北京大学学报(自然科学版) 第 57 卷 第 1 期 2021 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 57, No. 1 (Jan. 2021)
doi: 10.13209/j.0479-8023.2020.088

中文机器阅读理解的鲁棒性研究

李烨秋[1] 唐竑轩[1] 钱锦[1] 邹博伟[1,2] 洪宇[1,†]

1. 苏州大学计算机科学与技术学院, 苏州 215000; 2. 新加坡资讯通信研究院, 新加坡 138632;
† 通信作者, E-mail: tianxianer@gmail.com

摘要 为了更好地评价阅读理解模型的鲁棒性, 基于 Dureader 数据集, 通过自动抽取和人工标注的方法, 对过敏感、过稳定和泛化 3 个问题分别构建测试数据集。还提出基于答案抽取和掩码位置预测的多任务学习方法。实验结果表明, 所提方法能显著地提高阅读理解模型的鲁棒性, 所构建的测试集能够对模型的鲁棒性进行有效评估。

关键词 机器阅读理解; 鲁棒性; 中文语料库

Robustness of Chinese Machine Reading Comprehension

LI Yeqiu[1], TANG Hongxuan[1], QIAN Jin[1], ZOU Bowei[1,2], HONG Yu[1,†]

1. School of Computer Science and Technology, Soochow University, Suzhou 215000; 2. Institute for Infocomm Research, Singapore 138632; † Corresponding author, E-mail: tianxianer@gmail.com

Abstract In order to better evaluate the robustness of Machine Reading Comprehension (MRC) models, this paper builds three test sets from Dureader by automatically extracting and manually annotating, consisting of over-sensitivity, over-stability, and generalization. In addition, this paper proposes a multi-task learning framework with answer extraction task and masked position prediction task. Experimental results demonstrate that proposed method gains significant robustness improvements and show the effectiveness of the three test sets on evaluating the robustness of MRC models.

Key words machine reading comprehension; robustness; Chinese corpus

作为自动问答(question answering, QA)领域的一个重要子任务, 机器阅读理解(machine reading comprehension, MRC)旨在让模型理解段落内容后, 根据问题给出答案。机器阅读理解数据集主要包含 5 个类别: 抽取式(extractive)、生成式(generative)、选择题(multiple choice)、完形填空(cloze-type)和会话(conversation)。其中, 抽取式阅读理解由于问题的多样性及易于评价等特点, 近年来成为被研究频率最高的评测任务。随着大规模数据集(如 SQuAD[1])的发布, 基于神经网络的抽取式阅读理解模型研究得到快速发展, 特别是基于预训练语言模型[2]的阅读理解模型, 其水平甚至已超越人类[3]。

虽然抽取式阅读理解模型在大量数据集上不断刷新其最佳性能, 但在鲁棒性方面仍存在较大的缺陷, 即面对不同分布或干扰性强的数据时, 模型性能受到严重的影响[4–6]。目前, 抽取式阅读理解模型的鲁棒性问题主要体现在以下 3 个方面。

1) 过敏感问题。主要体现在回答复述问题(与原问题字面上不完全相同, 而表达相同含义且答案相同的问题)时, 模型通常输出与原问题不同的错误答案。存在过敏感问题的模型, 对复述问题在字面上的差异过度敏感, 导致其错误地理解问题的语义。如表 1 所示, 复述问题仅将原问题中“送”改为“送给”, 在语义没有发生变化的情况下, 模型便给出无关的答案。

2) 过稳定问题。与过敏感问题相反, 当段落中

国家自然科学基金(61703293, 61672368, 61672367)和江苏高校优势学科建设工程项目资助
收稿日期: 2020–06–08; 修回日期: 2020–08–14

