第 37 卷 第 2 期
2022 年 4 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 37 No. 2
Apr. 2022

DOI:10.13364/j.issn.1672-6510.20210146
数字出版日期：2021-11-12；数字出版网址：http://kns.cnki.net/kcms/detail/12.1355.N.20211111.1450.002.html

基于深度强化学习的文本生成研究综述

赵婷婷，宋亚静，李贵喜，王 嫄，陈亚瑞，任德华
（天津科技大学人工智能学院，天津 300457）

摘 要：文本生成任务需要对大量词汇或语句进行表征，且可将其建模为序列决策问题. 鉴于深度强化学习（deep reinforcement learning，DRL）在表征及决策方面的优良性能，DRL 在文本生成任务中发挥了重要的作用. 基于深度强化学习的文本生成方法改变了以最大似然估计为目标的训练机制，有效解决了传统方法中存在的暴露偏差问题. 此外，深度强化学习和生成对抗网络的结合进一步提高了文本生成质量，并已取得了显著的成果. 本综述将系统阐述深度强化学习在文本生成任务中的应用，介绍经典模型及算法，分析模型特点，探讨未来深度强化学习与文本生成任务融合的前景和挑战.

关键词：深度强化学习；自然语言生成；暴露偏差；生成对抗网络
中图分类号：TP391.1 **文献标志码**：A **文章编号**：1672-6510（2022）02-0071-10

Review of Text Generation Based on Deep Reinforcement Learning

ZHAO Tingting，SONG Yajing，LI Guixi，WANG Yuan，CHEN Yarui，REN Dehua
（College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China）

Abstract：Text generation tasks require representation of a large number of words or statements and can be modeled as sequential decision problems. In view of the excellent performance of deep reinforcement learning in representation and decision-making，it plays an important role in text generation tasks. The text generation method based on deep reinforcement learning changes the training mechanism aiming at maximum likelihood estimation and effectively solves the problem of exposure bias in traditional methods. In addition，the combination of DRL and generative adversarial networks has improved the quality of text generation and has achieved remarkable results. This review will elaborate the application of DRL in text generation tasks，introduce the classical models and algorithms，analyze the characteristics of the models，and discuss the prospects and challenges of the future integration of DRL and text generation tasks.

Key words：deep reinforcement learning；natural language generation；exposure bias；generative adversarial network

深度强化学习（deep reinforcement learning，DRL）集成了深度学习对复杂环境的感知能力，以及强化学习对复杂场景的决策能力，实现了端到端的学习模式[1]. 深度强化学习的出现使得强化学习技术真正走向实用，解决现实场景中的复杂问题[2]，在无人驾驶[3-4]、智能交通系统[5]、机器人系统[6-7]、游戏[8]等领域取得了突破性进展，被认为是最有希望实现通用人工智能目标的研究领域之一. 目前，更多的研究者开始把深度强化学习应用在各种不同领域，例如视觉导航[9]、细粒度图像分类[10]、商业游戏[11]、金融决策等[12]. 在自然语言处理（natural language processing，NLP）的文本生成领域中，有不少研究者尝试使用深度强化学习改进现有的网络模型结构或者网络训练流程，并取得了显著性成果[13].

文本自动生成是自然语言处理领域的一个重要研究方向，实现文本自动生成也是人工智能走向成熟的一个重要标志[14]. 文本生成问题是以文本、图像、数据等作为输入，通过计算机处理输出文本的过

收稿日期：2021-06-22；**修回日期**：2021-08-12
基金项目：国家自然科学基金资助项目（61976156）；天津市企业科技特派员项目（20YDTPJC00560）
作者简介：赵婷婷（1986—），女（蒙古族），内蒙古赤峰市人，副教授，tingting@tust.edu.cn

程. 文本生成技术作为各种生成任务的关键模块被广泛采用，包括机器翻译[15]、摘要总结[16-17]、图像字幕[18-19]、风格转换[20]等，文本生成模式根据各自应用场景而不同. 本文关注的是以已有文本为输入，输出相似类型文本的创作型文本生成任务.

