基于复述增广的医疗领域机器翻译

安 波* 龙从军

(中国社会科学院民族学与人类学研究所 北京 100081)

摘 要：医疗机器翻译对于跨境医疗、医疗文献翻译等应用具有重要价值。汉英神经机器翻译依靠深度学习强大的建模能力和大规模双语平行数据取得了长足的进步。神经机器翻译通常依赖于大规模的平行句对训练翻译模型。目前，汉英翻译数据主要以新闻、政策等领域数据为主，缺少医疗领域的数据，导致医疗领域的汉英机器翻译效果不佳。针对医疗垂直领域机器翻译训练数据不足的问题，该文提出利用复述生成技术对汉英医疗机器翻译数据进行增广，扩大汉英机器翻译的规模。通过多种主流的神经机器翻译模型的实验结果表明，通过复述生成对数据进行增广可以有效地提升机器翻译的性能，在RNNSearch, Transformer等多个主流模型上均取得了6个点以上的BLEU值提升，验证了复述增广方法对领域机器翻译的有效性。同时，基于MT5等大规模预训练语言模型可以进一步地提升机器翻译的性能。

关键词：神经机器翻译；汉英翻译；复述生成；数据增广；大规模预训练语言模型

中图分类号：TN912.3; TP393 **文献标识码**：A **文章编号**：1009-5896(2022)01-0118-09

DOI: 10.11999/JEIT210926

Paraphrase Based Data Augmentation For Chinese-English Medical Machine Translation

AN Bo LONG Congjun

(*Institute of Ethnology and Anthropology*, *Chinese Academy of Social Sciences*, *Beijing*, 100081, *China*)

Abstract: Medical machine translation is of great value for cross-border medical translation. Chinese to English neural machine translation has made great progress based on deep learning, powerful modeling ability and large-scale bilingual parallel data. Neural machine translation relies usually on large-scale parallel sentence pairs to train translation models. At present, Chinese-English translation data are mainly in the fields of news, policy and so on. Due to the lack of parallel data in the medical field, the performance of Chinese to English machine translation in the medical field is not compromising. To reduce the size of parallel data for training medical machine translation models, this paper proposes a paraphrase based data augmentation mechanism. The experimental results on a variety of neural machine translation models show that data augmentation through paraphrase augmentation can effectively improve the performance of medical machine translation, and has achieved consistency improvements on mainstream models such as RNNSearch and Transformers, which verifies the effectiveness of paraphrase augmentation method for domain machine translation. Meanwhile, the medical machine translation performances could be further improved based on large-scale pre-training language model, such as MT5.

Key words: Neural machine translation; Chinese to English translation; Paraphrase generation; Data augmentation; Large scale pre-traing language model

收稿日期：2021-09-01；改回日期：2021-11-30；网络出版：2021-12-29

*通信作者： 安波 anbo724@163.com

基金项目：国家自然科学基金(62076233)，中国社会科学院重大创新工程项目(2020YZDZX01-2)

Foundation Items: The National Natural Science Foundation of China (62076233), The Major Innovation Project of Chinese Academy of Social Sciences (2020YZDZX01-2)

1　引言

机器翻译(Machine Translation, MT)是利用计算机将源语言的文本翻译为目标语言的文本的技术，是自然语言处理的核心任务之一，对于实现跨语言交流等应用具有重要价值[1–3]。机器翻译按照发展阶段可以大致分为基于词典的机器翻译[4]、基于规则的机器翻译[5]、基于统计的机器翻译[6]和基于神经网络的机器翻译[3]。当前，随着深度神经网络在自然语言处理中的广泛应用，基于神经网络的机器翻译取得了较好的性能，成为当前机器翻译领域的主流方法[3,7]。医疗领域机器翻译在药品研发、跨境医疗等领域具有重要的应用价值，也得到了学界和企业界的广泛重视[8–10]。

基于神经网络的机器翻译通常需要较多的训练数据，目前的大规模平行语料主要以新闻、政策文档等领域的数据为主，缺少大规模开源医疗领域的汉英平行数据集[11–13],因此训练数据不足是制约医疗领域机器翻译的关键因素之一。针对训练数据不足的问题，研究者们提出无监督学习、半监督学习、数据增广等方法来减少模型对训练数据的依赖[13–16]。其中数据增广通过自动生成新的训练数据的方式来增加训练数据，具有较好的通用性，得到了学界的广泛关注[13]。常用的数据增广方法包括基于回译的数据增广、基于同义词替换的数据增广和基于复述生成的数据增广[16–20]。其中基于回译的数据增广通过两次不同方向的机器翻译实现[19]，如将汉语句子通过汉英翻译模型翻译为英文句子，然后通过英汉翻译模型将英文句子翻译为中文句子。该方法依赖于已有机器翻译模型的性能。基于词典替换的数据增广方法主要依赖于同义词词典对句子中的同义词进行替换，受限于同义词词典的规模和领域，并且句子语言的多样性变化较小[17]。

