

面向抽取式阅读理解的数据增强研究

胡新荣^{1,2}, 徐伟^{1,2}, 罗瑞奇^{1,2}, 刘军平^{1,2}, 朱强^{1,2}, 杨捷³, 李立军⁴

(1. 武汉纺织大学计算机与人工智能学院; 2. 湖北省服装信息化工程技术研究中心, 湖北 武汉 430200; 3. 伍伦贡大学计算机与信息技术学院, 伍伦贡 2522; 4. 宁波慈星股份有限公司, 浙江 宁波 315000)

摘要: 在抽取式阅读理解中, 语言模型在训练数据较少情况下的表现不佳, 使用EDA方法虽能有效增加数据量, 但会造成数据中语义信息损失, 导致模型训练效果不佳。针对上述问题, 结合EDA提出面向少样本情况下抽取式阅读理解的数据增强方法, 在保留数据中问题正确答案的前提下对数据进行单词级和句子级数据增强。同时, 为了对句子语义影响最小的单词进行数据增强, 使用基于语义相似度的数据增强方法(DASS)计算句子中某一个单词删除前后的语义相似度, 以判断该单词对句子语义的影响, 选择对语义影响最小的单词进行数据增强, 提升训练数据质量, 以提升语言模型鲁棒性。在HotpotQA数据集上的实验结果表明, DASS可以解决模型在样本数量较少时获取语义信息不足的问题, 在样本数量为500时, 模型预测的F1值提升23.94%, 在对整个数据集使用该方法时, 模型预测的F1值提升了2.54%。

关键词: 抽取式阅读理解; EDA; 数据增强; 语义相似度; 机器阅读理解

DOI: 10.11907/rjdk.231137

开放科学(资源服务)标识码(OSID):

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2024)006-0032-06



Research on Data Augmentation for Extractive Reading Comprehension

HU Xinrong^{1,2}, XU Wei^{1,2}, LUO Ruiqi^{1,2}, LIU Junping^{1,2}, ZHU Qiang^{1,2}, YANG Jie³, LI Lijun⁴

(1. School of Computer Science and Artificial Intelligence, Wuhan Textile University; 2. Hubei Provincial Engineering Research Center for Intelligent Textile and Fashion, Wuhan 430200, China; 3. School of Computing and Information Technology, University of Wollongong, Wollongong 2522, Australia; 4. Ningbo Cixing Co., Ltd, Ningbo 315000, China)

Abstract: In extractive reading comprehension, the performance of language model is poor in the case of less training data. Although EDA method can effectively increase the amount of data, it will cause the loss of semantic information in the data, resulting in poor training effect of the model. In response to the above problems, combined with EDA, a data augmentation method for extracting reading comprehension in the case of few samples is proposed. The data is enhanced at the word level and sentence level on the premise of retaining the correct answers to the questions in the data. At the same time, the data is enhanced for the single word with the least impact on sentence semantics. The data augmentation method based on semantic similarity (DASS) is used to calculate the semantic similarity of a word in a sentence before and after deletion to determine the impact of the word on sentence semantics. The word with the least impact on semantics is selected for data enhancement to improve the quality of training data, so as to improve the robustness of the language model. The experimental results on HotpotQA show that DASS can solve the problem of insufficient semantic information when the number of samples is small. When the number of samples is 500, the F1 value predicted by the model increases by 23.94%. When this method is used for the whole dataset, the F1 value predicted by the model increases by 2.54%.

