智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

探索低资源的迭代式复述生成增强方法

张琳, 刘明童, 张玉洁, 徐金安, 陈钰枫

引用本文:
张琳,刘明童,张玉洁,徐金安,陈钰枫. 探索低资源的迭代式复述生成增强方法[J]. 智能系统学报, 2022, 17(4): 680–687.
ZHANG Lin,LIU Mingtong,ZHANG Yujie,XU Jin’an,CHEN Yufeng. Explore the low-resource iterative paraphrase generation enhancement method[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(4): 680–687.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202106032

DOI: 10.11992/tis.202106032
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20220505.1608.002.html

探索低资源的迭代式复述生成增强方法

张琳，刘明童，张玉洁，徐金安，陈钰枫
（北京交通大学 计算机与信息技术学院，北京 100044）

摘　要：复述生成旨在同一语言内将给定句子转换成语义一致表达不同的句子。目前，基于深度神经网络的复述生成模型的成功依赖于大规模的复述平行语料，当面向新的语言或新的领域时，模型性能急剧下降。面对这一困境，提出低资源的迭代式复述生成增强方法，最大化利用单语语料和小规模复述平行语料迭代式训练复述生成模型并生成复述伪数据，以此增强模型性能。此外，提出了句子流畅性、语义相近性和表达多样性为基准设计的伪数据筛选算法，选取高质量的复述伪数据参与每轮模型的迭代训练。在公开数据集 Quora 上的实验结果表明，提出的方法仅利用 30% 的复述语料在语义和多样性指标上均超过了基线模型，验证了所提方法的有效性。

关键词：低资源；迭代式；复述生成；数据增强；筛选算法；神经网络模型；编码–解码框架；注意力机制
中图分类号：TP18　　**文献标志码**：A　　**文章编号**：1673−4785(2022)04−0680−08

中文引用格式：张琳，刘明童，张玉洁，等. 探索低资源的迭代式复述生成增强方法 [J]. 智能系统学报，2022，17(4)：680–687.
英文引用格式：ZHANG Lin，LIU Mingtong，ZHANG Yujie，et al. Explore the low-resource iterative paraphrase generation enhancement method[J]. CAAI transactions on intelligent systems，2022，17(4)：680–687.

Explore the low-resource iterative paraphrase generation enhancement method

ZHANG Lin, LIU Mingtong, ZHANG Yujie, XU Jin’an, CHEN Yufeng
(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Paraphrase generation aims to convert a given sentence into semantically consistent different sentences within the same language. At present, the success of deep neural network-based paraphrase generation models depends on large-scale paraphrase parallel corpora. When faced with new languages or new domains, the model’s performance drops sharply. We propose a low-resource iterative paraphrase generation enhancement method faced with this dilemma, which maximizes the use of monolingual and small-scale paraphrase parallel corpora to train the paraphrase generation model iteratively and generate paraphrase pseudo data to enhance the model performance. Furthermore, we propose a pseudo data screening algorithm based on fluency, semantic similarity, and expression diversity to select high-quality paraphrased pseudo data in each round of iterative training of the model. Experimental results on Quora, a public dataset, show that our proposed method exceeds the baseline model in semantic and diversity indicators using only 30% of the paraphrase corpus, which verifies the effectiveness of the proposed method.

Keywords: low-resource; iterative; paraphrase generation; data enhancement; screening algorithm; neural networks model; encoder-decoder; attention mechanism

复述生成（paraphrase generation）是自然语言处理中一个长期存在的问题[1]，旨在对给定文本转换为同语言中相同语义但不同表达的句子，例如给定原句“How far is Beijing from Shanghai?”，可以生成其他不同表达的句子（复述句）“What is the distance between Shanghai and Beijing?”。近年来，复述生成技术被广泛应用机器翻译[2]、自

收稿日期：2021−06−23.　**网络出版日期**：2022−05−06.
基金项目：国家自然科学基金项目 (61876198，61976015，61976016).
通信作者：张玉洁. E-mail：yjzhang@bjtu.edu.cn.

动问答[3]、文本摘要[4]、信息检索[5]和对话生成系统[6]等自然语言处理任务中, 进而提高系统的泛化性和鲁棒性。

目前, 深度神经网络的兴起愈来愈提高自然语言处理高资源复述模型的能力, 但仍存在以下两方面的问题。一方面, 现有的复述生成研究大多依赖于大规模的复述平行语料, 而可用的高质量复述资源十分匮乏, 严重阻碍了复述生成模型对新语言或新领域的适用性。因此, 如何在规模有限的训练数据上提高复述生成模型性能是该研究亟待解决的问题。另一方面, 以往基于数据增强的复述生成模型自动生成的复述伪数据伴随大量噪声[7], 直接使用这些数据不利于模型性能的提升。因此, 获取高质量的复述伪数据参与每轮模型的迭代训练是增强模型性能的重要环节。