自然语言生成问题通常是基于高维且稀疏的特征利用机器学习方法训练浅层模型[21]. 随着神经网络及其变体在诸多任务中展示出良好的应用前景，Bengio 等[22]提出了进行文本生成任务的神经网络语言模型，从语言模型的角度出发，将模型求解最优值的过程转换为求词向量预测的过程[23]. 然而，该方法不能捕捉单词之间的长期依赖关系，使得文本脱离了上下文. 为了解决此问题，Kombrink 等[24]提出了递归神经网络（recurrent neural network，RNN）语言模型，它是一种加入了马尔可夫特性的语言模型. 递归神经网络隐藏层之间的节点也是有连接的，且隐藏层接收来自输入层的输出和上一时刻隐藏层的输出，因此 RNN 模型能保留句子之间的依赖关系[24]. 然而，由于 RNN 模型的梯度消失问题，使得 RNN 语言模型更善于学习距离较近的依赖关系. 为了预测长距离的依赖关系，长短时记忆（long short-term memory，LSTM）[25]、门控循环单元（gated recurrent unit，GRU）[26]等被陆续提出. 训练 RNN 模型最常用的方法是使用最大似然估计（maximum likelihood estimation，MLE）[27]. 然而，由于训练阶段与推理阶段的内在差异，MLE 在理论上存在暴露偏差（exposure bias）问题，即模型在训练时基于真实样本前缀生成后续字符，而在推理时基于模型生成的字符前缀预测下一字符[28]. 这种差异随着序列长度的增加而累积，因此在长文本生成任务中效果不佳. 为了解决这一问题，计划抽样（scheduled sampling，SS）模型被提出，该模型以 ε 的概率选择真实样本前缀，以 $1-\varepsilon$ 的概率选择生成字符前缀，以此消除训练和推理阶段的差异[29]. SS 模型与 MLE 相比有明显改善，但 Huszár[30]从理论的角度证明了计划抽样是个不一致的策略，并不能从本质上解决暴露偏差问题.

为了有效解决暴露偏差问题，Guo[31]提出利用强化学习改变传统生成模型的训练方式. 随后，强化学习的 Actor-Critic 框架[32]也被用来和编码–解码器模型相结合应用于文本生成任务中[33]. 除了这种利用值函数求解改变生成模型训练方式的方法，还可将基于循环神经网络的文本序列生成模型看作是马尔可夫过程[34]. 如何获得精准的奖励函数设计指导生成模型的输出是将强化学习应用于文本生成任务的研究重点，Papineni 等[35]提出使用强化学习算法直接优化生成句子任务的评价指标，把测试时用的双语评估替换（bilingual evaluation understudy，BLEU）和基于召回率替换的二元主旨评价指标（recall-oriented understudy for gisting evaluation，ROUGE）[36]作为训练模型时的奖励[37]. 但这种使用静态奖励的方法计算量非常大，而且只能计算出真实文本与生成文本的 n-gram 相似性[38]，并不是一个完美的度量标准. Yu 等[39]成功将生成对抗网络[40]应用于自然语言处理的离散任务中，提出利用判别器为强化学习智能体提供动态奖励. 但基于二分类的判别器提供的信息有限，使生成模型在训练时存在奖励稀疏及模式崩溃的问题. RankGAN[41]、MaliGAN[42]、LeakGAN[43]算法通过设计不同的提供奖励信息的方式解决上述问题. 因此，奖励函数的设计是算法设计中的核心，这也是本文将要探讨的主要内容.

本文将以基于强化学习的文本生成任务为核心展开综述，首先介绍强化学习的背景知识及文本生成任务建模，然后综述强化学习方法在文本生成任务中的应用并分析各算法优缺点，最后总结全文并分析深度强化学习技术与自然语言生成任务相结合的研究趋势和应用前景.

1　强化学习背景知识

强化学习描述的是智能体为实现任务而连续作出决策控制的过程，其以试错机制与环境进行交互，最终找到适合当前状态的最优动作选择策略，取得整个决策过程的最大累积奖赏[44]，基本框架如图 1 所示.