出现干扰句(存在大量与原问题相同词语的句子)时，阅读理解模型无法区分干扰句与答案所在句，导致其在干扰句中寻找到错误的答案。现有的阅读理解模型虽然有能力通过词共现等浅层线索回答问题，但忽略了对关键语义的理解，从而使段落中的干扰句影响模型的预测结果。如表1所示，干扰问题与干扰句存在的相同词语为“光大”、“银行”、“信用卡”和“有效期”，而干扰问题与答案所在句相同的词语为“光大”、“信用卡”和“有效期”，因此模型被误导为从干扰句中抽取答案。

3) 泛化问题。现有阅读理解模型能够通过训练，在与训练集相同领域的测试集上取得高性能。如果将已经训练好的模型应用于不同领域或采用不同方法构建的数据集上，其性能会急剧地下降。如表1所示，由于训练集不包含数学领域的样本，模型无法学到“解方程”问题的相关表示，因此选择错误的答案。

1 相关工作

1.1 阅读理解数据集

随着众包服务模式的普及，大规模机器阅读理解数据集的标注成为可能。Rajpurkar等[1]采用众包方法，构建第一个大规模抽取式机器阅读理解数据集SQuAD，后续的数据集(如Trischler等[7]构建的NewsQA和Rajpurkar等[8]构建的SQuAD 2.0)中加入不可回答的问题，增加了数据集的难度。近年来，研究者开始在中文阅读理解领域构建大规模机器阅读理解数据集。Cui等[9]通过众包方式，构建中文抽取式阅读理解数据集CMRC。He等[10]基于百度搜索和百度知道的数据，构建面向中文领域的大规模数据集Dureader。以上数据集通过加入各种类型的问题，提高了抽取式阅读理解的难度，但大都忽略了针对阅读理解模型鲁棒性的评价。

最近，人们开始尝试构建针对阅读理解模型鲁棒性评价的数据集。Tang等[4]在Dureader数据集的基础上，针对鲁棒性问题进行标注，构建中文阅读理解数据集$\text{Dureader}_{\text{Robust}}$，但未针对各鲁棒性任务构建单独测试集，所以无法分别验证模型在过敏感、过稳定和泛化方面的性能。

1.2 阅读理解模型鲁棒性研究

阅读理解模型的鲁棒性通常集中在3个方面。

1) 过敏感问题在阅读理解模型上主要体现在当模型遇到复述问题时，无法给出正确答案。现有的研究主要通过构建复述问题来攻击阅读理解模型。Zhao等[11]提出利用生成对抗网络的方法，生成复述问题来攻击阅读理解模型。Ribeiro等[6]提出生成语义相同的对抗数据，并结合专家标注的方法，检测神经网络模型的过敏感问题。上述研究发现，现有的神经网络模型在面对复述问题时，性能远低于原问题，存在比较严重的过敏感问题。针对该问题，目前的主要方法是利用模型自动生成大量复述问题，但其数据集缺乏语言的自然性。因此，本文将真实应用环境下的用户数据作为复述问题的来源，确保评价数据集的真实性和自然性。

2) 存在过稳定问题的模型倾向于利用词语分布等浅层线索，而忽略了语义理解。Jia等[5]通过在SQuAD数据集的段落中加入干扰句的方法，干扰阅读理解模型，导致模型性能大幅下降。Kavumba等[12]认为现有的深度学习模型擅长浅层信息学习而非深层语义理解，通过改进COPA数据集[13]上词

表1　鲁棒性问题样例

Table 1　Examples of robustness issues

问题类型	段落	原问题及答案	鲁棒性问题及答案
过敏感	满天星花语是: 守望爱情、思念、清纯、梦境、真心喜欢、配角，但不可或缺，甘做配角的爱。赠送: 1、作玫瑰的衬材赠恋人，可以创造一个美丽的爱情故事; 2、配剑兰赠将毕业同学，为"大展鸿图"之意…	原问题: 满天星适合送什么人? 答案: 恋人 预测结果: 恋人	复述问题: 满天星适合送给什么人? 答案: 恋人 预测结果: 守望爱情
过稳定	光大银行信用卡积分有效期限为5年(含积分产生年)，卡片到期续卡后，积分可继续兑换礼品……因为光大信用卡有效期是三年，而积分有效期是五年……	原问题: 光大信用卡积分多久清零? 答案: 5年/五年 预测结果: 5年	过稳定问题: 光大银行信用卡有效期多少年? 答案: 三年 预测结果: 5年
泛化	$x/2–x/5=60$，两边同乘10，$10\cdot x/2–10\cdot x/5=600$，$5x–2x=600$, $3x=600$, $x=200$		问题: 2分之1x–5分之1x=60，解方程 答案: x=200 预测结果: $x/2–x/5=60$