Key Words: extractive reading comprehension; EDA; data augmentation; semantic similarity; machine reading comprehension

收稿日期: 2023-05-31

基金项目: 湖北省重点研发计划项目(2020BAB116); 宁波市科技创新重大专项(2021Z069)

作者简介: 胡新荣(1973-), 女, 博士, 武汉纺织大学计算机与人工智能学院教授、硕士生导师, 研究方向为自然语言处理、计算机图形图像处理、虚拟现实、机器学习; 徐伟(1997-), 男, 武汉纺织大学计算机与人工智能学院硕士研究生, 研究方向为自然语言处理; 罗瑞奇(1989-), 男, 博士, 武汉纺织大学计算机与人工智能学院讲师、硕士生导师, 研究方向为机器学习、视觉信息处理; 刘军平(1979-), 男, 博士, 武汉纺织大学计算机与人工智能学院副教授、硕士生导师, 研究方向为机器学习、自然语言处理; 朱强(1984-), 男, 博士, 武汉纺织大学计算机与人工智能学院讲师、硕士生导师, 研究方向为自然语言处理、生物信息学; 杨捷(1985-), 男, 博士, 伍伦贡大学计算机与信息技术学院讲师, 研究方向为自然语言处理; 李立军(1979-), 男, 硕士, 宁波慈星股份有限公司副总经理, 研究方向为自然语言处理。本文通讯作者: 刘军平。

0 引言

机器阅读理解(Machine Reading Comprehension, MRC)是自然语言处理领域的研究热点,其相关任务需要高水平的语言理解能力,因此阅读理解任务常用来评估各种用于语言理解任务的语言模型。抽取式阅读理解是机器阅读理解中的一个重要分支,对智能问答研究有着重要影响。在机器阅读理解中,大多数高性能的语言模型训练都需要大型数据集的支撑,以便模型能够学习到阅读理解文章中更多的语义信息。

在抽取式阅读理解任务中,语言模型需要学习给定数据中整段文本的语义,然后结合所提问题,从文本中选择出能回答问题的正确答案,且答案一定是包含在文本之中。高性能的抽取式阅读理解模型需要依赖庞大的数据集支撑,但也存在某些特殊情况,比如:无法提供足够的数据对模型进行训练。在这种情况下,为了更好地利用有限的数据集去训练语言模型,本文选择通过数据增强(Data Augmentation)解决该问题,以此帮助训练语言模型,提高模型性能。数据增强方法常用于计算机视觉领域,在自然语言处理领域,Wei等^[1]提出的简单数据增强方法(Easy Data Augmentation, EDA)用于提高文本分类任务的性能,通过使用同义词替换、随机交换、随机删除和随机插入对数据集进行增强扩充,可以有效地提升模型性能。但在抽取式机器阅读理解任务中尚未有在数据量较少时有效提升模型性能的方法。考虑到在抽取式阅读理解任务中,问题的答案必须包含在文本内容中这一特殊性,使用EDA方法会导致文本内容大量变化,使得问题在文本中找不到对应的答案,模型训练效果也会急剧下降。基于此,本文在原有EDA方法的基础上提出了适用于抽取式阅读理解的数据增强策略。在对句子中的单词进行数据增强时,先将包含正确答案的句子排除掉,不对包含有正确答案的句子进行EDA操作;对整个句子进行数据增强时,先将包含有正确答案的句子排除掉,对其他句子进行整句的随机交换和随机插入操作。同时,为了防止在数据增强后对文章语义产生较大影响,丢失原有的语义信息,本文提出了基于语义相似度的数据增强方法(Data Augmentation based on Semantic Similarity, DASS),通过计算句子中每个单词对整个句子语义的重要程度,选择出部分对语义影响较低的单词,再实施前两种数据增强策略。

1 相关工作

在自然语言处理中,机器阅读理解是评估机器语言理解能力的主要任务之一。然而,机器阅读理解模型训练需要庞大的数据集作支撑,为了满足机器阅读理解模型训练需求,Rajpurkar等^[2]提出一个大型的机器阅读理解数据集

SQuAD, Kwiatkowski等^[3]提出一个包含30万个自然产生的问题和对应回答注释的数据集 Natural Questions。

随着大规模机器阅读理解数据集的出现,大量的深度学习模型也相继被提出,如BERT^[4]等。BERT是一种用于下游任务微调前对transformer编码器^[5]进行预训练的自监督方法,其保留了transformer编码层,丢弃了预训练任务中最终用来预测的层,并添加了预测目标任务的层,最后对目标任务上的参数进行微调。基于此,Joshi等^[6]提出新的预训练模型SpanBERT,该模型拥有更好的mask方案,加入了Span Boundary Objective训练目标,提升了BERT在抽取式阅读理解上的性能。

