

文章编号: 1003-0077(2022)11-0121-10

中文开放域问答系统数据增广研究

杜家驹^{1,2,3},叶德铭^{1,2,3},孙茂松^{1,2,3}

(1. 清华大学 计算机科学与技术系,北京 100084;
2. 清华大学 人工智能研究院,北京 100084;
3. 清华大学 智能技术与系统国家重点实验室,北京 100084)

摘要: 开放域问答是自然语言处理中的重要任务之一。目前的开放域问答模型总是倾向于在问题和文章之间做浅层的文本匹配,经常在一些简单问题上出错。这些错误的原因部分是由于阅读理解数据集缺少一些真实场景下常见的模式。该文提出了几种能够提高开放域问答鲁棒性的数据增广方法,能有效减少这些常见模式的影响。此外,我们还构造并公开发布了一个新的开放域问答数据集,能够评估模型在真实场景下的实际效果。实验结果表明我们提出的方法在实际场景下带来了性能提升。

关键词: 开放域问答;鲁棒性;数据增广

中图分类号: TP391

文献标识码: A

Data Augmentation in Chinese Open-domain Question Answering

DU Jiaju^{1,2,3}, YE Deming^{1,2,3}, SUN Maosong^{1,2,3}

(1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. Institute of Artificial Intelligence, Tsinghua University, Beijing 100084, China;
3. State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

Abstract: Open-domain Question Answering (OpenQA) is an important task in natural language processing. However, OpenQA models tend to match texts on a superficial level between questions and documents and often make stupid errors on some easy questions. Part of the reason for these errors is that reading comprehension datasets lack some common patterns in the actual scenes. To eliminate the effects of these patterns, we propose several methods to improve the robustness of OpenQA models. Besides, we build a new dataset to evaluate the performance of models in the real world. The experimental results show that the proposed methods can improve the performance of OpenQA models on this dataset.

Keywords: open-domain question answering; robustness; data augmentation

0 引言

智能问答是自然语言处理中的重要任务,其目标是回答人类用自然语言形式提出的各种问题,涉及检索、语义匹配、推理等自然语言处理中的重要技术。与搜索引擎不同,它能够为用户直接提供答案,省去用户阅读文档的时间,拥有重要的实际价值。

开放域问答是智能问答的主要研究领域之一。它的目标是回答任意领域的问题,而不是把问题限

定在某个领域内。给定一个问题 Q 以及许多文档(例如维基百科的全部内容,或互联网上的所有网页),模型需要根据这些文档回答问题。开放域问答模型的一种常见实现方式由检索器和阅读器两部分构成。检索器需要从给定的文档集合中检索出可能相关的一些文档,之后阅读器需要阅读并综合处理这些文档,得出问题的答案。这两部分通常是分离的,检索器通常使用传统方法进行检索,如 TF-IDF 或 BM25,或者使用搜索引擎(如 Bing),阅读器一般使用标注好的阅读理解数据集训练,其文档通常来

收稿日期: 2020-12-31 定稿日期: 2021-03-04

基金项目: 国家重点研发计划项目(2020AAA0106500)

自于维基百科。阅读理解是智能问答的另一重要研究领域。目前,阅读理解模型通常在大规模预训练语言模型上微调(Fine-tune),在各类数据集上能够接近甚至超出人类水平。

但是,阅读理解和开放域问答是不同的场景,所以用阅读理解数据集训练得到的模型直接迁移到开放域问答往往会遇到很多问题。我们使用目前通行的方法训练了一个阅读理解模型,并在真实场景下进行了评测,发现了一些问题,如图1所示。在示例1中,检索器抽取网页的主要内容并进行分段处理,会出现形如“乔戈里峰”这样仅含一个实体的较短的段落。此时阅读器看到了问题要求回答某一座山峰,且段落中包含一座山峰后就直接输出了“乔戈里峰”。在示例2中,问题和文章中涉及的地点限定词是不同的,但模型仍然输出了属于“山峰”这一类型的一个实体。这两个例子说明目前的阅读理解模型没有真正地理解问题和文章之间的关系,只是学习到了浅层的文本匹配。

示例1:

问题:世界上最高的山峰是什么?

文章:乔戈里峰

答案:乔戈里峰

示例2:

问题:云南最高的山峰是什么?

文章:白云峰又名层岩,是长白山的主峰,位于长白山天池西侧。

海拔2691米,是中国东北地区最高的山峰。

答案:白云峰

图1 阅读理解模型错误输出示例

上述问题出现的原因本质上是因为目前开放域问答系统的阅读器使用阅读理解数据集训练。在开放域问答中,人们通常是先想到问题再去找相关文章。而在标注阅读理解数据集时,标注者在阅读文章后提出若干问题,文章内容的先入为主导致他们提出的问题通常与文章的句子比较相近。所以较短的段落以及与问题中不同的限定词这样的情况并不会在阅读理解数据集中出现,进而使开放域问答模型在这几种情况下出错。为了解决这些问题,我们提出了几种能在现实场景下增强问答系统鲁棒性的数

据增广方法。其中包括针对无上下文的答案,提出使用类似答案的文章;针对问题文章中限定词不匹配的问题,提出条件删除,构造不含问题中条件的文章;针对模型会受到与问题高度相似的句子影响,提出句子替换,用高度相关的句子替换含有答案的句子。这些数据增广方法能够帮助阅读器获得辨别这几种情况的能力,进而提升开放域问答系统的鲁棒性。