机器翻译数据是相同语义在不同语言下的表示，复述是相同语义在同一语言下的不同表述，因此基于已有的双语平行语料，通过在源语言/目标语言上进行复述生成，能够生成新的对齐数据，从而实现数据增广(如图1所示)[20]。基于高质量的复述数据可以训练较好的复述生成模型，生成语义一致但词汇、句法不同的数据[21]。因此基于复述生成的数据增广方法可以更好地处理语言的多样性，增强模型的鲁棒性、减少对训练数据的依赖[21,22]。

基于上述分析，本文提出基于复述增广的医疗机器翻译方法。该方法首先利用高质量的汉语复述数据训练汉语复述生成模型。其次，设计实现基于汉英双语医学电子书中抽取双语平行数据集，并在采集到的汉英医疗领域平行数据上利用复述生成方法进行数据增广，得到更大规模的汉英医疗机器翻译平行语料。最后，利用多种主流的神经机器翻译方法进行机器翻译的模型验证。实验结果表明，我们提出的方法能够有效地提升汉英医疗机器翻译的性能(平均提升6个点的BLEU值)，验证了基于复述增广的机器翻译方法的价值。需要说明的是，本文在数据增广时以汉语作为主要增广对象，主要目的是能够更好地实现汉语与其他语言的翻译，如汉英、汉日、汉韩等，汉语与这些语言之间均存在如跨境医疗的翻译需求。以汉语作为数据增广的对象，可以实现对汉语与其他多种语言之间机器翻译的性能。

本文的主要贡献包括以下3点：

(1) 本文设计实现了一种基于复述生成的方式提升医疗机器翻译性能的方法，该方法具有较好的通用性，能够提升多种主流的机器翻译模型；

(2) 通过对比基于同义词替换、基于深度学习的复述生成模型和基于大规模预训练语言模型的复述生成模型发现，基于大规模预训练语言模型(Bert, MT5)的复述生成方法能够更大程度地提升机器翻译的性能；

(3) 本文利用医疗领域著作、指南、病历等双语数据构建了一个汉英医疗机器翻译数据集。

2　相关工作

本文主要涉及机器翻译和基于数据增广的模型提升工作，本节将从这两个方面分别进行介绍。

2.1　机器翻译

机器翻译是自然语言处理的核心任务，因其具有非常强的应用价值和市场需求，一直是自然语言处理领域的研究热点[1,2]。机器翻译按照发展阶段可以大致分为：早期基于词典的机器翻译(Dictionary Based Machine Translation, DBMT)、融合词典和语言知识的规则翻译(Rule Based Machine translation, RBMT)、统计机器翻译(Statistic Machine Translation, SMT)和神经机器翻译(Neural Machine Translation, NMT)[3–6]。随着深度学习的快速发展和计算性能的爆炸式提升，基于深度学习的神经机器翻译成为当前研究和应用的主流方法[3]。

IBM在1954年在IBM-701计算机上首次实现了英俄机器翻译实验，验证了机器翻译的可行性，正式拉开了机器翻译研究的序幕[23]。这一时期由于军事、政治、文化的需求，各国对于外文资料均有较多的翻译需求，因此也对机器翻译研究提供了较多地支持，也产生了包含LMT等具有代表性的系统。但是由于翻译质量差、速度慢等特点，1966年

ALPAC对于机器翻译的负面评价导致机器翻译的研究出现了短暂的停滞。

20世纪70年代，基于规则的机器翻译逐渐成熟，机器翻译再一次得到较为广泛地应用。这类方法依赖于一定的规则对词法/句法等语言学信息进行转换实现机器翻译。这一时期的代表系统包括：Systran, Japanese MT systems和EUROTRA[24–26]等。这类方法的缺点也存在人工规则制定成本高、规则易冲突、不利于系统扩展等缺点。

统计机器翻译利用机器学习将机器翻译建模为从源语言到目标语言的生成问题，即求解最大化$p(\mathrm{t}|\mathrm{s})$，其中s为源语言句子，t为目标语言句子。统计机器翻译最早在1949年由瓦伦基于香农的信息论提出[27]。最早可行的统计机器翻译模型则是由IBM研究院提出，并实现IBM Model-1到IBM Model-5 5种统计机器翻译模型[28]。为了解决基于词翻译的语义单元过小的问题，研究者提出基于短语的机器翻译，得到了广泛地应用。目前爱丁堡大学维护的Moses[29]是统计机器翻译最为成功的开源实现。在国内，小牛翻译开源的NiuTrans也得到了较为广泛的关注[30]。