当训练样本数量较少时,机器阅读理解准确率会有明显下降。如何充分利用现有数据,扩充样本数量,使模型通过有限的数据也能达到较好的实验结果值得思考。数据增强是一种可以有效扩充样本数据数量的方法,其能有效地丰富数据集,解决数据集样本数量过少、样本质量较差的问题。由于分类任务的特殊性,数据增强方法可有效地在分类任务上得以体现,数据增强可生成大量结构和种类多样化的数据,能够有效地满足任务需求,减少过拟合现象,较好地提升相关任务的准确性。

数据增强最初用于图像和语音领域,并取得了较好效果。Szegedy等^[7]发现在图像和语音领域,数据增强可以有效地对样本进行增强。左航旭^[8]等在皮肤癌识别中采用过采样技术对少数类别的数据进行增强。但是由于文本数据的离散性,其对文本的处理依然具有一定挑战性。对于文本数据,本文的目标是使新生成的文本数据依然能够保持原有语义,在语法上不具有任何问题,且新生成的句子要通顺流畅。在文本数据增强方面,Wei等^[1]提出简单的数据增强方法(EDA),包括:对同一个句子随机选择一个单词进行同义词替换;随机删除某一个单词;随机交换句子中的某两个单词;随机选择一个单词,使用其同义词随机插入句子中的某个位置。Fadaee等^[9]在输入序列上对单词进行操作,利用稀有词替换常见词,从而为稀有词提供更多上下文信息。Sennrich等^[10]通过使用回译的方法,即使用自动翻译的数据和原始数据训练机器翻译系统。盛晓辉等^[11]使用半监督文本分类算法对大量无标记数据进行数据增强。Wang等^[12]使用同义词替换对推文进行分类。Xie等^[13]利用数据噪声,替换从单词频率分布中选择的单词或者插入下划线字符并将其作为占位符。Karimi等^[14]提出使用一种更为简单的数据增强方法,在原始文本中随机插入标点符号,这样可以保持单词顺序,同时改变了它们在句子中的位置,相较于EDA,该方法保持了更完整的文本信息。Hu等^[15]和Liu等^[16]利用条件语言模型进行强化学习,该模型在训练时将正确的标签附加到输入序列上以进行学习。根据Transformer模型,Sun等^[17]提出Mix-Transformer,通过将两个输入句子及其对应的标签进行线性化插值,以创建新的样本。

可以看出,目前关于机器阅读理解的数据增强方法较少,在少样本情况下对抽取式阅读理解的数据增强方法也尚未有研究。因此,在少样本情况下,对抽取式阅读理解的数据增强方法进行研究具有重要意义。

2 面向抽取式阅读理解的数据增强策略

由于抽取式阅读理解是给定的一段文章和一个问题,最后从文章中抽取出正确答案的过程,因此在对抽取式阅读理解数据集进行数据增强时,要保证文章中必须存在问题的正确答案。若使用简单的数据增强方法,可能会使得文章中的答案部分发生改变,从而导致文章中没有问题的正确答案,影响了模型训练效果。同时,随机操作还会导致文章中的一些句子丧失原本的语义信息,从而误导模型输出错误的答案。基于此,本文从单词层面和句子层面分别提出了适用于抽取式阅读理解的数据增强方法,保证文章中包含有问题的正确答案。并在此基础上,为了保证句子的语义连贯性,本文提出了基于语义相似度的数据增强方法 DASS,计算每个句子中每个单词对句子语义的影响分数,从而选择对句子语义影响最小的单词进行操作,极大程度地保证了原文的语义信息,提高了所训练数据的质量。