考虑到阅读理解数据集和实际场景的巨大差异,我们构造了一个开放域问答数据集OpenCQA用于评测。与以往阅读理解数据集的区别是:OpenCQA给出的文章是从网页中提取出来的,且没有做相关的过滤处理,更接近人类阅读网页时的情景,具有更强的干扰性。实验结果表明本文提出的几种数据增广方法都在这个数据集上取得了一定的效果提升。

本文的贡献主要包含:

(1)发现目前的阅读理解模型不能够处理真实场景下的文章,提出了几种能够增强鲁棒性的数据增广方法,并在真实场景下获得了性能提升。

(2)为了修正目前阅读理解数据集不能有效评估实际场景下模型效果的问题,构造并发布一个开放域问答数据集。

1 相关工作

智能问答近年来已经有了许多进展,性能有了很大的提升。这些进展得益于许多数据集的出现。在中文领域,阅读理解数据集有以下几种形式:完型填空形式,给出的文本有若干个词被删除,需要根据上下文恢复这些词,如CMRC 2017^[1];抽取式,答案是给出的文本中的一个区间,如CMRC 2018^[2],DRCD^[3],WebQA^[4],XQA^[5];生成式,需要根据问题和文章生成一段文本作为答案,如DuReader^[6],Gaokao History^[7];以及多项选择形式,从四个答案中选出最合适的一项,如Gaokao Challenge^[8-9],MCQA 2017^[10],ChID^[11],C3^[12]。总体上来说,上述中文的数据集的规模远远小于英文数据集,如SQuAD^[13],TriviaQA^[14],Natural Questions^[15]等。

当前问答模型遇到的一个重要问题是模型总是倾向于做简单的文本匹配。为了解决这一问题,Jia和Liang^[16]通过对文章进行一定的修改,可以误导模型输出错误的答案,并通过加入对抗样本训练缓解这一问题。Zhu^[17]等人提出通过神经网络模型生

成若干不可回答的问题来帮助训练问答模型。Welbl^[18]等人发现了问答系统的不敏感性(Under-sensitivity),即对文章做出一定的更改后,模型仍然会输出原有的答案,然后用了对抗训练的方法降低不敏感性。Back^[19]等人提出了NeurQuRI,能够检测出问题中的一些条件能否被答案所满足,但这一模型仅使用答案的表示作为输入,没有显式地对文章中的片段和问题中的条件进行匹配。

除阅读理解外,近年来开放域问答也有了许多进展。Chen^[20]等人提出了两阶段的“检索+阅读”框架,在此基础上,有一些研究专注于提升这一类模型在某方面的效果,如多文章训练^[21]、文章排序^[22]等。这类方法在检索阶段需要使用一个传统的检索器,因此有一些工作尝试使用神经网络模型来做检索。Lee^[23]等人提出用问题和文章编码后的向量做检索,同时优化检索任务和阅读任务,形成一个端到端的模型。在此基础上,Guu^[24]等人为这种端到端的模型提出了一种预训练方法,提升了其性能。

此外,还有一些特殊技巧被应用于开放域问答中。如Seo^[25]等人提出了PIQA,把维基百科中所有区间都编码为向量,用向量相似度直接从这些区间中检索出答案,避免了阅读大量文本带来的性能开销。还有一些工作把知识图谱融入检索和阅读中^[26-27]。

2 模型

在本节,我们首先介绍作为基准模型的开放域问答系统,包含目前抽取式阅读理解的常见做法和对某些特殊情况的处理,然后介绍几种能够改善系统鲁棒性的数据增广方法。

2.1 基准模型

基准模型采用检索+阅读的流水线形式,其中检索器使用搜索引擎返回的结果,阅读器在抽取式阅读理解数据集上训练。抽取式阅读理解是阅读理解的一种特殊形式,可以形式化为如下问题:给定一个问题Q,以及若干篇文章 D_1, D_2, \dots, D_n ,从这些文章中选择一个片段a作为答案输出,或者输出“无答案”。目前,抽取式阅读理解问题最好的解决方法是利用大规模的预训练模型。

以BERT^[28]为例,如图2所示,对于问题Q和文章 D_i ,我们使用WordPiece^[29]切分它们,得到符号序列 $Q = q_1, q_2, \dots, q_m, D_i = d_1, d_2, \dots, d_l$ 。把问

题和文章拼起来,得到输入序列[CLS] $q_1 q_2 \dots q_m$ [SEP] $d_1 d_2 \dots d_l$ [SEP],其中[CLS]和[SEP]是两个特殊符号,分别用于输入的开头和分隔问题与文章。然后,BERT模型将会处理这个序列,经过embedding层和若干self-attention及全连接层后,输出一组向量 $\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_n \in R^H$,与输入序列中的符号一一对应。之后,这些向量会分别通过两个线性层及softmax层,得到每个符号是答案的开始位置或结束位置的概率:

$$P_{\text{start}}(i) = \text{softmax}(\mathbf{W}_{\text{start}} \mathbf{h}_i + \mathbf{b}_{\text{start}}) \quad (1)$$

$$P_{\text{end}}(i) = \text{softmax}(\mathbf{W}_{\text{end}} \mathbf{h}_i + \mathbf{b}_{\text{end}}) \quad (2)$$