图 1　强化学习基本框架

Fig. 1　Framework of reinforcement learning

强化学习任务通常建模为马尔可夫决策过程（Markov decision process，MDP）[45]，由状态集合 S、动作集合 A、状态转移函数 P、初始状态概率密度 P_0 和奖励函数 R 这 5 个基本元素组成. 强化学习的核心是找到能够产生最优动作的策略 π，π 可定义为状态空间到动作空间的映射. 智能体在当前状态 s_t 下

根据策略 π 选择动作 a_t 作用于环境，接收到环境反馈的奖励 r_t，并以转移概率 $P^a_{s_t s_{t+1}}$ 转移到下一个状态 s_{t+1}. 强化学习的目的是通过不断调整策略使长期累积奖励 $R_T=\sum_{k=0}^{\infty}\lambda^k r_{t+k}$ 最大化，$\lambda\in[0,1]$ 表示折扣因子. 为了预测累计奖励的期望大小，有两种类型的价值函数：状态值函数 $V^{\pi}(s)$ 和状态–动作值函数 $Q^{\pi}(s,a)$. 状态值函数在遵循策略 π 下描述某个状态的期望奖励. 状态–动作值函数在遵循策略 π 下描述某个状态下执行某个动作的期望奖励. 随后，可以根据 $\pi^*=\arg\max\ V^{\pi}(s)$ 或者 $\pi^*=\arg\max\ Q^{\pi}(s,a)$ 得到最优策略 π^*.

求解强化学习问题主要可通过基于值函数的策略迭代与基于策略的策略搜索两大算法. 基于值函数的策略迭代根据上述的值函数贪婪地选择值函数最大的动作，有效地解决离散状态动作空间问题. 基于策略的策略搜索直接对策略建模并学习，此类算法适用于解决具有连续动作空间的复杂决策任务.

1.1　基于值函数的方法

基于值函数的策略迭代方法通常使用线性或者非线性的函数逼近器近似表示状态值函数或者动作值函数，其通过选择最大值函数的动作从而获得策略. 基于值函数的策略迭代方法的核心是对状态值函数或者动作值函数进行近似估计，其中，时序差分学习[46]和 Q 学习[47]是分别用于求解状态值函数和动作值函数的经典算法. Mnih 等[48]提出了深度 Q 网络（deep Q-network，DQN）模型，该模型创新性地将卷积神经网络和 Q 学习相结合，可以直接将游戏的原始图像作为输入，不依赖于手动提取特征，实现了端到端的学习方式. 自 DQN 被提出后，出现了各种改进方法，其中包括对训练算法的改进、网络结构的改进、学习机制的改进以及算法的改进等[49]. Schaul 等[50]提出了一种带有优先级经验回放的 DQN 模型，Van Hasselt 等[51]提出了 Double DQN 模型，Wang 等[52]提出了基于 DQN 的竞争网络模型，Hausknecht 等[53]提出了 DRQN 模型，Fortunato 等[54]提出了 Noisy DQN 模型，Bellemare 等[55]提出了分布式 DQN 模型.

基于值函数的策略学习方法需要计算所有状态–动作值函数，再从中选择值函数最优的对应动作. 此类方法可以有效解决离散状态空间问题，但是由于值函数的极度非凸性，因此难以在每一个时间步骤都通过最大化值函数选择动作.

1.2　基于策略的方法

基于策略的策略搜索方法直接对策略进行建模学习，适用于解决具有连续动作空间的复杂决策任务. 最具代表性的传统策略搜索算法包括 PEGASUS[56]、策略梯度[57–58]、自然策略梯度[59]、EM[60]及 NAC 等[32]. 其中，策略梯度算法是最实用、最易于实现且被广泛应用的一种策略搜索方法.

相比于基于值函数的方法，基于策略的方法直接在策略空间中搜索最优策略，省去了求解值函数的繁琐环节. 基于策略的策略搜索方法能够有效解决高维度连续动作空间问题. 然而，由于所处理问题的复杂性，基于策略的方法容易陷入局部最优；此外，由于梯度估计方差过大，导致算法不稳定且收敛慢.

1.3　基于 Actor-Critic 的方法

基于策略的策略搜索方法根据累计期望回报指导策略参数调整幅度，使用蒙特卡罗采样估计期望回报时需要完整的状态序列以积累多步的回报，因此会导致方差大的问题. Bahdanau 等[33]结合了基于值函数及基于策略的方法，提出 Actor-Critic（AC）算法框架. Actor 即为策略函数，其与环境交互生成动作；Critic 通过神经网络拟合值函数指导 Actor 进行更新. 相比基于值函数的算法，AC 算法借鉴了策略梯度的做法，使其能够处理具有连续或者高维动作空间的决策任务. 相比传统的策略梯度算法，AC 算法能进行单步更新而不是以轨迹为单位的更新. 然而，AC 算法框架属于在策略（on-policy）算法，其无法使用经验回放提升学习效率.