为了解决以上问题, 本文提出一种低资源下的迭代式复述生成方法来增强模型性能。其基本思路是最大化利用单语语料和已有小规模复述平行语料迭代式训练复述生成模型并生成复述伪数据, 以此构建高质量的复述生成模型。与此同时, 为了在模型每轮训练过程中获取高质量的伪复述, 本文提出句子流畅性、语义相近性和表达多样性的 3 种计算方法, 并以此为基准设计伪数据筛选算法, 渐进地增益模型性能。在公开数据集 Quora 上实验结果表明, 仅利用 30% 的复述数据在语义和多样性评测指标上均超过了基线模型, 验证该方法的有效性。

1　相关研究

早期研究人员致力于使用人工定义的规则[8]和复述模板[9]来进行复述生成, 通常效果不佳且需要大量的专家知识。之后, 基于统计机器翻译的方法[10-11]在一定程度上提升了复述生成模型效果, 但受语料规模和模型学习能力的限制, 复述生成质量仍有待提高。

近年来, 研究者们发现数据增强技术是扩充复述数据和提高模型性能的有效策略[12]。例如, 基于同义词[13]和近义词[14]的词汇替换数据增强方法, 通过替换原句词汇来生成不同的复述, 但就其所能提供的多样性而言, 在句法多样性方面仍受限制。随后, Wieting 等[15]借助机器翻译中的回译方法, 将捷克语-英语平行语料中捷克语句子翻译为英语, 利用译文与原英语句合并构建复述数据, 但机器翻译模型需要大规模的捷克语-英语平行语料来训练。后来, Iyyer 等[16]提出基于对抗训练的语言模型方法显示出了优势, 通过向样本中注入噪声以达到增强模型鲁棒性的效果, 但该方法在确定注入扰动的方向上需要昂贵的计算成本[17]。

此外已有研究表明, 迭代式模型的性能在很大程度上取决于伪数据质量[18], 大规模低质量的伪数据容易导致模型学习效果差, 生成语义偏差且句式单一的句子。为此, 本文提出伪数据筛选算法选取高质量的复述句对参与模型的迭代训练。

2　迭代式复述生成增强方法

本节将详细阐述我们提出的方法, 包括其基本组成部分及模型工作机制。

2.1　方法概述

迭代式复述生成增强方法主要分为 3 个阶段: 初始训练阶段、预训练–微调阶段、迭代式训练阶段。其整体框架如图 1 所示。

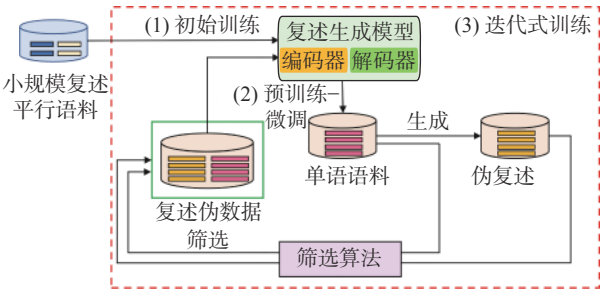

图 1　迭代式复述生成增强方法的整体框架

Fig. 1　Framework of iterative paraphrase generation enhancement method

1) 初始训练阶段。给定小规模复述平行语料 $D_{\text{para}}=\{(X_i,Y_i)\}_{i=1}^{n}$, 其中 $X=\{x_1,x_2,\cdots,x_p\}$ 表示长度为 p 的输入句, $Y=\{y_1,y_2,\cdots,y_q\}$ 表示对应长度为 q 的复述句, 训练初始的复述生成模型 M^0。随后该模型对大规模单语语料 $S=\{s_1,s_2\cdots,s_m\}$ 生成对应的伪复述语料 $T=\{t_1,t_2,\cdots,t_m\}$, 这里 m 表示句子数量。之后, 将伪复述平行数据 (T,S) 作为新的训练数据进行模型迭代再训练。

2) 预训练–微调阶段。在获取成对的复述伪数据后, 我们使用 2.2 节设计的筛选算法抽取出高质量的复述伪数据 D_{filter}, 其余的复述伪数据记为 D_{remain}。其中, 我们仅使用 D_{filter} 来预训练模型, 直到验证集上没有观察到改进, 最后利用小规模复述语料 D_{para} 对模型进行微调。

3) 迭代训练阶段。我们使用筛选后的复述伪数据 D_{filter} 和小规模复述平行语料 D_{para} 训练一轮模型的过程记为第一次迭代。接下来, 我们采用微调后的模型对余下单语语料 $D_{\text{remain}}(S)$ 生成伪复述, 并与上一轮筛选后保留的数据进行合并, 作