2.1 适应于抽取式阅读理解的单词增强策略

在自然语言处理领域的相关任务中,由于数据资源的匮乏,通常需要对有限的文本数据进行数据增强。常见方法是 Wei 等^[1]提出的 EDA 方法,其通过对一句话中的部分单词进行同义词替换、随机插入、随机交换和随机删除 4 种操作以获取新的数据。EDA 方法可以有效地对数据进行扩充,同时可以引入新的词汇以增强模型泛化,还可以加入部分噪声以防止模型过拟合。

由于抽取式阅读理解的特殊性,在对文本进行数据增强时,需要保证不对问题的正确答案进行随机操作,因此在选择句子时,将包含有正确答案的句子排除掉,不对包含有正确答案的句子进行操作。选择完句子后,再对句子进行以下 4 种操作:①在句子中选择一个单词然后使用其同义词进行替换;②随机选择一个单词,将其同义词随机插入句子;③随机选择句子中的两个单词,交换单词位置;④随机选择一个单词进行删除。对数据集中的样本实施单词增强策略前后对比如图 1 所示。

采用适应于抽取式阅读理解的单词增强策略可以选择那些不与正确答案相关的单词进行操作,防止对答案相关的单词进行操作从而导致文本语义产生较大变化,影响模型输出结果。

2.2 适应于抽取式阅读理解的句子增强策略

由于抽取式阅读理解所使用的语言模型都较大,数据集样本数量不够会导致模型训练效果偏差,仅使用针对单词的数据增强方法在样本数量过少的情况下可能仍会存

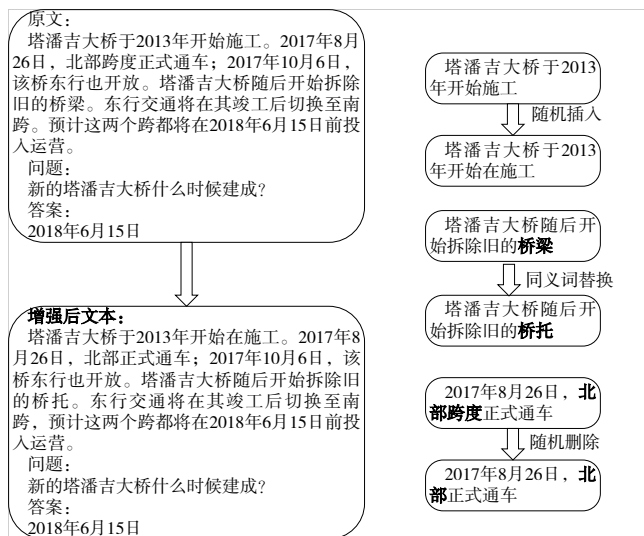


Fig. 1 Example of comparison before and after implementing word enhancement strategies on text

图1 对文本实施单词增强策略前后对比示例

在数据量不足的情况。基于此,本文提出两种针对句子的数据增强方法:①随机交换文章中两个句子的位置以生成新的文本内容;②从相邻样本中随机选取一个句子以替换样本中的一个句子。在对句子进行随机选择时,先检查选择出的句子中是否包含有正确答案,如果包含有正确答案,则重新选择句子,主要目的是防止随机交换导致问题的正确答案丢失。此方法可以有效地对样本数量进行扩充,同时保证样本有效性。针对句子的数据增强可以在一定程度上制造噪声但不影响文章的整体结构,对数据集中的样本实施句子策略前后对比如图 2 所示。

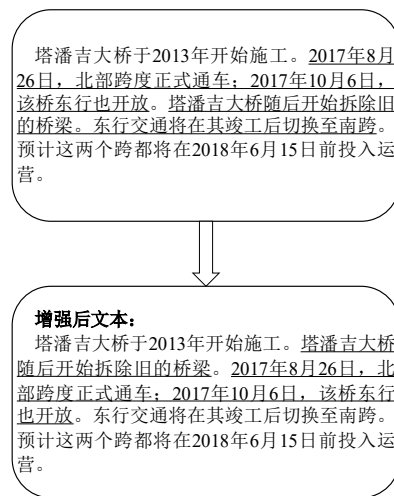


Fig. 2 Example of comparison before and after implementing text enhancement strategies on text

