基于相似度进行句子选择的机器阅读理解数据增强

聂双， 叶正， 覃俊， 刘晶
中南民族大学　计算机科学学院　湖北省　武汉市　430074
819258834@qq.com　yezheng@scuec.edu.cn

摘要

目前常见的机器阅读理解数据增强方法如回译，单独对文章或者问题进行数据增强，没有考虑文章、问题和选项三元组之间的联系。因此，本文探索了一种利用三元组联系进行文章句子筛选的数据增强方法，通过比较文章与问题以及选项的相似度，选取文章中与二者联系紧密的句子。同时为了使不同选项的三元组区别增大，我们选用了正则化Dropout的策略。实验结果表明，在RACE数据集上的准确率可提高3.8%。

关键词： 多项选择 ； 长文本数据增强 ； 正则化Dropout

Machine reading comprehension data Augmentation for sentence selection based on similarity

Shuang Nie,　Zheng Ye ,　Jun Qin,　Jing Liu
School of Computer Science, South-Central University for Nationalities, Wuhan 430074
819258834@qq.com　yezheng@scuec.edu.cn

Abstract

At present, the commonly used data augmentation methods for machine reading comprehension, such as back translation, enhance the data of articles or questions alone, without considering the relationship among articles, questions and option triples. Therefore, this paper explores a data augmentation method for article sentence screening by using triplet connection. By comparing the similarity among the article and the questions and options, and select the sentences closely related to the two in the article. At the same time, in order to increase the difference between triples of different options, we use the strategy of regularizing dropout.The experimental results show that the accuracy can be improved by 3.8%.

Keywords: Multiple choice , Long text data augmentation , Regularized Dropout

1 引言

机器阅读理解是自然语言处理中重要的一环，与其他领域息息相关。机器阅读理解的目的是为了让计算机像人一样对文本进行理解 (Seo M et al., 2016)，进而能够实现阅读和推理。为

第二十一届中国计算语言学大会论文集，第569页-第579页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

了了解计算机对文本的掌握能力，就需要进行计算机对问题回答的测试 (Tang M et al., 2019)。计算机能够回答的问题的难度和回答问题的正确率在一定程度上能够反映出计算机对文本的了解程度。

阅读理解可分为四类，填空式、选择式、抽取式和生成式 (Liu S et al., 2019)。多项选择式的阅读理解既有需要精读的细节题，也有需要总结的概括题，如图1所示，加粗的选项为问题的正确答案，其中的细节题需要找到答案位置，而总结题需要纵观全文，因此，多项选择阅读理解同时测试了计算机关注细节与整体推理能力，是一个综合性的任务，更具有挑战性。

随着预训练模型 (DEVLIN J et al., 2018) 的出现，机器阅读理解任务得到了快速发展，但是这种大型神经网络为了能够得到充分的训练，需要大量的数据来训练，训练的数据越多，往往训练得到的效果也就更好。因此如何得到更多高质量的数据就成为了关注点，然而多项选择式的阅读理解进行数据增强有两个挑战，一是阅读理解的目的是为了考察学生的快速阅读能力，所以文章设置多数很长，二是为了测试学生是否真正理解了文章而不是基于表面，设置的答案不一定是原文片段 (Zhu H et al., 2018)。这两个难点使得进行人工扩充的人员需要一定的知识储备，人工构建难度高。所以大家将目光放在了自动数据增强方法上，且预训练模型输入有长度限制，长文章无法全部输入模型，而目前常用的数据增强方法主要是针对能够输入模型的部分进行的。这种方法的缺点是忽略了被删掉文章部分，会导致有的问题找不到答案，因为通常设置的问题及其对应答案在文章中是均匀分布的。