为第二次迭代。经过 N 轮迭代训练后，得到最终复述生成模型 $\hat{M}$。

综上，我们设计迭代式复述生成增强的方法，算法步骤如下所示。

算法 迭代式复述生成增强方法

输入 复述平行语料 $D_{\text{para}}=\{(X_i,Y_i)\}_{i=1}^{n}$，单语语料 $\{S_i\}_{i=1}^{m}$，迭代轮数 N；

输出 训练完成的复述生成模型 $\hat{M}$。

1) 使用 D_{para} 训练初始化复述生成模型 M^0；

2) 使用初始单语语料 $\{S_i\}_{i=1}^{m}$ 作为第 0 轮单语语料；

3) 当迭代轮数 $r\in\{0,1,\cdots,N\}$：获取伪复述 $\{T_i^r\}_{i=1}^{m}$，利用筛选算法获取 D_{filter}^r 和 D_{remain}^r，利用 D_{filter}^r 训练下一轮模型 M^{r+1}，对模型 M^{r+1} 进行微调，从 D_{remain}^r 中获取单语语料 $\{S_i^r\}_{i=1}^{l}$，利用 $\{S_i^r\}_{i=1}^{l}$ 替换为单语语料 $\{S_i^{r+1}\}_{i=1}^{m}$；

4) 输出最终迭代 N 轮的复述生成模型 $\hat{M}$。

值得注意的是，模型自动生成的复述伪数据含有大量的噪声数据在迭代训练过程将被不断积累和放大，进而影响模型效果。对此，我们将生成句 T 作为复述伪数据的原句，原句 S 作为复述伪数据的参考句合并配对，使得模型在解码过程中更多关注原句的语义信息，提升复述生成效果。

2.2 筛选算法

为了避免模型性能受到噪声数据影响。我们联合句子流畅性、语义相近性和表达多样性 3 种计算方法选取高质量的复述句对，逐步受益于复述生成模型，满足最终的筛选目标。

1) 句子流畅性 (fluency)

采用典型的长短期记忆神经网络 (long short-term memory, LSTM) 预训练语言模型计算困惑度 (perplexity, PPL) 得分，评估生成句表达自然流畅程度[19]。该方法主要根据词的概率来估计句子的自然流畅程度，困惑度的具体计算过程为

$$\text{PPL}(T)=P(w_1w_2\cdots w_N)^{-\frac{1}{N}}=\sqrt[N]{\prod_{i=1}^{N}\frac{1}{P(w_i|w_1w_2\cdots w_{i-1})}}$$

式中：T 为当前句子，$T=\{w_1,w_2,\cdots,w_N\}$；N 表示该句子长度；$P(w_i)$ 是第 i 个词的概率，第 1 个词是 $P(w_1|w_0)$；w_0 为句子的起始。

困惑度越小，说明该句子更加自然流畅，故我们保留困惑度较低的复述句。

2) 语义相近性 (semantics)

其次，由于语义越相近的句对在向量空间中越接近，可以通过余弦相似度，衡量两个句子的语义差异。为此，我们采用 Sentence-BERT[20] 方法（SBERT）获取动态词向量，继而计算句对的余弦相似度。具体地，首先将复述伪数据输入到两个共享的 BERT 模型，然后将输出的字向量进行求平均值操作，将均值向量作为整个句子的句向量。之后，通过计算复述句向量和参考句向量之间的余弦相似度，进行语义相近性的评估，余弦值越大，语义越相近，其具体计算方法如式 (1) 所示。

$$\text{Cosine}(\boldsymbol{V}_X,\boldsymbol{V}_Y)=\frac{\boldsymbol{V}_X\cdot\boldsymbol{V}_Y}{\|\boldsymbol{V}_X\|\cdot\|\boldsymbol{V}_Y\|}\tag{1}$$

式中：$\boldsymbol{V}_X$ 表示复述伪数据中原句的句向量；$\boldsymbol{V}_Y$ 表示复述伪数据中生成句的句向量。

3) 表达多样性 (diversity)

最后，基于句子流畅性和语义相近性筛选方法后的复述伪数据，进一步获取句式丰富的复述句对。首先，需要去除原句和生成句相同的句子。之后，利用 BLEU 指标[21] 比较复述句和原始句的 N-gram 重合程度，重合程度越低，则两个句子结构具有明显差别，多样性越好。因此，在余下的句子中选取 BLEU 得分低的句子加入训练。

综上，筛选算法的具体执行过程如下：首先，进行句子流畅性筛选，保留小于困惑度阈值的伪复述句及对应输入句。其次，进行语义相近性筛选，保留大于余弦相似度阈值的复述伪数据。最后，进行表达多样性筛选，保留小于 BLEU 阈值的伪复述句对。最终，通过上述筛选算法得到高质量的复述伪数据。筛选方法的阈值设置如表 1 所示。