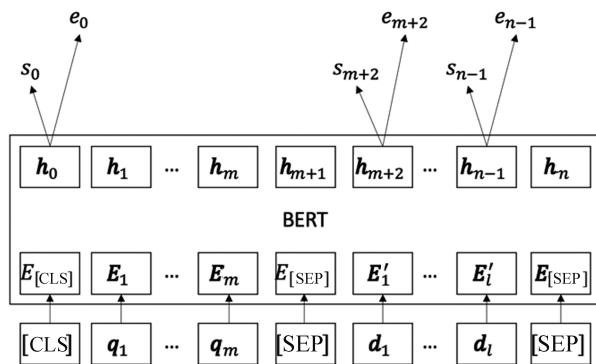


图2 BERT阅读理解模型示例

这里假定答案的开始位置和结束位置的概率分布是相互独立的。我们把 $\mathbf{W}_{\text{start}} \mathbf{h}_i + \mathbf{b}_{\text{start}}, \mathbf{W}_{\text{end}} \mathbf{h}_i + \mathbf{b}_{\text{end}}$ 称为*i*位置作为初始位置和结束位置的分数,记为 s_i 和 e_i 。答案是某一区间 $[a, b]$ 的概率为:

$$P([a, b]) = P_{\text{start}}(a)P_{\text{end}}(b) \propto \exp(s_a + e_b) \quad (3)$$

可以认为文章中每个答案区间的分数为开始位置和结束位置分数之和。

由于输入序列包含问题和文章,所以只需要考虑起始位置和结束位置都在文章对应位置的区间。一般来说,答案通常不会太长,所以在进行预测时,还会剔除掉那些太长的区间。另外,文章中不一定含有回答问题所必要的信息,所以模型需要针对这种情况给出“无答案”的预测。我们把包含[CLS]这个符号的区间作为一个特殊的“答案”。如果这个答案的概率是最高的,就认为模型输出了“无答案”。

在实际情况中,有的文章的长度会超出预训练模型能处理的长度上限(如512个符号)。在这种情况下,需要采用滑动窗口的形式把文章划分为有重叠的若干段,如图3所示。其中会有一些段落不含有答案,在训练时需要让这些段落预测“无答案”。在阅读完所有的文章后,需要把所有文章或段落预

测出的答案合并起来,每个答案的分数是它在所有文章段落中被预测的分数的最大值。在下文中,本小节提到的模型记为 Baseline。

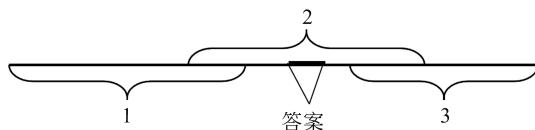


图3 文章长度超过预训练模型能处理的上限时的滑动窗口机制(仅第2个窗口含有答案,用其他窗口训练时训练目标是输出“无答案”。)

2.2 提升鲁棒性的数据增广

为了缓解上文提到的问答模型的问题,我们提出了若干种数据增广的方法,期望模型在学习过这些增广数据之后能够避免上述的问题。

2.2.1 针对无上下文的答案的增广

在实际场景中,如果某一个段落仅包含一个实体,且这个实体的类型和问题询问的类型是匹配的,那么模型有很大的可能性会直接把这个实体预测为答案。但是这种情况是不合理的,因为文章没有提供任何与问题有关的信息。

为了避免这种情况,可以手工构造出这样的情况,并要求模型不预测这个实体。假定有一个实例,包含问题 Q ,一些文章 D_1, \dots, D_n ,以及一些答案 a_1, \dots, a_m 。我们构造了这样的实例,它的问题是 Q ,文章是 a_1, \dots, a_m ,每篇文章都是一个原有的答案,答案为空。图 4(1)展示了一个替换的例子。对已有的所有数据做这样的处理后,把所有新实例加入到原有的数据集中。这种数据增广方法记为类似答案的文章(Answer-like Context, AC)。

2.2.2 针对缺少条件的文章的增广

人类在阅读文章并回答问题时通常会快速找到问题所隐含的条件,然后在文章中寻找能够匹配上所有隐含条件的地方。如果缺少一个条件或条件错误,一般会认为无法回答此问题。例如,在问题“云南最高的山峰是什么?”中,“云南”“最高”“山峰”可以认为是隐藏的条件,如果“云南”或其近义词没有在文章中出现,就无法回答问题。目前的模型无法满足这种需求,所以我们针对这种情况提出了一种增广数据的方法。

假定一个实例含有问题 Q ,文章 D_1, \dots, D_n ,以及答案 a_1, \dots, a_m 。首先使用 Stanford CoreNLP^[30]对 Q 进行分词、命名实体识别和依存语法分析。然后,抽取出问题中所有的命名实体、名词和所有形如“第

……”“最……”的词,这些词被视为回答问题必需的条件,构成条件集合 $P = \{p_1, \dots, p_l\}$ 。一般来说,一个问题通常可以抽取出 2~4 个条件,如果一篇文章中不含有任意一个条件,那么它几乎不可能含有回答问题所需要的信息。把文章中每个条件都删除掉,得到一个新的实例,包含问题 Q ,文章 D_1-P, \dots, D_n-P ,以及空答案。然后新实例加入到已有的数据集中,记为条件删除(Condition Deletion, CD)。图 4(2)给出了条件删除的一个数据增广示例。