针对上述问题，研究者从异步、离散策略、稳定性方面改进，提出了具体改进算法，如异步优势动作评价（asynchronous advantage actor-critic，A3C）[62]、深度确定性策略梯度（deep deterministic policy gradient，DDPG）[62]、置信域策略优化（trust region policy optimization，TRPO）[63]等经典算法.

2　基于强化学习的文本生成方法

将文本生成任务建模为强化学习可以很好地解决传统文本生成方法所存在的暴露偏差问题. 基于强化学习的文本生成方法主要分为通过值函数的求解改变编码–解码模型训练方式的方法以及直接求解策略得到生成模型的方法.

2.1　基于值函数的强化学习文本生成方法

Guo[31]将深度强化学习应用到文本生成任务中，

提出了一种基于深度 Q 网络的序列生成框架以解决文本生成任务中词汇空间过大的难题. 该模型的状态为某一时刻的输入词汇和输出词汇，奖励为评价文本相似性的双语评估替换指标，其利用传统的编码–解码语言模型中的解码器为深度 Q 网络生成动态的候选动作空间，并使用双向长短期记忆网络作为深度 Q 网络的网络模型. 该模型极大地减小了深度 Q 网络需要处理的动作空间，从上万的词汇空间减少至数十个候选词汇. 此外，文中选取了 10 000 条句子进行自然语言再生任务实验，即尽量使基于深度 Q 网络改进的解码器的输出和编码器的输入一致. 实验结果表明，此模型比使用长短期记忆网络模型的解码器生成的句子获得的平均平滑双语评估替换指标更高.

上述基于值函数的文本生成方法都需要值函数的求解，通常是利用深度 Q 网络将文本生成任务建模为序列决策问题，状态和动作都是自然语言的形式. 然而，由于需要单独求解值函数，且在文本生成任务中状态空间和动作空间都很庞大，此类方法在训练时往往不稳定，其性能还有待改进.

2.2　基于策略的强化学习文本生成方法

基于循环神经网络的文本序列生成模型将文本生成任务建模为马尔可夫过程[34]，通过最大化生成文本的奖励期望获得最优文本生成策略，利用强化学习方法直接求解模型中的参数. 将生成模型作为强化学习中的智能体与环境进行交互，并将已生成文本序列作为当前状态 s_t，将要生成的单词或字符作为动作 a_{t+1}，在选择了动作 a_{t+1} 后，进而转移到下一个状态 s_{t+1}，状态转移函数 P 为确定性转移函数，具体模型框架如图 2[34]所示.

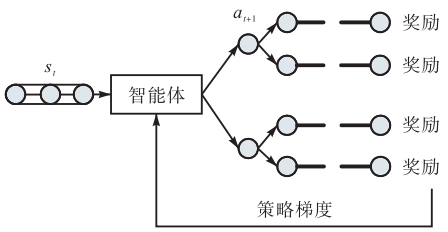

图 2　基于策略的强化学习文本生成模型

Fig. 2　Policy-based reinforcement learning text generation model

在基于策略的文本生成方法中，如何设计奖励函数是此类方法的核心. 关于奖励函数的设计一直都是强化学习在各个领域应用的研究要点. 为了设计出合适的奖励函数指导文本生成模型，提出了如直接使用测试标准作为奖励[35]、通过神经网络学习奖励函数[64]、结合其他模型设计奖励函数[65]等方法.

Ranzato 等[37]提出使用强化学习算法直接优化生成句子任务的评价指标，把测试用的双语评估替换指标和基于召回率替换的二元主旨评价指标（ROUGE-2）作为训练模型的奖励，并利用 REINFORCE 算法对模型进行训练. 然而，强化学习方法往往存在训练难的问题，尤其是面对文本生成的大规模动作空间问题，其每次搜索都面向整个动作空间，其训练初期的随机搜索模式使得模型很难取得有效的提升. 针对上述问题，文献[37]提出了混合增量式交叉熵强化学习（mixed incremental cross-entropy reinforce，MIXER）算法提高模型训练效果，该算法前 s 步按照原有文本生成模型进行预训练，优化目标是最小化生成文本和真实文本之间的交叉熵，s 步后直接将预训练后的循环神经网络模型作为深度强化学习的策略网络模型，再使用 REINFORCE 算法进行训练. 通过在图像描述、机器翻译任务上计算双语评估替换指标和在文本摘要任务上计算二元主旨评价指标表明，MIXER 算法相较于以往方法有不同程度的提升.