表 1 筛选方法阈值设置

Table 1 Threshold settings of the filter methods

筛选方法	筛选指标	阈值
Fluency	LSTM-PPL	300
Semantics	Cosine	0.6
Diversity	BLEU	0.6

2.3 模型介绍

采用融合注意力机制[22]、复制机制[23] 和覆盖机制[24] 的编码–解码复述生成模型，其中复制机制解决模型解码时无法生成未登录词、低频词的问题，覆盖机制解决模型解码时生成的词汇重复问题，以此提高模型生成复述句的质量。

复述生成模型的编码器采用双向长短时记忆循环神经网络 Bi-LSTM，解码器采用单向 LSTM。给定复述语料 $D_{\text{para}}=\{(X_i,Y_i)\}_{i=1}^{n}$，模型编码时，编码器首先顺序读取输入句 $X=\{x_1,x_2,\cdots,x_p\}$ 每个词进行语义编码。解码时，注意力机制动态计算当前 t 时刻编码器的隐状态 $\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_p\}$ 与解码状态 $\boldsymbol{s}_t$ 的对齐分数 e_i^t。然后将 softmax 激活函数应用

于对齐分数的计算, 获得归一化 t 时刻所有词的注意力权重$\boldsymbol{a}^t$, 注意力权重越高的词, 对目标词预测的影响越大, 计算公式如式 (2) 和式 (3) 所示。

$$e_i^t = \boldsymbol{v}^{\mathrm{T}} \tanh(\boldsymbol{W}_h \boldsymbol{h}_i + \boldsymbol{W}_s \boldsymbol{s}_t + \boldsymbol{b}_{\mathrm{attn}}) \tag{2}$$

$$\boldsymbol{a}^t = \mathrm{softmax}(\boldsymbol{e}^t) \tag{3}$$

其中$\boldsymbol{v}, \boldsymbol{W}_h, \boldsymbol{W}_s, \boldsymbol{b}_{\mathrm{attn}}$表示可学习的参数。

随后, 整合编码器隐藏层状态$\boldsymbol{h}_i$和输入句中每个词语注意力权重a_i^t, 得到 t 时刻输入句的上下文语义向量$\boldsymbol{c}_t$, 计算公式为

$$\boldsymbol{c}_t = \sum_{i=1}^{p} a_i^t \boldsymbol{h}_i$$

为解决模型无法生成未登录词的问题, 我们在注意力机制的基础上引入复制机制。融合复制机制模型预测词语 w 的概率$P(w)$的计算公式为

$$P(w) = P_{\mathrm{gen}} P_{\mathrm{vocab}}(w) + (1 - P_{\mathrm{gen}}) P_{\mathrm{copy}}(w)$$

其中P_{gen}表示当前词选择生成模式进行预测的概率, 取值范围为 [0,1], 计算公式如式 (4) 所示; $P_{\mathrm{vocab}}(w)$表示对词汇表中 w 的生成概率, 计算公式如式 (5) 所示; $1 - P_{\mathrm{gen}}$表示当前词选择复制模式进行预测的概率, 取值范围同为 [0,1], $P_{\mathrm{copy}}(w)$表示原句中词w的复制概率, 计算公式如式 (6) 所示。当词w为未登录词时, 即选择复制模式获取目标词, 令$P_{\mathrm{vocab}}(w) = 0$; 当词$w$不是未登录词时, 令$P_{\mathrm{copy}}(w) = 0$。

$$P_{\mathrm{gen}} = \sigma\left(\boldsymbol{w}_{h^*}^{\mathrm{T}} \boldsymbol{c}_t + \boldsymbol{w}_s^{\mathrm{T}} \boldsymbol{s}_t + \boldsymbol{w}_{y_{t-1}}^{\mathrm{T}} \boldsymbol{y}_{t-1} + \boldsymbol{b}_{\mathrm{gen}}\right) \tag{4}$$

$$P_{\mathrm{vocab}}(w) = f(\boldsymbol{U}(\tanh(\boldsymbol{V}[\boldsymbol{y}_{t-1}, \boldsymbol{s}_t, \boldsymbol{c}_t] + \boldsymbol{b}_v)) + \boldsymbol{b}_u) \tag{5}$$

$$P_{\mathrm{copy}}(w) = \sum_{i:w_i=w} a_i^t \tag{6}$$

式中: σ表示 sigmoid 函数; f表示 softmax 函数; $\boldsymbol{c}_t$表示 t 时刻的上下文语义向量; $\boldsymbol{y}_{t-1}$表示上一个时刻 $t-1$ 生成词语的向量表示; $\boldsymbol{s}_t$表示 t 时刻解码器隐状态; $\boldsymbol{w}_{h^*}, \boldsymbol{w}_s, \boldsymbol{w}_{y_{t-1}}, \boldsymbol{b}_{\mathrm{gen}}, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{b}_u, \boldsymbol{b}_v$表示可学习参数。