(1) 类似答案的文章: ..
问题:世界上最高的山峰是什么?..
文章:珠穆朗玛峰是世界上最高的山峰。(答案:珠穆朗玛峰)..
新文章:珠穆朗玛峰(无答案)..
(2) 条件删除: ..
问题:世界上最早的报纸诞生于?..
文章:...北宋末年(公元 11, 12 世纪)出现的印刷报纸,不仅是中国新闻史上最早的印刷报纸,也是世界新闻史上最早的印刷报纸。...(答案:中国)..
新文章:...北宋末年(公元 11, 12 世纪)出现的印刷,不仅是中国新闻史上的印刷,也是新闻史上的印刷。...(无答案)..
(3) 句子替换: ..
问题:范廷硕是于何时何地出生的?..
文章:范廷硕于 1919 年 6 月 15 日在越南宁平省天主教发艳教区出生,童年时接受良好教育后...(答案:1919 年 6 月 15 日在越南宁平省天主教发艳教区)..
新文章:著名历史学家黄仁宇出生于 1920 年;童年接受良好教育之后...(无答案)..

图4 类似答案的文章示例

2.2.3 针对与问题高度相关的句子的增广

如果文章中含有与问题高度相似的句子但此句子又不含有真正的答案,那么这个句子中的一些区间(或词语)就很容易被预测为答案。为了解决这一问题,我们提出一种利用句子替换增广数据的方法。

假定数据集中已经有了若干问题 Q_1, \dots, Q_n ,我们首先找到与这些问题语义高度相关的一些句子。具体来说,在搜索引擎中检索这些问题,得到一些文章,并拆分为许多句子 S_1, \dots, S_m ,构成候选句子集合 S 。之后训练一个模型,使之能够辨别出哪些句子是和问题高度相似的。模型用预训练模型(如 BERT)编码问题或句子,得到一些向量 $Q_1, \dots, Q_n, S_1, \dots, S_m$ (例如使用[CLS]符号对应的向量)。定义 Q_i 和 S_j 之间的相似度为 $s = Q_i^T S_j$ 。为方便优化,从 S 中随机采样出若干个句子作为负例,计算与问题相似度 s_1, \dots, s_K ,优化目标为交叉熵损失函数:

$$L = -\log \frac{\exp s}{\exp s + \sum_{i=1}^K \exp s_i} \quad (4)$$

训练完成后,就可以检索出与问题相似的所有句子。我们用 Faiss^[31],一个十分高效的开源向量相似度检索与聚类库,进行基于向量内积的检索,为每个问题找到相似度最高的 100 个句子。

只要把文章中含有答案的句子替换成与问题高度相似且不含答案的句子,就可以认为得到的新文章不足以回答问题。模型在用这样的(问题,文章)训练之后就应当能够避免直接用问题匹配与问题高度相关的句子。这种处理方式记为句子替换(Sentence Replacement,SR)。图 4(3)给出了句子替换的一个数据增广示例。

但是,上述处理方法可能存在以下问题:如果一个句子含有与答案重合度较高的一些片段,如答案为“古埃及人”,但原文章中有一个句子含有“古代埃及人”,这个句子就不会作为无答案的句子被替换掉,导致构造的文章仍然能够回答问题。为了处理这种情况,我们引入一种启发式的匹配方法。假设答案长度为 n ,统计它的 $n(n+1)/2$ 个子串有哪些在句子中出现。如果在句子中出现的子串数量不少于 $2n-1$,就认为句子和答案是匹配的,应当被替换掉。这种处理方式记为近似句子替换(Approximate Sentence Replacement,ASR)。

最后,以上几种数据增广方法分别针对了几种不同的问题,在实际场景中可以混用这几种增广方法,记为集成(Ensemble)。

3 实验

3.1 数据集

本文主要专注于抽取式的阅读理解和开放域问答,相关的数据集主要有:

(1) CMRC2017^[1]是一个填空式的中文阅读理解数据集,但人工标注了少量抽取式的问题。其语料主要来源于《人民日报》和《格林童话》。

(2) CMRC2018^[2]是第一个标准的抽取式阅读理解数据集,其文章来源于中文维基百科。

(3) DRCD^[3]同样是一个抽取式数据集,所有问题和答案都是繁体中文,文章取自繁体中文维基百科。我们使用 OpenCC^① 进行繁简转换。

(4) WebQA^[4]是一个大规模的真实场景下的问答数据集,其问题主要来源于百度知道中的事实

性的问题,都是在非受限的场景下提出的。用搜索引擎检索问题,得到若干文章,并人工标注了答案。

(5) DuReader^[6]包含了许多从搜索引擎日志中获得的高频问题,包括事实型问题、观念型问题和是否型问题。在百度搜索和百度知道中检索这些问题,得到一些文章,并人工标注答案。与其他数据集不同的是 DuReader 给出的是完整的文章,而不是单个段落。

为了能够充分利用各个数据集,本文把这些数据集转化为了统一的格式。同时,为了与抽取式的问答模型兼容,我们删除了答案没有在给出的文章中出现的问答对。由于部分数据集没有提供测试集,因此在实验中统一把验证集和测试集合并为验证集。最终得到的数据集规模统计如表 1 所示。

表 1 数据集规模统计信息

数据集	阅读理解		开放域问答
	训练集	验证集	验证集
CMRC2017	—	5 000	—
CMRC2018	10 142	4 221	4 217
DRCD	26 936	7 017	7 003
WebQA	36 181	11 930	12 068
DuReader	135 341	4 437	4 006
合计	208 600	32 605	27 294

