面向垂直领域的阅读理解数据增强方法

吕政伟，杨雷，石智中，梁霄，雷涛，刘多星
汽车之家/ 中国，北京
{lvzhengwei, yanglei, shizhizhong,
liangxiao12030, leitao, liuduoxing}@autohome.com.cn

摘要

阅读理解问答系统是利用语义理解等自然语言处理技术，根据输入问题，对非结构化文档数据进行分析，生成一个答案，具有很高的研究和应用价值。在垂直领域应用过程中，阅读理解问答数据标注成本高且用户问题表达复杂多样，使得阅读理解问答系统准确率低、鲁棒性差。针对这一问题，本文提出一种面向垂直领域的阅读理解问答数据的增强方法，该方法基于真实用户问题，构造阅读理解训练数据，一方面降低标注成本，另一方面增加训练数据多样性，提升模型的准确率和鲁棒性。本文用汽车领域数据对该方法进行实验验证，其结果表明该方法对垂直领域阅读理解模型的准确率和鲁棒性均能有效提升。

关键词： 阅读理解 ； 数据增强 ； 问答系统

Method for reading comprehension data enhancement in vertical field

Zhengwei Lv,Lei Yang,Zhizhong Shi,Xiao Liang,Tao Lei,Duoxing Liu
Autohome Inc. / Beijing,China
{lvzhengwei,yanglei,shizhizhong,
liangxiao12030,leitao,liuduoxing}@autohome.com.cn

Abstract

Reading comprehension question answering system uses natural language processing technologies such as semantic understanding to analyze unstructured documents and generate answers, which has important theory value and vast application prospect. However, the costs of obtaining training samples for reading comprehension model are expensive and the user questions are complex and diverse in the vertical field, which leads to the poor accuracy and robustness. In response to the problem, this paper proposes a data enhancement method for reading comprehension question answering in the vertical field, which constructs training samples based on real user questions. So that it can reduce the cost of annotation and increase the diversity of training data. The experiments are carried out with the data in the automobile field and the results show that the method can effectively improve the accuracy and robustness of reading comprehension model in the vertical field.

Keywords: Reading comprehension , Data enhancement , Question answering system

第十九届中国计算语言学大会论文集，第257页-第266页，海南，中国，2020年10月30日至11月1日。

1 引言

随着近几年智能问答的高速发展，阅读理解问答作为其重要发展方向之一，也逐渐成为了各领域的研究和应用热点。不同于传统问答系统中利用知识表示和检索方式获取答案(Qu et al., 2018; 安波et al., 2018)，基于阅读理解的问答利用模型直接对非结构化文档进行认知，从而获取给定问题的答案(Wang et al., 2017; Chen et al., 2017; Yu et al., 2018)。这种方式减少了知识的收集和表示过程，具有重要研究和应用价值。

阅读理解问答根据答案的产生方式，分为选择式、抽取式、生成式等类型，其中抽取式阅读理解根据问题从文档中抽取一个连续关片段作为答案，不用考虑答案的序列生成问题，答案获取方式直接，标注相对方便，难度适中，因此对抽取式阅读理解的研究相对较多。同时一系列大规模高质量评测数据集的发布，如SQUAD数据集(Rajpurkar et al., 2016; Rajpurkar et al., 2018)、DuReader数据集(He et al., 2017)、CMRC2018数据集(Cui et al., 2018)等，进一步促进了阅读理解问答的研究。但是这些数据集偏向于通用领域或百科知识，内容广而泛，针对垂直领域专业性的知识少，因此，面向垂直领域的抽取式阅读理解数据集标注和应用研究是十分必要。

在抽取式阅读理解数据集的标注过程中，标注人员提出的问题容易出现标注数据模式化，其表达方式单一、多样性不足，从而导致在应用中造成模型的准确性和鲁棒性较差。数据增强通常被用来解决这一问题，其原理是通过无监督、半监督或者有监督的方法构造新的训练样本，对原始的训练数据进行扩充，增加训练数据的量级和多样性，从而提升模型的准确性和鲁棒性。在机器阅读理解中，常见的数据增强方法有以下几种。

（1）远程监督方法，利用外部知识库自动对语料进行标注(白龙et al., 2019)，构造训练数据，增加数据的量级。然而，这种方法会引入很大的噪声，影响模型的语义理解，如图1，当问题的答案“日本”在一篇文档中多次重复出现，答案的标注位置不能很好的确定，将会影响模型整体的语义理解；