近年来，随着深度学习、神经网络在自然语言处理领域的广泛应用，基于神经网络的机器翻译(NMT)也得到广泛的关注。神经机器翻译同样将机器翻译建模为从源语言到目标语言的生成问题。2013年Kalchbrenner等人[31]提出了基于编码器-解码器结构的神经机器翻译方法，该方法使用卷积神经网络(Convolution Neural Network, CNN)[32]作为源语言的编码器，使用迭代神经网络(Recursive Neural Network, RNN)[33]作为目标语言的解码器。为了解决RNN带来的梯度爆炸和梯度消失问题，基于长短时记忆网络(Long Short Time Memory, LSTM)[34]的模型被引入机器翻译的编解码模型，并提出了在机器翻译领域著名的Seq2Seq框架[35]。受到图像领域启发，注意力机制(Attention)被引入到机器翻译模型中，该机制动态的建模在生成目标词时所应当关注的源语言词的信息，能够更好地指导目标词的生成，因此得到了广泛地应用[36]。近期，谷歌将基于自注意力机制(self attention)的Transformer结构引入到机器翻译模型中，取得了非常好的效果，成为当前神经机器翻译的主流方法[37]。

因为有着强烈的市场需求，神经机器翻译得到了学界和企业界的广泛重视，在各大自然语言处理、人工智能的顶级会议中均为较多数量的神经机器翻译的研究工作。目前，谷歌、百度、搜狗、有道、小牛等公司也在神经机器翻译上投入了大量的资源。

2.2 基于复述的数据增广

与其他基于深度学习的模型类似，神经机器翻译通常需要大量的训练数据才能训练得到较好的模型，然而对于小语种或垂直领域而言，往往缺乏大规模的训练数据，如汉藏翻译、医疗机器翻译等。针对数据稀缺的问题，基于无监督的机器翻译、基于迁移学习的机器翻译和基于数据增广的机器翻译受到学者们的广泛关注。

数据增广在图像处理、自然语言处理等领域具有广泛地应用。在自然语言处理领域，数据增广的方法主要包括：基于同义词替换的方法、基于回译(back translation)的方法和基于复述生成的数据增广方法。基于同义词替换的方法借助于已有的同义词词典或词向量来获取词汇的同义词，通过同义词替换的方式生成新的句子，以达到数据增广的目的。然而，基于同义词替换的方法主要受限于高质量的同义词词典，并且仅在词汇级别上进行替换难以生成具有多样性的句子。随着机器翻译等技术的提升，基于回译的机器翻译越来越多地用于数据增广。然而，基于回译的机器数据增广方法严重依赖于已有的机器翻译模型，且已有的商用机器翻译服务(百度、谷歌)均为通用领域的机器翻译，在医疗文本翻译方面不能进行有效的翻译。基于复述的数据增广是利用复述生成的方法对数据进行增广，复述生成也成为自然语言处理领域数据增广的常用方法[20,22]。

通常机器翻译的训练数据为语义对齐的双语句子，而复述是相同语义在同种语言下的不同表达，因此通过复述生成的方法对机器翻译训练句对中的一个句子进行复述，得到的复述句与训练句对中的另外一个句子天然的形成新的机器翻译训练句对。基于复述的数据增广方法主要涉及复述数据集和复述生成方法，在汉语环境下已经有了多种公开的复述数据集，如BQ Corpus[38], Chinese PPDB[1)], PKU paraphrase bank[39], Phoenix Paraphrasing dataset[2)]等，为本文的研究提供了语料库支撑。复述生成方法主要可以分为：基于词典与规则的复述生成、基于统计学习的复述生成和基于神经网络的复述生成。随着训练数据规模的提升，深度学习依赖其强大的建模能力，在复述生成领域取得了较好的效果。包括基于迭代神经网络的复述生成，基于长短是记忆网络的复述生成和基于Transformer的复述生成[40]。近期，大规模预训练语言模型在自然

1) https://github.com/casnlu/Chinese-PPDB
2) https://ai.baidu.com/broad/subordinate?dataset=paraphrasing

语言处理领域得到了广泛的应用，如Bert[41], MT5[42]等，这些模型利用其较强的文本表示与文本生成能力能够在一定程度上提升模型的泛化能力和生成文本的多样性。

3　基于复述增广的医疗机器翻译方法

本文的基本思路是在已有汉英医疗机器翻译平行句对的基础上，利用复述生成技术对平行句对中的汉语句子进行复述，进而生成具有与英文句子相同语义的汉语新句子构建新的平行句对，从而达到复述数据扩充的目的，如图1所示。本文的方法主要包含以下3个步骤：(1)首先基于已有的汉语复述语料集构建汉语复述生成模型；(2)然后利用中文复述生成模型对采集的汉英医疗机器翻译数据集进行数据增广；(3)最后在增广后的双语平行数据集上进行神经机器翻译模型的训练，得到医疗机器翻译模型。本节将从中文复述生成模型、医疗汉英平行语料采集和复述增广的神经机器翻译方法3个方面分别进行介绍。

3.1　汉语复述生成模型

复述生成模型能够产生与给定文本字面不同但语义相同的文本，按照复述粒度的不同，可以分为词级复述(即同义词)、短语级复述、句子级复述和文档级复述。本文针对机器翻译双语平行语料库数据增广的需要，仅涉及句子级复述。我们使用复述生成来实现汉语句子的复述，复述生成模型的训练依赖于高质量的复述数据集，本文通过融合BQ Corpus, Chinese PPDB, PKU paraphrase bank和Phoenix Paraphrasing dataset 4个数据集，形成一个较大规模的中文复述数据集。