另一方面，Shi 等[64]将文本生成中奖励函数的设计任务视为逆强化学习（inverse reinforcement learning，IRL）[66]问题，试图通过神经网络动态拟合单步奖励函数. 如图 3[66]所示，将 IRL 用于文本生成任务中有两个迭代步骤：首先，通过神经网络学习奖励函数解释真实文本数据；其次，以奖励期望最大为目标，学习生成文本的最优策略. 生成模型采用 LSTM 网络表示，奖励函数逼近器依据最大熵逆强化学习[67]求得. 与使用文本评价指标只能在生成完整序列后提供奖励相比，此方法通过拟合即时奖励为模型提供更密集的信息；此外，该方法在生成模型目标函数中加入熵正则项提高生成的多样性. 然而，由于自然语言的复杂性，能够精确拟合奖励函数依然极具挑战.

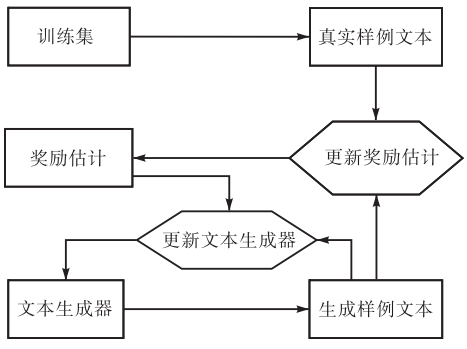

图 3　用于文本生成任务的 IRL 框架

Fig. 3　IRL framework for text generation tasks

Chen 等[65]认为将强化学习用于语言生成会带来方差梯度高、奖励信息少和训练薄弱的问题. 为了解决这些问题，他们对强化学习和最优运输（optimal transport，OT）学习的不同机制进行分析，提出了一种集成了 RL 和 OT 正则化的退火调度学习策略——最优运输强化学习（OTRL），利用 OT 损失自适应地调节 RL 序列生成时在策略空间的探索，从而稳定整个训练过程. OTRL 算法的目标函数主要包括三部分：生成模型的最大似然目标 L_{MLE}、最优运输距离目标 L_{OT} 和基于 RL 训练的目标函数 L_{RL}，其中最大似然用于序列生成模型的预训练，最优传输帮助稳定训练，同时鼓励语义一致性，而强化学习帮助捕捉长短语的一致性. 使用 RL 的方法进行序列生成虽然可以获取长序列的信息，然而梯度差异会很大；只使用 OT 的方法尽管解决了梯度问题，目前却只限于 1-gram 匹配，会造成大量信息流失，如果简单地将其扩展到 k-gram，会极大程度地增加其复杂度. OTRL 算法结合了两种不同方法的优点，从而互补了对方的缺点，获得了较好的效果.

2.3　基于 Actor-Critic 的强化学习文本生成方法

基于 Actor-Critic 的强化学习文本生成方法融合了基于值函数和策略两种方法的优点. Actor 网络通过策略梯度的方法选择动作，Critic 网络通过评估的值函数优化 Actor 网络. 在结合了基于值函数的方法后，该类方法可实现策略梯度的单步更新.

为了解决使用最大似然方法训练生成模型所产生的暴露偏差问题，Bahdanau 等[33]提出了与 Ranzato 等[37]不同的评价指标优化方法，该方法将 Critic 网络引入结构化输出[68]的监督学习问题，使用 Actor-Critic 框架改变传统生成模型的训练方式，如图 4[33]所示.

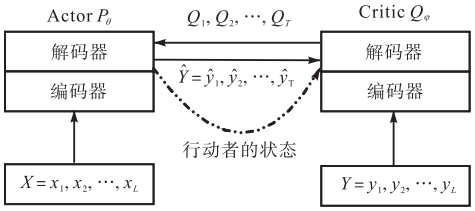

图 4　基于 Actor-Critic 算法的文本生成框架

Fig. 4　Text generation framework based on Actor-Critic algorithm

Actor 和 Critic 都采用典型的编码器–解码器网络结构，其中，Actor 网络接收长度为 L 的真实文本序列 X 作为输入，然后输出预测文本序列 $\hat{Y}$；Critic 网络接收真实的标签序列 Y 和 Actor 在 t 时刻生成的词语，最后输出状态–动作值去训练 Actor 网络. 以强化学习的角度来看，Actor 的状态则为解码器输出的部分序列，动作为下一个生成词，之后使用 Critic 输出状态–动作值对这个动作进行评价. 作者将此方法应用在一个合成任务和一个真实的机器翻译任务上，结果证实了对最大似然方法的改进效果.