（2）问题生成方法，利用模型生成标注数据中问题的同义复述(Kim et al., 2019; Zhao et al., 2018)，实现增加问题表达的多样性。但是，目前序列生成技术相对不够成熟、且缺乏适当的评测指标，生成数据的质量难以控制，最终会造成构造数据的误差大，阅读理解模型效果差。

（3）完全生成方法，给定未标注文档，首先利用模型从文档中获取适合作为答案的片段，再根据文档内容和该片段生成相关问题，这种方法不需要已有阅读理解标注数据即可构造数据，能极大的提升构造数据的量级和覆盖范围。但是该方法引入的误差较大，除了问题生成环节的误差，在答案片段选取环节、问题和答案相关性等方面也会引入误差，形成误差的累积，最终影响构造数据的质量。

上述这些方法都是针对通用领域的研究，忽略了数据增强与实际应用数据的结合，造成构造数据与应用数据之间的语义偏差，影响模型应用效果。另外，在垂直领域中，领域术语多，问题更为专业，衍生出的表达方式更多样化，用远程监督或模型生成方式构造的数据，很难满足专业性和多样化性，容易造成模型应用中准确率低、鲁棒性差。

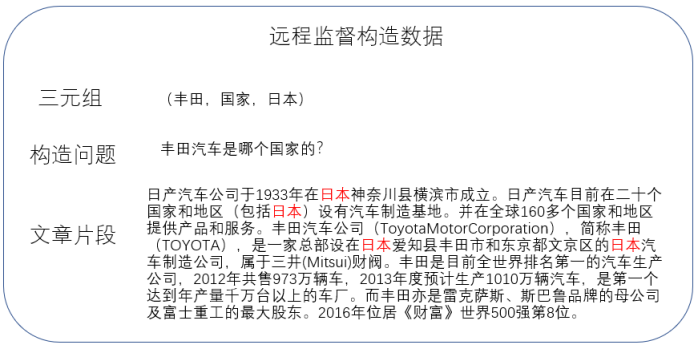


图 1. 远程监督方法构造训练数据

针对以上问题，本文提出了一种垂直领域中基于真实用户问题的数据增强方法，该方法也是对训练数据中的问题产生复述，以增加数据多样性，但不采用序列生成的方式，而是基于用户问题的表达形式进行构造，避免了序列生成模型的训练，增加数据的可控性，同时构造数据是基于真实数据产生的，增加了数据的一致性。该方法首先通过实体识别构建问题的语义原型库，并利用相似度计算获取当前问题的相似原型，然后对相似原型进行语义泛化，构造出包含真实语义结构的同义问句，增加问题的多样性，从而实现增加整个训练数据的量级和多样性。我们将本文提出的方法在真实汽车领域数据上进行实验，其结果表明该方法有效的提升了问答模型的准确率和鲁棒性。综上所述，本文的主要贡献包括：（1）提出了一种垂直领域中基于真实问题的数据增强方法，提升了模型的准确率和鲁棒性。（2）在汽车领域数据上对多个模型和数据增强方法进行了对比实验，实验结果证明了该方法的有效性。

2 相关工作

SQUAD等大规模评测数据的出现，引起学术界和工业界对抽取式阅读理解的深入研究，R-Net(Wang et al., 2017)、DrQA(Chen et al., 2017)、QANET(Yu et al., 2018)等一大批深度学习模型被相继提出。随着BERT(Devlin et al., 2018)、Roberta(Liu et al., 2019)、Albert(Lan et al., 2019)等预训练模型的提出，抽取式阅读理解取得了突破性进展，多种基于预训练模型的方法，在SQUAD2.0数据集上的评价指标超过了人类水平。本文根据各种深度学习模型和预训练模型，对垂直领域抽取式阅读理解的数据增强方法进行研究，以提升各种模型在垂直领域中的准确率和鲁棒性。在机器阅读理解任务中，常用的数据增强方法有远程监督方法、问题生成方法和完全生成方法三种。

远程监督的方法利用外部知识库自动对语料进行标注，构造训练数据，如Chen等(2017)利用QA问答对作为知识库，通过检索得到相关文档片段，构造训练数据。Zhang等(2018)通过知识三元组(E_1, R, E_2)，用实体E_1和关系R构造问题，实体E_2作为答案，用问题和答案检索无标注文档，构造训练数据，从而增加数据量级，提升模型性能。