近期，基于深度学习的文本生成方法取得了显著地提升，本文在Seq2Seq框架下实现了3种常用的复述生成模型，包括基于RNNSearch的复述生成模型、基于BiLSTM的复述生成模型和基于Transformer[36]的复述生成模型。同时，为了能够更好地实现对医疗专有名词的翻译(疾病词、症状词、药品名、手术名等)，本文引入了Copy机制来实现高质量的专有名词的翻译。大规模预训练语言模型通过在大规模文本数据上的训练，可以增强模型的泛化能力，也能提升文本生成的多样性。因此，我们在Bert和MT5[43]的基础上进行微调，训练得到复述生成模型。复述生成的整体框架如图2所示。

如图2所示，其中基于深度学习的复述生成模型(BiLSTM, Transformer)的表示层使用预训练的词向量，本文使用腾讯发布中文预训练词向量[3)]，中文分词采用北京大学开源的pkuseg[4)]。编码层和解码层采用对应的模型，如Transformer的编码层和解码层均使用Transformer，分类层采用Softmax。基于预训练语言模型的复述生成模型(Bert,MT5)均以汉字为单位作为输入，表示层和编码层均采用语言模型的文本表示方法。其中基于Bert的方法在编码层和解码层为两个单独的Bert模型，共享词表但是分别训练。由于MT5本身为文本生成模型，因此

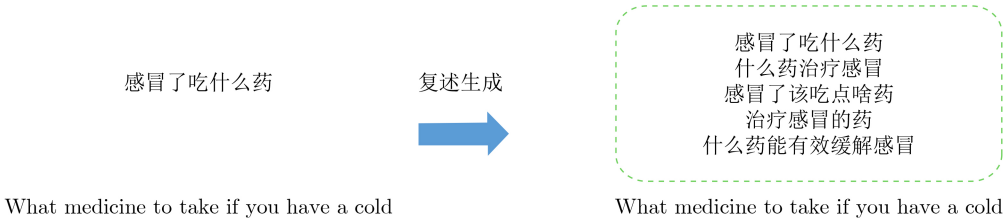

图 1 基于复述生成的机器翻译数据增广示意图

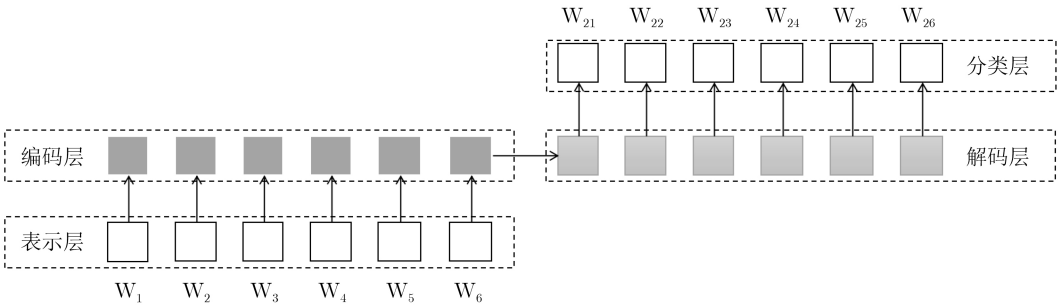

图 2 复述生成整体框架图

3) https://ai.tencent.com/ailab/nlp/zh/embedding.html
4) https://github.com/lancopku/pkuseg-python

只需要在汉语复述数据上进行微调(fine-tuning)即可得到复述生成模型。

3.2 汉英医疗机器翻译数据采集

目前缺少开源的大规模医疗领域汉英机器翻译数据[44]。针对这种现状，本文通过对医疗领域的双语电子书进行双语平行语料的抽取(包括：医学著作、指南、病历中英双语版本)，构建了一个包含10万句对的医疗领域汉英机器翻译数据集。具体的构建流程如图3所示。其中“汉语书”和“英语书”指的是相同电子书的不同语言的版本，通过扫描后形成对齐的电子书。OCR模块将图片格式的数据转换为汉语和英语的文本数据，本文使用百度开源的OCR识别接口实现字符识别[5)]。在得到文本数据之后，通过章节编号、标题等信息实现章节的切分和对齐。在章节内部，使用Giza++[45]实现词级别的对齐。利用词对齐的信息，找到双语数据中的锚点句(双语句子中的词完全对齐)，然后使用动态规划算法来实现双语章节内部的句子对齐。之后，通过谷歌翻译[6)]将英文翻译为英文，并通过SentenceBert[46]计算句子的语义相似度，过滤掉相似度低于一定阈值的句子对。最后，对得到的双语对齐数据进行去重，去掉中英文完全一致的句子对。

通过上述步骤，本文从医学著作、指南、双语病历等数据中抽取出了约10万条句子对，形成了一个较大规模的医疗机器翻译数据集。本文通过随机抽取的方式将数据分为训练集、验证集和测试集，具体的统计信息如表1所示。