3　基于强化学习和生成对抗网络结合的文本生成方法

生成对抗网络（generative adversarial network，GAN）是由 Goodfellow 提出的一种对抗性网络，由生成器 G 和判别器 D 两个核心部分组成[40]. 生成器模型以随机噪声作为输入，试图拟合真实数据分布；判别器模型以真实数据和生成数据作为输入，并试图对两类数据加以区分. 两个模型通过对抗训练的方式进行逐步更新，进而使生成器能够生成接近真实的数据. 生成对抗网络被广泛用于计算机视觉领域的图像生成任务，并取得了很好的效果. 图像作为连续型数据，生成对抗网络可以直接进行梯度求导和反向传播，进而可以达到判别器指导生成器的效果. 将 GAN 应用于自然语言处理领域中的文本生成任务时，文本作为离散标记序列，在生成过程中存在采样过程，导致梯度无法回传. 此外，判别器只能对生成的完整序列进行评分，而无法评价部分序列的好坏. 因此，将 GAN 应用到文本生成领域具有一定难度.

3.1　基于序列生成对抗网络模型的文本生成方法

Yu 等[39]结合强化学习和生成对抗网络提出了序列生成对抗网络模型（SeqGAN），该模型将使用循环神经网络的文本生成模型视为强化学习任务中的智能体，当前状态 s_t 为已经生成的词语，动作 a_{t+1} 定义为下一时刻将要生成的词语，当选定下一个动作后，当前状态以确定性转移到下一状态. 判别器模型以真实文本数据和生成文本数据作为输入，输出数据为真实数据的概率，如图 5[39]所示.

为了解决将 GAN 应用到文本生成时所存在的梯度无法回传的问题，SeqGAN 提出使用强化学习策略梯度的方法对生成器进行更新，生成器的目标是最大化序列的累积奖励，其目标函数定义为

$$J(\theta)=E\left[R_T\middle|s_0,\theta\right]=\sum_{a_1\in A}G_\theta\left(a_1\middle|s_0\right)\cdot Q_{D_\varphi}^{G_\theta}\left(s_0,a_1\right) \tag{1}$$

其中，R_T 是一条完整序列的累积奖励. $Q_{D_\varphi}^{G_\theta}\left(s_t,a_{t+1}\right)$ 是

序列的状态–动作值函数，表示在当前状态 s_t 下，选定动作 a_{t+1} 的好坏程度. SeqGAN 模型将判别器 D 的输出概率作为强化学习中的奖励函数为

$$Q_{D_\varphi}^{G_\theta}(a=y_t, s=Y_{1:T-1}) = D_\varphi(Y_{1:T}) \tag{2}$$

其中，$Y_{1:T}$ 为长度为 T 的完整序列. 由此可见，判别器只能对完整序列进行评价，针对此问题，可采用蒙特卡罗方法对部分序列进行补全为完整序列，并近似求得中间状态的动作价值函数. 判别器 D 以迭代的训练方式进行更新，对生成器 G 提供动态指导，其目标函数可表示为

$$\min_\varphi -E_{Y\sim p_{\text{data}}}[\log D_\varphi(Y)] - E_{Y\sim G_\theta}[\log(1-D_\varphi(Y))] \tag{3}$$

其中，$Y\sim p_{\text{data}}$ 是真实数据，$Y\sim G_\theta$ 是生成数据.