问题生成的方法利用模型生成新的问题，构建训练数据，包括生成相关问题和生成同义复述问题两种。Zhu等(2019)通过生成相关但不可回答的问题，提升模型的语义理解能力，在SQUAD2.0数据集上取得1.9个F1点的提升。Gan等(2019)提出引导式的生成方法，利用seq2seq模型生成同义问题，增加问题的多样性，提升模型的准确性和鲁棒性。

完全生成的方法是给定文档，直接利用模型根据文档内容生成相关问题和答案，构造训练数据。如Subramanian等(2017)利用模型先从文档中提取关键短语，并以该短语为参考答案生成相关的问题，从而构造训练数据。Puri等(2020)先用BERT从文档中提取答案片段，再将答案和文档进行拼接，利用GPT2(Radford et al., 2019)模型生成相关问题，构造训练数据。

以上几种数据增强方法都是针对通用领域的研究，忽略了数据增强与实际应用数据的结合，会造成在垂直领域应用中构造数据与实际数据之间的语义偏差，从而影响模型应用效果。另外，远程监督的方法容易引入数据噪音，问题生成的方法其数据质量难以控制并且需要训练序列生成模型，同时垂直领域中数据专业性程度高，领域实体数量多，表达更多样，因此以上方法在垂直领域中不适用。借鉴其它自然语言处理任务中利用替换的方式进行数据增强的思想(Wei and Zou, 2019; Fadaee et al., 2017)，本文提出了一种垂直领域中基于真实用户问题的数据增强方法。该方法利用真实用户数据，对训练数据中的问题产生复述，以增加数据多样性，避免了序列模型的训练，增加数据的可控性，同时构造数据是基于真实数据产生的，增加了数据的一致性。最后在汽车领域数据集上，本文通过实验证明该方法对模型的准确率和鲁棒性均能有效提升。

3 数据增强方法

本文提出的数据增强方法是基于真实用户问题，该数据来源于问答系统的日志记录。该方法首先通过实体识别对用户问题进行处理，构建语义原型库；然后利用相似度计算方法，从原型库中获取当前问题的若干相似原型；最后对相似原型进行语义原型泛化，构造出包含真实用户问题语义结构的同义问题。

第十九届中国计算语言学大会论文集，第257页-第266页，海南，中国，2020年10月30日至11月1日。

3.1 问题预处理

问题预处理，是将用户问题进行实体识别，从而获取问题语义原型的过程。将问句抽象为字符序列$Q = (c_1, c_2, c_3, \ldots, c_{(n-1)}, c_n)$，对序列Q进行实体识别，得到序列$Q^T = [c_1, \ldots E_1(c_i, c_{(i+1)}), \ldots E_2(c_{i+k}), \ldots c_n]$，其中$E_i$为识别出的实体，$c_i$为问句中的字符，$Q^T$称为问句的语义原型。将真实用户数据进行预处理，构建问句的语义原型库，从而可以获取大量的表达多样的语义原型数据来构造训练数据的同义问题。

问句预处理过程中的实体识别是指将文本中具有特定含义的文字片段作为一个整体识别出来。在通用领域，实体的类型主要有人名、地名、机构名称、专用名词等，在汽车垂直领域，实体的类型有车系、车型、品牌、车身参数、配置等等。

3.2 语义原型相似度计算

从原型库中找到与训练数据问题相似的问句原型，是构造同义问题的重要环节，相似语义原型的挑选既要考虑问句中已识别的实体序列的相关性，也要考虑字符序列的语义相关性。假设语义原型Q_1^T,Q_2^T，其相似度计算方法如式1：

$$P(Q_1^T, Q_2^T) = w_1 R_1 + w_2 R_2 + w_3 R_3 \tag{1}$$

其中w_1、w_2、w_3为权重参数；R_1为实体类型相关因子，代表两个原型实体类型的相关性；R_2为实体顺序相关因子，代表两个原型实体类型的先后顺序一致性；R_3为语义相关因子，代表两个原型的语义相关性。