3.3 基于复述增广的机器翻译方法

通过上述步骤，本文得到了汉语复述生成模型和汉英医疗机器翻译数据集。本节介绍通过复述生成模型对双语平行句对中的汉语句子进行复述生成。新生成的句子与原句子对应的英文句子构成新的双语对齐数据。通过上述方法实现了对双语平行语料的增广。该方法的整体框架如图4所示。

表 1 汉英医疗机器翻译数据集

训练集	验证集	测试集	中文平均字数	英文平均词数
85000	7500	7500	14.3	11.2

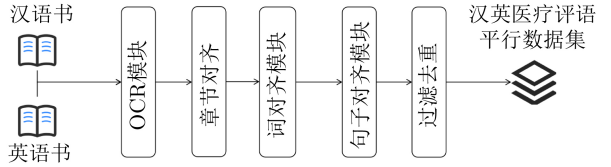

图 3 基于双语电子书的汉英医疗机器翻译数据抽取方法

本文的主要目的是验证基于复述生成的增广方法是否能够有效地提升神经机器翻译的性能，因此本文复现了几种主流的机器翻译模型作为基础模型，包括Seq2Seq, RNNSearch和Transformer。本文在这3种模型先开展实验，来验证方法的有效性。

4 实验

4.1 实验设置

本节主要介绍复述生成模型、神经机器翻译模型的实验模型设置。本文基于Transformer实现复述生成模型，word embedding dim地址为300、beam设置为50，batch size设置为64、句子长度设置为256、learning rate设置为0.01、optimizer设置为Adam。神经机器翻译包含Seq2Seq, RNNSearch和Transformer3种模型，模型的超参数设置如表2所示。本文使用BLEU值作为模型的评价指标。本文的所有实验均为在训练集上进行训练，在验证集上找到最优的超参和epoch次数，在测试集上得到结果。本文所有实验均在一台GPU服务器上进行，其基本配置如下：CPU 2*AMD 霄龙7742、512G DDR4内存、4* Nvidia RTX 24G显卡。本文使用BLUE值作为评价不同模型翻译结果的主要指标。

4.2 对比实验

为了能够验证复述增广方法对于汉英医疗机器翻译的作用，本文设置了多组对比实验，包括：(1)在采集的机器翻译语料上直接使用基础机器翻译模型(Seq2Seq, RNNSearch和Transformer)进行训练；(2)使用基于同义词替换(WordRep)的方法对机器翻译数据进行增广，然后使用机器翻译模型进行训练；(3)使用本文提出的几种复述生成方法对数据进行增广，然后使用机器翻译模型进行训练。其中基于同义词替换的方法，本文使用哈尔滨工业大学的同义词词典[7)]作为同义词数据源。

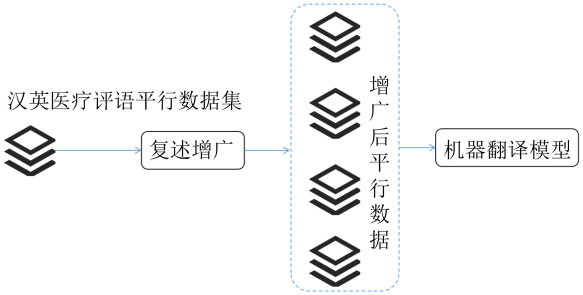

图 4 复述增广的机器翻译方法框架图

5) https://github.com/PaddlePaddle
6) https://translate.google.cn/
7) http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

4.3 实验结果

由于本文主要为了验证基于复述生成的数据增广方法对于医疗机器翻译的增强效果，因此主实验为3种模型在没有数据增广和有数据增广之后的效果的对比，该实验设置生成的复述句子为4个，新生成的训练数据集是原始训练数据的5倍数据量。该部分的实验结果如表3所示，其中“-para”表示增广之后的训练集得到的模型。

从表3我们可以得到以下结论：

(1)基于复述生成的数据增广方法能够显著地提升医疗领域机器翻译的性能，验证了复述增广的方法在机器翻译领域具有一定的通用性。

(2)基于同义词替换的方法(WordRep)基本不能提升机器翻译的性能，在RNNSearch模型下甚至降低了模型的性能，我们推测是可能是由于同义词词典为通用领域的同义词，在医疗领域缺少相关的词汇导致的。

(3)基于语言模型的复述生成方法(Bert, MT5)能够更大程度地提升模型的性能，说明通过这种方法生成的复述句子能够更好地提升机器翻译的性能。

(4)基于MT5的复述生成方法相对于基于Bert的复述生成方法能够更大程度地提升机器翻译的性能，说明MT5在复述生成任务上具有更好的性能和多样性。

为了更清晰地展示本文提出的方法训练得到医疗领域机器翻译的性能，本文在表4中使用一个例子来直观地展示本文提出的方法与百度和谷歌的机器翻译的对比。医疗专家的人工评价也认为本文提出的方法能够较好地保持汉语句子的语义，翻译的结果比较符合常见的病例描述方式，同时在医疗词汇的翻译上也更加准确(如“并持续加重”翻译为“progressive worsening”)。