此外,为了评估模型在实际场景中的性能,我们利用这些数据集提供的问答对构造了一个开放域问答的数据集 OpenCQA,类似于 Chen^[20] 等人提出的做法。首先忽略数据集给出的所有文章,对于数据集中的所有问题答案对 (Q, a) ,在 Bing^② 搜索引擎中用 Q 检索并抓取前十位的网页,抽取出其主要内容作为文章 D ,与原有的问题 Q 和答案 a 合并构造一个新的 (Q, D, a) 三元组。如果原有的答案 a 没有在这些文章 D 中出现,则把 a 替换为“无答案”。问题 Q 无法检索出相关结果时直接丢弃此问答对。文章如果存在大量不可读字符或中文字符占比小于一半,也会直接丢弃。为促进中文问答系统的研究,我们公开发布了 OpenCQA^③ 数据集,包含约 20 万问题、答案,以及每个问题的参考文章。此外,为方便其他研究者,还把所有的阅读理解数据集整合在了一起,并统一成相

① <https://github.com/BYVoid/OpenCC>

② <https://www.bing.com/>

③ <https://github.com/jiajudu/openCQA>

同的格式,也同时公开发布。

3.2 评测指标

封闭域阅读理解和开放域问答都使用 EM 和 F_1 两种指标评测。假设问题有若干个可能的答案 a_1, \dots, a_n , 模型给出的预测为 a 。EM 和 F_1 的计算方法如式(5)、(6)所示。

$$EM = \max_i I(a = a_i) \quad (5)$$

$$F_1 = \max_i \frac{2 \times \text{lcs}(a, a_i).length}{a.length + a_i.length} \quad (6)$$

其中 lcs 为两个字符串的最长公共子串。计算前需要先去除这些答案中的标点符号。

3.3 实验设置

我们使用阅读理解数据集训练了若干模型, 分别使用了不同的数据增广策略。然后, 分别在阅读理解验证集和 OpenCQA 验证集上评测了性能。所有的实验都借助 Transformers 库^[32]完成, 预训练语言模型^①使用中文维基百科、新闻、问答等数据训练, 利用了全词 Mask(Whole Word Masking)技术, 区分大小写。微调时使用的学习率为 3e-5, 其中前 10% 的时间学习率由 0 线性上升至最大值, 随后线性下降至 0。使用的优化器为 Adam, 共训练 2 轮。模型使用了 8 张 RTX 2080 Ti 显卡, batch size 设置为 48。其余参数均采用常见的默认值。

此外, 为了让模型能够处理是否型问题, 我们在每个段落前都添加两个特别的符号“Yes”和“No”。如果应该输出是/否, 就要求模型预测含有这两个符号的区间。在训练句子和问题的相似度模型时, 每个问题都随机采样 5 个负例。在评估集成增广方法时, 使用三种增广方法分别构造了三组额外的数据, 每组数据都随机采样出 1/3 的数据, 加入原有的训练集。表 2 给出了各种增广策略对应的数据规模以及在原数据集的基础上增加的比例。

表 2 数据增广规模

增广策略	数据规模	增广比例/%
Baseline(基准)	208 600	—
AC(类似答案的文章)	399 295	91.42
CD(条件删除)	674 531	223.36
SR(句子替换)	395 944	89.81
ASR(近似句子替换)	395 944	89.81
Ensemble(集成)	489 923	134.86

3.4 开放域问答实验结果

在进行开放域问答的评测时, 一个问题会对应十篇文章, 但是这些文章中可能会存在特别长的段落。如果直接阅读这些段落, 就会消耗大量的计算资源。由于计算资源有限, 我们把所有的段落用滑动窗口切分为片段, 仅取出 512 个片段阅读。片段数量超过 512 时, 优先选取每个段落靠前的片段, 丢弃长段落靠后的内容。这样模型阅读的内容会覆盖所有段落, 而又不在极长段落上花费过多资源。

与阅读理解不同, OpenCQA 中有许多无答案的例子。但是在使用不同的数据训练之后, 不同的模型输出“无答案”的概率是不同的。为了保证比较的公平性, 我们引入了“分数差”^②这一概念。假设已经获得了文章中所有区间的最高得分 s_m 和“无答案”的分数 s_{null} , 可以仅在 $s_m > s_{\text{null}} - \tau$ 时输出一个非空的答案。这里 τ 是一个可以任意调整的变量。显然 τ 减小时, 输出“无答案”的概率 P_{null} 增大。我们可以适当地取一些 τ 值, 得到 $EM-P_{\text{null}}$ 曲线(见图 5)和 F_1-P_{null} 曲线(见图 6)。曲线的最高点代表每个模型在调整 τ 后能达到的最优性能, 如表 3 左侧两列所示。曲线的最左侧代表 $P_{\text{null}}=0$ (即模型必须输出一个非空答案)时的性能, 如表 3 右侧两列所示。

表 3 开放域问答性能比较

(单位: %)

	最优性能		必须输出答案时的性能	
	EM	F_1	EM	F_1
Baseline	50.49	57.51	39.44	49.05
AC	55.18	62.26	40.63	50.48
CD	55.48	62.48	41.60	51.36
SR	55.11	60.57	40.78	49.27
ASR	55.73	60.71	40.69	49.06
Ensemble	58.38	64.40	41.83	50.80

