联合自编码任务的多机制融合复述生成模型

刘明童[1] 张玉洁[1,†] 张姝[2] 孟遥[3] 徐金安[1] 陈钰枫[1]

1. 北京交通大学计算机与信息技术学院，北京 100044； 2. 富士通研究开发中心，北京 100027；3. 联想研究院人工智能实验室，北京 100085

† 通讯作者，E-mail: yjzhang@bjtu.edu.cn

摘要 复述生成指同一种语言内意思相同的不同表达之间的转换。目前，基于神经网络编码-解码框架的复述生成模型成为主流方法，但是仍然存在两方面的问题。一方面，生成的复述句中存在实体词不准确、未登录词和词汇重复生成的问题。对此，提出在解码过程中融合注意力机制、复制机制和覆盖机制的多机制复述生成模型，利用复制机制从原句复制词语以解决实体词和未登录词生成问题；利用覆盖机制建模学习注意力机制的历史决策信息以规避词汇重复生成问题。另一方面，复述平行语料的有限规模限制了已有模型中编码器的语义学习能力，成为性能提升的瓶颈。对此，基于多任务学习框架，提出在复述生成任务中联合自编码任务，两个任务共享一个编码器，同时利用平行复述语料和原句子数据，共同增强复述生成编码器的语义学习能力。在 Quora 复述数据集上的实验结果表明，提出的多机制融合复述生成模型可以有效解决实体词、未登录词和词汇重复生成问题；联合自编码任务的复述生成模型可以增强复述编码器的语义学习能力，进一步提高生成复述句的质量。

关键词 复述生成；自编码；多任务学习；多机制融合；注意力机制；复制机制；覆盖机制

中图分类号 TN914

A Multi-Mechanism Fused Paraphrase Generation Model with Joint Auto-Encoding Learning

LIU Mingtong[1]，ZHANG Yujie[1†]，ZHANG Shu[2]， MENG Yao[3]， XU Jinan[1]， CHEN Yufeng[1]

1. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044; 2. Fujitsu Research and Development Center, Beijing 100027; 3. Lenovo Research, AI Lab, Beijing 100085

†Corresponding Author, E-mail: yjzhang@bjtu.edu.cn

Abstract Paraphrase generation refers to generate different expressions with the same meaning in the same language. Neural network encoder-decoder framework has became the popular method for paraphrase generation, but there are still two problems. On the one hand, there are such problems as inaccurate entity words, unknown words and word repetition in the generated paraphrase sentences. To solve these problems, this paper proposed a mul-ti-mechanism fused paraphrase generation model that improves the decoder. The copy mechanism was used to copy words form input sentence for improving the generation of entity and unknown words. The coverage mechanism was used to model historical attention information to avoid word repetition. On the other hand, the limited-scale parallel paraphrase corpus limits the learning ability of the encoder. This paper proposed to jointly learn auto-encoding task, which shares one encoder with paraphrase generation task. The joint auto-encoding task enhances the learning ability of the encoder. Experimental results on Quora paraphrase dataset show that the multi-mechanism fused paraphrase generation model with joint auto-encoding task can effectively improve the performance of paraphrase generation.

Key words paraphrase generation; auto-encoding; multi-task learning; multi-mechanism fusion; attention mechanism; copy mechanism; coverage mechanism

复述生成是自然语言处理的关键技术之一，广泛应用于机器翻译[1]、文本摘要[2]和自动问答[3]等任务中。

国家自然科学基金(61876198,61370130,61473294)，中央高校基本科研业务费专项资金(2018YJS025,2015JBM033)，科学技术部国际科技合作计划（K11F100010）资助

收稿日期: 2010-12-03; 修回日期: 2010-04-12；网络出版时间：

网络出版地址：

在机器翻译和文本摘要任务中，利用复述可以生成多个参考译文和文摘，使得自动评测系统的性能得到有效提升。在自动问答中，利用复述可以扩展检索文本，提高检索系统性能。主流的复述生成框架主要包括基于统计机器翻译和基于神经网络的两种模型。近年来，基于神经网络的复述生成方法成为重要研究方向之一。本文主要针对基于神经网络编码-解码框架的复述生成方法展开研究。