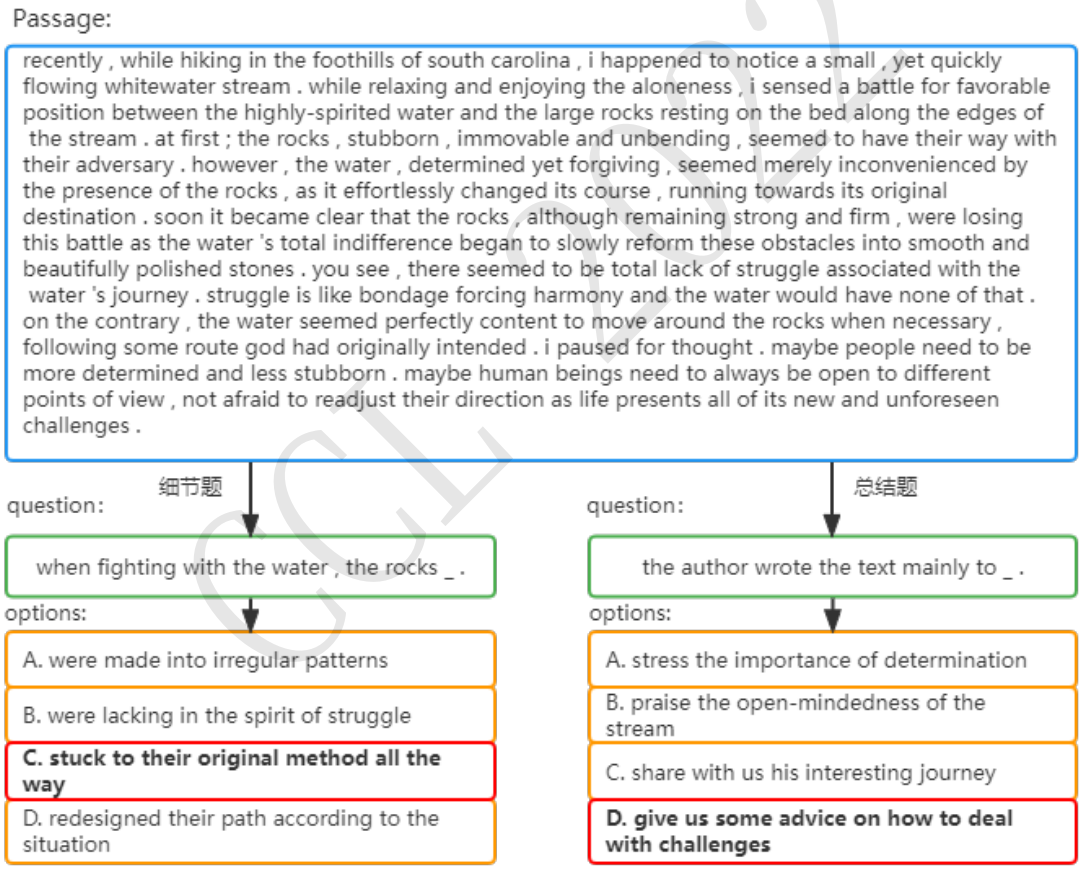

Figure 1: 机器阅读理解中的细节题与总结题比较

为了解决多项选择机器阅读理解数据集人工扩充难度大的问题，同时充分利用全部文本信息，本文提出了一个基于句子选择的自动数据增强方法。该方法首先利用文本相似度计算文章与问题以及选项的相似度，挑选出文章中与问题以及选项相关的句子，保持相对顺序不变的前提下形成新的文章。这种数据增强方法不仅能扩充数据集，而且能在保证原文语义不变的条件下让模型学到更多的内容。同时大型神经网络会采用Dropout来防止过拟合，但是Dropout在训

练时和推理时使用的策略不一样，即训练时会以概率P随机删除一些神经元，而在推理时将所有的神经元加入，因此会导致训练和推理的模型差异。为了消除两者之间的差异，我们引入了简单的R-Drop正则化 (Wu L et al., 2021)。

2 相关工作

机器阅读理解是机器通过阅读文章来回答问题的技术，机器阅读理解任务可分为基于规则的传统方法 (Smith E et al., 2015) 和基于深度学习的方法 (Seo M et al., 2016)。基于规则的传统方法由于受到数据集的限制和需要特征构建而不能取得较好的效果。随着CNN/DailyMail (Hermann K M et al., 2015)大型数据集的出现，深度学习在机器阅读理解任务得到了迅速的发展，已经成为当前的主流方法。基于深度学习的机器阅读理解要求训练数据足够多，训练的数据越多，往往训练得到的效果也就更好。而数据增强是可以解决数据匮乏的一种有效方法，且能够提高模型的准确率。因此许多学者开始研究基于文本的数据增强方法。

传统的文本数据增强方法主要可分为三类，基于字符层面的数据增强、基于词层面的数据增强和基于句子层面的数据增强。基于字符层面的数据增强一般是在文本中加入噪声，常用的有拼写错误注入、键盘错误注入，使所训练的模型对扰动具有鲁棒性。基于词层面的数据增强方法最常见的是词汇替换，词汇替换方法可分为基于词典的替换 (Zhang et al., 2015) 、基于词向量的替换 (Jiao X et al., 2020)和基于TF-IDF的词替换 (Xie Q et al., 2019)。这些方法都是将文本句子中的某个词替换为另一个相近词，基于词典的替换是随机将句子中的一个单词使用同义词词典替换为同义词，基于词向量的替换使用预先训练好的单词嵌入，如Word2Vec (Mikolov T et al., 2013)、GloVe等，并使用嵌入空间中最近的相邻单词替换句子中的某些单词，以此来提高语言模型在下游任务上的泛化能力。而基于TF-IDF的词替换的思想是分数较低的单词不能提供信息，因此可以在不影响句子的基本真值标签的情况下替换它们。随着预训练语言模型的发展，研究发现MLM（遮盖语言模型）通过预训练也能进行词的替换 (JM Tapia-Téllez and Escalante H J, 2020)，由于MLM是基于上下文来推测出遮盖词，因此替换的词拥有上下文语境，同一个词在不同的语境中可能会生成不同的同义词，从而对解决歧义问题有帮助。基于句子层面的数据增强比较常用的是回译方法 (Lee S et al., 2021)，回译的方式就是将句子翻译成另一种语言，然后再翻译成原来的语言。这种数据增强方式的优点是尽量保证了在原文意思不变的基础上生成了新的补充版本。还可以同时使用多种不同的语言来进行回译以生成更多的文本变体。最近通过对语言的语义分析，出现了在不改变语义的情况下进行语态转变的数据增强方法 (Dehouck M and Gómez-Rodríguez C, 2020)，这主要是通过语法分析建立依赖树，转换依赖树后生成意思相同的句子。

