 对输入文本进行词嵌入,将输入文本的每个字转化为向量表示;之后,用对抗训练模块在每个字的向量上加入微小扰动,原字向量和加入扰动的字向量先后送入 BiLSTM 进行前向传播和反向传播获取序列特征;最后,接入全连接层进行二分类得到答案的起始位置和终止位置。基于 MacBERT 与对抗训练的机器阅读理解模型结构如图 1 所示。

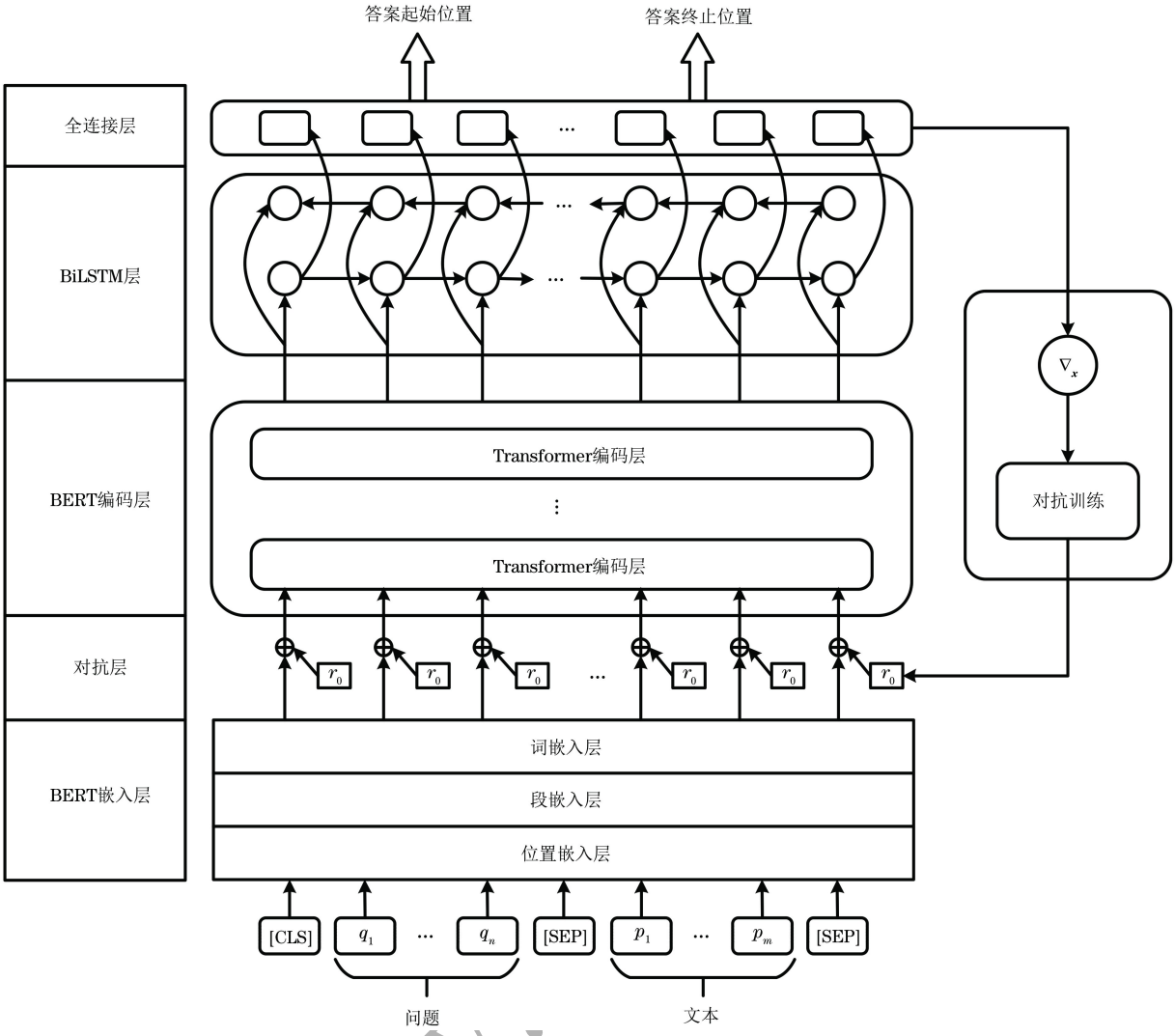

图 1 基于 MacBERT 与对抗训练的机器阅读理解模型结构

Fig. 1 Structure of machine reading comprehension model based on MacBERT and adversarial training

1.1 文本嵌入

通过词嵌入将输入文本转换成向量表示是所有 NLP 任务的第 1 步，选择 MacBERT 预训练模型来进行此操作。以 BERT 为基础的 MacBERT 为动态词向量，相较于静态词向量的 Word2Vec 和 GloVe，MacBERT 的词向量是随着上下文输入的不同而返回不同的词向量表达，弥补了静态词向量无法表述一词多义的缺陷。另外，相较于 BERT 和 RoBERTa，MacBERT 在预训练时对掩码语言模型(MLM)掩码训练任务进行了改进，使用全词 mask 以及 *N*-gram mask 策略来替代随机 mask，减轻了 BERT 模型预训练阶段和微调阶段存在差异的问题。