注: AC: 类似答案的文章; CD: 条件删除; SR: 句子替换; ASR: 近似句子替换。

从表 3、图 5 和图 6 可以观察到:

(1) $P_{\text{null}}=1$ 时, 所有模型的 EM 和 F_1 都相同, 这是因为数据集中无答案数据的比例是固定的。

(2) $P_{\text{null}}=0$, 即模型必须给出非空答案时, 我

① <https://huggingface.co/hfl/chinese-roberta-wwm-ext>

② https://github.com/huggingface/transformers/blob/v3.4.0/examples/question-answering/run_squad.py#L542

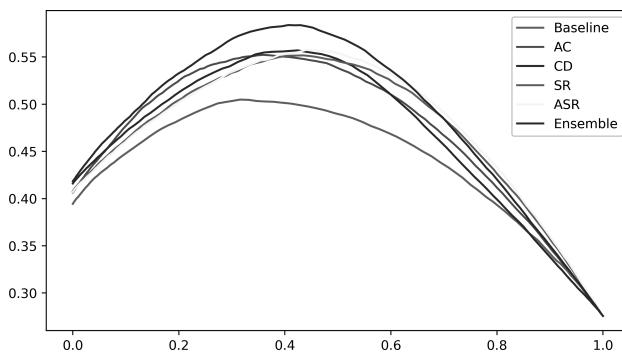


图 5 开放域问答各模型曲线

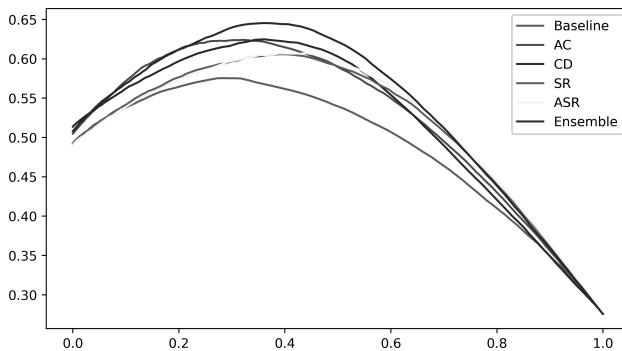


图 6 开放域问答各模型曲线

们提出的方法都能带来一些提升,在 EM 上的提升更明显(+1.2~2.2,集成方法+2.4),只有句子替换(SR/ASR)在 F_1 指标上提升较小,可能是因为近义词没有被替换时问题仍然是可回答的。

(3) P_{null} : $0 \rightarrow 1$ 时,EM 和 F_1 都先上升后下降。可以看到 P_{null} 处于(0,1)之间时,各个模型相比于基准模型的提升更大(+3.0~5.3,集成方法+6.9~7.9,见表 3 左侧两列)。这说明数据增广不仅

能够提升在可回答问题上的效果,还可以提高模型判断问题是否可回答的能力。

(4) AC 的 EM/ F_1 相比于基线模型增长了 4.69/4.75,我们对性能的增长做了来源分析。性能的增长来源于三方面:①问题有答案且 AC 和基线模型均预测出了答案的情况下,AC 贡献了 1.14/1.23 的性能增长。②在 11.45% 的问题上,AC 把基线模型预测出的有答案改成了无答案,其中有 5.22% 的问题 AC 是正确的,贡献了 2.92/1.91 的性能提升。尽管正确率不到一半,但由于发生错误的那部分问题基线模型也没有完全回答正确,所以总体贡献仍然是正的。③在 8.16% 的问题上,AC 把基线模型预测出的无答案改成了有答案,其中有 6.08% 的问题 AC 是正确的,贡献了 0.63/1.61 的提升。因此我们同样可以得到 AC 不仅提高了预测“无答案”的比例,在其他方面也有所贡献。其他策略也有相似的贡献分布,在此不再列出。

(5) 综合 EM 和 F_1 指标,集成三种方法的数据增广能够取得最好的效果,条件删除(CD)是最好的单一数据增广方法。

我们还分析了各种数据增广方法在 OpenCQA 各个子数据集上的效果,如表 4 所示。实验结果表明,无论在哪个子数据集上,模型在可回答的问题上的 EM 和 F_1 都有所提升,其中条件删除(CD)仍然是最好的方法。在除 WebQA 外的数据集中,模型识别问题是否可回答的成功率也都有所提升。在 WebQA 上数据增广会导致识别成功率下降,其原因主要是 WebQA 中几乎的所有问题都有答案,数据增广导致模型把部分问题预测为了“无答案”。

表 4 各种数据增广方法在 OpenCQA 的子数据集上的表现

(单位: %)

	所有数据集	CMRC2018	DRCD	DuReader	WebQA
Baseline	73.33/54.39/67.64	38.27/43.10/62.45	63.07/41.88/56.56	56.14/35.33/49.98	97.25/63.99/75.65
AC	78.80/56.03/69.60	64.33/40.08/59.11	68.64/43.50/58.38	67.97/35.43/51.27	93.34/66.52/78.48
CD	72.86/57.37/70.83	67.11/48.49/67.46	64.66/43.40/57.96	57.79/36.54/52.41	84.64/67.49/79.39
SR	74.79/56.24/67.95	57.60/46.76/65.44	66.37/40.73/54.56	59.14/34.91/48.42	90.88/67.12/76.79
ASR	74.41/56.12/67.66	58.41/46.50/64.66	65.64/40.26/53.97	57.11/33.71/47.72	90.84/67.34/76.75
Ensemble	73.33/57.68/70.06	75.05/46.89/66.76	64.20/44.51/59.38	61.43/35.52/50.14	81.97/68.02/78.08