图2 对文本实施句子增强策略前后对比示例

2.3 基于语义相似度的数据增强(DASS)

抽取式阅读理解的特点在于模型需要学习所给出的整段文本,在文章篇幅较大的情况下,可能会包含较多与问题无关的文本信息,这会给模型的学习增加较多噪声。

而针对单词和句子的数据增强只是在数据集的样本数量上进行了扩充,随着数据量的增多,可能产生过多的噪声从而降低了模型的学习效率。基于此,为了提升模型的学习效率和数据集中样本的质量,本文提出采用语义相似度计算的方法,将原文句子与数据增强后的句子进行语义相似度比较,选择数据增强后对原句语义影响较小的句子进行操作,从而提高模型预测准确率。

为了保证新生成的样本具有语义连贯性,本文需要选择那些对文本语义影响较小的单词进行操作。对此,DASS通过计算句子中每个单词删除前后的语义相似度,对相似度分数进行一个排序,选择删除单词前后语义相似度分数最高的单词进行删除、插入或同义词替换操作,具体操作过程如图 3 所示。

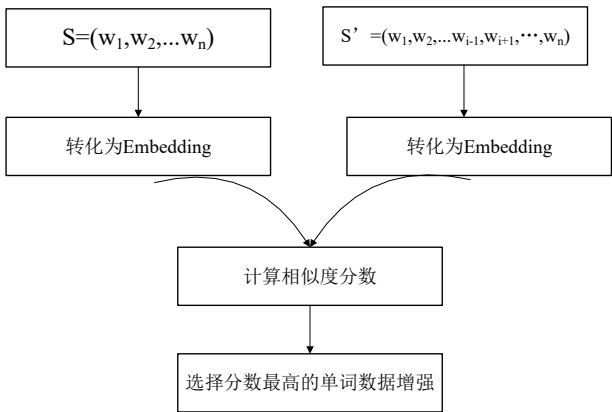


Fig. 3 Operation process of word selection

图 3 单词选择操作过程

在进行数据增强时,会选择句子中的部分单词进行操作。在此之前,先对随机选中的句子作预处理,排除停用词列表中出现的单词,防止在单词选择过程中选中停用词,增强对选中的单词进行数据增强的有效性。预处理完之后,对句子中的每个单词计算一个得分,这个得分定义为对这个单词进行删除前后整个句子语义相似度的变化。例如,原句子 $S=(w_1, w_2, w_3, \dots, w_n)$,在对计算得分的这个单词 w_i 进行删除后得到一个新的句子 $S'=(w_1, w_2, w_3, \dots$

$w_{i-1}, w_{i+1}, \dots, w_n)$,将这两个句子输入到编码器中转换为 embedding,然后通过式(1)计算单词 w_i 的得分。

$$Score(w_i) = 1 - \arccos\left(\frac{s \cdot s'}{\|s\| \cdot \|s'\|}\right) / \pi \quad (1)$$

其中, w_i 表示句子中的第 i 个单词, S 表示原句子, S' 表示将第 i 个单词删除后的句子。 $Score(w_i)$ 表示第 i 个单词的相似度得分。首先计算两个句子的余弦相似度,然后使用 $\arccos$ 将其转化为角距离,句子相似度越高,其角距离就越小,然后通过 1 减去角距离除以 π ,则语义相似度越高, $Score(w_i)$ 的值越大。删除该单词之后的句子语义与删除之前的句子语义越接近,语义相似度则越高, $Score$ 则越大,可证明该单词对语句的语义影响越小。将句子中每个单词的 $Score$ 计算出来后进行排序,分数越高,证明删除这个单词对整个句子的语义影响越低。为了尽量降低数据增强后对句子语义的影响,选择分数排名最高的单词进行操作。基于语义相似度选择后再实施单词增强策略的效果如图 4 所示。