注: 预测模型采用$\text{RoBERTa}_{\text{large}}$。

语的分布和词性的重叠等浅层线索，证明现有模型在失去浅层线索后会导致性能损失。上述研究通过增加或消除文本中的浅层线索，误导模型的答案，但这种修改方式易使段落语义受到干扰，从而为模型的预测引入不可预知的变量。为避免该问题，本文首先选取存在干扰句的段落，然后根据该干扰句和答案，人工标注问题，在保证段落语义不被改变的同时，对模型进行干扰。

3) 阅读理解模型的泛化对开放域问答和现实应用(如搜索引擎)的发展有不可忽视的作用。Fisch 等[14]在 EMNLP 中设置针对阅读理解泛化能力的工作站任务 MRQA①，整合 SQuAD 和 NewsQA 等 18 个领域各异的数据集，将其中 6 个领域作为训练集，12 个领域作为泛化测试集。受该研究及 Tang 等[4]的启发，本文选用与训练集不同的教育和金融领域，采用人工标注方法构建泛化测试集。

2 数据集构建

本文基于 $\text{Dureader}_{\text{Robust}}$ 构建 3 个测试数据集②，分别评估阅读理解模型在过敏感、过稳定和泛化 3 个方面的性能。此外，以 $\text{Dureader}_{\text{Robust}}$ 的验证集作为领域内测试集，用来对比模型在非鲁棒性任务上的性能。数据格式参照 SQuAD 数据集，数据样本为<p, q, a>三元组形式，其中 q 为问题，p 为供模型阅读的段落，p 中包含答案 a。如表 1 中过稳定样例所示，答案 a 可能存在多种形式，数据集对所有答案均标注起始位置。参照 $\text{Dureader}_{\text{Robust}}$ 的设置，本数据集的数据类型为实体类问题，即答案为实体，段落为描述该实体的一段文字。构建的测试集数据规模如表 2 所示。

2.1 过敏感测试集

将一个复述问题输入机器阅读理解系统，应当得到与原问题同样的答案，否则，该系统对字面上的差异过度敏感，影响其对语义的理解。为构建过敏感测试集，本文采用复述问题查询和人工校验的方法产生测试样本。首先，通过“问题–段落”二元组匹配的方法，从 $\text{Dureader}_{\text{Robust}}$ 测试集和 Dureader 2.0 训练集中取出问题与段落分别相同的二元组<p, q>集合；然后，将 Dureader 2.0 中存在多个复述问题{q_1, …, q_m}的样例<p,{q_1, …, q_m}, a>作为候选样本；最后，通过人工校验的方法，去除语言不自然和语义不一致的样例，生成过敏感样本集合{<p, q', a>}。过敏感测试集划分为原问题和过敏感问题两个测试集，分别包含 246 和 2703 个样本。

表 2 测试集数据统计

Table 2 Statistics of test sets

测试问题	测试集	统计量			
		问题数	段落数	问题平均长度/字	答案平均长度/字
过敏感	原问题	246	246	9.63	5.15
	鲁棒性问题	2703	246	10.11	5.48
过稳定	原问题	456	456	9.40	5.99
	鲁棒性问题	496	456	16.21	5.92
泛化	领域内问题	1417	1417	9.41	6.44
	跨领域问题	1036	664	12.16	11.16