早期复述生成主要采用基于规则[4]、基于复述模板[5]和基于词典[6]的方法，这些方法依赖于人工定义的规则，通常效果不佳且局限在特定领域；之后提出的基于统计机器翻译的方法[7,8]在一定程度上提高了复述生成的质量，但受模型学习能力的限制，复述生成质量仍有待提高。

近年来，神经网络模型在特征学习方面展示出了明显优势，研究者[9,10]利用神经网络编码-解码框架建模复述生成。在该框架中，编码器用于获取原句子的语义表示，解码器依赖于原句子的语义表示生成目标复述句，学习语义表示成为了复述生成的研究重点。文献[10]在基于编码-解码框架的模型中加入注意力机制，改进长句生成质量。但该模型生成的复述句子存在实体词不准确、未登录词和词汇重复问题。此外，由于复述平行语料资源十分有限，已有方法对复述语料的利用不够充分，在语义学习方面仍有改进的空间。

为了解决以上问题，本文提出多机制融合的复述生成模型和联合自编码任务的复述生成模型。在多机制融合的模型中，我们在编码-解码框架中的解码端融合注意力机制、复制机制和覆盖机制。其中，复制机制从原句中复制词汇生成在目标句中，以解决实体词和未登录词（低频词）的生成问题；覆盖机制建模注意力机制的历史决策信息，约束当前注意力机制的决策，改善词汇重复生成的问题。同时，为了充分利用有限的复述语料，提高模型的语义学习能力，本文基于多任务学习框架，提出在复述生成任务中联合自编码任务，两个任务共享同一个编码器。在复述生成任务中，模型利用平行复述语料学习编码器，从而生成对应复述句；在自编码任务中，模型利用原句子学习编码器，并解码生成原句子本身。

本文剩余部分组织如下：第一节介绍多机制融合的复述生成模型；第二节详细介绍联合自编码任务的复述生成模型；第三节介绍实验设计和模型实施细节；第四节介绍评测实验和结果分析；第五节介绍复述生成相关研究工作；最后对本文工作进行总结。

1 多机制融合的复述生成模型

复述生成任务是同语言内输入句子到输出句子间的转换问题，即给定长度为 n 的输入原句 $X=[x_1,\ldots,x_n]$，生成长度为 m 的复述句 $Y=[y_1,\ldots,y_m]$，要求生成句 Y 与输入原句 X 语义一致，且在表现形式上有所差别。我们采用基于深度神经网络的编码-解码框架[11,12]构建复述生成模型，在解码时，采用注意力机制学习当前生成词最相关的上下文信息。引入复制机制从原句中复制词汇，提高命名实体、未登录词和低频词的生成质量；并融合覆盖机制，对注意力机制的历史决策信息建模，减少词汇重复生成问题。

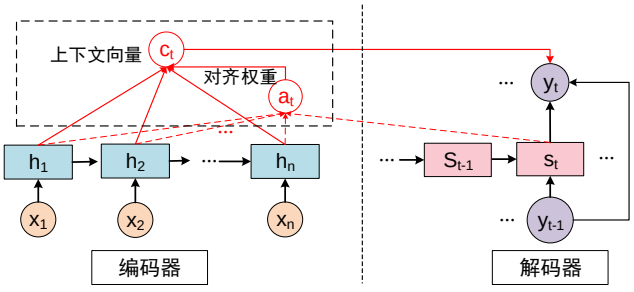

图 1 基于注意力机制的复述生成模型

Fig. 1 Paraphrase generation model based on attention mechanism

1.1 基于注意力机制的复述生成模型

基于注意力机制的复述生成模型如图 1 所示，我们采用循环神经网络 LSTM 作为编码器和解码器。其中，编码器依次读取原句 $X=[x_1,\ldots,x_n]$ 中的每个词生成原句语义表示，解码器根据原句语义表示逐词生成目标复述句。我们在解码过程中引入注意力机制[11,12]，根据当前解码器状态 s_t 和编码器状态 $H=[h_1,\ldots h_n]$ 动态计算上下文语义向量 c_t，其中注意力权重越大的词，在当前位置被给予越多的关注。