还有介于字符与词之间的数据增强方法，即使用正则表达式的简单的模式匹配的转换，文本表面转换 (Coulombe C, 2018)是其中常见的一种。是将动词形式由简写转化为完整形式或者反过来的方法。这种方法在扩展模棱两可的动词形式时可能会出现错误，为了避免出现这种问题，提出了允许模糊收缩，但跳过模糊展开的方法。

以上是自然语言处理领域通用的数据增强方法，在机器阅读理解任务中，常用的数据增强方法也可以分为以下几种:

(1)传统的简单数据增强方法EDA (Wei J and Zou K, 2019)，也是基于词层面的方法，利用对词的简单操作，同义词替换，随机插入，随机互换和随机删除来扩充数据，然而对于是使用预训练的模型来说，效果并没有什么提高。

(2)基于问题生成的数据增强，问题生成又进一步可以分成基于规则和模板 (Mitkov R, 2003)以及基于深度学习 (Mirshekari M et al., 2021; Yu A W et al., 2018)两种方式，前者是使用设定的规则或者模板来生成问题，效率低且泛化能力差。后者是通过将文章和答案放入生成模型中训练来形成新问题。这种数据增强方法尽量使生成的问题贴近原问题，以此达到原义相近的条件下问题增多的目的。然而这种方法当答案在文章中多次出现时，无法判断是哪个位置，因此可能会出现与原问题背道而驰的情况。而后者仅依赖文章来生成问题-答案对，这种生成新数据的方法生成的问题-答案对质量不高，会造成冗余的数据。

(3)无监督数据增强，在英语机器阅读理解中常使用的是英法回译方法 (Fabbri A R et al., 2021)，这样可以保留原义而生成不同的意译。然而对于问题中的关键字并不一定能保留下来，影响找答案的位置。

以上的数据增强方法遇到输入序列过长时，采取的是简单的截断处理。这么做的缺点是忽

视了截断文章部分，对于多项选择型的阅读理解任务来说，文章的问题在文章中是平均分布的，只截取前半部分文章会导致后半部分问题在输入的文本中找不到答案，因此生成新的数据也难以提高模型的准确性。

针对以上缺点，本文引出了一种同时兼顾机器阅读理解的长文本与数据量不够的数据增强方法，该方法是对数据集中的文章进行扩充，以增加文章样本多样性。但是本方法也考虑了长文本问题。通过机器阅读理解中的三元组信息，比较多个计算文本相似度将文章总与问题以及选项相关的重要信息提取出，构造出既与输入文本不完全一致，能够起到补充作用，又能不改变原文意思的数据增强方法。

3 方法

多项选择的机器阅读理解过程可以化为以下的三元组形式$< P, Q, A >$：给定的一篇文章P，根据文章内容提出的问题Q以及相对应的几个选项A。多个答案选项中只有一个选项为正确答案，其余的为迷惑的错误答案，目的就是为了选出正确答案。其中问题$Q = \{Q_1, Q_2, Q_3...Q_m\}$（m表示文章的问题数量），选项$A = \{A_1, A_2, A_3...A_n\}$（n表示每个问题的选项数量）。本文将运行过程分成了三个部分：第一部分是利用三元组中三者关系的相似性，计算文本相似度来抽取重要的句子生成新的文章来进行数据增强。第二部分将原来的数据和新生成的数据一起放入基线模型即双向匹配模型中进行训练。第三部分是利用自身的数据来进行样本间对比的R-Drop正则化来选出最终的答案。模型的整体架构如图2所示。