为解决生成的词汇重复问题, 我们引入了覆盖机制。具体地, 通过添加覆盖向量$\boldsymbol{g}^t$记录历史决策信息, 使得注意力决策避免重复关注同一词的位置, 让原句中未解码的词得到关注, 其计算过程为

$$\boldsymbol{g}^t = \sum_{t'=0}^{t-1} a^{t'}$$

其中$a^{t'}$表示时间步t'时刻原句中各个词语的注意力对齐权重, 其计算公式与式 (3) 保持一致。在注意力机制的基础上添加覆盖向量的对齐分数计算公式为

$$e_i^t = \boldsymbol{v}^{\mathrm{T}} \tanh(\boldsymbol{W}_h \boldsymbol{h}_i + \boldsymbol{W}_s \boldsymbol{s}_t + \boldsymbol{w}_g \boldsymbol{g}_i^t + \boldsymbol{b}_{\mathrm{attn}})$$

其中$\boldsymbol{v}$、$\boldsymbol{w}_g$、$\boldsymbol{W}_h$、$\boldsymbol{W}_s$、$\boldsymbol{b}_{\mathrm{attn}}$表示可学习参数。

最后, 我们基于注意力机制复述生成模型融合复制机制和覆盖机制目标函数为

$$\mathrm{Loss}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \left(-\ln P(w_t) + \lambda \sum_i \min(a_i^t, g_i^t)\right)$$

式中: θ表示模型参数, $-\ln P(w_t)$为生成词语的损失函数, 覆盖机制的损失函数为括号内的第 2 项, 最小化重复关注同一个词的注意力决策, a_i^t和g_i^t表示在t时刻对词x_i的注意力决策。当根据上下文历史决策对词x_i的注意力权重变大, 则当前时刻t对该单词注意力变小; 反之, 对该单词注意力权重变大。超参数λ用来对覆盖损失进行加权平衡, 有效规避词汇重复生成的问题。

3　实验设计和实验设置

3.1　实验数据与对比模型

我们在英文 Quora 问句复述语料上进行了评测实验, 训练集、验证集和测试集分别为 3×10^4、3000 和 3×10^4 对句子。

本文采用的单语语料来自 Quora 和 WikiAnswers 两个数据集, 共选取 5×10^5 对问题句。其中, 从 WikiAnwsers 数据集中的 2.5×10^6 对不同问题句中选取 1.62×10^5 对, 从 Quora 数据集的非复述句对中, 通过拆分、去重等方法选取 3.38×10^5 对问题句。另外, 为了验证单语数据领域对模型性能影响, 我们从 WikiAnwsers 语料中随机获取了 3×10^4 对复述句对作为域外测试集。

我们与近年来使用 Quora 数据集中 1×10^5 数据规模的复述生成模型进行了比较, 这些模型主要包括:

1) Residual LSTM[25] 重构 Seq2Seq 编码–解码网络, 在 LSTM 网络层间加入残差网络来训练更深的网络模型结构生成复述句;

2) DiPS[13] 模型在解码阶段采用子模块化函数生成不同的复述句;

3) DNPG[26] 将复述句分解为句子级和短语级, 并对每个级别使用单独的 Transformer;

4) Seq2Seq-att[27] 基于注意力机制的 Seq2Seq 模型;

5) SCSVED[27] 提出一种提取目标句的句法变量和输入句语义变量的变分编码–解码模型。

3.2　实验设置

实验预处理部分, 本文首先过滤长度大于 40 个词的句子, 然后选择词频高的前 1.5×10^4 个词作为词典, 使用<UNK>符号表示未登录词。

模型实验参数具体设置如下: 编码器和解码器中 LSTM 的隐藏层大小设置为 256 维, 词向量维度设置为 128 维。设置 Batch 大小为 64, 采用 Adam

优化算法训练模型，Beam Size 设置为 4，设置训练阶段学习率为 1×10^{-3}，微调阶段学习率为 5×10^{-4}。

我们采用 ROUGE-1(R-1)、ROUGE-2(R-2)、ROUGE-L(R-L)[28]、BLEU、METEOR[29] 和利用 Stanford-corenlp 工具包计算句法编辑距离 (Syn-TED) 自动评测指标进行评测。