R_1为实体类型相关因子，来衡量两个原型实体类型的相关性。定义：函数$E_set(Q^T)$为语义原型中实体类型的集合，$|E_set(Q^T)|$为实体类型个数，$|E_in| = |E_set(Q_1^T) \cap E_set(Q_2^T)|$为两个原型实体集合交集的实体个数。则$R_1$的计算方法如式2所示：

$$R_1 = 2(\frac{|E\_in|}{|E\_set(Q_1^T)|}\frac{|E\_in|}{|E\_set(Q_2^T)|})/(\frac{|E\_in|}{|E\_set(Q_1^T)|} + \frac{|E\_in|}{|E\_set(Q_2^T)|}) \tag{2}$$

R_2为实体顺序相关因子，顺序一致为1，否则为0。

R_3为两个原型的语义相似度值，本文语义相似度的计算采用SBERT(Reimers and Gurevych, 2019)模型，首先将原型中的实体词替换为实体名称，得到新的问题表示，利用孪生网络对问题中的字符进行向量化表示，通过计算向量的余弦值得到问题的相似度。网络的训练和推理如图2，训练阶段，问题1和问题2输入到BERT模型，经过平均池化，输出得到向量u和v。向量u、v及两个向量内部元素的差值$|u - v|$进行拼接，输入到Softmax分类器中进行训练。在推理阶段，直接计算u和v的余弦值，得到R_3值。

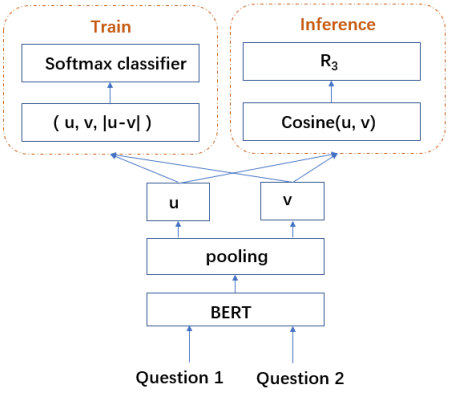


图 2. 语义相似度计算网络

3.3 语义原型泛化

语义原型泛化是对相似原型进行处理获取同义问题的过程，利用问题原型中的实体内容，替换相似原型对应的实体内容，改变了相似原型问句表达的内容主体，但是相似原型的语义结构保持不变，从而构造出主体内容一致，但表达形式多样的同义问题，从而能有效增强构造数据中问题的多样性表达。

语义原型的泛化过程如图3所示，通过对当前问题进行处理，从原型库选取与当前问题语义原型相似的若干原型，用当前原型的实体，替换相似原型中同类别实体，例如：用“宝马X3”替换“奥迪Q3”，用“价格”替换“钱”等，保留相似原型中的其它字符不变，从而得到当前问题的同义问题。

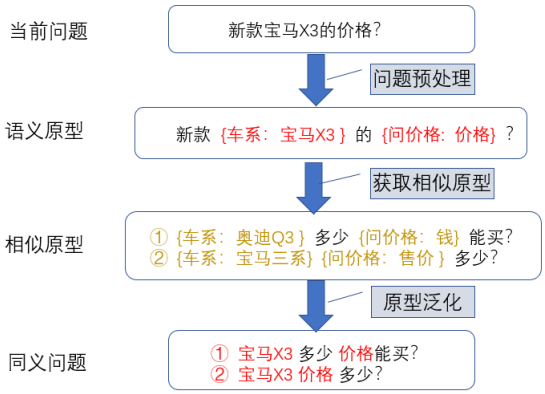


图 3. 基于语义原型的同义问题构造

4 实验

4.1 实验数据

本文将提出的数据增强方法在汽车领域数据集上进行验证，该数据集通过人工标注获取，对给定的每篇资讯文章提出3至5个相关问题并标出答案位置。该数据集共包含905篇汽车类资讯文章和2746个相关的问题，分为训练集和测试集两部分。同时为了验证鲁棒性，对测试集中的问题进行人工复述，每个问题生成若干个同义表达，产生鲁棒性测试集，共包含2312个同义表达的问题，数据样例见表1，具体细节见表2。

数据样例	文档	..... 今年11月昂希诺纯电动正式在国内上市，补贴后售价为**17.28-19.88万元**，共三款车型，今天我们测试的是它的顶配车型TOP悦享版。昂希诺纯电动的电池容量为64.2kWh，电芯为宁德时代NCM配比为523的方壳三元锂电芯，NEDC续航里程为500km。.....
	问题	北京现代昂希诺纯电动补贴后售价是多少?
	答案/索引	17.28-19.88万元/(*start_index*: 713)
鲁棒性样例	人工复述1	北京现代昂希诺纯电动补贴后多少钱?
	人工复述2	现代昂希诺纯电动版补贴后的价格是多少?