注:根据问题最初来源于哪个数据集进行划分。每个单元格的内容包含三项:模型正确预测问题是否可回答概率,模型在有答案问题上的 EM、 F_1 。AC: 类似答案的文章;CD: 条件删除;SR: 句子替换;ASR: 近似句子替换。

最后,我们人工评测了本文提到的三种错误情况(无上下文的答案;问题文章中限定词不匹配;与问题高度相似的句子)的出现频率。从基线模型判

断出错的例子中随机选取了 100 个,手工标注了每个错误例子是否与三种错误情况相关。三种错误情况分别对应了其中 17%、23%、9% 的例子,总计占

所有错误例子的一半左右。

3.5 阅读理解实验结果

表 5 展示了各种模型在 3.1 节中构造的阅读理解数据集上的结果。模型有可能对于某些问题输出“无答案”，但由于此数据集中所有问题都有答案，所以我们强制所有模型忽略“无答案”，必须输出一个答案。从表 5 可见，各种数据增广方法在阅读理解这一任务上对性能影响不大，或者会使性能略微下降。这说明阅读理解和开放域问答是两个差别较大的任务，OpenCQA 将有助于更准确地评估未来模型的效果。

表 5 阅读理解性能比较（单位：%）

	EM	F_1
Baseline	71.48	82.68
AC	70.94	82.09
CD	70.86	81.91
SR	71.43	82.52
ASR	71.21	82.52
Ensemble	70.44	81.89

3.6 案例分析

在这一部分，我们首先给出从验证集中选出的若干例子，证明数据增广能够有效地解决上述几种问题。图 7 列出了这些例子以及模型给出的答案和得分。在示例 1 中，与基线模型相比，加入由答案作为文章的实例(AC)后，模型不再把“乔戈里峰”预测为答案，原有的一些高分答案的分数也大大下降。在示例 2 中，条件删除(CD)模型能够识别出“云南”和“东北”是不同的条件，降低了“白云峰”的得分。在示例 3 中，句子替换(SR)模型降低了作为干扰项的某抑制剂的得分。总之，对模型输出的答案分数的观察说明数据增广达到了目的。

4 结论

本文针对阅读理解模型在实际场景中出现的几种问题，提出了能够增强中文开放域问答鲁棒性的数据增广方法。实验结果表明这几种方法能够提升模型在实际场景中的性能。本文还发布了一个开放域问答的数据集，用于评估中文问答系统的性能。

虽然在使用本文提出的几种方法后，模型能够

示例 1：无上下文的答案

问题：世界上最高的山峰是什么？

文章：乔戈里峰

Baseline：乔戈里峰 (14.88), 乔戈里 (10.34), 无答案 (10.27), 戈里峰 (7.72)

AC：无答案 (9.64), 乔戈里峰 (-4.05), 峰 (-7.76), 乔戈里 (-7.80)

示例 2：缺少条件的文章

问题：云南最高的山峰是什么？

文章：白云峰又名层岩，是长白山的主峰，位于长白山天池西侧。海拔 2691 米，是中国东北地区最高的山峰。

Baseline：白云峰 (17.09), 白云峰又名层岩，是长白山 (14.59), 长白山 (11.74), 白云 (11.32)

CD：无答案 (5.43), 白云峰 (4.45), 白云峰又名层岩，是长白山 (0.82), 白云峰又名层岩，是长白山的主峰，位于长白山天池西侧。 (0.73)

示例 3：与问题高度相似的句子

问题：与重度抑郁症、创伤后应激障碍有间接相关性的为哪种营养因子？

文章：脑源性神经营养因子减少与自杀有直接相关性，跟重度抑郁症、创伤后应激障碍、精神分裂症、强迫症有间接相关性。一些解剖研究发现自杀者的海马体和前额叶皮质中的脑源性神经营养因子有所减少，不论死者是否患有精神疾病都一样。

选择性 5-羟色胺再摄取抑制剂 (SSRI) 是现今最常用的一类抗抑郁药。

Baseline：选择性 5-羟色胺再摄取抑制剂 (10.55), 脑源性神经营养因子 (9.87)

SR：脑源性神经营养因子 (12.48), 脑源性神经营养因子减少 (10.88)

图 7 案例分析(括号内是答案的分数)

避免犯某些错误，但距离人类阅读理解水平还有一定距离，仍然有很多错误情况没有解决。所以我们下一步将会进一步地分析目前问答系统的弱点，并针对问题进行改进。

参考文献

- [1] CUI Y, LIU T, CHEN Z, et al. Consensus attention-based neural networks for Chinese reading comprehension[C]//Proceedings of COLING, the 26th International Conference on Computational Linguistics: Technical Papers, 2016: 1777-1786.
- [2] CUI Y, LIU T, CHEN Z, et al. Dataset for the first evaluation on Chinese machine reading comprehension [C]//Proceedings of the 11th International Conference on Language Resources and Evaluation, 2018: 2721-2725.
- [3] SHAO C C, LIU T, LAI Y, et al. DRCD: A Chinese machine reading comprehension dataset[J]. arXiv preprint arXiv: 1806.00920, 2018.
- [4] LI P, LI W, HE Z, et al. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering[J]. arXiv preprint arXiv: 1607.06275, 2016.
- [5] LIU J, LIN Y, LIU Z, et al. XQA: A cross-lingual open-domain question answering dataset[C]//Proceed-