4　评测实验和结果分析

4.1　评测结果

本实验以大量复述数据 (1×10^5 个句对) 训练的复述生成模型作为基线模型，实验结果如表 2 所示。

表 2　迭代式复述生成增强方法评测结果

Table 2　Evaluation results of iterative paraphrase generation enhancement method

模型	迭代次数	BLEU	R-1	R-2	R-L	METEOR	Syn-TED
基线(1×10^5)	N=0	28.74	59.18	35.12	61.39	34.36	11.32
本文模型(3×10^4)	N=0	24.55	55.30	31.04	57.99	31.18	11.12
本文模型(3×10^4)	N=1	28.78	61.59	37.58	63.84	35.13	10.48
本文模型(3×10^4)	N=2	30.27	62.99	38.22	64.71	36.20	11.01
本文模型(3×10^4)	N=3	**32.03**	**63.41**	**39.12**	**65.77**	**37.06**	**11.43**

实验结果表明，在各项指标上，使用 3×10^4 个句对 (30%) 的复述数据迭代训练的模型效果明显优于利用 1×10^5 个句对 (100%) 复述数据训练的基线模型。

相较于基线模型，迭代 1 轮的模型性能在各指标上已具有显著的提升，R-1 和 R-2 指标分别提升了 2.41 和 2.46，但在 BLEU 指标上提升效果不明显。随着迭代次数增加，当模型迭代 3 轮时，单语数据利用率约达到 70%，复述生成模型性能在所有评测指标上均有大幅度的提升，BLEU、R-1、R-2、R-L、METEOR 和 Syn-TED 分别提升了 3.29、4.23、4.00、4.38、2.70 和 0.11，充分验证了迭代式复述生成增强方法对改进复述生成质量的有效性，生成句法结构更为丰富的句子。究其原因在于微调模型辅助增加了训练侧的数据语义信息和多样性，有利于模型发挥更为准确的预测能力。

进一步，表 3 是我们与近年来经典的复述生成模型比较的实验结果，从中可以得出以下结论：

表 3　复述生成模型评测结果

Table 3　Evaluation results of paraphrase generation models

模型	BLEU	R-1	R-2	R-L	METEOR	Syn-TED
Residual LSTM	24.56	59.69	32.71	—	29.39	—
DiPS	25.37	59.79	31.77	—	29.28	—
DNPG	25.03	**63.73**	37.75	—	—	—
Seq2Seq-att	24.02	57.56	31.47	57.16	—	—
SCSVED	26.04	58.76	33.77	58.93	—	—
本文模型(3×10^4, N=0)	24.55	55.30	31.04	57.99	31.18	11.12
本文模型(3×10^4, N=3)	**32.03**	63.41	**39.12**	**65.77**	**37.06**	**11.43**

1) 从表中第 6 行的数据可以看出，本方法在 BLEU、R-1、R-2 指标数值虽然相较于最好模型有一定差距，但是结果在正常的数值范围内，说明其生成的复述句虽然语义一致性较低，但是能够生成没有句法和语法错误的正确句子，反映了本文设计的基于多机制融合的复述生成模型能够满足基本的复述要求，为后续迭代训练奠定基础。

2) 从表中第 7 行的数据可以看出，随着迭代轮数的增加，复述生成模型性能有所提升。相较于性能较好的模型 SCSVED 而言，本方法在 BLEU、R-2 和 R-L 指标上分别提升 5.99、5.35 和 6.84。与 DNPG 模型相比，在 R-2 指标上，本方法提升了 1.37。以上实验结果表明本方法生成的句子能够与输入句保持较高的语义相近性，验证了所提方法的有效性。

4.2　消融实验

为了更好的探究不同筛选计算方法对复述生成模型性能的影响，本文在 3×10^4 复述数据训

练获得的初始复述生成模型M^0上进行迭代一轮消融实验, 没有进行预训练–微调阶段。此外, 为验证筛选算法的有效性并非通过增加数据规模对模型影响, 我们随机选取 (*) 与筛选算法同规模的复述伪数据进行迭代再训练, 评测结果如表 4 所示, 其中*表示随机选取。

表 4　消融实验

Table 4　Ablation experiments

模型	BLEU	R-1	R-2	R-L	METEOR	Syn-TED
本文 (*N*=0)	24.55	55.30	31.04	57.99	31.18	11.12
+All (*N*=1)	24.46	55.57	31.12	58.32	31.36	11.27
+Fluency (*N*=1)	25.46	**56.65**	**32.40**	**59.36**	31.68	11.14
+Fluency(*N*=1, *)	25.14	56.07	31.65	58.79	31.45	11.10
+Semantics (*N*=1)	**25.47**	56.44	32.16	59.10	**31.76**	11.24
+Semantics (*N*=1, *)	24.99	55.95	31.68	58.67	31.43	11.15
+Diversity (*N*=1)	24.49	55.53	31.22	58.35	31.07	**11.30**
+Diversity(*N*=1, *)	25.21	56.23	32.04	58.30	31.48	11.05
No	24.86	55.92	31.72	58.71	31.29	11.20