表 1. 标注数据样例

本文在3.1节问题预处理部分，实体识别是用汽车领域专用的实体识别算法，能够识别出车系、车型、品牌、车身参数、配置等领域实体。在3.1节语义原型相似度计算部分，SBERT语义相似度模型需要数据进行训练，为了避免人工标注，本文在网络上爬取百度知道中的提问和相关提问数据，用汽车领域的关键词进行筛选，最终得到约20万组相关问题，约100万条数据。同组问题组合标记为正样本，不同组数据组合标记负样本，构造模型的训练数据和测试数据，训练SBERT语义相似度模型。

数据集	文档数目	问题数目	问题长度	答案长度	文档长度
Train	610	1998	19	8	832
Test	295	748	20	12	920
Robust_test	2312	1998	20	11	920

表 2. 实验数据

4.2 对比实验

为了验证本文提出的数据增强方法的有效性，本文用BERT_base模型作为基准模型进行实验，其中Batch_size为6，Epoch为4，其它超参数保持不变，对比以下各种数据增强方法：

简单数据增强方法EDA(Wei and Zou, 2019)：对原始训练数据集中的问题进行处理（同义词替换、插入、删除、交换位置）得到新问题，随机抽出新问题与原始训练数据中的文档进行组合，构造训练数据。

远程监督增强方法DS(Zhang et al., 2018)：将汽车领域新闻资讯文章按段落进行切分，构建Elasticsearch索引，用汽车领域知识图谱中3万个知识三元组数据进行搜索，将检索到的段落作为文档D，用知识三元组(E_1, R, E_2)中的实体E_1和关系R构建问题Q，实体E_2作为答案A，构建训练数据(Q, D, A)。

语义原型泛化增强方法（PG）：本文所提数据增强方法。

以上三种方法在测试集和鲁棒性测试集上的实验结果如图4和图5所示。横坐标N_{aug}表示添加构造数据的数量，$N_{aug} = 0$表示没有添加构造数据，$N_{aug} = 1$表示添加了原始训练数据1倍数量的构造数据。实验结果表明，在汽车领域数据测试集和鲁棒性测试集中PG方法效果要优于其他两种方法。

如图4在测试集中，PG方法构造的数据对测试集的EM和F1值均有2个点以上的提升，N_{aug}在2至8时，效果最好，N_{aug}超过16时，提升效果有所下降；其他两种方法效果相当，对测试集几乎没有提升效果。对于远程监督方法，汽车领域知识三元组数据量大，但是种类相对较少，构造出来的数据形式相对单一，另外数据构造过程中也会引入较大的噪声，这些因素可能对构造数据质量产生影响，从而影响实验结果；对于EDA构造方法，形式上相对简单，在分类任务中有效果，在阅读理解任务中表现不明显。

如图5在鲁棒性测试集中，三种方法对F1指标均有提升效果，PG的提升效果明显高于其他两种方法，EDA的方法略高于DS的方法。对于EM指标，PG和EDA方法优于DS方法，并且DS方法随着数据量的增加，EM指标呈下降趋势，有此可以看出DS方法构造的数据引入的噪声相对较大，对原始训练数据造成了干扰。

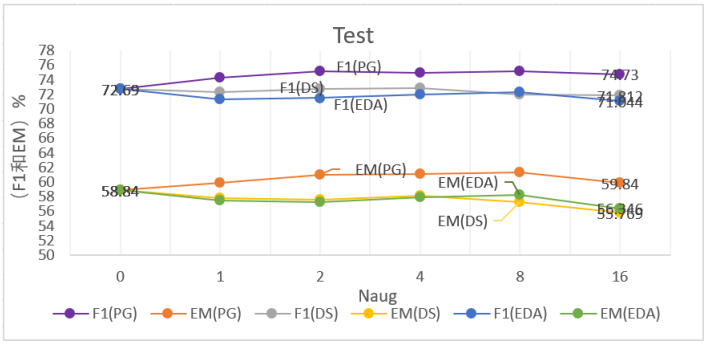


图 4. 在测试集上三种方法对比实验。PG方法提升效果明显优于EDA和DS方法；PG方法在EM和F1指标上均有2个点以上的提升；N_{aug}大于16时，三种方法效果均有下降趋势。