MacBERT 首先对输入的文本进行分词操作，将完整的一句话分为一个个单词和字符，然后输入词嵌入层。词嵌入层包括 3 种嵌入表示，分别为词嵌入、段嵌入和位置嵌入。词嵌入是将输入文本的每一个字和起始位置与两句之间插入的特殊符号转化为 768 维的向量表示；段嵌入是对输入文本在句子级上进行标记，每个句子中的所有字符按句子交替顺序转化为 768 维向量；位置嵌入是对输入文本不同位置的字符进行向量表示，文本中每一个位置的字符都会对应一个特定的 768 维向量。最后将生成的词嵌入、段嵌入和位置嵌入相加得到了输入文本维度为(1,512,768)的嵌入表示，将词嵌入所得的向量输入编码层进行训练。MacBERT 的编码层是由 Transformer 模型改进而来，由 12 层 Transformer 的 Encoder 模块堆叠而成，整体的 MacBERT 模型结构如图 2 所示。

1.2 对抗训练

对抗训练最早应用于计算机视觉(CV)领域，先通过对图像添加噪声来生成对抗样本，再利用对抗样本攻击模型，最终提高模型的鲁棒性。对抗样本的概念由 SZEGEDY 等[18]提出，在图像上添加微小扰动生成的对抗样本即使在视觉上与原图区别不

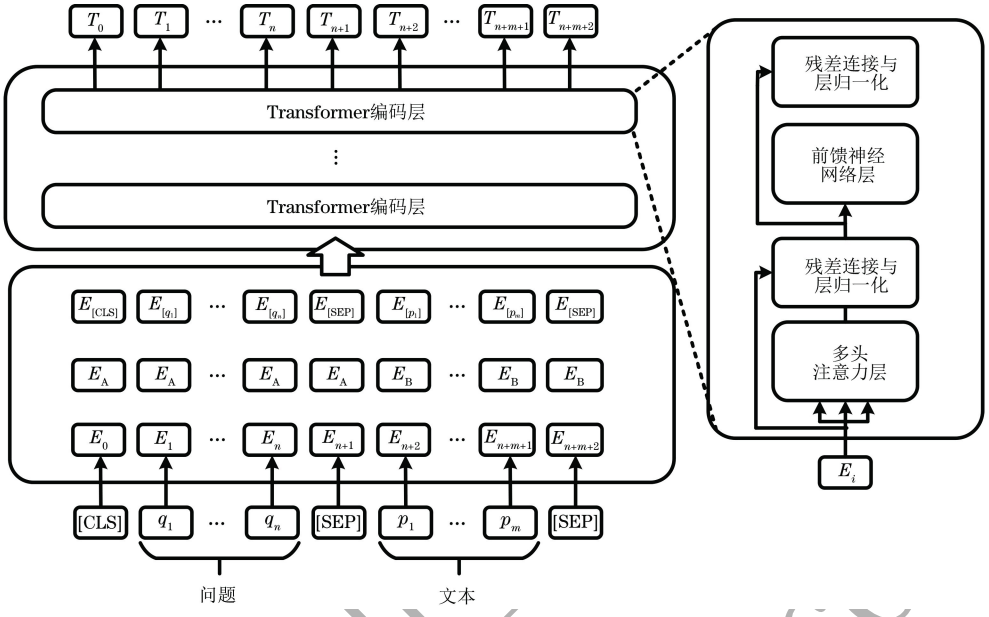

图 2 BERT 文本嵌入模型结构

Fig. 2 Structure of BERT text embedding model

大，但是会使模型对其进行错误分类。据此，GOODFELLOW 等[19]提出对抗训练的概念，并验证了对抗训练在 CV 领域可以提高模型的鲁棒性。根据对抗训练在 CV 领域的应用，MIYATO 等[20]将对抗训练引入 NLP 领域，提出在词嵌入生成的词向量上添加微小扰动生成对抗样本来进行对抗训练，尽管生成的对抗词向量无法映射到具体单词，但是经验证该方法提高了模型的鲁棒性以及泛化能力。在此基础上，MADRY 等[21]提出了投影梯度下降(PGD)模型，采用多次迭代的方式来找到最优的对抗扰动，但是所带来的时间成本过高，每增加 1 轮迭代就增加了 1 倍的训练时间。由于 PGD 模型只将最后 1 次迭代的对抗梯度加到原始梯度上，ZHU 等[22]提出了名为 FreeLB 的模型，将全部迭代的对抗梯度的平均值加到原始梯度上，但是同样有着时间成本过高的问题。为了使扰动效果以及时间效率最佳，选择在 Loss 增加最多的方向上添加扰动，计算公式如式(1)所示：

$$\max(-\ln p(\boldsymbol{y} \mid \boldsymbol{x}+\boldsymbol{r};\theta)) \tag{1}$$

式中：$\boldsymbol{y}$ 为真实标签；$\boldsymbol{x}$ 为原始词向量；$\boldsymbol{r}$ 为添加的微小扰动；θ 为模型常量。

因此，使 Loss 增加最多的扰动 $\boldsymbol{r}_{\text{adv}}$ 的计算公式如式(2)、式(3)所示：

$$\boldsymbol{r}_{\text{adv}}=\varepsilon\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \tag{2}$$

$$\boldsymbol{g}=\nabla_x \ln p(\boldsymbol{y} \mid \boldsymbol{x};\theta) \tag{3}$$