2.2 过稳定测试集

当阅读理解段落中出现干扰句时，过稳定问题尤其明显，表现为模型无法区分段落中包含正确答案的句子和仅与原问题存在较多共同词语的句子。

本文利用自然段落中存在的干扰性语句，在标注问题时尽量与干扰句保持更多的共同词语，在保证语言自然性的同时，也使段落语义不受干扰句影响。过稳定数据集的标注过程如算法 1 所示。

算法 1 过稳定测试样本构建算法

1. INPUT: 原问题 <p, q, a>
2. OUTPUT: 过稳定样例<p, q', a'>或 null
3. 识别 p 中所有实体，标记同类实体{e_1,…, e_m}
4. IF m≥2 THEN
5. 定位{e_1, …, e_m}所在句{s_1, …, s_m}
6. 从{s_1, …, s_m}选择一对相似句 s_i 和 s_j
7. s_i 作为答案句，s_j 作为干扰句，e_i 记为 a'
8. 标注人员根据 s_i 标注问题 q'
9. 返回<p, q', a'>
10. ELSE 返回 null

对 $\text{Dureader}_{\text{Robust}}$ 中的样本<p, q, a>，识别段落 p 中的所有实体及实体类型，若 p 中存在两个以上个类型相同的实体，将其标记为{e_1, …, e_m}，组成候选样本<p, {e_1, …, e_m}>，并进行下一步标注，否则跳过此样例。对每个候选样例<p, {e_1, …, e_m}>，标注人员定位所有实体所在句{s_1, …, s_m}，从中选出一对相似句 s_i 和 s_j，将 s_i 作为答案句，s_i 对应的实体 e_i 作为答案，记为 a'，将 s_j 作为干扰句。标注人员根

① https://mrqa.github.io/; ② https://github.com/unlimitedaki/Chinese-MRC-Robust-Dataset

据 s_i 和 s_j 标注问题 q'，使 q' 尽可能与干扰句保持更多的共同词。最终得到一组过稳定测试样例<p, q', a'>。过稳定数据集同样分为原问题和过稳定问题两个测试集，分别有456和496个样例。

2.3 泛化测试集

机器阅读理解模型的泛化能力通常指该模型在与训练集数据分布不同的测试数据集上，能够取得接近领域内测试集上的性能[14]。根据 $\text{Dureader}_{\text{Robust}}$ 数据集的领域设置，本文采用21个教育和金融领域的关键词，从 $\text{Dureader}_{\text{Robust}}$ 测试集中检索出 q 中至少包含一个关键词的候选样本<p, q>；然后人工筛选符合要求的候选样本，并根据 p 和 q 人工标注答案 a；最终获得测试样本<p, q, a>。泛化测试集包含1036个样本，与其对比的领域内测试集包含1417个样本。

3 机器阅读理解鲁棒性模型

3.1 中文 RoBERTalarge 抽取式问答模型

Liu等[15]基于BERT模型架构提出改进预训练方法的RoBERTa，该模型利用更大的模型参数量、更多的训练数据和新的动态掩码模式，在很多自然语言处理任务上取得比BERT更好的性能。本文采用Liu等[15]和Cui等[16]在来自中文维基百科的13.6 M条数据上预训练的 $\text{RoBERTa}_{\text{large}}$ 模型为基线模型，该模型在CMRC等中文数据集上取得了较好的性能。对输入样本<p, q, a>，首先将段落 p 和问题 q 作为输入(input)，接受输入的RoBERTa模型输出字符级隐状态hidden_state，最后hidden_state通过全连接层，输出段落中每个位置作为开始和结束位置的概率:

$$\text{input} = \{[\text{CLS}], q, [\text{SEP}], p, [\text{SEP}]\}, \quad (1)$$

$$\text{hidden\_state}=\text{RoBERTa (input)}, \quad (2)$$

$$\text{start\_logit, end\_logit} = \text{FC (hidden\_state)}。 \quad (3)$$

3.2 基于答案抽取和掩码位置预测的多任务学习模型

针对阅读理解的过敏感问题，本文结合答案抽取与掩码位置，预测进行多任务学习的方法。掩码语言模型(masked language model)为Devlin等[2]提出的预训练方法，通过随机遮蔽，输入数据中的一些词，利用上下文预测被遮蔽的词。在掩码位置模型训练过程中，模型获得从上下文中理解语义的能力，因此，针对语义相同的复述问题产生的扰动则更为健壮。本文参考Liu等[17]的研究，共享RoBERTa的编码层，在解码层进行基于答案抽取和掩码位置预测的多任务学习。

针对答案抽取任务，本文采用开始位置预测和结束位置预测的交叉熵损失的平均值作为损失，答案起始位置start_position由训练集样本提供，结束位置end_position通过起始位置与答案 a 的长度计算得到:

$$\text{start\_loss}=\text{CrossEntropyLoss (start\_logit, start\_position)}, \quad (4)$$

$$\text{end\_loss}=\text{CrossEntropyLoss (end\_logit, end\_position)}, \quad (5)$$

$$L_{\text{MRC}}(\theta)=\frac{1}{2}(\text{start\_loss}+\text{end\_loss})。 \quad (6)$$

针对掩码位置预测任务，本文将掩码预测结果mlm_prediction与掩码位置标签mlm_label的交叉熵作为模型损失，损失函数如下所示:

$$L_{\text{MLM}}(\theta)=\text{CrossEntropyLoss (mlm\_prediction, mlm\_label)}。 \quad (7)$$

答案抽取与预测掩码位置多任务学习的过程如算法2所示。对每个批次的样本，按1:2的比例随机分配答案抽取和掩码位置预测任务，并对每个批次样本单独计算损失，计算梯度并更新参数。

算法2 多任务学习算法

1. 划分训练集 D 为 m 个批次 $\{D_1, \ldots, D_m\}$
2. FOR D_i IN D DO
3. 1:2 的比例随机分配答案抽取和掩码位置预测任务
4. IF 答案抽取任务
5. 计算答案抽取损失 $L_{\text{MLM}}(\theta)$
6. IF 掩码位置预测
7. 遮蔽 15%的 token
8. 计算掩码位置预测损失 $L_{\text{MLM}}(\theta)$
9. 计算梯度 $\nabla(\theta)$
10. 更新模型参数 $\theta = \theta - \varepsilon\nabla(\theta)$
11. END

3.3 多轮微调机制

为进一步提高模型的鲁棒性，本文采用多轮微调预训练模型的方法。Conneau等[18]以及McCann等[19]的研究表明，从大规模同类数据集进行迁移学习可以有效提高预训练语言模型在目标任务上的性能。Liu等[20]和Li等[21]进一步证明，基于多轮微调的预训练模型可以取得更好的泛化能力。因此，本文在上述抽取式问答模型和多任务学习中都利用大

规模中文阅读理解数据集 CMRC 和 Dureader, 预微调 $\text{RoBERTa}_{\text{large}}$ 模型, 然后在 $\text{Dureader}_{\text{Robust}}$ 训练集上进行微调。

4 实验

4.1 实验设置

将 $\text{Dureader}_{\text{Robust}}$ 的训练集作为训练集, 将其验证集作为领域内测试集, 鲁棒性测试集采用本文构建的 3 个数据集。在多轮微调实验中, 采用 CMRC 和 Dureader 2.0 作为辅助数据集。参照 $\text{Dureader}_{\text{Robust}}$ 训练集的构造方法, 对 Dureader 2.0 训练集数据进行清洗, 仅使用其中单实体类的问题。各数据集的规模如下: $\text{Dureader}_{\text{Robust}}$ 为 14520 个样本, CMRC 2018 为 10142 样本, Dureader 2.0 (单实体)为 15519 个样本。