$$e_i^t = V^T \tanh\left(W_h h_i + W_s s_t + b_{attn}\right) \tag{1}$$

$$a^t = softmax\left(e^t\right) \tag{2}$$

$$c_t = \sum_{i=1}^{n} a_i^t h_i \tag{3}$$

其中 e_i^t 表示当前解码器状态 s_t 和编码器状态 h_i 之间的对齐得分，a_i^t 表示归一化后词的注意力权重。在解码时，解码器利用 softmax 分类器预测下一个词的生成概率：

$$P_{vocab}\left(w\right) = softmax\left(U\left(\tanh\left(V\left[y_{t-1}, s_t, c_t\right] + b_v\right)\right) + b_u\right) \tag{4}$$

其中 $s_t = LSTM\left(y_{t-1}, s_{t-1}\right)$，$s_t$ 表示当前解码器隐藏层状态，y_{t-1} 表示上一个已生成词的词向量，U、V、b_v、b_u 表示可学习参数。

给定复述句对（X,Y），基于注意力机制的复述生成模型的训练目标，即最大化如下对数似然函数：

$$Loss_{seq2seq}\left(\theta\right) = \sum_{t=1}^{m} \log p_\theta\left(y_t | Y_{1:t-1}, X\right) \tag{5}$$

1.2 融合复制机制和覆盖机制的复述生成模型

本文在基于注意力机制的复述生成模型上，融合复制机制[13-14]和覆盖机制[15]，如图 2 所示。我们对原句中的未登录词采用不同于集内词的解码方式。在解码时，我们设置含有生成模式和复制模式两种模式的解码器。生成模式对集内词进行预测，概率计算公式同公式（4）。复制模式利用复制机制，计算原句中未登录词的复制概率，解决未登录词无法生成的问题。如图 2 所示，在解码 t=3 时刻，词 x_3 在集内词表中不存在，解码器选择复制模式，利用复制机制从原句中复制词 x_3，作为当前时刻的生成词。

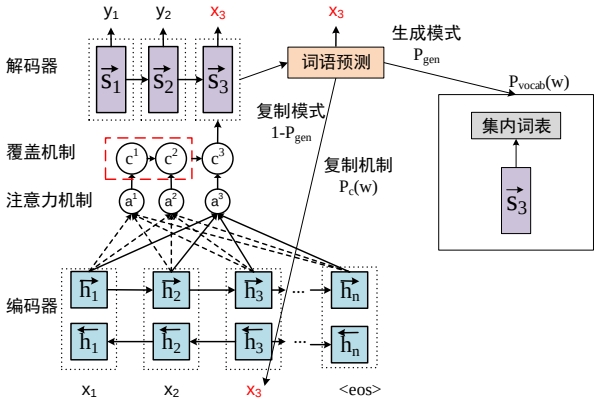

图 2 多机制融合的复述生成模型

Fig. 2 Multi-mechanism fused paraphrase generation model

融合复制机制的解码模型在预测词语时，同时考虑原句中的词和目标词典中的词，计算公式如下：

$$\mathrm{P}\left(\mathrm{w}\right) = P_{gen} P_{vocab}\left(w\right) + \left(1 - P_{gen}\right) P_c\left(w\right) \tag{6}$$

$$P_c\left(w\right) = \sum_{i:w_i=1} a_i^t \tag{7}$$

其中，P_{gen} 表示当前词选择生成模式预测的概率，计算公式如公式（8），$1-P_{gen}$ 表示当前词选择复制模式预测的概率。$P_{vocab}\left(w\right)$ 表示集内词 w 的生成概率，$P_c\left(w\right)$ 表示原句中词 w 的复制概率，词 w 的复制概率由公式（2）计算，若词 w 在原句中多次出现，我们加和每一个位置的注意力权重作为复制概率。当词 w 为未登录词时，$P_{vocab}\left(w\right) = 0$；当词 w 在原句中不存在时，$P_c\left(w\right) = 0$。

$$P_{gen} = \sigma\left(w_h^T c_t + w_s^T s_t + w_{y_{t-1}}^T y_{t-1} + b_{gen}\right) \tag{8}$$