- ings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2358-2368.
- [6] HE W, LIU K, LIU J, et al. DuReader: A Chinese machine reading comprehension dataset from real-world applications[C]//Proceedings of the Workshop on Machine Reading for Question Answering, 2018: 37-46.
- [7] ZHANG Z, ZHAO H. One-shot learning for question answering in Gaokao history challenge[C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018: 449-461.
- [8] CHENG G, ZHU W, WANG Z, et al. Taking up the Gaokao challenge: An information retrieval approach [C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016: 2479-2485.
- [9] GUO S, ZENG X, HE S, et al. Which is the effective way for Gaokao: Information retrieval or neural networks? [C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017: 111-120.
- [10] GUO S, LIU K, HE S, et al. IJCNLP-2017 Task 5: Multi-choice question answering in examinations [C]//Proceedings of the IJCNLP, Shared Tasks, 2017: 34-40.
- [11] ZHENG C, HUANG M, SUN A. ChID: a large-scale Chinese IDiom dataset for cloze test[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 778-787.
- [12] SUN K, YU D, YU D, et al. Investigating prior knowledge for challenging Chinese machine reading comprehension[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 141-155.
- [13] 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, 2018: 784-789.
- [14] JOSHI M, CHOI E, WELD D S, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1601-1611.
- [15] KWIATKOWSKI T, PALOMAKI J, REDFIELD O, et al. Natural questions: A benchmark for question answering research[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 453-466.
- [16] JIA R, LIANG P. Adversarial examples for evaluating reading comprehension systems [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 2021-2031.
- [17] ZHU H, DONG L, WEI F, et al. Learning to ask unanswerable questions for machine reading comprehension[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4238-4248.
- [18] WELBL J, MINERVINI P, BARTOLO M, et al. Undersensitivity in neural reading comprehension [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing: Findings, 2020: 1152-1165.
- [19] BACK S, CHINTHAKINDI S C, KEDIA A, et al. NeurQuRI: Neural question requirement inspector for answerability prediction in machine reading comprehension[C]//Proceedings of the International Conference on Learning Representations, 2019.
- [20] CHEN D, 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, 2017: 1870-1879.
- [21] WANG Z, NG P, MA X, et al. Multi-passage BERT: a globally normalized BERT model for open-domain question answering[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 5881-5885.
- [22] WANG S, YU M, GUO X, et al. R³: Reinforced reader-ranker for open-domain question answering [J]. arXiv preprint arXiv: 1709.00023, 2017.
- [23] LEE K, CHANG M W, TOUTANOVA K. Latent retrieval for weakly supervised open domain question answering [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 6086-6096.
- [24] GUU K, LEE K, TUNG Z, et al. Realm: Retrieval-augmented language model pre-training [J]. arXiv preprint arXiv: 2002.08909, 2020.
- [25] SEO M, LEE J, KWIATKOWSKI T, et al. Real-time open-domain question answering with dense-sparse phrase index[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4430-4441.
- [26] MIN S, CHEN D, ZETTLEMOYER L, et al. Knowledge guided text retrieval and reading for open domain question answering[J]. arXiv preprint arXiv: 1911.03868, 2019.
- [27] XIONG W, YU M, CHANG S, et al. Improving question answering over incomplete KBs with knowledge-aware reader[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4258-4264.
- [28] 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, 2019: 4171-4186.
- [29] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: bridging the gap between human and machine translation[J]. arXiv preprint arXiv: 1609.08144, 2016.
- [30] MANNING C D, SURDEANU M, BAUER J, et al. The Stanford CoreNLP natural language processing
- toolkit[C]//Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014: 55-60.
- [31] JOHNSON J, DOUZE M, JÉGOU H. Billion-scale similarity search with GPUs[J]. IEEE Transactions on Big Data, 2019.
- [32] WOLF T, CHAUMOND J, DEBUT L, et al. Transformers: State-of-the-art natural language processing[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020: 38-45.



杜家驹(1997—),硕士研究生,主要研究领域为智能问答。

E-mail: djj18@mails.tsinghua.edu.cn



孙茂松(1962—),教授,博士生导师,主要研究领域为自然语言处理。

E-mail: sms@tsinghua.edu.cn



叶德铭(1995—),博士研究生,主要研究领域为智能问答。

E-mail: ydm18@mails.tsinghua.edu.cn

(上接第 120 页)

- [24] DZIKOVSKA M O, NIELSEN R D, BREW C, et al. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge[R]. North Texas State Univ Denton, 2013.
- [25] DZIKOVSKA M O, ISARD A, BELL P, et al. BEE-TLE II: An adaptable tutorial dialogue system[C]// Proceedings of the SIGDIAL Conference, 2011: 338-340.
- [26] FOLTÝNEK T, MEUSCHKE N, GIPP B. Academic plagiarism detection: A systematic literature review[J]. ACM Computing Surveys, 2019, 52(6): 1-42.



陈爽(1994—),硕士研究生,主要研究领域为自然语言处理、机器学习。

E-mail: chen60423351@email.swu.edu.cn



李莉(1967—),通信作者,博士,教授,博士生导师,主要研究领域为机器学习、人工智能及其应用。

E-mail: lily@swu.edu.cn