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。

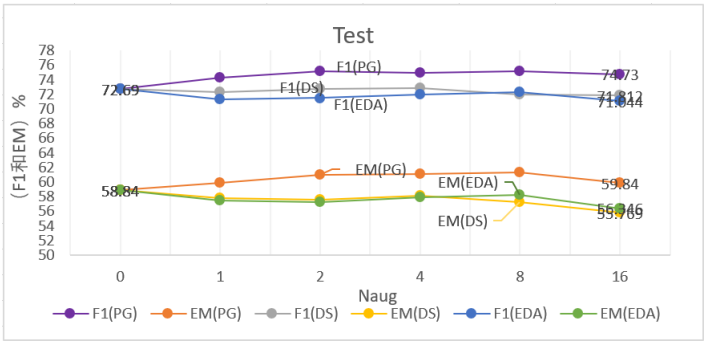


图 4. 在测试集上三种方法对比实验。PG方法提升效果明显优于EDA和DS方法；PG方法在EM和F1指标上均有2个点以上的提升；N_{aug}大于16时，三种方法效果均有下降趋势。

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

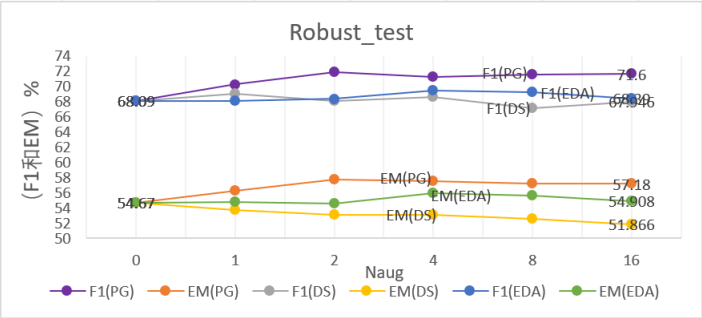
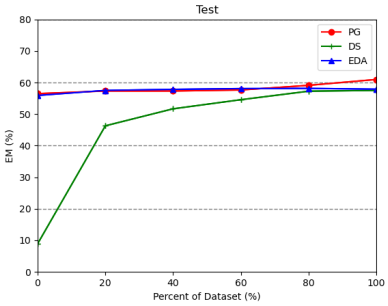
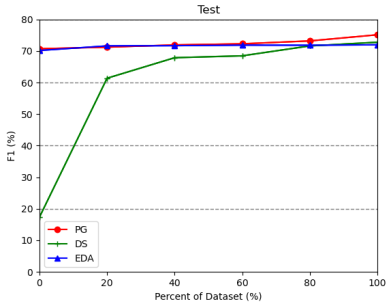


图 5. 在鲁棒性测试集上三种方法对比实验。三种方法对F1值均有提升效果，PG的提升效果明显高于EDA和DS；对于EM指标，PG和EDA方法要优于DS方法，并且DS方法随着数据量的增加，EM指标呈明显下降趋势。

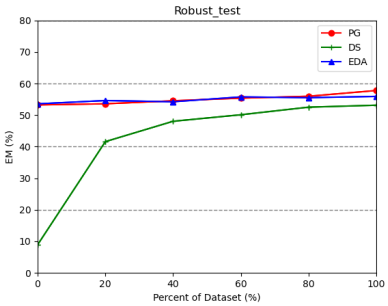
用构造数据时，DS效果明显低于PG和EDA。



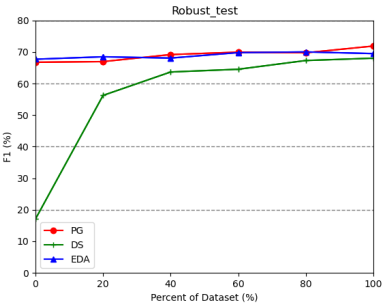
(a) 测试集EM变化



(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.