其中 w_h，w_s，$w_{y_{t-1}}$ 和 b_{gen} 表示可学习的参数，σ 表示 sigmoid 函数，c_t 表示当前上下文向量，s_t 表示解码器状态，y_{t-1} 表示上一个已生成词的词向量。

针对词汇重复生成问题，我们引入覆盖机制。覆盖机制建模历史注意力决策信息，对重复关注同

一个词的决策进行惩罚，使当前决策更多关注原句中未被解码的词，从而避免词汇的重复生成。覆盖机制利用注意力机制中的对齐向量，引入覆盖向量 c^t，记录历史注意力信息，计算如公式（9）所示。

$$c^t = \sum_{t'=0}^{t'=t-1} a^{t'} \tag{9}$$

其中 $a^{t'}$ 表示 t' 时刻原句中各个词语的注意力对齐权重，计算公式如公式（2）所示。

覆盖机制将覆盖向量 c^t 输入到注意力计算公式（1）中，使得注意力决策，会受到历史决策信息的约束。公式（1）更改为公式（10），其中 w_c 表示可学习参数。

$$e_i^t = V^T \tanh\left(W_h h_i + W_s s_t + w_c c_i^t + b_{attn}\right) \tag{10}$$

我们利用覆盖损失函数[14]惩罚重复关注同一个词的注意力决策，计算公式如（11）。

$$loss_{coverage}^t = \sum_i \min\left(a_i^t, c_i^t\right) \tag{11}$$

最终，融合复制机制和覆盖机制学习目标函数如下：

$$Loss_{copy+coverage}\left(\theta\right) = \frac{1}{T}\sum_{t=1}^{T} loss_t \tag{12}$$

$$loss_t = -\log P\left(w_t\right) + \lambda \sum_i \min\left(a_i^t, c_i^t\right) \tag{13}$$

其中 $-\log P\left(w_t\right)$ 表示生成词 w_t 的损失函数， $P\left(w_t\right)$ 通过公式（6）获得。超参数 λ 用来对覆盖损失进行加权平衡。

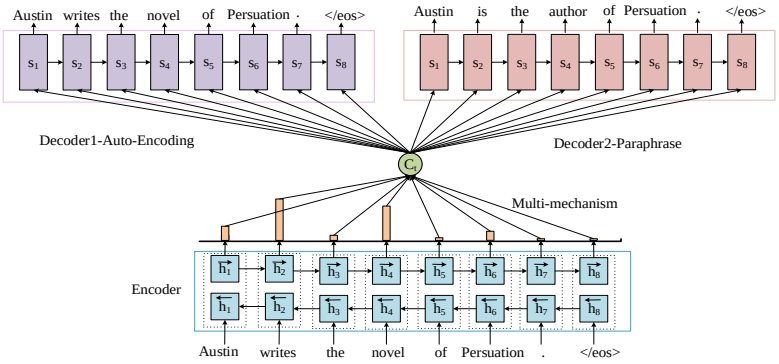

图 3 联合自编码任务的复述生成模型

Fig. 3 Paraphrase generation model with joint auto-encoding learning task

2 联合自编码任务的复述生成模型

2.1 多任务学习：复述生成任务和自编码任务

本文基于多任务学习框架[16, 17]，提出联合自编码任务的复述生成模型，在多机制融合的编码-解码复述生成模型中，联合学习自编码任务和复述生成任务，两个任务共享一个编码器，整体框架如图 3 所示。

复述生成任务：复述生成任务利用平行复述语料 D_p （$\left(\mathrm{X}_n, \mathrm{Y}_n\right) \in D_p$ ）对模型进行训练学习。在给定原句 X_n 时，解码器（Decoder2-Paraphrase）可以生成与原句语义一致且表现形式不同的复述句 Y_n。复述生成任务（文本转换 $\mathrm{X}_n \rightarrow \mathrm{Y}_n$ ）的目标即最大化概率 $p(\mathrm{Y}_n \mid \mathrm{X}_n)$，以生成高质量的复述句 Y_n。如图 3 中的示例，给定原句“Austin writes the novel of Persuation.”，编码器编码学习原句的语义表示，之后经过复述生成解码器生成对应的复述句“Austin is the author of Persuation.”。