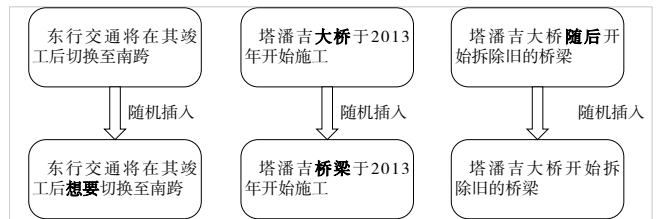


Fig. 4 Comparison before and after selecting words based on semantic similarity and implementing word enhancement strategies

图 4 基于语义相似度选择单词后再实施单词增强策略前后对比

3 实验与结果分析

3.1 数据集

实验使用的数据集为来自于 MRQA 共享任务的 HotpotQA 数据集^[18-19],它将抽取式问答的文本内容、问题和答案整理成统一的格式,数据集示例如表 1 所示。此数据集拥有约 11 万个例子,每个例子中都包含 1 个问题,1 个答案。

Table 1 An example of data

表 1 数据示例

数据集	示例
文本	塔潘吉大桥于 2013 年开始施工。2017 年 8 月 26 日,北部跨度正式通车;2017 年 10 月 6 日,该桥东行也开放。塔潘吉大桥随后开始拆除旧的桥梁。东行交通将在其竣工后切换至南跨。预计这两个跨都将在 2018 年 6 月 15 日前投入运营
问题	新的塔潘吉大桥什么时候建成
答案	2018 年 6 月 15 日

为了创造样本数量较少时的训练环境,本文选择从原始训练集中选取一部分数据构成较小的数据集,选择数据量的大小分别为 50、100、200、500、1 000。测试集依旧使用原数据集提供的测试集,不进行扩充。

3.2 基线模型

BERT 是由 Google 提出的建立在 Transformer Encoder

基础上的语言模型,在此之后, Joshi 等^[6]提出了 SpanBERT,一个主要关注跨度表示的 BERT 风格的语言模型。SpanBERT 的训练方法是掩蔽连续的一段文本并且进行两种优化:一种是掩蔽语言模型,从模型自己的向量表示中预测连续的掩蔽内容;另一种是跨度边界目标,它通过掩蔽跨度开始的前一个单词和结束的后一个单词以预测掩

蔽跨度中的每个单词。为了更好地获取上下文信息,Seonwoo等^[20]提出了基于SpanBERT的针对上下文的块注意力BLANC。本文选择BLANC作为本文实验的基线模型。

3.3 实验参数设置

本文在有限的参数设置上进行实验,将训练的批次大小设置为8,学习率设置为 2×10^{-5} ,epoch设置为30,Transformer编码器的最大序列长度设置为384,并且使用了HotpotQA数据集原论文中的优化器。

3.4 评价指标

本文使用常用于QA模型预测的F1值作为评价指标,可以有效地反映模型预测能力。F1值计算如式(2)所示。

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

3.5 实验结果分析

表2显示了在训练数据分别为50、100、200、500、1 000时,原模型性能和进行数据增强后模型的性能。可以看出,在样本数量较少时,模型预测能力会受到较大影响,当训练数据只有50条时,模型的F1值仅有8.26%。在使用了EDA方法后,数据量增加到200条,模型预测能力有了明显提升,模型的F1值达到18.52%,提升较为明显。当使用了本文提出的数据增强方法后,数据量也扩充至200条,模型的F1值达到了26.56%。可以发现,本文提出的数据增强方法可以通过选择特定的目标进行增强以有效地降低EDA方法对数据进行操作后造成数据语义的大量损失,帮助模型更好地学习数据中的语义信息,从而提升模型性能。当训练数据量达到100条时,使用本文提出的数据增强策略后模型的F1值(46.43%)也远高于未进行数据增强时的F1值(22.71%)和使用EDA方法进行数据增强后的F1值(27.92%)。由图5可以看出,随着数据量的增加,数据增强的效果会逐渐降低,但是依然能保证数据增强后模型的能力高于未进行数据增强时。