当添加 Fluency 筛选方法后, R-1、R-2 和 R-L 相对于基线模型分别提高 1.35、1.36 和 1.37。当添加 Semantics 筛选方法后, 在 BLEU 和 METEOR 指标上均超过基线模型。当添加 Diversity 筛选方法后, 在 Syn-TED 指标上超过基线模型, 直观反映生成句式变化的多样性。与没有加入筛选方法(No)训练的模型相比, 本方法在多样性指标 Syn-TED 上表现较好。最后, 当添加所有筛选方法后, 各项评测指标相较于基线模型性能有所提升。

另外, 由于本方法采用的单语数据是来自 Quora 和 WikiAnswers(记为 Wiki) 两个不同领域的数据集。对此, 表 5 比较分析了这些单语数据领域迭代一轮后模型的领域适应能力的影响。

以 3×10^4 句对的 Quora 数据训练初始复述生成模型作为基线模型。与加入 Quora 单语数据结果相比, 加入 Wiki 单语数据训练的模型在各项评测指标上均有显著提升, 表明我们的方法具有较强的领域适应能力, 并且在域外测试集上显现出较好的泛化性, 使得模型跨领域能力有所增强。

表 5　单语数据领域评测结果

Table 5　Monolingual data domain evaluation results

数据集	BLEU	R-1	R-2	R-L	METEOR
Train	4.67	27.72	7.51	30.73	16.36
+Quora	4.53	28.25	7.69	31.32	16.67
+Wiki	**5.56**	**31.93**	**9.00**	**34.99**	**19.30**

4.3　实例分析

此外, 我们还对本方法所生成的实例进行了分析, 具体如表 6 所示, 粗体为生成的最佳复述句。

表 6　生成的复述实例

Table 6　Examples of generated paraphrases

输入句	What happens to a question on quora if it is marked as needing further improvement ?
复述参考句	Do quora users still see questions that are marked as needing improvement ?
N=0	Why my answer is marked as needing imrovement ? Why my question was marked as needing imrovement ? What happens to a question on quora if it was marked as needing further improvement ? **What happens to a question on quora if it is marked as needing further improvement ?**
N=3	What is the solution for quora marking your every question on quora ? What happens to a question on quora if it is marked as needing improvement ? What happens to a question on quora if it is marked as needing further improvement ? **Why does quora mark my questions as needing improvement ?**

从表中可以看出，初始复述生成模型 (N=0) 生成的复述句“What happens to a question on quora if it is marked as needing further improvement ?”和原句相比，在句法和词汇上没有变化。然而，模型经过 3 轮迭代训练后，所生成的复述句“Why does quora mark my questions as needing improvement ?”在表达结构上由“What happens to”转换为“Why does”，以及根据句法信息将动词的被动形式“marked”转为主动形式“mark”。

由此可见，本方法所生成的复述句在句子表达结构和词性转换上均有较大的变化，进一步表明本方法可以生成自然流畅、语义相近且表达丰富的复述句。

5　结束语

本文提出了一种探索低资源的迭代式复述生成增强方法，主要思想是最大化利用单语语料和小规模复述平行语料，迭代式训练复述生成模型并生成复述伪数据，以此扩充复述资源，进而增强模型性能。为了去除生成的复述伪数据中存在的噪声，本文设计了伪数据筛选算法，以获取高质量的复述伪数据参与每轮迭代的模型训练，从而渐进地提升模型性能。在公开数据集 Quora 上的实验结果表明，本方法仅使用 30% 的复述数据集在语义和多样性指标上均超过基线模型，从而验证了所提方法的有效性。本方法的提出有利于在低资源环境下更容易构建出高质量的复述生成模型，减少对人工标注复述数据的依赖。在今后的工作中，如何利用丰富的外部复述知识进一步提升模型性能，将作为我们未来主要的研究方向。

参考文献：

[1] BARZILAY R, MCKEOWN K R. Extracting paraphrases from a parallel corpus[C]//ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. New York: ACM, 2001: 50−57.

[2] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1724−1734.

[3] YIN Jun, JIANG Xin, LU Zhengdong, et al. Neural generative question answering[C]//IJCAI'16: Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York: ACM, 2016: 2972−2978.

[4] ZHANG Chi, SAH S, NGUYEN T, et al. Semantic sentence embeddings for paraphrasing and text summarization[C]//2017 IEEE Global Conference on Signal and Information Processing. Montreal: IEEE, 2017: 705−709.

[5] GUPTA R, ORĂSAN C, ZAMPIERI M, et al. Improving translation memory matching and retrieval using paraphrases[J]. Machine translation, 2016, 30(1/2): 19–40.