为进一步分析各种方法构造出的训练数据的区别，本文使用原始数据量4倍的构造数据，分别按比重$(0, 0.2, 0.4, 0.6, 0.8, 1)$加入原始训练数据，进行实验。实验结果如图6（a-d），在测试集和鲁棒性测试集中，PG和EDA效果相当，仅使用构造数据就能达到与训练数据接近的效果。DS方法随着原始训练数据的增加，效果逐步提升。DS方法构造的数据完全没有使用原始训练数据，PG和EDA方法构造的数据是在对原始训练数据微调的基础上获取的，因此在仅使

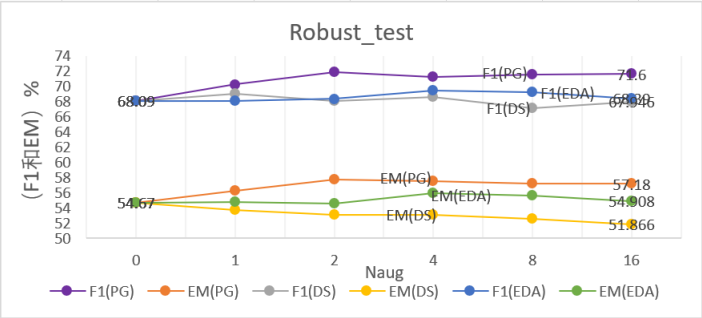
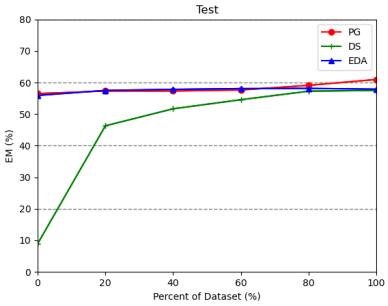
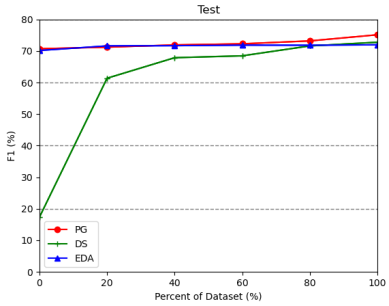


图 5. 在鲁棒性测试集上三种方法对比实验。三种方法对F1值均有提升效果，PG的提升效果明显高于EDA和DS；对于EM指标，PG和EDA方法要优于DS方法，并且DS方法随着数据量的增加，EM指标呈明显下降趋势。

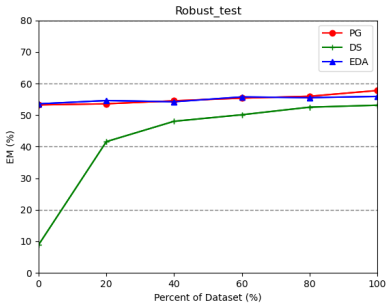
用构造数据时，DS效果明显低于PG和EDA。



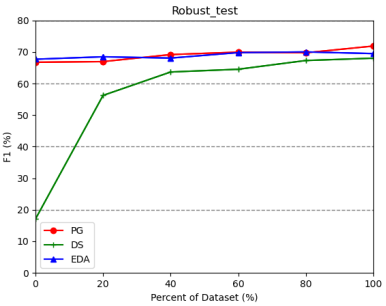
(a) 测试集EM变化



(b) 测试集F1变化



(c) 鲁棒性测试集EM变化



(d) 鲁棒性测试集EM变化

图 6. 训练数据占比变化图。图a、b是在测试集上随着训练数据比重增加F1和EM指标的变化，图c、d是在鲁棒性测试集上随着训练数据比重增加F1和EM指标的变化。在测试集和鲁棒性测试集中，PG和EDA效果相当，仅使用构造数据就能达到与训练数据接近的效果。DS方法随着原始训练数据的增加，效果逐步提升；在仅使用构造数据时，DS效果明显低于PG和EDA。