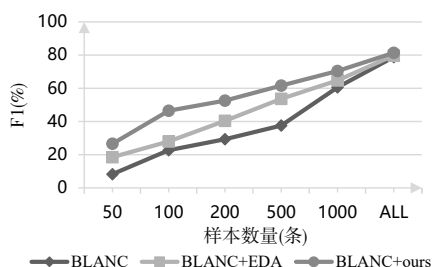


Fig. 5 Experimental results with different sample numbers

图5 不同样本数量时实验结果

表2显示了本文对完整的数据集进行数据增强后的模型性能,当数据量达到几万条时,数据增强后模型的F1值依然比原始模型的F1值提升了2.54%。该结论表明,在抽取式阅读理解任务中,当可用于训练的样本数量较少时,本文提出的针对抽取式阅读理解的数据增强策略能够有效地提升模型预测能力。

Table 2 Experimental results with different sample numbers

表2 不同样本数量时实验结果 (%)

模型	样本数量					
	50	100	200	500	1000	ALL
BLANC	8.26	22.71	29.31	37.63	60.56	78.84
+EDA	18.52	27.92	40.47	53.72	64.81	79.77
+DASS	26.56	46.43	52.59	61.57	70.37	81.38

为了验证本文提出的数据增强方法在其他抽取式问答数据集下依然有效果,本文在NaturalQ数据集和SearchQA数据集进行了实验,结果如表3所示。可以看出,在进行数据增强后,模型在NaturalQ数据集上的F1值相较于之前提升了4.33%,在SearchQA数据集上的F1值也比之前提升了2.41%。实验证明,本文提出的方法可以有效地提升模型在多个抽取式阅读理解数据集上的效果。

Table 3 Experimental results of DA on different data sets

表3 数据增强方法在不同数据集上的实验结果 (%)

模型	NaturalQ	SearchQA	HotpotQA
BLANC	80.04	81.50	78.84
BLANC+DASS	84.37	83.91	81.38

本文对该实验方法进行消融研究,以更好地理解针对单词的数据增强、针对句子的数据增强以及使用相似度计算筛选后再进行数据增强3种方法的贡献。首先,只单独使用针对单词和针对句子的数据增强方法;然后,分别对同时使用两种方法和相似度计算方法进行测试。不同数据增强方法的实验结果如表4所示。

Table 4 Experimental results of different DA

表4 不同数据增强方法的实验结果 (%)

模型	样本数量					
	50	100	200	500	1 000	All
BLANC	8.26	22.71	29.31	37.63	60.56	78.84
BLANC+单词	16.33	33.14	46.32	52.26	64.61	79.85
BLANC+句子	10.41	36.40	38.52	56.17	65.43	80.63
BLANC+单词+句子	24.07	40.79	49.65	59.60	68.84	80.99
BLANC+DASS	26.56	46.43	52.59	61.57	70.37	81.38

实验结果表明,本文使用的3种数据增强方法在抽取式阅读理解任务中的效果均有一定提升。针对数据集的单词进行数据增强,可以有效扩充样本数量,同时设置单词的增强比例仅为10%,可以很大程度地保证文本语义在数据增强后不会发生较大变化,使得模型预测能力得到较大提升,在样本数量较少时提升效果更为明显。

针对数据集中的句子进行数据增强后,可以得到更多的增强样本,而且在交换句子顺序时不会影响单句的语义,模型可以更好地获取文本信息。但是,可能会对文章整体上下文语义产生影响,这也是模型预测能力出现波动的原因。