本文采用抽取式阅读理解中通常采用的指标作为评价标准。完全匹配值(exact match, EM): 以预测结果与正确答案是否完全匹配作为衡量系统性能的评价指标, EM 值为答案完全匹配的样本数 n 与总样本数 m 的比值。F1 值(F1-Score): F1 值是准确率(precision, P)和召回率(recall, R)的加权调和平均值, 评价预测结果字符串与答案字符串的匹配程度。准确率为预测片段和正确答案重叠字符数 o (overlap)与预测片段字符数 p 的比值, 召回率为重叠字符数 o 与答案字符数 g 的比值。两类评价指标的计算公式如下:

$$\text{EM}=\frac{n}{m},\ P=\frac{o}{p},\ R=\frac{o}{g},\ \text{F1}=\frac{2PR}{P+R}。\tag{9}$$

本文进行的 3 组实验使用同一组超参数, 学习率为 3×10^{-5}, 批次大小为 32, 最大文档长度设置为 512, 最大答案长度为 20。训练轮次安排如下: 基线模型在训练集上微调轮次为 2 轮; 多轮微调在 CMRC, Dureader 2.0 数据集上各微调 1 轮, 在训练集上微调 2 轮; 多任务学习训练在 CMRC, Dureader 2.0 数据集上各微调 3 轮, 在训练集上微调 6 轮。

4.2 实验结果与分析

4.2.1 不同模型在机器阅读理解鲁棒性任务中的性能比较

基线模型 中文 $\text{RoBERTa}_{\text{large}}$ 预训练模型已在多个中文机器阅读理解数据集上取得最好的结果[16], 本文在训练集上直接微调中文 RoBERTa 模型作为基线模型。如表 3 所示, 其在领域内数据集上的 F1 值和 EM 值分别为 85.98%和 74.38%, 高于 $\text{Dureader}_{\text{Robust}}$ 官方提供的最佳模型 $\text{ENRIE2.0}_{\text{large}}$[22] 的 84.68%和 72.74%。然而, $\text{RoBERTa}_{\text{large}}$ 在 3 个鲁棒性测试集上的性能, 与其在领域内测试集上的性能相比, F1 值降低 18%~25%。由此可见, 现有的机器阅读理解模型, 即使是在领域内取得较高性能的 RoBERTa, 在鲁棒性测试上仍存在很大的缺陷。

多阶段微调 旨在让模型借助其他大规模同类数据集进行迁移学习, 从而提高模型的鲁棒性。如表 3 所示, 多阶段微调的模型在 3 个鲁棒性测试集上的性能比基线模型都有明显的提升, F1 值分别提高 8.8%, 9.2%和 3.6%。说明通过迁移自其他大规模数据集的预训练模型, 学习了更多的语言现象, 语义理解的能力获得提升, 因此, 复述问题和干扰句产生的扰动对其产生的影响相对较小。尤其在过稳定测试集上, 多阶段微调的方法使模型减少对浅层线索的依赖, 模型性能显著提高。尽管可以在一定程度上克服鲁棒性问题, 多阶段微调模型的鲁棒性测试结果与领域内测试结果之间仍然存在较大的差距。

基于答案抽取和掩码位置预测的多任务学习能够有效地减轻模型过拟合的情况, 提高模型的泛化

表 3 基线模型在原问题和鲁棒性测试集的实验结果(%)

Table 3 Experimental results on origin question and robust test sets (%)

测试问题	测试集	模型性能(F1/EM)		
		RoBERTa 基线模型	多轮微调	多轮微调+多任务学习
过敏感	原问题	71.00/55.28	75.33/58.94	77.72/63.41
	鲁棒性问题	64.47/46.76	73.28/56.46	77.30/62.41
过稳定	原问题	82.61/67.54	86.38/73.03	88.19/75.88
	鲁棒性问题	60.35/41.14	69.55/52.82	65.23/46.77
泛化	领域内问题	85.98/74.38	88.21/76.85	87.96/76.99
	跨领域问题	67.26/48.17	70.84/52.90	70.24/51.64