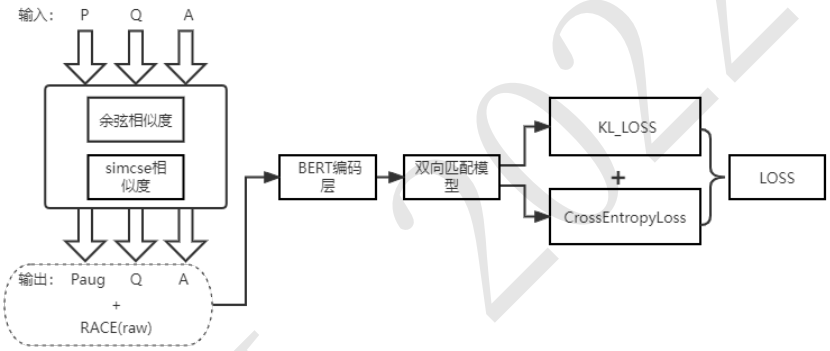

Figure 2: 结构图

3.1 数据增强模块

目前多项选择的机器阅读理解任务的训练数据量相对较少且人工扩充数据集较难，因此找到有效的数据增强方法很重要。多项选择阅读理解任务中问题取材于文章，答案也在文章中寻找。所以受解题信息来源于文章的启发，本文提出建立以文章为中心的数据增强方法。考虑到预训练模型长度限制，就首要挑选文章最重要的内容来进行数据增强。因此，本文研究了两种短文本相似度的方法来获取文章的核心句子，分别是余弦相似度和SimCSE相似度 (Gao T et al., 2021)。余弦距离是通过算出两个向量的夹角余弦值来衡量两者相似程度的。对于文本来说余弦距离是通过利用两个短文本词频向量来计算相似性的。余弦相似度由于通常用于正空间，因此规定夹角余弦取值范围为[0,1]。通过余弦值可以判断两个向量指向相似度。0度为重合，余弦值为1，90度余弦值为0。两个向量越相似，余弦值越大。短文本余弦相似度的计算方法为：

$$similarity = Cos\theta = \frac{\sum_{i=1}^{n} A_i.B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{1}$$

其中的A_i,B_i分别表示两个句子的第i个词频向量。文章中每个句子与问题、每个句子与选项的相似度可计算出：

$$S^{pq} = Cosine(p_i, q) \tag{2}$$

$$S^{pa} = Cosine(p_i, a) \tag{3}$$

$$Score^{pq} = W_1 S^{pq} \tag{4}$$

$$Score^{pa} = W_2 S^{pa} \tag{5}$$

$$M_p = K_{max}(Score^{pq}, Score^{pa}) \tag{6}$$

其中，S^{pq}是问题和文章中第i个句子余弦相似度，S^{pa}是每个选项和文章中第i个句子的余弦相似度。P_i表示文章中的第i个句子，q表示问题，a表示选项。其中W_1和W_2表示可调节的参数。$Score^{pq}$，$Score^{pa}$分别表示文章与问题，文章与选项之间的相似度倒序排列。K_{max}表示将$Score^{pq}$，$Score^{pa}$序列的前K个值，M_p是从原文章中按照K的取值抽取的句子。

SimCSE相似度是利用对比学习的方式来进行模型训练，通过拉近相似的数据推远不相似的数据的方式来进行对比。

$$l_i = -log\frac{e^{sim(h_i, h_i^+)/\tau}}{\sum_{j=1}^{N} e^{sim(h_i, h_j)/\tau}} \tag{7}$$

其中，τ表示温度超参数，h_i，h_i^+，h_j表示第i个样本、其对应的正样本以及其对应的负样本经过模型编码得到的向量。sim表示相似度，l_i表示第i个样本的损失函数。

本文采取的相似度方法是分别提取文章句子与问题，文章句子与每个选项之间的相似度的前k个句子，这是为了让挑选出的句子尽可能与问题和每个选项有关联。同时为了最大保持原始文本的语义信息，我们将抽取出来的句子仍然按照原始文本的顺序采集出来，不破坏原文的相对位置。

3.2 多项选择的基础模型

本小节将介绍多项选择型的阅读理解模型的基本架构，本文选用DCMN(Dual Co-matching network) (Zhang S et al., 2020)作为基线模型，该基线模型主要由三部分组成，首先使用预训练模型对三元组进行编码，获得编码信息后将三元组信息通过双向匹配模块相互进行融合和交互，最后再进行答案选择的输出。同时在此模型上采用R-Drop作为辅助策略。

3.2.1 三元组编码

对三元组的编码需要获取上下文信息，所以我们选择使用预训练模型（BERT）作为编码器。BERT把三元组即文章P,问题Q,选项A中的每一个词编码成固定长度的向量，编码之后可以获得:

$$H^p = Encode(M^p) \tag{8}$$

$$H^q = Encode(M^q) \tag{9}$$

$$H^a = Encode(M^a) \tag{10}$$

H^P,H^q,H^a分别表示文章、问题、选项在预训练模型中最后一层的输出表示。

3.2.2 双向匹配模块

为了充分提取三元组两两之间的信息，我们需要用到前面预训练模型作为编码器的输出表示H^p,H^q,H^a。注意力机制方式本文采用了双向匹配机制，分别获得文章-问题表示，文章-选项表示，问题-选项表示。