4.4 不同的复述数量对翻译性能的影响

为了进一步地验证复述增广对机器翻译性能地提升作用，本节通过设置不同的复述数量来观察对于复述模型的提升效果。本部分实验以Transformer作为基础模型，然后通过不同的增广数量来开展实验。该部分的实验结果如图5所示，其中横坐标为1表示仅使用原始训练数据，横坐标为2时复述生成数量设置为1，即使用2倍的数据进行训练，以此类推。

从图5可知，不同的复述数量对于机器翻译的性能有较大影响，在初期阶段通过增加训练数据可

表 2　模型参数设置

机器翻译模型	参数	参数值
Seq2Seq	Embedding size	300
	Beam size	50
	Batch size	64
	Sentence length	256
	Learning rate	0.01
	Optimizer	Adam
	RNN cell	LSTM
	Drouput	0.2
RNNSearch	Embedding size	300
	Beam size	50
	Batch size	64
	Sentence length	256
	Learning rate	0.01
	Optimizer	Adam
	RNN cell	LSTM
	Drouput	0.2
Transformer	Embedding size	300
	Beam size	50
	Batch size	64
	Sentence length	256
	Learning rate	0.1
	Optimizer	Adam
	RNN cell	LSTM
	Drouput	0.2
	Num head	8

表 3　汉英医疗机器翻译结果

机器翻译模型	数据增广模型	BLEU	提升(%)
Seq2Seq	–	31.99	–
	WordRep	32.12	0.41
	BiLSTM-para	33.45	4.56
	Transformer-para	35.23	10.13
	Bert-para	35.28	10.28
	MT5-para	35.74	11.72
RNNSearch	–	41.28	–
	WordRep	40.98	-0.73
	BiLSTM-para	43.25	4.77
	Transformer-para	44.12	6.88
	Bert-para	44.67	8.21
	MT5-para	44.97	8.94
Transformer	–	48.21	–
	WordRep	48.29	0.17
	BiLSTM-para	49.86	3.42
	Transformer-para	51.32	6.45
	Bert-para	51.36	6.53
	MT5-para	51.97	7.80

表 4 汉英医疗机器翻译例子

汉语句子	患者男，31岁，因中重度反复头痛18天入院，表现为枕部至双额部逐渐发作，呈搏动性，发作持续超过4h，并持续加重。
百度	The patient, a 31 year old male, was hospitalized for 18 days due to moderate and severe recurrent headache. He showed a gradual attack from the occipital part to the double frontal part, which was pulsatile. The attack lasted for more than 4 hours and continued to worsen.
谷歌	A 31-year-old male patient was admitted to the hospital for 18 days with moderate to severe recurrent headaches. The manifestations were pulsatile attacks from the occiput to the forehead. The attacks lasted more than 4 hours and gradually worsened.
本文	A 31-year-old man was admitted with an 18-day history of a moderate to severe recurrent headache, presenting gradual onset from occipital to bifrontal regions, pulsatile, in episodes lasting beyond four hours, and progressive worsening.

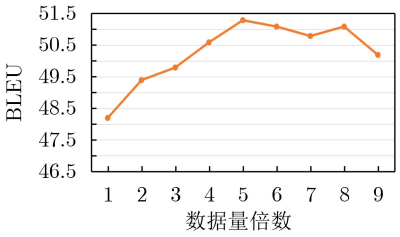

图 5 不同复述数量对机器翻译性能的影响

以快速提升机器翻译的BLEU值，并且当使用5倍的数据进行训练时达到最优的效果。当训练数据超过5倍的数据时，性能开始下降，我们推测是因为复述模型引入了更多的噪音且多样性不足等原因，导致机器翻译性能的下降。

综上所述，实验结果表明基于单语复述增强的方式能够较好地提升医疗机器翻译的性能。我们认为这是由于机器翻译在理解愿语言文本和生成目标语言文本的时候均需要处理语言多样性的问题。在训练数据不足的情况下，单语复述能够提升模型应对一种语言多样性的能力，进而优化机器翻译的性能。

5 结束语

针对医疗领域机器翻译训练数据不足的问题，本文提出一种基于复述生成进行数据增广的方法来增强医疗领域机器翻译的性能的方法。该方法借助于大规模单语复述数据集构建复述生成模型。同时，本文设计实现了一种从医疗领域电子书中抽取汉英医疗机器翻译数据的方法，构建了一个10万句级别的医疗领域机器翻译数据集。最后，利用复述生成模型对医疗机器翻译的训练数据进行增广，得到更大规模的训练数据。在3种不同的神经机器翻译方法的实验结果表明，基于复述增广的机器翻译方法能够有效地提升医疗机器翻译的效果。同时，实验结果表明基于大规模预训练语言模型的复述方式能够最大程度地提升机器翻译的性能。但从实验结果中也可以看出，复述生成仍然会引入一部分噪音，因此针对机器翻译如何生成更高质量的复述句子，避免引入噪音是未来工作的重点。

参 考 文 献

[1] 刘群. 统计机器翻译综述[J]. 中文信息学报, 2003, 17(4): 1–12. doi: 10.3969/j.issn.1003-0077.2003.04.001.
LIU Qun. Survey on statistical machine translation[J]. *Journal of Chinese Information Processing*, 2003, 17(4): 1–12. doi: 10.3969/j.issn.1003-0077.2003.04.001.