式中：ε 为强度因子；$\boldsymbol{g}$ 为损失函数 Loss 对 $\boldsymbol{x}$ 的偏导，即梯度。

原始词向量 $\boldsymbol{x}$ 的维度为(1,512,768)，在此上添加扰动进行对抗训练的步骤如下：

1)对每一次送入网络的原始词向量 $\boldsymbol{x}$ 计算前向传播的 Loss 和反向传播的梯度。

2)根据梯度计算出所要添加的扰动 $\boldsymbol{r}_{\text{adv}}$，与 $\boldsymbol{x}$ 相加得到对抗样本 $\boldsymbol{a}=\boldsymbol{x}+\boldsymbol{r}_{\text{adv}}$。

3)计算对抗样本 $\boldsymbol{a}$ 的前向传播 Loss 和反向传播梯度，将生成的对抗梯度累加到步骤 1)中原始词向量反向传播生成的梯度上。

4)将词向量恢复为原始词向量 $\boldsymbol{x}$。

5)根据步骤 3)生成的累加梯度对网络参数进行更新。

经过对抗训练后的网络模型鲁棒性更高、泛化能力更强，对问答任务中可能出现的错别字、简写、语法错误等问题有了很好的抵抗能力。

1.3 BiLSTM 网络

BiLSTM 网络是双向 LSTM 网络，由前向 LSTM 网络和反向 LSTM 网络组合而成。单向的 LSTM 网络只能提取输入信息的上文信息而忽略了下文信息，BiLSTM 网络则通过前向 LSTM 网络和反向 LSTM 网络可以提取到输入信息的上下文信息。LSTM 网络模型如图 3 所示，具体计算公式如式(4)～式(6)所示：

$$\begin{cases}\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{z}_t]+\boldsymbol{b}_f)\\ \boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},\boldsymbol{z}_t]+\boldsymbol{b}_i)\\ \boldsymbol{o}_t=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{h}_{t-1},\boldsymbol{z}_t]+\boldsymbol{b}_o)\\ \widetilde{\boldsymbol{C}}_t=\tanh(\boldsymbol{W}_C\cdot[\boldsymbol{h}_{t-1},\boldsymbol{z}_t]+\boldsymbol{b}_C)\end{cases} \tag{4}$$

$$\boldsymbol{C}_t=\boldsymbol{f}_t*\boldsymbol{C}_{t-1}+\boldsymbol{i}_t*\widetilde{\boldsymbol{C}}_t \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t * \tanh(\boldsymbol{C}_t) \tag{6}$$

式中：$\boldsymbol{f}_t$、$\boldsymbol{i}_t$、$\boldsymbol{o}_t$ 为 t 时刻的遗忘门、记忆门和输出门；$\boldsymbol{W}_f$、$\boldsymbol{W}_i$ 和 $\boldsymbol{W}_o$ 为对应的权重系数；$\boldsymbol{b}_f$、$\boldsymbol{b}_i$ 和 $\boldsymbol{b}_o$ 为对应偏置量；$\boldsymbol{h}_{t-1}$ 为 $t-1$ 时刻的隐藏层状态；$\boldsymbol{z}_t$ 为 t 时刻的输入；$\tilde{\boldsymbol{C}}_t$ 为临时细胞状态；$\boldsymbol{C}_t$ 为当前细胞状态；$\boldsymbol{h}_t$ 为 t 时刻隐藏层状态；σ 为 Sigmoid 激活函数。

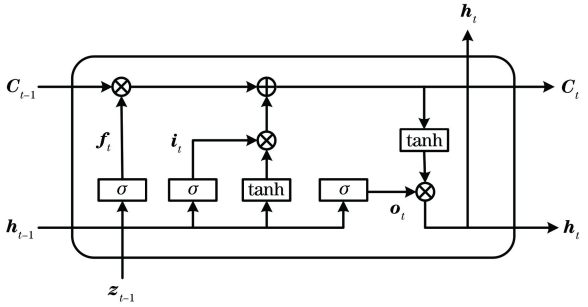

图 3 单个 LSTM 模型结构

Fig.3 Structure of single LSTM model

2 实验与结果分析

2.1 数据集

为了全面评估对抗训练对模型性能的影响，选择简体中文数据集 CMRC2018 和繁体中文数据集 DRCD 进行实验分析，并加入英文数据集 SQuADv1.1 来评估所提模型在多种语言上的表现，这 3 个数据集皆属于抽取式机器阅读理解数据集且结构相似。CMRC2018 和 DRCD 数据集皆是从中文维基百科中整理得到，分别包括约 20 000 和 30 000 个问题，SQuADv1.1 从英文维基百科上的 536 篇文章中整理得到，包含了约 100 000 个问题，具体数据集规模如表 1～表 3 所示，其中，短文数为数据集的文章数量，问题数为对文章所提问题的数量，答案数为每个问题所给的标准答案数量，每个标准答案之间会有一定偏差，模型会以与预测答案匹配度最高的标准答案来计算匹配度，其中，“—”表示无该数据集。