[6] SAHAY S, OKUR E, HAKIM N, et al. Semi-supervised interactive intent labeling[C]//Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances. Stroudsburg: Association for Computational Linguistics, 2021: 31−40.

[7] WEI J, ZOU Kai. 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. Stroudsburg: Association for Computational Linguistics, 2019: 6382−6388.

[8] SHAKEEL M H, KARIM A, KHAN I. A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts[J]. Information processing & management, 2020, 57(3): 102204.

[9] MCKEOWN K. Paraphrasing questions using given and new information[J]. American journal of computational linguistics, 1983, 9(1): 1–10.

[10] 刘圆圆, 王忠建. 基于模板的对几种特殊结构句子的语句改写 [J]. 现代电子技术, 2009, 32(3): 157–159,166.
LIU Yuanyuan, WANG Zhongjian. Paraphrasing of several special sentence structure based on templates[J]. Modern electronics technique, 2009, 32(3): 157–159, 166.

[11] 胡金铭, 史晓东, 苏劲松, 等. 引入复述技术的统计机器翻译研究综述 [J]. 智能系统学报, 2013, 8(3): 199–207.
HU Jinming, SHI Xiaodong, SU Jinsong, et al. A survey of statistical machine translation using paraphrasing technology[J]. CAAI transactions on intelligent systems, 2013, 8(3): 199–207.

[12] SERAJ R. Paraphrases for statistical machine translation [D]. Burnaby Campus: Simon Fraser University, 2015.

[13] KUMAR A, BHATTAMISHRA S, BHANDARI M, et al. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 3609−3619.

[14] KOBAYASHI S. Contextual augmentation: data augmentation by words with paradigmatic relations[EB/OL].

New York: arXiv, 2018. (2018−05−16)[2021−06−19]. https://arxiv.org/abs/1805.06201.

[15] WIETING J, GIMPEL K. ParaNMT-50M: pushing the limits of paraphrastic sentence embeddings with millions of machine translations[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 451−462.

[16] IYYER M, WIETING J, GIMPEL K, et al. Adversarial example generation with syntactically controlled paraphrase networks[EB/OL]. New York: arXiv, 2018. (2018−04−17)[2021−06−19].https://arxiv.org/abs/1804.06059.

[17] CHENG Yong, JIANG Lu, MACHEREY W, et al. AdvAug: robust adversarial augmentation for neural machine translation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 5961−5970.

[18] ABDULMUMIN I, GALADANCI B S, ALIYU G. Tagless back-translation[J]. Machine translation, 2021, 35(4): 519–549.

[19] 薛佳奇, 杨凡. 基于交叉熵与困惑度的 LDA-SVM 主题研究 [J]. 智能计算机与应用, 2019, 9(4): 45–50.
XUE Jiaqi, YANG Fan. Research on LDA-SVM subject based on cross entropy and perplexity[J]. Intelligent computer and applications, 2019, 9(4): 45–50.

[20] REIMERS N, 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. Stroudsburg: Association for Computational Linguistics, 2019: 671−688.

[21] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]//ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. New York: ACM, 2002: 311−318.

[22] MENG F, LU Z, LI H, et al. Interactive attention for neural machine translation[C]//Proceedings of the 26th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016: 2174−2185.

[23] GU Jiatao, LU Zhengdong, LI Hang, et al. Incorporating copying mechanism in sequence-to-sequence learning[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016: 1631−1640.

[24] TU Zhaopeng, LU Zhengdong, LIU Yang, et al. Modeling coverage for neural machine translation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016: 76−85.

[25] PRAKASH A, HASAN S, LEE K, et al. Neural paraphrase generation with stacked residual LSTM networks[C]//Proceedings of the 26th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016: 2923−2934.

[26] LI Zichao, JIANG Xin, SHANG Lifeng, et al. Decomposable neural paraphrase generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 3403−3414.

[27] CHEN Wenqing, TIAN Jidong, XIAO Liqiang, et al. A semantically consistent and syntactically variational encoder-decoder framework for paraphrase generation[C]//Proceedings of the 28th International Conference on Computational Linguistics. Stroudsburg: International Committee on Computational Linguistics, 2020: 1186−1198.

[28] LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]//In Proceedings of Workshop on Text Summarization Branches Out. Barcelona: ACL, 2004: 74−81.

[29] LAVIE A, AGARWAL A. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments[C]//StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation. New York: ACM, 2007: 228−231.

作者简介:

张琳, 硕士研究生, 主要研究方向为复述生成和机器翻译。

刘明童, 博士, 主要研究方向为依存句法分析、句子匹配、复述生成、机器翻译和自然语言处理。

张玉洁, 教授, 主要研究方向为机器翻译、多语言信息处理、句法分析和自然语言处理。发表学术论文 30 余篇。