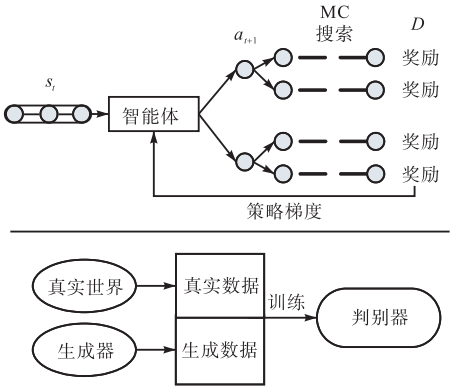

图 5　序列生成对抗网络结构图

Fig. 5　Structure of SeqGAN

SeqGAN 模型摒弃了传统基于强化学习的文本生成任务中采用静态奖励函数的机制，首次结合生成对抗网络提供动态的奖励函数. 此外，通过与强化学习的结合，解决了生成对抗网络无法应用到文本生成任务的两大难题，从而为生成对抗网络应用到自然语言生成任务构建了一种通用框架. SeqGAN 在合成数据及真实场景如中文诗歌生成、奥巴马演讲生成、音乐生成等具体应用场景均取得了较好的结果. 然而，SeqGAN 模型在训练时存在梯度消失和模式崩塌两大问题，业界就如何解决这两个问题对 SeqGAN 模型提出了进一步改进，下面对模型存在的问题及改进方法做详细讨论.

3.2　梯度消失问题

在 SeqGAN 模型训练过程中，由于判别器作为一个二分类器提供的奖励值稀疏，生成器在训练时很难有所提高，其所有生成实例都会被评分为 0，无法进行实质性的更新，导致生成器无法生成多样、符合现实逻辑的文本. 该现象被称作梯度消失，通常以重新设计能够提供更多信息的奖励函数缓解此问题.

RankGAN 利用一个排序器替代判别器，即由序列生成器 G 和排序器 R 组成[40]. 其中，排序器 R 在给定参考时可以对真实数据和生成数据进行相对排序，以相对排序信息作为奖励指导生成器. 排序奖励计算的具体步骤为：首先，通过计算余弦相似度表示输入序列在给定一个参考时的相关性得分；然后据此使用 softmax 公式计算某序列在给定比较集的排序分数.

从某种意义上说，RankGAN 将二元分类器替换为基于多个句子的排序分数，可以缓解梯度消失问题，在改善 SeqGAN 的收敛性能方面显示了良好的结果. 但是，由于它需要对参考集进行额外的采样，因此其计算成本高于其他模型.

除了上述使用排序器增强奖励信息，重新设定分数作为奖励函数是另一种解决方案. 其中，经典的工作是 Che 等[42]提出的最大似然增强的离散生成对抗网络（maximum-likelihood augmented discrete generative adversarial networks，MaliGAN）. MaliGAN 的生成器采用了新的优化目标，其利用重要性抽样，结合判别器的输出重新计算获得的分数作为奖励，即

$$R_{D(x)} = \frac{D(x)}{1-D(x)} \tag{4}$$

MaliGAN 使训练过程更接近自回归模型的最大似然训练，从而使梯度更稳定. 此外，为了降低方差，MaliGAN 采用了两个技巧：第一个是使用蒙特卡罗方法搜索，第二个是使用 MLE 进行训练，逐步向 MaliGAN 方法进行过渡. 实验表明，该网络不仅缓解了梯度消失问题，而且在一定程度上提高了生成器的多样性.

3.3　模式崩溃问题

除了梯度消失外，SeqGAN 模型存在的另一个问题是模式崩溃，即在训练过程中，生成器通过只拟合目标分布的特定部分以欺骗判别器获得高分，往往只能生成简单且短的重复性语句，这极大地降低了生成文本的多样性. 因此，诸多学者通过增强生成器的多样性缓解模式崩溃问题.

与传统直接采用判别器输出作为指导不同，LeakGAN 模型通过判别器泄露自身的提取特征以进一步指导生成器[43]. 同时，生成器建模为层次强化学习问题[69]，包含高阶的 Manager 模块和低阶的 Worker 模块，这两个模块均采用长短时记忆网络构

建. 在每一个时间步，Manager 模块以从判别器接收到高维特征表征作为输入，输出指导目标向量. Worker 模块把当前已生成的单词经过长短时记忆网络编码，将其输出和目标向量用矩阵乘积的方式结合起来，以确保能够综合依据 Manager 的指导和当前状态生成一个合适的新单词.

通过上述过程，判别器使用目标嵌入向量的方式为序列生成提供单步奖励信息，指导生成器如何改进. 其首次通过泄露内部特征的方式训练生成器，并结合层次化强化学习解决以往生成模型在生成长文本中存在的问题.