[2] 李亚超, 熊德意, 张民. 神经机器翻译综述[J]. 计算机学报, 2018, 41(12): 2734–2755. doi: 10.11897/SP.J.1016.2018.02734.
LI Yachao, XIONG Deyi, and ZHANG Min. A survey of neural machine translation[J]. *Chinese Journal of Computers*, 2018, 41(12): 2734–2755. doi: 10.11897/SP.J.1016.2018.02734.

[3] STAHLBERG F. Neural machine translation: A review[J]. *Journal of Artificial Intelligence Research*, 2020, 69: 343–418. doi: 10.1613/jair.1.12007.

[4] TRIPATHI S and SARKHEL J K. Approaches to machine translation[J]. *Annals of Library and Information Studies*, 2010, 57: 388–393.

[5] CHAROENPORNSAWAT P, SORNLERTLAMVANICH V, and CHAROENPORN T. Improving translation quality of rule-based machine translation[C]. Proceedings of the 2002 COLING workshop on Machine translation in Asia, Taipei, China, 2002. doi: 10.3115/1118794.1118799.

[6] LIU Shujie, LI C H, and ZHOU Ming. Statistic machine translation boosted with spurious word deletion[C]. Proceedings of Machine Translation Summit, Xiamen, China, 2011.

[7] GOODFELLOW I, BENGIO Y, and COURVILLE A. Deep Learning[M]. Cambridge: MIT Press, 2016.

[8] ECK M, VOGEL S, and WAIBEL A. Improving statistical machine translation in the medical domain using the Unified Medical Language System[C]. Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 2004. doi: 10.3115/1220355.1220469.

[9] DUŠEK O, HAJIČ J, HLAVÁČOVÁ J, *et al.* Machine translation of medical texts in the Khresmoi project[C]. Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, USA, 2014. doi: 10.3115/V1/W14-

3326.

[10] WOLK K and MARASEK K P. Translation of Medical Texts Using Neural Networks[M]. HERSHEY P A. Deep Learning and Neural Networks: Concepts, Methodologies, Tools, and Applications. IGI Global, 2020: 1137–1154.

[11] ZOPH B, YURET D, MAY J, *et al*. Transfer learning for low-resource neural machine translation[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, USA, 2016. doi: 10.18653/v1/D16-1163.

[12] PARK C, YANG Y, PARK K, *et al*. Decoding strategies for improving low-resource machine translation[J]. *Electronics*, 2020, 9(10): 1562. doi: 10.3390/electronics9101562.

[13] FADAEE M, BISAZZA A, and MONZ C. Data augmentation for low-resource neural machine translation[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 2017. doi: 10.18653/v1/P17-2090.

[14] LAMPLE G, CONNEAU A, DENOYER L, *et al*. Unsupervised machine translation using monolingual corpora only[J]. arXiv: 1711.00043, 2017.

[15] ARTETXE M, LABAKA G, and AGIRRE E. An effective approach to unsupervised machine translation[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019: 194–203. doi: 10.18653/v1/P19-1019.

[16] CHENG Yong. Semi-supervised Learning for Neural Machine Translation[M]. CHENG Yong. Joint Training for Neural Machine Translation. Singapore: Springer, 2019: 25–40. doi: 10.1007/978-981-32-9748-7_3.

[17] DUAN Sufeng, ZHAO Hai, ZHANG Dongdong, *et al*. Syntax-aware data augmentation for neural machine translation[J]. arXiv: 2004.14200, 2020.

[18] PENG Wei, HUANG Chongxuan, LI Tianhao, *et al*. Dictionary-based data augmentation for cross-domain neural machine translation[J]. arXiv: 2004.02577, 2020.

[19] SUGIYAMA A and YOSHINAGA N. Data augmentation using back-translation for context-aware neural machine translation[C]. Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), Hong Kong, China, 2019. doi: 10.18653/v1/D19-6504.

[20] FREITAG M, FOSTER G, GRANGIER D, *et al*. Human-paraphrased references improve neural machine translation[J]. arXiv: 2010.10245, 2020.

[21] GANITKEVITCH J, VAN DURME B, and CALLISON-BURCH C. PPDB: The paraphrase database[C]. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, USA, 2013: 758–764.

[22] BERANT J and LIANG P. Semantic parsing via paraphrasing[C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, USA, 2014: 1415–1425. doi: 10.3115/v1/P14-1133.

[23] STIX G. The Elusive goal of machine translation[J]. *Scientific American*, 2006, 294(3): 92–95. doi: 10.1038/scientificamerican0306-92.