表 1 CMRC2018 数据集规模

Table 1 CMRC2018 dataset size 单位：个

数据集	短文数	问题数	答案数
训练集	2 659	11 144	1
开发集	848	3 219	3
测试集	1 718	6 895	3

表 2 DRCD 数据集规模

Table 2 DRCD dataset size 单位：个

数据集	短文数	问题数	答案数
训练集	8 016	26 936	1
开发集	1 000	3 524	2
测试集	1 000	3 493	2

表 3 SQuADv1.1 数据集规模

Table 3 SQuADv1.1 dataset size 单位：个

数据集	短文数	问题数	答案数
训练集	19 125	87 599	1
开发集	2 085	10 570	3
测试集	—	—	—

2.2 评价指标

评价抽取式机器阅读理解模型的性能主要是计算模型预测出的答案与数据集所给答案的相似度，常用评价指标为精准匹配(EM)值和 F1 值，其中，EM 值为精确匹配度，计算模型预测答案与所给答案完全相同所占的百分比，F1 值是模糊匹配度，计算模型预测答案与所给答案的相似度，并加入 2 个指标的均值进行综合分析。

2.3 实验配置与参数设置

实验操作系统为 Ubuntu 18.04，GPU 为 NVIDIA GTX 2080Ti 11 GB，深度学习框架为 PyTorch 1.10.2，MacBERT 的版本选择 base 版，具体的实验参数如表 4 所示。

表 4 超参数设置

Table 4 Hyperparameter setting

参数	CMRC2018	DRCD	SQuADv1.1
Epoch	3	4	5
Batch Size	4	6	4
Learning rate	0.000 03	0.000 03	0.000 03
Max Answer Length	100	100	70
Dropout	0.1	0	0
Warmup Step	0.1	0.1	0.1

2.4 结果分析

为了验证对抗训练和 BiLSTM 在机器阅读理解任务中的有效性，在 CMRC2018 数据集上进行实验，并将实验结果与其他先进的机器阅读理解模型进行对比，实验结果如表 5 所示(其中最优指标值用加粗字体标示，下同)，对比模型具体如下：

$\text{BERT}_{\text{base}}$ 是 BERT 模型的基础版本，主干为 12 层编码器层，隐藏层大小为 768，在多头注意力中的头数量为 12。

RoBERTa-wwm-ext-base 是 RoBERTa 模型的基础版本，相较于 BERT 模型使用了更多的训练数据。

中文机器阅读理解模型 Hybrid-Attention[23] 首先在利用 BERT 进行词嵌入后加入混合注意力机制来提取文本中与答案有关的关键信息，然后对多层次的信息进行融合，最后用 BiLSTM 建模后输出正确答案。

中文预训练模型 ChineseBERT[24] 在传统的预

训练方法中加入汉字的字形和拼音信息，不但能够从视觉特征中捕获字符语义，而且还处理了中文中普遍存在的同义词现象。

基于话头话体共享结构信息的机器阅读理解模型 BERT-NTC[25]将小句复合体结构自动分析与机器阅读理解任务融合为一个整体，首先利用小句复合体结构自动分析模型将每个标点句扩充为话头自足句，将问题要素与答案要素归置于同一标点句中，辅助机器阅读理解模型更好地理解远距离成分共享的语义信息。

预训练模型 ERNIE2.0 的基础版为 ERNIE 2.0_{base}[26]，ERNIE2.0 与 BERT 模型相比不但进行了无监督的预训练，而且加入了多种有监督任务进行训练来使模型学习词汇、句法和语义信息，并且多任务同时训练学习，防止遗忘先前任务所学习到的信息。

明确跨度-句子预测阅读器(ESPReader)[27]在对数据集进行词嵌入的同时提出了一个新的句子级的子任务，即跨度-句子预测，先将文本按照标点符号划分为自然句子，再对包含答案的句子进行标记，最后训练模型在提取答案跨度时准确定位该答案对应的句子，这样 BERT 模型在对文本进行词嵌入时会进行 4 种词嵌入，即词嵌入、段嵌入、位置嵌入和跨度-句子嵌入。

$MacBERT_{base}$ 是 MacBERT 模型的基础版本，该实验的基线模型。

$MacBERT_{base}$+BiLSTM 是在 $MacBERT_{base}$ 的基础上加入了 BiLSTM。

$MacBERT_{base}$+AT 是在 $MacBERT_{base}$ 的基础上加入了对抗训练。

$MacBERT_{base}$+BiLSTM+AT 是在 $MacBERT_{base}$ 的基础上同时加入了 BiLSTM 和对抗训练。

表 5 CMRC2018 数据集上的实验结果

Table 5 Experimental results on CMRC2018 dataset %

模型	EM 值	F1 值	均值
人类表现	91.08	97.35	94.22
$BERT_{base}$	63.60	83.90	73.75
RoBERTa-wwm-ext-base	67.40	87.20	77.30
$MacBERT_{base}$	68.50	87.90	78.20
Hybrid-Attention	69.84	88.04	78.94
ChineseBERT	70.70	88.09	79.40
BERT-NTC	71.92	88.46	80.19
$ERNIE2.0_{base}$	69.10	88.60	78.85
ESPReader	71.80	88.70	80.30
$MacBERT_{base}$+BiLSTM	71.45	87.93	79.69
$MacBERT_{base}$+AT	71.82	88.96	80.39
$MacBERT_{base}$+BiLSTM+AT	**72.35**	**89.29**	**80.82**