$$C = [M^{pq}; M^{pa}; M^{qa}] \tag{11}$$

M^{pq}、M^{pa}、M^{qa}分别是文章-问题对、文章-选项对、问题-选项对的匹配表示，C是最终每个问题对应的三元组表示。

3.2.3 正则化Dropout思想

第一个句子选择模块对多项选择的机器阅读理解中的文章提取了多方面且精准的内容。而模型的双向匹配机制提供了两两配对，将三元组的内容通过提取融合在了一起。多项选择机器阅读理解采用是交叉熵损失函数，Dropout两次就能得到两个不同的子模型分布，即：

$$L^{E}_{P,Q,A_i} = -logp_1(y_i|x_i) - logp_2(y_i|x_i) \tag{12}$$

正则化Dropout就是控制使用两个不同的子模型预测的结果能尽量保持相同，达到模型优化的目的。通过最小化两个分布之间的双向KL散度，减小Dropout带来的训练和测试时带来的不同。

$$L^{KL}_{P,Q,A_i} = \frac{1}{2}[KL(P_2(y_i|x_i)|P_1(y_i|x_i)) + KL(P_1(y_i|x_i)|P_2(y_i|x_i))] \tag{13}$$

所以最终的损失函数为两者的加权和：

$$L = L^{E}_{P,Q,A_i} + \alpha.L^{KL}_{P,Q,A_i} \tag{14}$$

4 实验与分析

4.1 实验数据集及评价指标

本文研究的是机器阅读理解型的特定任务的数据增强方法，本文选择在多项选择型的机器阅读理解任务数据集RACE (Lai G et al., 2017)上进行实验设计并进行分析，来验证本文所论述的数据增强方法的有效性。RACE 是一个来源于初高中学生英语考试题目的大规模多项选择型的阅读理解数据集，RACE的形式是给定一个三元组<文章，问题，选项>，通过阅读并理解文章，对提出的问题从四个选项中选择正确的答案。该数据集由初中阅读RACE-MIDDLE和高中阅读RACE-HIGH组成，表1显示了这两个子集的训练集中文章长度分析，在5个长度区间统计了包含的文章数量以及占的总比、文章长度的平均值。如表1所示，训练集中文章长度在500以下的仅占全部数据集的2.1%,有98%左右的文章是不能全部放入模型中的。

passage_len	**0-500**	**500-1000**	**1000-1500**	**1500-2000**	>2000
RACE_M	493	2453	2519	816	128
RACE_H	34	608	3873	9829	4384
SUM	527	3061	6392	10645	4512
% ON RACE	2.10%	12.18%	25.43%	42.34%	17.95%

Table 1: RACE训练集文章长度统计

多项选择机器阅读理解的评价指标采用的是准确率指标ACC(Accuracy):

$$ACC = \frac{right}{all} \tag{15}$$

right表示的是模型预测正确的数量，all表示问题总数。

4.2 实验参数设置

为了证明该阅读理解任务数据增强方法的有效性，本文选择了双向匹配策略模型(DCMN)作为本实验的基线模型，本文采用深度学习框架PyTorch对相关内容进行编码实现，并在Ubuntu系统上采用GPU进行模型的训练和测试。本文采用12层的预训练模型BERT作为三元组的编码器，其中batch_size为16，gradient_accumulation_steps设置为2，训练的epochs的值是10，learning_rate为1e-05，优化函数采用Adam。

4.3 实验结果及分析

4.3.1 不同数据增强方法与基线模型的实验结果比较

本文在RACE数据集上进行了数据增强实验，将DCMN基线模型和四种不同的数据增强方法进行比较。在此实验中，回译数据增强选用的方法是英语-德语-英语的方式，词向量替换选用词向量Word2Vec。而生成问题的数据增强方法，使用unilm生成模型，以RACE数据集中文章截断部分作为生成模型输入部分的文章，将正确答案在此截断文章中的数据提取出来，正确答案所在片段截取出来作为生成模型输入部分的答案，这两者一起来生成问题。对于选项中的错误选项部分，本文采取与正确答案相似的处理，将错误选项所在文章片段提取出来作为迷惑选项。而我们的方法选择的是用SimCSE相似度计算方法来选择与问题选项相关TOP2的句子，并且其中R-Drop的α值设置为1。