[24] GERBER L and YANG Jin. Systran MT dictionary development[C]. Machine Translation: Past, Present, and Future. In: Proceedings of Machine Translation Summit VI, 1997.

[25] NAGAO M, TSUJII J, MITAMURA K, *et al*. A machine translation system from Japanese into English: Another perspective of MT systems[C]. Proceedings of the 8th Conference on Computational Linguistics, Tokyo, Japan, 1980: 414–423. doi: 10.3115/990174.990250.

[26] JOHNSON R, KING M, and DES TOMBE L. Eurotra: A multilingual system under development[J]. *Computational Linguistics*, 1985, 11(2/3): 155–169. doi: 10.5555/1187874.1187880.

[27] WEAVER W. Translation[J]. *Machine Translation of Languages*, 1955, 14: 15–23.

[28] PETER F B, PIETRA S A D, PIETRA V J D, *et al*. The mathematics of statistical machine translation: Parameter estimation[J]. *Computational Linguistics*, 1993, 19(2): 263–311.

[29] KOEHN P, HOANG H, BIRCH A, *et al*. Moses: Open source toolkit for statistical machine translation[C]. Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Prague, Czech Republic, 2007: 177–180. doi: 10.5555/1557769.1557821.

[30] XIAO Tong, ZHU Jingbo, ZHANG Hao, *et al*. NiuTrans: An open source toolkit for phrase-based and syntax-based machine translation[C]. Proceedings of the ACL 2012 System Demonstrations, Jeju Island, Korea, 2012: 19–24. doi: 10.5555/2390470.2390474.

[31] KALCHBRENNER N and BLUNSOM P. Recurrent continuous translation models[C]. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, USA, 2013: 1700–1709.

[32] TRAORE B B, KAMSU-FOGUEM B, and TANGARA F. Deep convolution neural network for image recognition[J]. *Ecological Informatics*, 2018, 48: 257–268. doi: 10.1016/j.ecoinf.2018.10.002.

[33] İRSOY O and CARDIE A. Deep recursive neural networks for compositionality in language[C]. Proceedings of the 27th International Conference on Neural Information Processing

Systems, Montreal, Canada, 2014: 2096–2104. doi: 10.5555/2969033.2969061.

[34] HOCHREITER S and SCHMIDHUBER J. Long short-term memory[J]. *Neural Computation*, 1997, 9(8): 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

[35] CHEN M X, FIRAT O, BAPNA A, *et al*. The best of both worlds: Combining recent advances in neural machine translation[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018. doi: 10.18653/v1/P18-1008.

[36] LUONG T, PHAM H, and MANNING C D. Effective approaches to attention-based neural machine translation[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Porzgal, 2015: 1412–1421. doi: 10.18653/v1/D15-1166.

[37] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6000–6010. doi: 10.5555/3295222.3295349.

[38] CHEN Jing, CHEN Qingcai, LIU Xin, *et al*. The BQ corpus: A large-scale domain-specific Chinese corpus for sentence semantic equivalence identification[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018: 4946–4951. doi: 10.18653/v1/D18-1536.

[39] ZHANG Bowei, SUN Weiwei, WAN Xiaojun, *et al*. PKU paraphrase bank: A sentence-level paraphrase corpus for Chinese[C]. 8th CCF International Conference on Natural Language Processing and Chinese Computing, Dunhuang, China, 2019: 814–826. doi: 10.1007/978-3-030-32233-5_63.

[40] EGONMWAN E and CHALI Y. Transformer and seq2seq model for paraphrase generation[C]. Proceedings of the 3rd Workshop on Neural Generation and Translation, Hong Kong, China, 2019: 249–255. doi: 10.18653/v1/D19-5627.

[41] DEVLIN J, CHANG Minfwei, LEE K, *et al*. 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, Minneapolis, Minnesota, 2019: 4171–4186. doi: 10.18653/v1/N19-1423.

[42] RAFFEL C, SHAZEER N, ROBERTS A, *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. *JMLR*, 2019, 21(140): 1–67.

[43] XUE Linting, CONSTANT N, ROBERTS A, *et al*. mT5: A massively multilingual pre-trained text-to-text transformer[C]. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico, USA, 2020: 483–498. doi: 10.18653/v1/2021.naacl-main.41.

[44] LIU Boxiang and HUANG Liang. NEJM-enzh: A parallel corpus for English-Chinese translation in the biomedical domain[J]. arXiv: 2005.09133, 2020.

[45] CASACUBERTA F and VIDAL E. GIZA++: Training of statistical translation models[J]. *Retrieved October*, 2007, 29: 2019.

[46] REIMERS N and GUREVYCH I. Sentence-BERT: Sentence embeddings using siamese BERT-networks[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), Hong Kong, China, 2019: 3982–3992. doi: 10.18653/v1/D19-1410.

安 波：1986年生，男，副研究员，硕士生导师，研究方向为自然语言处理、机器翻译.

龙从军：1978年生，男，副研究员，硕士生导师，研究方向为民族语言处理、自然语言处理.

责任编辑：陈 倩