表 4　领域内、过敏感、过稳定和泛化测试集实验结果(%)

Table 4　Experiment results on in-domain, over-sensitivity, over stability and generalization test sets (%)

模型	领域内		过敏感		过稳定		泛化	
	F1	EM	F1	EM	F1	EM	F1	EM
ERNIE2.0 基线模型	84.68	72.74	N/A	N/A	N/A	N/A	N/A	N/A
RoBERTa 基线模型	85.98	74.38	64.47	46.76	60.35	41.13	67.26	48.17
多轮微调	**88.21**	76.85	73.28	56.46	**69.55**	**52.82**	**70.84**	**52.90**
多轮微调+多任务学习	87.97	**76.99**	**77.30**	**62.41**	65.23	46.77	70.24	51.64

说明: 粗体数字表示最佳效果, N/A 表示未进行实验。

能力[17]。同时, 掩码语言模型的训练可以增强模型的上下文理解能力, 提高模型的语义理解能力。从表 3 可以看出, 利用遮蔽语言模型的多任务学习方法, 在过敏感测试集上取得的性能比较高, F1 值比基线模型提高 12.8%, 说明提高模型上下文理解能力可以让模型减少对复述句扰动的敏感性。在过稳定数据集上, 虽然多任务学习模型的 F1 值也比基线模型高 4.9%, 但比多阶段微调模型低 4.3%, 可能是由于在利用遮蔽语言模型提高上下文理解能力的同时, 干扰句中出现的共同词对语义的影响变大, 导致在过稳定测试集上性能的下降。

4.2.2 机器阅读理解鲁棒性测试集有效性

本文在构建过敏感与过稳定测试集时, 还标注了原问题。为了验证本文构建的 3 个数据集标注的有效性, 对上述 3 个模型在原问题集和测试集进行对比实验, 其中泛化测试集与领域内测试集的比较。实验结果如表 4 所示。

过敏感测试集　该数据集包含 2703 个样本, 来源于 246 个原问题。本文将基线模型用于原问题测试集和过敏感测试集进行测试, 实验结果如表 4 所示。基线模型在过敏感测试集上的 F1 和 EM 值比原问题测试集分别降低 6.5%和 8.5%, 多轮微调和多任务学习的方法降低了模型对复述句中微小差异的敏感性, 但与原问题仍存在差距。可见预训练模型对复述问题与原问题之间字面上的微小差异过度敏感, 导致性能下降, 同时也证明了该数据集在测试阅读理解模型的过敏感问题时的有效性。

过稳定测试集　该测试集包含 496 个样本, 来源于 456 个原问题样本。从表 4 可以看出, 所有模型在过稳定测试集上的 F1 和 EM 值明显下降。以基线模型为例, 相较于原问题, 测试集分别降低 22.3%和 26.4%。由此可见, 在遇到干扰句时, 预训练模型倾向于根据共现的词语去选择干扰句中的同类实体作为答案, 体现现有的神经网络模型倾向于通过浅层线索, 而非通过深入理解语义去解决自然语言处理任务。同时证明本文构建的过稳定测试集能够有效地反映模型在稳定性方面的性能。

泛化测试集　该数据集包含 1036 个样本。实验结果(表 4)表明, 与领域内数据集的性能相比, 基线模型的 F1 和 EM 值分别降低 18.7%和 26.2%, 多轮微调和多任务学习方法也存在同样的差距, 表明在缺少领域知识时, 预训练模型在领域外数据上的性能远低于领域内。本文构建的泛化测试集可用来有效地评价机器阅读理解模型的泛化性能。