由表 5 可以看出：$MacBERT_{base}$ 在加入 BiLSTM 后 EM 值为 71.45%，F1 值为 87.93%，相较于基线模型分别提高了 1.95 和 0.03 个百分点，说明利用 BiLSTM 融合上下文特征提高了模型的性能；$MacBERT_{base}$ 在加入对抗训练后 EM 值为 71.82%，F1 值为 88.96%，相比性能最优的 ESPReader 模型高出 0.02 和 0.26 个百分点，说明对抗训练使得模型在对抗样本的攻击中不断优化性能，并且提高了模型的鲁棒性和泛化能力；$MacBERT_{base}$ 在同时加入 BiLSTM 和对抗训练之后 EM 值为 72.35%，F1 值为 89.29%，较基线模型高出 3.85 和 1.39 个百分点，较 ESPReader 高出 0.55 和 0.59 个百分点，模型的性能进一步提升。

为了验证所提模型的通用性，在 DRCD 数据集上同样进行实验，实验结果如表 6 所示，其中，连接了多粒度语义信息的中文预训练语言模型 CLOWER[28]对词和字符表示进行对比学习，捕获了它们的语义关系，并将粗粒度的语义信息编码为细粒度的 mask，不仅增强了粗粒度语义信息，而且在细粒度上具有灵活性。

表 6 DRCD 数据集上的实验结果

Table 6 Experimental results on DRCD dataset %

模型	EM 值	F1 值	均值
$BERT_{base}$	83.10	89.90	86.50
RoBERTa-wwm-ext-base	86.60	92.50	89.55
CLOWER	88.27	93.44	90.86
$MacBERT_{base}$	88.30	93.50	90.90
Hybrid-Attention	89.05	94.14	91.60
$ERNIE2.0_{base}$	88.50	93.80	91.15
ESPReader	89.60	94.70	92.15
$MacBERT_{base}$+BiLSTM	88.96	94.04	91.50
$MacBERT_{base}$+AT	89.96	94.64	92.23
$MacBERT_{base}$+BiLSTM+AT	**90.01**	**94.72**	**92.37**

由表 6 可以看出：在 DRCD 数据集上基线模型 $MacBERT_{base}$ 在加入 BiLSTM 后的 EM 值为 88.96%，F1 值为 94.04%，比基线模型高出 0.66 和 0.54 个百分点；$MacBERT_{base}$ 在加入对抗训练后的 EM 值为 89.96%，F1 值为 94.64%，比基线模型高出 1.66 和 1.14 个百分点；$MacBERT_{base}$ 在同时加入 BiLSTM 和对抗训练之后 EM 值为 90.01%，F1 值为 94.72%，较基线模型高出 1.71 和 1.22 个百分点，并且优于对比模型，由于 DRCD 为繁体中文数据集，因此对比模型数量要少于 CMRC2018 数据集。综上，引入的 BiLSTM 和对抗训练在其他数据集上同样对模型的性能有提高，具有通用性。但

是，在与其他先进模型的对比中表现不如在CMRC2018数据集上优势明显，比如在加入BiLSTM和对抗训练后所提模型仅比ESPReader模型的EM值高出0.41个百分点，F1值高出0.02个百分点，主要是由于数据集语言是繁体中文，在一些字词上与简体中文有一定差异，并且模型对繁体中文的训练量不够，因此提升效果不理想。

为了验证所提模型在多种语言上的通用性，在SQuADv1.1数据集上进行实验，该实验基线模型为$BERT_{base}$，实验结果如表7所示，其中，基于维基百科的问答系统DrQA[29]先使用二元散列和词频-逆向文件频率(TF-IDF)匹配的模块在给定问题时返回相关的文章，再使用多层递归神经网络机器理解模型，最后在返回的相关文章中定位到具体的答案。

表7 SQuADv1.1数据集上的实验结果

Table 7 Experimental results on SQuADv1.1 dataset %

模型	EM值	F1值	均值
BiDAF	68.00	77.30	72.65
DrQA	68.70	78.16	73.43
$BERT_{base}$	68.49	81.09	74.79
ESPReader	**82.50**	**85.40**	**83.95**
$BERT_{base}$+BiLSTM	69.25	81.18	75.22
$BERT_{base}$+AT	71.47	82.55	77.01
$BERT_{base}$+BiLSTM+AT	71.85	82.86	77.36

由表7可以看出，在SQuADv1.1数据集上基线模型$BERT_{base}$在加入BiLSTM后的EM值和F1值分别比基线模型高出0.76和0.09个百分点，在加入对抗训练后的EM值和F1值分别比基线模型高出2.98和1.46个百分点，在同时加入BiLSTM和对抗训练后的EM值和F1值分别较基线模型高出3.36和1.77个百分点。这说明所提模型对英文数据集依旧有优化能力，适用于多种语言，但是与ESPReader模型相比还有一定差距，主要是由于中英文之间存在着语法结构等差异，并且英文按单词分词来嵌入表示与中文按汉字分词来嵌入表示也存在着显著差异。

为了更加清晰地验证所提模型引入的BiLSTM和对抗训练模块对模型性能的影响，将所提对抗训练模块与常见对抗训练模型PGD进行对比，消融实验结果如表8所示，其中，PGD对抗训练迭代次数为3，CMRC2018和DRCD数据集的baseline为$MacBERT_{base}$，SQuADv1.1数据集的baseline为$BERT_{base}$。