自编码任务：自编码任务利用原句子集（无复述句对） D_u （$\mathrm{X}_n \in D_u$）训练模型，学习原句的语义表示，通过与复述任务的联合学习，共同提升编码器的语义表示能力。使得给定原句 X_n 时，自编码器可以生成原句本身 $\mathrm{X}_n^{'}$。和复述任务类似，自编码任务（文本转换 $\mathrm{X}_n \rightarrow \mathrm{X}_n^{'}$ ）通过最大化条件概率 $p(\mathrm{X}_n^{'} \mid \mathrm{X}_n)$ 来对自编码任务的模型参数进行优化。如图 3 中的示例，给定原句“Austin writes the novel of Persuation.”，编码器编码学习原句语义表示，之后经过自编码解码器生成原句本身“Austin writes the novel of Persuation.”。

2.2 联合学习的目标函数

本文融合自编码任务的复述生成模型采用交叉熵损失函数作为训练目标函数，模型的目标函数如下。

$$L(\theta)=-\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{T_y}\log p\left(y_{i,n}|y_{1:i-1,n},\mathrm{X}_n,\theta\right)-\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{T_x}\log p\left(x'_{i,n}|x'_{1:i-1,n},\mathrm{X}_n,\theta\right) \tag{14}$$

其中第一项表示复述生成任务的损失，第二项表示自编码任务的损失；$\theta=(\theta_{enc},\theta_{dec_p},\theta_{dec_{auto}})$ 表示模型所有参数集合，θ_{enc} 表示编码器网络的参数集合，θ_{dec_p} 表示复述生成任务的解码器参数集合，$\theta_{dec_{auto}}$ 表示自编码任务的解码网络参数集合；N 表示输入训练集样本个数，T_y 表示复述句中含有 T_y 个词，T_x 表示生成的原句中含有 T_x 个词，$y_{i,n}$ 表示复述句 Y_n 中第 i 个词 y_i。

3 实验设计和实施细节

3.1 实验数据

本文在 Quora 问句复述语料上进行了评测实验。Quora 包含 150k 复述句对，其中 140k 复述句对作为训练集，验证集和测试集各 5k。在多任务联合学习中，复述生成任务利用平行复述句对（X,Y）训练模型，自编码任务仅利用原句子集（X）。Quora 数据集在复述生成和自编码任务上具体数据统计信息如表 1 所示。

表 1　复述生成任务和自编码任务数据统计信息

Table 1 Data statistical information for paraphrase generation task and auto-encoding task

任务	训练集	验证集	测试集	词汇表(源/目标)
复述生成	140k（对）	5k（对）	5k（对）	20k/20k
自编码	140k	5k	5k	20k/20k

3.2 模型实施细节

在预处理阶段，我们过滤长度大于40的句子，选择出现概率高的前20k词作为词典，其中未登录词用UNK表示。模型参数具体细节实施如下：编码器采用两层双向Bi-LSTM网络，两个解码器采用两层单向LSTM网络，其中编码器中Bi-LSTM和和解码器中LSTM的隐藏层大小设置为500维，词向量维度设置为500维。我们设置Batch大小为64，在层之间采用了dropout正则化技术，设置drop率大小为0.3。我们采用Adam[18]优化算法训练模型，设置初始学习率为0.001，每一轮迭代中学习率以0.88的频率衰减。我们没有使用预训练词向量的方法，所有参数均采用随机初始化的方法，然后跟模型一起学习。

我们采用 ROUGE-1，ROUGE-2[19]，BLEU[20]和 METEOR[21]自动评测指标对复述生成进行评测。

4 评测实验与结果分析

我们设置评测实验分别验证多机制融合的复述生成模型和联合自编码任务的复述生成模型在改进复述生成质量方面的有效性。

4.1 多机制融合复述生成模型的评测

我们采用基于注意力机制的编码-解码模型作为本实验的基线模型，为了验证各个机制的性能以及多机制融合方法的有效性，我们设置了三组对比实验，分别是在基线模型中融合复制机制、在基线模型中融合覆盖机制、在基线模型中同时融合复制机制和覆盖机制。实验结果如表 2 所示。