5 结语

针对机器阅读理解中模型的鲁棒性评价问题, 本文构建过敏感、过稳定和泛化 3 个中文阅读理解鲁棒性测试数据集以及对应的原问题测试集, 以便比较模型性能。同时, 本文采用目前最好的中文机器阅读理解模型作为基准系统, 其在鲁棒性测试数据集上的性能大幅下降表明本文鲁棒性测试集的有效性。此外, 本文还提出基于答案抽取和掩码位置预测的多任务学习模型和多阶段微调策略, 提高了中文机器阅读理解模型的鲁棒性。未来的工作中, 将在阅读理解模型的训练阶段尝试加入更多类型的扰动, 进一步提高模型在鲁棒性任务中的性能。

参考文献

[1] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100,000+ questions for machine comprehension of text // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, 2016: 2383–2392

[2] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Con-

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019: 4171–4186

[3] Lan Z, Chen M, Goodman S, et al. Albert: a lite bert for self-supervised learning of language representations // 8th International Conference on Learning Representations. Addis Ababa, 2020: 1–14

[4] Tang H, Liu J, Li H, et al. DuReaderrobust: a Chinese dataset towards evaluating the robustness of machine reading comprehension models [EB/OL]. (2020–04–23)[2020–08–06]. https://arxiv.org/abs/2004.11142

[5] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017: 2021–2031

[6] Ribeiro M T, Singh S, Guestrin C. Semantically equivalent adversarial rules for debugging nlp models // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018: 856–865

[7] Trischler A, Wang T, Yuan X, et al. NewsQA: a machine comprehension dataset // Proceedings of the 2nd Workshop on Representation Learning for NLP. Vancouver, 2017: 191–200

[8] Rajpurkar P, Jia R, Liang P. Know what you don't know: unanswerable questions for SQuAD // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018: 784–789

[9] Cui Y, Liu T, Che W, et al. A span-extraction dataset for Chinese machine reading comprehension // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, 2019: 5886–5891

[10] He W, Liu K, Liu J, et al. DuReader: a Chinese machine reading comprehension dataset from real-world applications // Proceedings of the Workshop on Machine Reading for Question Answering. Melbourne, 2018: 37–46

[11] Zhao Z, Dua D, Singh S. Generating natural adversarial examples [EB/OL]. (2018–02–23)[2020–08–06]. https://arxiv.org/abs/1710.11342

[12] Kavumba P, Inoue N, Heinzerling B, et al. When choosing plausible alternatives, clever hans can be clever // Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing. Hong Kong, 2019: 33–42

[13] Roemmele M, Bejan C A, Gordon A S. Choice of plausible alternatives: an evaluation of commonsense causal reasoning // Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI. Stanford, 2011: 90–95

[14] Fisch A, Talmor A, Jia R, et al. MRQA 2019 shared task: evaluating generalization in reading comprehension // Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Hong Kong, 2019: 1–13

[15] Liu Y, Ott M, Goyal N, et al. Roberta: a robustly optimized bert pretraining approach [EB/OL]. (2019–07–26)[2020–08–06]. https://arxiv.org/abs/1907.11692

[16] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinese bert [EB/OL]. (2019–10–29)[2020–08–06]. https://arxiv.org/abs/1906.08101

[17] Liu X, He P, Chen W, et al. Multi-task deep neural networks for natural language understanding // Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence, 2019: 4487–4496

[18] Conneau A, Kiela D, Schwenk H, et al. Supervised learning of universal sentence representations from natural language inference data // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017: 670–680

[19] McCann B, Bradbury J, Xiong C, et al. Learned in translation: contextualized word vectors // Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. Long Beach, 2017: 6294–6305

[20] Liu C, Yu D. BLCU-NLP at COIN-Shared Task1: stagewise fine-tuning BERT for commonsense inference in everyday narrations // Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing. Hong Kong, 2019: 99–103

[21] Li X, Zhang Z, Zhu W, et al. Pingan smart health and SJTU at COIN-Shared Task: utilizing pre-trained language models and common-sense knowledge in machine reading tasks // Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing. Hong Kong, 2019: 93–98

[22] Sun Y, Wang S, Li, et al. ERNIE 2.0: a continual pre-training framework for language understanding // Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence. New York, 2020: 8968–8975