表8 消融实验结果

Table 8 Results of ablation experiment %

数据集	模型	EM值	F1值	均值
CMRC 2018	baseline	71.35	87.89	79.62
	baseline+BiLSTM	71.45	87.93	79.69
	baseline+AT(our)	71.82	88.96	80.39
	baseline+PGD	71.69	**89.31**	80.50
	baseline+BiLSTM+AT(our)	**72.35**	89.29	**80.82**
DRCD	baseline	88.87	93.95	91.41
	baseline+BiLSTM	88.96	94.04	91.50
	baseline+AT(our)	89.96	94.64	92.23
	baseline+PGD	89.78	94.56	92.17
	baseline+BiLSTM+AT(our)	**90.01**	**94.72**	**92.37**
SQuAD v1.1	baseline	68.99	81.01	75.00
	baseline+BiLSTM	69.25	81.18	75.22
	baseline+AT(our)	71.47	82.55	77.01
	baseline+PGD	71.22	82.41	76.82
	baseline+BiLSTM+AT(our)	**71.85**	**82.86**	**77.36**

由表8可以看出：

1)对于CMRC2018数据集，baseline在加入BiLSTM后EM值提高了0.10个百分点，F1值提高了0.04个百分点；baseline在加入对抗训练后EM值提高了0.47个百分点，F1值提高了1.07个百分点；baseline在同时加入BiLSTM和对抗训练后EM值提高了1.00个百分点，F1值提高了1.40个百分点。这说明所提模型在加入BiLSTM后提高了上下文信息融合能力，在加入对抗训练后提升了泛化能力和鲁棒性。

2)对于DRCD数据集：baseline在加入BiLSTM后EM值提高了0.09个百分点，F1值提高了0.09个百分点；baseline在加入对抗训练后EM值提高了1.09个百分点，F1值提高了0.69个百分点，baseline在同时加入BiLSTM和对抗训练后EM值提高了1.14个百分点，F1值提高了0.77个百分点。这说明所提模型在其他数据集上依旧具有较好的性能，通用性较强。

3)对于SQuADv1.1数据集：baseline在加入BiLSTM后EM值提高了0.26个百分点，F1值提高了0.17个百分点；baseline在加入对抗训练后EM值提高了2.48个百分点，F1值提高了1.54个百分点；baseline在同时加入BiLSTM和对抗训练后EM值提高了2.86个百分点，F1值提高了1.85个百分点。这说明所提模型在英文数据集上依旧具有较好的性能，适用于多种语言，并且性能提升水平

较高，主要原因为该数据集规模较大，对抗训练的对抗样本较多，从而明显提高了模型的泛化能力和鲁棒性。

将所提 AT 模块与常见对抗训练模型 PGD 进行对比，由表 8 还可以看出：

1）对于 CMRC2018 数据集：baseline 在加入 AT 后 EM 值提高了 0.47 个百分点，F1 值提高了 1.07 个百分点；baseline 在加入 PGD 后 EM 值提高了 0.34 个百分点，F1 值提高了 1.42 个百分点，整体提高效果更好。这主要是因为对模型性能影响最大的对抗扰动有阈值设置，对抗扰动在超过该阈值后模型性能会下降，PGD 在多次迭代后比所提 AT 模块更加接近该阈值。

2）对于 DRCD 数据集：baseline 在加入 AT 后 EM 值提高了 1.09 个百分点，F1 值提高了 0.69 个百分点；baseline 在加入 PGD 后 EM 值提高了 0.91 个百分点，F1 值提高了 0.61 个百分点，在 EM 值和 F1 值上的表现皆不如所提 AT 模块。这说明 PGD 多次迭代并不能在所有数据集上均保证比所提 AT 模块更接近扰动阈值。

3）对于 SQuADv1.1 数据集：baseline 在加入 AT 后 EM 值提高了 2.48 个百分点，F1 值提高了 1.54 个百分点；baseline 在加入 PGD 后 EM 值提高了 2.23 个百分点，F1 值提高了 1.40 个百分点，在 EM 值和 F1 值上的表现皆不如所提 AT 模块。这进一步说明了 PGD 并不适用于所有数据集，具体数据集需要选择具体的对抗训练模型，并且 PGD 训练时间较长，3 次对抗训练迭代使原本的训练时间增加了 3 倍，而所提 AT 模块仅增加了 1 倍。

综上，所提 AT 模块优于常用的 PGD 模型。

进一步在 CMRC2018 数据集上进行实验，分析对抗扰动程度对模型性能的影响，以本文对抗训练扰动 $\boldsymbol{r}_{adv}$ 为基线，实验验证 0.5 倍、2.0 倍、3.0 倍、4.0 倍和 6.0 倍 $\boldsymbol{r}_{adv}$ 对抗训练对模型的性能影响，结果如表 9 所示。

表 9 对抗扰动程度对模型的影响

Table 9 Impact of adversarial disturbance on the model %

对抗扰动程度	EM 值	F1 值	均值
baseline	71.35	87.89	79.62
baseline+AT	71.82	88.96	80.39
baseline+AT_0.5	71.45	88.83	80.40
baseline+AT_2.0	71.94	89.23	80.40
baseline+AT_3.0	71.45	88.91	80.50
baseline+AT_4.0	71.26	88.52	80.23
baseline+AT_6.0	70.11	87.89	79.79