表 2　多机制融合复述生成评测结果

Table 2 Experiment results on multi-mechanism fused paraphrase generation model

模型	ROUGE-1	ROUGE-2	BLEU	METEOR
基线模型	57.45	32.25	40.93	28.29
+ 复制	61.12	36.04	**45.13（4.20↑）**	31.45
+ 覆盖	56.92	31.50	41.19	28.22
+ 复制 + 覆盖	**61.63（4.18↑）**	**36.50（4.25↑）**	45.01（4.08↑）	**31.48（3.19↑）**

从表 2 可以看出，加入复制机制和覆盖机制有效改进了模型性能。其中，融合复制机制的模型相较于基线模型在所有评测指标均至少提升 3 个点，在 ROUGE-1（ROUGE-2）上提高了 3.67（3.79），BLEU 上提高了 4.20，METEOR 上提高了 3.17，验证了复制机制对改进复述生成质量的有效性。仅融合覆盖机制的模型在 ROUGE、BLEU 和 METEOR 三个评测指标上提升效果不明显。相比于仅融合复制机制的模型，同时融合复制机制和覆盖机制的模型进一步改进了模型性能，在 ROUGE 值上提高 0.5 个点，在 BLEU 和 METEOR 值上也有一定的改进。相比于基线模型，多机制融合模型在各个指标提升明显，表明了本文提出的多机制（注意力机制+复制机制+覆盖机制）融合模型在提升复述生成质量整体性能方面的有效性。

为了进一步验证覆盖机制的有效性，我们分别统计了基线模型和融合覆盖机制模型中存在词汇重复现象的复述句的个数，得到基线模型存在词汇重复的有 127 个句子，融合覆盖机制的模型存在 64 个句子，融合覆盖机制和复制机制的模型有 45 个句子存在词汇重复。可以看出覆盖机制有效的减少了词汇重复生成的现象，改进了复述生成质量。具体统计数据如图 4 所示。

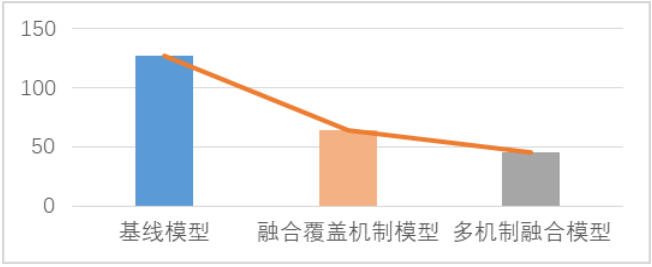

图 4 覆盖机制减少词汇重复的统计信息

Fig. 4 Statistical information of reducing repeated words with coverage mechansim

4.2 联合自编码任务的复述生成模型的评测

为了更好的学习编码器，我们在自编码任务中采用了正序解码和逆序解码两种不同的解码方式。其中，正序解码是按照原句编码的顺序进行解码，逆序解码是按照原句编码的相反顺序进行解码。联合自编码任务的复述生成模型的评测结果如表 3 所示，本实验中我们采用第二节的多机制融合复述生成模型作为基线模型，其中 AutoEncoder 表示自编码输出为正序解码，ReversedAutoEncoder 表示自编码输出为逆序解码。

表 3　联合自编码任务的复述生成评测结果

Table 3 Experiment results on paraphrase generation with joint auto-encoding task

模型	ROUGE-1	ROUGE-2	BLEU	METEOR
基线模型	61.63	36.50	45.01	31.48
AutoEncoder	62.57	37.80	45.83	32.02
ReversedAutoEncoder	**62.95（1.32↑）**	**38.54（2.04↑）**	**46.13（1.12↑）**	**32.10（0.62↑）**
Zichao Li et al., 2017[22]	63.71	37.55	42.33	31.54
Gupta et al., 2017[23]	--	--	38.3	33.6