MODEL	RACE
DCMN	65.05
DCMN+回译	63.99
DCMN+词向量替换	64.56
DCMN+生成问题	66.24
OURS	68.84

Table 2: 不同的数据增强方法实验结果对比

如表2结果所示，DCMN基线模型结果为65.05%，相比于DCMN的基线模型，对多项选择的机器阅读理解的长文本文章进行截断处理，对截断部分使用回译和词向量替换的数据增强方法，其准确率反而下降。原因之一是预训练模型本身强大，对模型中的文章部分进行数据增强学到的新内容有限，而且这两种数据增强的方法学习到了很多与问题无关的内容。原因之二是回译与词向量替换的数据增强方法有过多的替换，无法保留解决问题的关键词，因此影响在文章中找到问题的答案。基于生成问题的数据增强方法所得到的效果相比基线模型有所提升，是由于挑选的是能够在截断部分找到答案的数据来进行数据增强，因此生成的新数据中问题也能找到答案，是有效数据。而我们的数据增强方法使模型学习了截断之外的文章内容，其次，我们的方法并没有改变原始的文章内容，即保留了指示答案位置关键词。从实验结果来看，对于长文本的数据增强方法来说，我们的数据增强方法更能提升模型的准确性。

为了能够更加直接的感受各种数据增强方法之间的不同，我们将基线模型与回译、数据增强以及本文对文章的不同选择来进行对比。如图3所示，基线模型采取的是截断方式，基线模型中的文章部分为画线部分。而回译和词向量替换是基于基线模型的文章进行的。而本文数据增强的方法提取的句子为红色字体部分。可以看出，当相对应的问题是涉及到截断部分之外的内容，基线模型和其他两种数据增强方法都无法找到对应的答案，会造成冗余数据。

4.3.2 文本相似度对比实验

4.3.3 消融实验

为了进一步分析本文的数据增强方法中各部分的作用，研究利用文本相似度进行的句子选择数据增强模块和R-Drop模块各自的作用，本文设置了保留其中一个模块，去掉另一个模块的消融实验的对比。

MODEL	ACC
DCMN	65.05
DCMN_SENAUG	67.34
DCMN_RDROP	67.23
DCMN_SENAUG_RDROP	68.86

Table 3: 消融实验比较

表3中DCMN表示数据增强模块和R-Drop模块都去除。DCMN_SENAUG表示只保留数据增强模块，去除R-Drop模块。DCMN_RDROP表示保留R-Drop模块，去除数据增强模块。表