由表 9 可以看出：在使用 0.5 倍 $\boldsymbol{r}_{adv}$ 进行对抗训练后，相较于所提对抗训练模块的 EM 值降低了 0.37 个百分点，F1 值降低了 0.13 个百分点，说明了在一定程度上对抗扰动的增加会进一步提高模型性能；在使用 2.0 倍 $\boldsymbol{r}_{adv}$ 进行对抗训练后，相较于使用 0.5 倍 $\boldsymbol{r}_{adv}$ 进行对抗训练的 EM 值提升了 0.49 个百分点，F1 值提升了 0.40 个百分点，进一步论证了上述结论；在使用 3.0 倍 $\boldsymbol{r}_{adv}$ 进行对抗训练后，相较于使用 2.0 倍 $\boldsymbol{r}_{adv}$ 进行对抗训练的 EM 值降低了 0.49 个百分点，F1 值降低了 0.32 个百分点，说明了对抗扰动程度达到阈值后，模型性能的提升幅度会下降；在使用 4.0 倍 $\boldsymbol{r}_{adv}$ 进行对抗训练后，相较于使用 3.0 倍 $\boldsymbol{r}_{adv}$ 进行对抗训练的 EM 值进一步降低了 0.19 个百分点，F1 值同样进一步降低了 0.39 个百分点，说明随着对抗扰动程度的增加，模型性能的提升幅度进一步下降；在使用 6.0 倍 $\boldsymbol{r}_{adv}$ 进行对抗训练后，相较于使用 4.0 倍 $\boldsymbol{r}_{adv}$ 进行对抗训练的 EM 值降低了 1.15 个百分点，F1 值降低了 0.63 个百分点，并且 EM 值低于不进行对抗训练的模型，说明了随着对抗扰动程度的不断增加，模型性能会不断下降。

2.5 具体问答样例对比

表 10 以 CMRC2018 中文数据集的部分问答为例，对比展示了基线模型和所提模型对同一个问题的答案预测情况。在第 1 个问答中问题为“宝山镇位于哪个省?”，所提模型预测的答案与正确答案完全相同，而基线模型预测的答案则是对问题“宝山镇位于哪个市?”的回答，2 个问题仅有 1 字之差，说明了所提模型相较于基线模型而言对自然语言的理解能力更强。在第 2 个问答中问题为“海军十字勋章创立至今大约授予了多少枚?”，所提模型预测的答案同样与正确答案完全相同，而基线模型预测的答案遗漏了“以上”这个词，使得答案的含义出现了变化，说明了所提模型相较于基线模型而言可以更加准确完善地进行答案预测，避免了信息遗漏。在第 3 个问答中问题为“伊维菌素在哪一年被人发现?”，在文本中“伊维菌素”、“发现”等词语附近有 2 个年份，分别为“1981 年”和“1975 年”，对答案的预测产生了干扰，而基线模型因为干扰产生了误判，但是所提模型做出了正确的答案预测，说明在加入了对抗训练后，所提模型能够更好地适应噪声和干扰，具有较高的准确性和鲁棒性。

为了模拟真实问答场景中可能出现的简写、语法错误等现象，对原有的问题进行更改，并对比展示了基线模型和所提模型对更改后问题的答案预测情况，如表 11 所示。在第 1 个问答中原问题为“《战国

表 10 模型答案预测对比

Table 10 Comparison of predicted answers of models

文本	问题	正确答案	基线模型	所提模型
宝山镇是中国辽宁省凤城市下辖的一个镇……	宝山镇位于哪个省?	中国辽宁省	辽宁省凤城市	中国辽宁省
海军十字勋章(Navy Cross)建立于 1919 年 2 月 4 日……,自创立以来,授予了 6 300 枚以上……	海军十字勋章创立至今大约授予了多少枚?	6 300 枚以上	6 300 枚	6 300 枚以上
伊维菌素(Ivermectin)也称为爱获灭……,伊维菌素被发现于 1975 年,并在 1981 年被列为药物,它是世界卫生组织基本药物标准清单中的一种……	伊维菌素在哪一年被人发现?	1975 年	1981 年	1975 年

无双 3》是由哪 2 个公司合作开发的?",将原问题简写为"《战国无双 3》哪 2 个合作?"后进行测试,所提模型预测的答案与正确答案完全相同,而基线模型答案预测为"系列作品外传作品",说明了在进行问题简写后,基线模型无法对问题进行正确预测,而所提模型面对问题中的干扰依旧可以进行正确预测。在第 2 个问答中原问题为"大莱龙铁路在哪?",更改问题语序并进行简写后为"在哪大莱龙铁路?",所提模型预测的答案同样与正确答案完全相同,但基线模型答案预测错误,说明了所提模型面对问题中语法错误的干扰同样可以正确预测出答案。

表 11 更改问题后模型答案预测对比

Table 11 Comparison of predicted answers of models after changing the question

文本	原问题	更改后的问题	正确答案	基线模型	所提模型
《战国无双 3》是由光荣和 ω-force 开发的……	《战国无双 3》是由哪 2 个公司合作开发的?	《战国无双 3》哪 2 个合作?	光荣和 ω-force	系列作品 外传作品	光荣和 ω-force
大莱龙铁路位于山东省北部环渤海地区,西起益羊铁路的潍坊大家洼车站……	大莱龙铁路在哪?	在哪大莱龙铁路?	山东省北部 环渤海地区	1997 年 11 月 批复立项	山东省北部 环渤海地区