相较于基线模型，联合自编码正序解码的复述生成模型，在所有评测指标上均有所提升，在ROUGE-1（ROUGE-2）、BLEU和METEOR上分别提高了0.94（1.3）、0.82和0.54，验证了联合自编码的多任务学习在改进复述生成质量方面的有效性。联合自编码逆序解码的复述生成模型在各个指标上均取得了比正序解码更好的性能，进一步验证了提高编码器语义表示能力是改进复述生成的有效方法。我们将联合自编码逆序解码的实验结果与前人最好工作[22,23]进行了比较，在ROUGE-2和BLEU上分别提高了0.99和3.8，进一步表明联合自编码任务的复述生成模型在提高复述生成质量方面的有效性。

表 4 给出了联合自编码复述生成模型的实例分析。通过与原句和参考复述句的对比，可以看出本文提出的多任务联合学习复述生成模型，可以生成语义一致、语法正确、表达流畅且表达形式较为多变的复述句，如复述句对“how do you cook chicken gizzards ?”和“what is the best way to cook chicken gizzards ?”在表达结构上由副词“how”转换为名词性短语“what is the best way to”；以及在复述句对“why do people

from different races or region have different facial features ?”和“why do people from different races differ in facial features ?”中，将形容词短语“different facial features”转换为动词短语“differ in facial features ?”，在句子表达结构和词性转换上均有较大的变化，进一步验证了本文所提模型的有效性。

表 4　多任务学习生成的复述实例

Table 4 Paraphrase examples generated by multi-task learning model

1. Reference	what can I do with chicken gizzards ?
Input	how do you cook chicken gizzards ?
AutoEncoder	what is the best way to cook chicken gizzards ?
ReversedAutoEncoder	how do you cook chicken gizzards ?
2. Reference paraphrase	how do I do affiliate marketing without a website?
Input	can we do affiliate marketing without having a website ?
AutoEncoder	can we do affiliate marketing without having a website ?
ReversedAutoEncoder	Is there a way to get affiliate marketing without a website ?
3. Reference	why is there a facial difference in people with different races ?
Input	why do people from different races or region have different facial features ?
AutoEncoder	why do people have different facial features in different races ?
ReversedAutoEncoder	why do people from different races differ in facial features ?

5　相关工作

传统的复述生成方法，主要采用基于模板和基于统计机器翻译的方法。文献[5]通过抽取复述模板进行模板匹配生成复述，此类方法能够生成含有复杂句法变化的复述句，但模板的覆盖率较低，无法保证生成任意句子的复述句。在基于统计机器翻译（SMT）的复述生成中，文献[8]提出了一种基于短语 SMT 的框架，利用大量单语可比新闻摘要作为训练数据，文献[5]结合了多种资源学习复述生成。

近年来，复述生成的研究方法主要集中在基于神经网络的编码-解码模型。文献[9]通过深度网络来提高模型性能，在编码-解码模型的 LSTM 网络层间加入残差网络。文献[23]在编码-解码模型的基础上，加入了变分机制，引入随机采样学习多样性的复述生成。文献[23]将强化学习应用于编码-解码框架的复述生成任务中，模型可以利用其它目标函数优化复述生成。文献[24]在神经机器翻译编码-解码框架中利用迁移学习，将训练好的文本蕴含任务的模型参数迁移到复述生成任务中，缓解复述语料不足导致模型训练不充分的问题。文献[25]利用句法结构信息作为输入变量约束生成的复述句，学习可控制的复述生成。

虽然前人方法取得了不错的成绩，但利用编码-解码框架进行复述生成时，复述句存在实体词无法准确生成、未登录词和词汇重复等问题。对此，本文提出多机制融合的复述生成模型，有效解决了实体词和未登录词的生成问题，缓解了词汇重复生成。同时，本文提出联合自编码任务的复述生成模型，基于多任务学习框架，联合学习自编码任务和复述生成任务，充分利用了复述语料，有效改进了复述生成质量。