由实验结果比较可知,在使用增强目标选择之后,模型预测能力均得到有效提升,这表明在进行数据增强时必须考虑被增强的文本对文章整体语义的影响,在尽可能保证文章语义的前提下对数据进行增强可以有效地提升模型预测能力。

4 结语

本文结合传统的EDA方法提出了适应于抽取式阅读理解的数据增强方法,并在此基础上采用基于语义相似度的数据增强方法DASS以选择对文本语义影响较小的目标进行数据增强。研究结果表明,在少样本情况下,对抽取式阅读理解进行研究具有重要意义,其可以有效帮助模型在训练数据受限时学习到更多的语义信息。本文提出的方法只适用于抽取式阅读理解任务,后续可在此基础上针对其他阅读理解任务展开研究,也可探索更多数据增强方法并将其应用于抽取式阅读理解任务。

参考文献:

- [1] WEI J, ZOU K. EDA: easy data augmentation techniques for boosting performance on text classification tasks [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 6382–6388.
- [2] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100, 000+ questions for machine comprehension of text [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 2383–2392.
- [3] 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.
- [4] KENTON J D M W C, TOUTANOVA L K. BERT: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171–4186.
- [5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [DB/OL]. <https://arxiv.org/abs/1706.03762>, 2023.
- [6] 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.
- [7] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]//Boston: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
- [8] ZUO H X, LIAO B, CHEN X K, et al. Application of SC-Net model integrated with transfer learning and data augmentation in skin cancer classification [J]. Application Research of Computers, 2022, 39(8): 2550–2555, 2560.
左航旭,廖彬,陈小昆,等.融合迁移学习和数据增强的SC-Net模型在皮肤癌识别中的应用[J].计算机应用研究,2022,39(8):2550–2555, 2560.
- [9] FADAEI M, BISAZZA A, MONZ C. Data augmentation for low-resource neural machine translation [C]//Vancouver: the 55th Annual Meeting of the Association for Computational Linguistics, 2017.
- [10] SENNRICH R, HADDOW B, BIRCH A. Improving neural machine translation models with monolingual data [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 86–96.
- [11] SHENG X H, SHEN H L. Semi-supervised text classification algorithm with data augmentation and similar pseudo-labels [J]. Application Research of Computers, 2023, 40(4): 1019–1023, 1051.
盛晓辉,沈海龙.基于数据增强和相似伪标签的半监督文本分类算法[J].计算机应用研究,2023,40(4):1019–1023,1051.
- [12] WANG W Y, YANG D. That's so annoying!!!: a lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using#petpeeve tweets [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 2557–2563.
- [13] XIE Z, WANG S I, LI J, et al. Data noising as smoothing in neural network language models [DB/OL]. <https://arxiv.org/abs/1703.02573>. 2017.
- [14] KARIMI A, ROSSI L, PRATI A. AEDA: an easier data augmentation technique for text classification [C]//Proceedings of Findings of the Association for Computational Linguistics, 2021: 2748–2754.
- [15] HU Z, TAN B, SALAKHUTDINOV R, et al. Learning data manipulation for augmentation and weighting [C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019: 15764–15775.
- [16] LIU R, XU G, JIA C, et al. Data boost: text data augmentation through reinforcement learning guided conditional generation [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020: 9031–9041.
- [17] SUN L, XIA C, YIN W, et al. Mixup-transformer: dynamic data augmentation for NLP tasks [C]//Proceedings of the 28th International Conference on Computational Linguistics, 2020: 3436–3440.
- [18] FISCH A, TALMOR A, JIA R, et al. MRQA 2019 shared task: evaluating generalization in reading comprehension [C]//Hong Kong: Proceedings of the 2nd Workshop on Machine Reading for Question Answering, 2019.
- [19] YANG Z, QI P, ZHANG S, et al. HotpotQA: a dataset for diverse, explainable multi-hop question answering [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 2369–2380.
- [20] SEONWOO Y, KIM J H, HA J W, et al. Context-aware answer extraction in question answering [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020: 2418–2428.
(责任编辑:孙娟)