3 结束语

本文针对实际问答时存在不同的人对同一问题的表述有偏差以及文本中包含简写、语法错误等现象,提出一种基于 MacBERT 和对抗训练的机器阅读理解模型。该模型首先使用 MacBERT 对输入的自然语言进行词嵌入生成向量表示;然后使用对抗训练根据前向传播的梯度变化在词向量上添加微小扰动生成对抗样本;最后将原始样本和对抗样本输入 BiLSTM 网络进一步提取文本上下文特征,利用对抗样本攻击模型来提高模型的鲁棒性和泛化能力。在 CMRC2018、DRCD 和 SQuADv1.1 数据集上的实验结果表明,该模型相较于其他模型性能有所提高,并且在具体的问答样例上进行了模型答案预测对比,验证了所提模型相较于基线模型对于自然语言的理解能力更强,回答更加准确,同时在加入对抗训练后模型的抗干扰性和鲁棒性更强。

本文分析了抗扰动程度对模型性能的影响,得出对抗训练没有生成使模型性能提升最多的对抗扰动,因此在接下来的工作中将研究更有效的对抗训练方法,使生成的对抗扰动可以进一步提升模型性能。

参考文献

[1] HERMANN K M, KOČISK Ý T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[EB/OL]. [2023-07-07]. https://arxiv.org/abs/1506.03340.

[2] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1611.01603.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of NIPS'17. Cambridge, USA: MIT Press, 2017: 5998-6008.

[4] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional Transformers for language understanding[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1810.04805.

[5] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1907.11692.

[6] CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[EB/OL]. [2023-07-07]. http://arxiv.org/abs/2004.13922.

[7] YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1906.08237.

[8] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1909.11942.

[9] JOSHI M, CHEN D Q, LIU Y H, et al. SpanBERT: improving pre-training by representing and predicting spans[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 64-77.

[10] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100, 000+ questions for machine comprehension of text[C]// Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2016: 2383-2392.
[11] TRISCHLER A, WANG T, YUAN X, et al. NewsQA: a machine comprehension dataset[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1611.09830.
[12] JOSHI M, CHOI E, WELD D S, et al. TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1705.03551.
[13] LAI G, XIE Q, LIU H, et al. RACE: large-scale reading comprehension dataset from examinations[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1704.04683.
[14] RAJPURKAR P, JIA R, LIANG P. Know what you don't know: unanswerable questions for SQuAD[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, USA: Association for Computational Linguistics, 2018: 784-789.
[15] CUI Y, LIU T, CHEN Z, et al. Dataset for the first evaluation on Chinese machine reading comprehension[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1709.08299.
[16] CUI Y M, LIU T, CHE W X, et al. A span-extraction dataset for Chinese machine reading comprehension[C]//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, USA: Association for Computational Linguistics, 2019: 5883-5888.
[17] SHAO C C, LIU T, LAI Y, et al. DRCD: a Chinese machine reading comprehension dataset[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1806.00920.
[18] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1312.6199.
[19] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1412.6572.
[20] MIYATO T, DAI A M, GOODFELLOW I. Adversarial training methods for semi-supervised text classification[EB/OL]. [2023-07-07]. http://arxiv.org/abs/1605.07725.
[21] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. [2023-07-07]. https://arxiv.org/abs/1706.06083.
[22] ZHU C, CHENG Y, GAN Z, et al. FreeLB: enhanced adversarial training for natural language understanding[EB/OL]. [2023-07-07]. https://arxiv.org/abs/1909.11764.
[23] 刘高军,李亚欣,段建勇. 基于混合注意力机制的中文机器阅读理解[J]. 计算机工程, 2022, 48(10): 67-72, 80.
LIU G J, LI Y X, DUAN J Y. Chinese machine reading comprehension based on hybrid attention mechanism[J]. Computer Engineering, 2022, 48(10): 67-72, 80. (in Chinese)
[24] SUN Z J, LI X Y, SUN X F, et al. ChineseBERT: Chinese pretraining enhanced by glyph and Pinyin information[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, USA: Association for Computational Linguistics, 2021: 2065-2075.
[25] 韩玉蛟,罗智勇,张明明,等. 基于话头话体共享结构信息的机器阅读理解研究[C]//第二十一届中国计算语言学大会论文集.[出版地不详]: 中国中文信息学会计算语言专业委员会, 2022: 634-643.
HAN Y J, LUO Z Y, ZHANG M M, et al. Research on machine reading comprehension based on shared structure information between naming and telling[C]//Proceedings of the 21st Chinese National Conference on Computational Linguistics. [S. l.]: Computational Language Professional Committee of the Chinese Information Society, 2022: 634-643. (in Chinese)
[26] SUN Y, WANG S H, LI Y K, et al. ERNIE2.0: a continual pre-training framework for language understanding[EB/OL]. [2023-07-07]. https://arxiv.org/abs/1907.12412.
[27] WANG J W, ZHAO H, ZHAO Y G, et al. What if sentence-hood is hard to define: a case study in Chinese reading comprehension[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg, USA: Association for Computational Linguistics, 2021: 2348-2359.
[28] CHEN B, TANG H, WANG J, et al. CLOWER: a pre-trained language model with contrastive learning over word and character representations[EB/OL]. [2023-07-07]. http://arxiv.org/abs/2208.10844.
[29] CHEN D Q, FISCH A, WESTON J, et al. Reading Wikipedia to answer open-domain questions[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, USA: Association for Computational Linguistics, 2017: 1870-1879.

编辑 陆燕菲