6　结语

本文基于神经网络编码-解码框架的复述生成模型，提出在解码过程中有效融合注意力机制、复制机制和覆盖机制的复述生成模型，进一步改进了复述生成的质量。复制机制解决了生成复述句中实体词不准确和未登录词的问题；覆盖机制有效缓解了生成复述句中词汇重复的问题。同时，本文提出联合自编码任务的复述生成模型，探索了通过联合学习改进复述生成编码器学习能力对提高复述生成质量的方法，充分利用了平行复述语料和原句子数据，增强复述生成编码器的语义编码能力。我们在公开的 Quora 问句复述数据集上进行了评测试验，评测结果显示，本文提出的复述生成模型有效提高了复述生成质量。今后的研究工作，我们将考虑在模型中融入已有的复述知识，如复述短语，以进一步提高复述生成句子的多样性。

参考文献

[1]　Zhang L, Weng Z, Xiao W, Extract Domain-specific Paraphrase from Monolingual Corpus for Automatic Evaluation of

Machine Translation, In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, 2016, Vol. 2: 511~517.

[2] Sanqiang Zhao, Rui Meng, Daqing He, et al. Integrating Transformer and Paraphrase Rules for Sentence Simplification[J]. arXiv preprint arXiv:1810.11193, 2018.

[3] Dong L, Mallinson J, Reddy S, et al. Learning to paraphrase for question answering[J]. arXiv preprint arXiv:1708.06022, 2017.

[4] K. R. McKeown. Paraphrasing Questions Using Given and New Information[J]. Computational Linguistics, 1983, 9(1):1–10.

[5] 赵世奇. 基于统计的复述获取与生成技术研究[D]. 哈尔滨工业大学. 2009.

[6] R. Kozlowski, K. F. McCoy, and K. Vijay-Shanker. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. Proceedings of IWP. 2003:1-8.

[7] C. Quirk, C. Brockett, W. Dolan. Monolingual Machine Translation for Paraphrase Generation. Proceedings of EMNLP. 2004:142-149.

[8] S. Wubben, A. van den Bosch, and E. Krahmer. Paraphrase Generation As Monolingual Translation: Data and Evaluation. In Proceedings of INLG, 2010: 203–207.

[9] Prakash A, Hasan S A, Lee K, et al. Neural Paraphrase Generation with Stacked Residual LSTM Networks[J]. 2016.

[10] Hasan S A, Liu B, Liu J, et al. Neural Clinical Paraphrase Generation with Attention[J]. ClinicalNLP 2016, 2016: 42.

[11] Kyunghyun Cho, Bart Van Merri ¨enboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yosua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. EMNLP, 2014.

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

[13] Gu J, Lu Z, Li H, et al. Incorporating copying Mechanism in Sequence-to-Sequence Learning[J]. 1631-1640(2016).

[14] See A, Liu P J, Manning C D. Get To The Point: Summarization with Pointer-Generator Networks[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1073-1083.

[15] Tu Z, Lu Z, Liu Y, et al. Modeling Coverage for Neural Machine Translation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 76-85.

[16] Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multitask sequence to sequence learning. CoRR, abs/1511.06114. (2015).

[17] Caruana, Rich. Multitask learning. Machine Learning, 28(1):41–75, 1997.

[18] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[19] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop, volume 8. Barcelona, Spain, (2004).

[20] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02), Philadelphia, Pennsylvania. (2002).

[21] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In Proceedings of the Second Workshop on Statistical Machine Translation, pp. 228–231. Association for Computational Linguistics, (2007).

[22] Li Z, Jiang X, Shang L, et al. Paraphrase Generation with Deep Reinforcement Learning[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3865-3878.

[23] Gupta A, Agarwal A, Singh P, et al. A Deep Generative Framework for Paraphrase Generation[J]. arXiv preprint arXiv:1709.05074, 2017.

[24] Florin Brad and Traian Rebedea. Neural paraphrase generation using transfer learning. In Proceedings of the 10th Internationa Conference on Natural Language Generation, pages 257–261, Santiago de Compostela, Spain. Association for Computational Linguistics. 2017.

[25] Iyyer M, Wieting J, Gimpel K, et al. Adversarial Example Generation with Syntactically Controlled Paraphrase Networks[C]//Proceedings of NAACL-HLT. 2018: 1875-1885.