4.3 多模型验证实验

为了验证本文提出的数据增强方法在各种模型上的通用性，本文选择近期在阅读理解任务中表现突出的多个模型进行实验。

BERT模型：BERT模型在阅读理解任务取得突破性的成绩，它采用多层Transformer结构堆叠而成，层数的不同，模型大小不同，本文采用层数为12的BERT_base模型进行微调，验证方法的有效性，其中Batch_size为6，Epoch为4，其他参数不变。

Albert模型：Albert模型在BERT模型基础上进行了改进，通过词嵌入矩阵的分解和隐藏层参数共享，减小模型的参数，提升模型的性能。本文选择与BERT_base模型参数量相当

的Albert_xlarge模型进行实验，其中Batch_size为6，Epoch为4，其他参数不变。

DrQA模型：DrQA模型是一个完整的端到端的阅读理解问答系统，包含文档检索和文档阅读两个模块，本文仅使用文档阅读模块，验证方法的有效性。在实验中，数据预处理采用CoreNLP(Manning et al., 2014)进行分词和实体识别，使用腾讯中文词向量(Song et al., 2018)进行词嵌入，训练参数与原模型一致。

如表3，实验结果表明本文提出的数据增强方法在三个模型上均有效果，测试集和鲁棒性测试集的F1和EM指标都有2个点以上的提升。从模型之间的对比可以看到：Bert、Albert预训练语言模型在阅读理解任务中表现突出，DrQA是非预训练模型，没有经过大量无监督数据的预训练，因此效果较差；在参数量相当的时候，经过改进的Albert模型效果比Bert更好。

模型	数据集	数据增强	F1	EM
Albert	Test	N	74.12	59.01
		Y	76.97	62.30
		Δ	2.85	3.29
	Robust_test	N	70.56	55.79
		Y	73.49	58.80
		Δ	2.93	3.01
BERT	Test	N	72.69	58.84
		Y	74.94	61.08
		Δ	2.25	2.24
	Robust_test	N	68.09	54.67
		Y	71.19	57.49
		Δ	3.10	2.82
DrQA	Test	N	61.80	44.55
		Y	66.00	49.62
		Δ	4.20	5.07
	Robust_test	N	56.45	39.19
		Y	60.35	43.71
		Δ	3.90	4.52

表 3. 数据增强方法在多个模型上的实验结果，其中N表示不使用数据增强，Y表示使用数据增强，Δ表示增加量。

5 结束语

本文提出了一种垂直领域中基于真实用户问题的数据增强方法，该方法对真实用户问题的语义原型进行泛化，构造同义表达问题，从而增强问题的多样性，同时提升构造数据和应用场景中数据的一致性，从而提升模型的准确率和鲁棒性。该方法结合了垂直领域的数据特点和相关技术方法，如：领域实体识别技术，在汽车领域数据集上，验证多种模型，F1和EM指标均能取得2至5个百分点的提升。本文面向垂直领域的数据增强方法对其它各垂直领域都有借鉴作用，具有很大的普适性，下一步将结合本方法，在通用领域数据上进行分析和研究。

参考文献

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. A span-extraction dataset for chinese machine reading comprehension. *arXiv preprint arXiv:1810.07366*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*.

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. Improving neural question generation using answer separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6602–6609.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Raul Puri, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. *arXiv preprint arXiv:2002.09599*.

Yingqi Qu, Jie Liu, Liangyi Kang, Qinfeng Shi, and Dan Ye. 2018. Question answering over freebase via attentive rnn with similarity matrix based cnn. *arXiv preprint arXiv:1804.03317*, 38.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Yoshua Bengio, and Adam Trischler. 2017. Neural models for key phrase detection and question generation. *arXiv preprint arXiv:1706.04560*.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Hongzhi Zhang, Xiao Liang, Guangluan Xu, Kun Fu, Feng Li, and Tinglei Huang. 2018. Factoid question answering with distant supervision. *Entropy*, 20(6):439.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3901–3910.

Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. 2019. Learning to ask unanswerable questions for machine reading comprehension. *arXiv preprint arXiv:1906.06045*.

安波, 韩先培, and 孙乐. 2018. 融合知识表示的知识库问答系统. 中国科学:信息科学, 48(11):59–70.

白龙, 靳小龙, 席鹏弼, and 程学旗. 2019. 基于远程监督的关系抽取研究综述. 中文信息学报, 33(10):10–17.