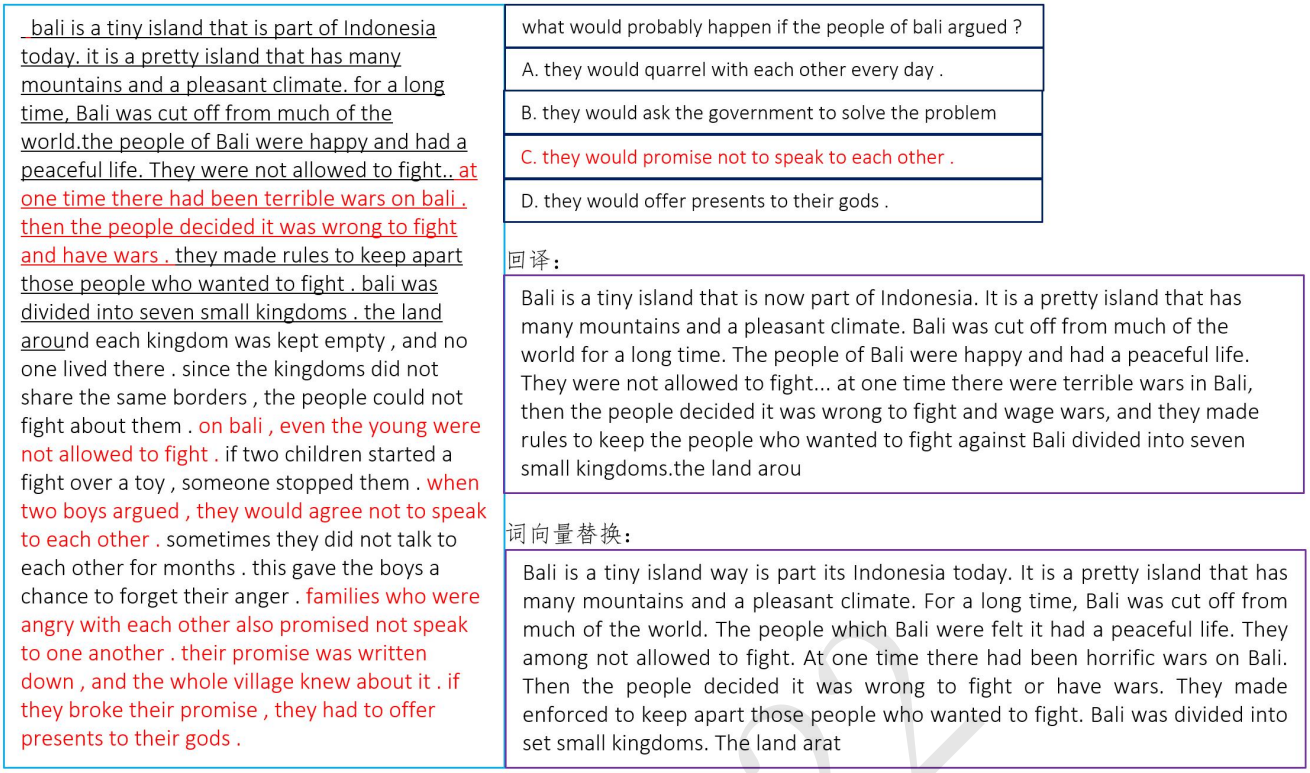

Figure 3: 不同数据增强的文章部分

中最后一行表示两个模块都保留。如表3实验结果所示，基于句子选择的部分做数据增强的方法在DCMN基线模型上提高了2.3%，而R-Drop模块在基线模型上提高了近2.2%，这表明这两部分都使模型得到了正确的数据训练，都是有效的，而最后组合在一起的结果比在R-Drop的基础上得到的结果提高了1.63%，说明了两者学习到的内容并不是完全一样的，都是分别有效果的。

4.3.4 文本相似度对比实验

本文选择通过数据增强的方式提高模型的准确率，为了证明通过文本相似度的方式选择的句子质量对实验结果的影响。我们选择了两种文本相似度方法进行句子的挑选。为了比较余弦相似度和simcse相似度挑选句子的效果，本文设置了两者的两组比较值，分别是在不同K值下和两种相似度方法在RACE数据集及其子集上的对比实验。为了使其不受R-Drop的影响，通过设置α=0使得R-Drop模块不起作用。

如图4结果所示，两种余弦相似度的方法做数据增强得到的结果趋势都是先上升然后下降迅速，再缓慢上升。这种准确率振荡的原因有以下几点：第一是由于本文采用的是对问题和每个选项与文章的相似度赋予同等权重，并且取各自相似度的Top k按照相对顺序进行组合，SimCSE相似度在k为2时取得最大值，表明SimCSE相似度能够在各自挑选2个句子的时候找到最关键的句子，后面k增加1急速下降说明各自新增加的句子具有干扰作用，学习到了无关的句子。而后慢慢上升表示挑选出的句子增多时，由于句子内容逐渐丰富，能逐渐消除带来的歧义。因为SimCSE相似度更敏锐所以相对余弦相似度曲线表现得更明显。使用SimCSE相似度得到的结果比使用余弦相似度得到的结果效果更好，表明不同相似度挑选出的句子是有很大区别的。挑选出更相关的句子进行训练对模型提升准确度更有帮助。

4.3.5 R-Drop中α值的选择

为了比较R-Dropα值的效果，我们设置了四组实验，分别设置为α为1到4，α等于1时得到的结果是最好的。α是KL散度损失函数的权重系数，值越大表示越重要，在α为1时取得最好的效果，说明我们不需要太关注KL散度。只需要给一点KL散度正则化就可以。

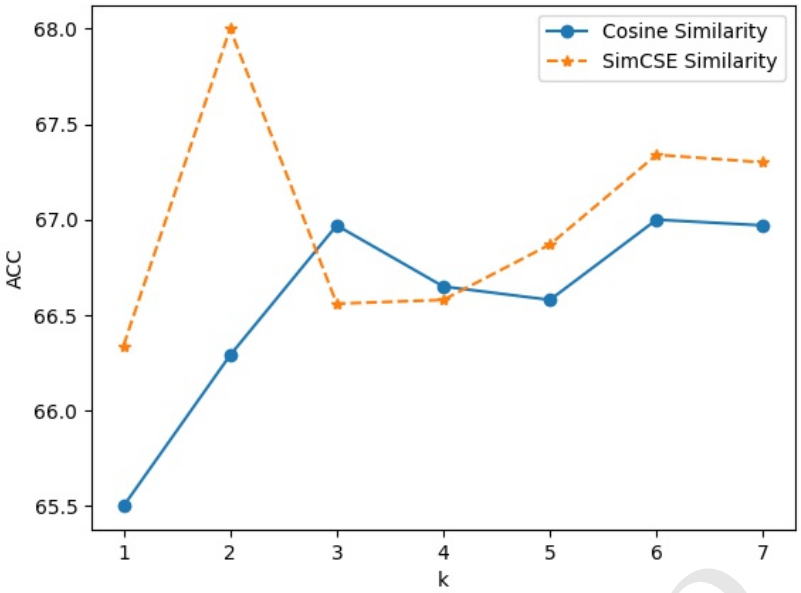

Figure 4: 两种相似度方法在k值不同时的准确率

α	**1**	**2**	**3**	**4**
ACC	68.86	68.59	68.43	68.08

Table 4: α值的选择

4.3.6 长文本文章长度影响

为了探索不同文章长度进行基于句子选择数据增强方法对于结果的影响，我们将以4.1数据集的划分方式将RACE_HIGH数据集分为四段分别进行数据增强。比较不同长度段之间的效果。

passage_len	**ACC**
DCMN	61.69
500-1000	62.41
1000-1500	63.22
1500-2000	63.66
>2000	63.14

Table 5: RACE_HIGH数据集不同长度区间进行数据增强后的结果比较

从表5结果可以看出，这四个区间段进行分别的数据增强都是有效果的，但是效果是有些微差别的。原因一是RACE_HIGH数据集每个区间段数量相差较大，500-1000区间仅有2028条数据，而1000-1500为11422条，1500-2000数据量最多，达到了31481，而大于2000以上的有17410条，数量差别影响了带来效果提升的区别。原因二是不同长度的难度不一样，这也会导致结果的差异。

5 总结

本文通过对多项选择的机器阅读理解进行简单有效的数据增强，比较两种相似度的方法来提取长文本文章中对问题和选项有用的句子来进行数据增强，又使用了R-Drop来消除训练和推理时的差异。实验结果表明使用不同的相似度方法挑选出的句子对提高模型效果是不一样的，

如何能够找到完全切合问题的句子有待进一步探究，甚至可以不再拘泥于挑选整个句子，而是将挑选的句子关键字进行整合，形成新的句子再组成新的文章进行数据增强。

参考文献

Coulombe C. 2018. *Text Data Augmentation Made Simple By Leveraging NLP Cloud APIs.* Prentice-Hall, Englewood Cliffs, NJ.

Dehouck M and Gómez-Rodríguez C. 2020. *Data augmentation via subtree swapping for dependency parsing of low-resource languages.* Proceedings of the 28th International Conference on Computational Linguistics. 2020: 3818-3830.

DEVLIN J, CHANG M-W, LEE K, et al. 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Computation and Language (cs.CL).

Fabbri A R, Han S, Li H, et al. 2021. *Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation.* Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 704-717.

Gao T, Yao X, Chen D. 2021. *SimCSE: Simple Contrastive Learning of Sentence Embeddings.* Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 6894-6910.

Hermann K M , Tomávs Kovcisk, Grefenstette E , et al. 2015. *Teaching Machines to Read and Comprehend.* NIPS.

Jiao X , Y Yin, Shang L , et al. 2020. *TinyBERT: Distilling BERT for Natural Language Understanding.* Findings of the Association for Computational Linguistics: EMNLP 2020.

JM Tapia-Téllez, Escalante H J . 2020. *Data Augmentation with Transformers for Text Classification.*

Lai G, Xie Q, Liu H, et al. 2017. *RACE: Large-scale ReAding Comprehension Dataset From Examinations.* Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 785-794.

Lee S, Lee D B, Hwang S J. 2021. *Contrastive Learning with Adversarial Perturbations for Conditional Text Generation.* Ninth International Conference on Learning Representation, ICLR 2021. The International Conference on Learning Representations.

Liu S , Zhang X , Zhang S , et al. 2019. *Neural Machine Reading Comprehension: Methods and Trends.* Applied Sciences.

Mikolov T, Sutskever I, Chen K, et al. 2013. *Distributed representations of words and phrases and their compositionality.* Advances in neural information processing systems, 2013, 26.

Mirshekari M, Gu J, Sisto A. 2021. *ConQuest: Contextual Question Paraphrasing through Answer-Aware Synthetic Question Generation.* Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021). 2021: 222-229.

Mitkov R. 2003. *Computer-aided generation of multiple-choice tests.* Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing. 2003: 17-22.

Seo M , Kembhavi A , Farhadi A , et al. 2016. *Bidirectional Attention Flow for Machine Comprehension.*

Smith E , Greco N , Bosnjak M , et al. 2015. *A Strong Lexical Matching Method for the Machine Comprehension Test.* Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.

Tang M , Cai J , Zhuo H H . 2019. *ulti-Matching Network for Multiple Choice Reading Comprehension.* Proceedings of the AAAI Conference on Artificial Intelligence. 33:7088-7095.

Wei J, Zou K. 2019. *EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks.* 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..

Wu L, Li J, Wang Y, et al. 2021. *R-drop: regularized dropout for neural networks.* Advances in Neural Information Processing Systems.

Xie Q , Dai Z , Hovy E , et al. 2019. *Unsupervised Data Augmentation for Consistency Training.*

Yu A W, Dohan D, Luong M T, et al. 2018. *QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension.* nternational Conference on Learning Representations.

Zhang, Xiang, Junbo Zhao, and Yann LeCun. 2015. *Character-level convolutional networks for text classification.* Advances in neural information processing systems 28 (2015): 649-657.

Zhu H, Wei F, Qin B, et al. 2018. *Hierarchical attention flow for multiple-choice reading comprehension.* Proceedings of the AAAI Conference on Artificial Intelligence. 32(1).

Zhang S , Zhao H , Wu Y , et al. 2020. *DCMN+: Dual Co-Matching Network for Multi-Choice Reading Comprehension.* Proceedings of the AAAI Conference on Artificial Intelligence 34(5):9563-9570.