LeakGAN 模型中的判别器依然是一个二分类器，Xu 等[70]认为现有的基于分类器的判别器存在饱和性的问题，即其只能区分句子真假，不能判断新句子的新颖程度，从而导致文本生成模型倾向于生成一些重复、无意义的文本. 因此，Xu 等[70]提出了 DP-GAN（diversity-promoting generative adversarial network，DP-GAN）模型，采用基于语言模型的单向长短时记忆神经网络作为判别器，并且使用模型的输出交叉熵作为奖励. 生成器是一个两层的长短时记忆神经网络解码器，底层对句子表示进行解码，顶层根据底层的输出对每个单词进行解码. 另外，DP-GAN 采用两种奖励方式，即局部的单词级别的奖励（word-level reward）和全局的句子级别奖励（sentence-level reward）. 单词级别奖励是当前状态的立即奖励，可以直接根据当前的词给出，采用的是语言模型的交叉熵输出，即

$$R\left(y_{t,k}\left|y_{t,<k}\right.\right)=-\log D_{\varphi}\left(y_{t,k}\left|y_{t,<k}\right.\right) \tag{5}$$

句子级别奖励，则是简单地对整个句子的单词级别奖励取平均值，即

$$R\left(y_t\right)=-\frac{1}{k}\sum_{k=1}^{K}\log D_{\varphi}\left(y_{t,k}\mid y_{t,<k}\right) \tag{6}$$

DP-GAN 对重复文本的奖励较低，对新颖流畅的文本奖励较高，鼓励生成者生成新颖多样的文本. 随着多样性的提高，DP-GAN 生成的数据分布能够更接近真实数据分布. 然而，若只注重生成文本的新颖性，DP-GAN 等文本生成器模型不足以生成跨多个句子的长格式文本，主要原因是缺乏一个有效的机制衡量和控制模型生成文本的局部一致性和全局一致性.

受深度结构化语义模型（DSSM）[71]的启发，Cho 等[72]将语义相似性扩展到长文本的连贯和衔接性，提出了一种新的神经语言模型，其包含连贯判别器和衔接判别器，分别在句子（衔接）和段落（连贯）层面提供反馈信号. 连贯判别器通过计算文本块编码后的余弦相似度测量一个段落中所有句子之间的相容性，衔接判别器通过计算两条相邻句子的余弦相似度得到的不同分值区分真实或生成的相邻句子对. 生成器是一个基于注意力的双向 Seq2Seq 模型[73]，通过最大化训练数据的对数似然度进行预训练，并采用了负样本估计其奖励基线的策略梯度方法，因此无需单独的批评函数. 通过上述方法，使用 TripAdvisor 酒店英语评论数据集[74]和 Yelp 英语评论集在长文本生成任务上进行测试，测试结果与人工评价结果一致，说明上述方法在判别器的帮助下生成的文本局部和全局一致程度更高. 但为了生成更有意义、逻辑性更强的长文本，所提出的方法还有待改进.

Zhou 等[75]借鉴 AlphaGo 中使用的自我博弈（self-play）机制，提出一种新的自对抗学习（SAL）范式改进生成对抗网络在文本生成任务中的表现[76]. SAL 的核心思想是：如果发现当前生成的样本比先前生成的更好，则奖励生成器. 自对抗学习中采用的是基于比较思想的判别器，假设其输入是两个文本序列 A 和 B，输出标签包含 3 类，分别对应序列 A 的质量比 B 高（＞）、低（＜）和无法区分（≈）. 与 SeqGAN、MaliGAN 等文本生成对抗网络模型一样，自对抗学习通过 REINFORCE 算法训练生成器. 在训练期间，SAL 通过比较判别器，将生成器当前生成的文本序列与其自身先前生成的文本序列进行比较. 若当前生成的序列比其先前生成的序列质量更高时，生成器得到正奖励，反之奖励为负，两者质量无法区分时奖励为 0. 通过这种自我完善的奖励机制，生成器更易于获得非稀疏奖励，并且在训练后期，SAL 防止重复性样本取得较高的分数，从而能够缓解生成对抗网络奖励稀疏和模式崩溃的问题，使训练更加稳定. 生成的文本序列在质量、多样性、低方差上也都有很好的表现.

4　总结与展望

在社会逐步迈向智能化的时代，文本生成作为实现人工智能的重要标志之一，一直是科技领域研究的热点. 由于人类自然语言的丰富性，提高生成文本的流畅度及多样性是一项很大的挑战. 本文对现有的基于深度强化学习的文本生成方法进行了综述，从提出的背景、基本概念、算法的思想及优缺点等方面进行了详细的分析. 强化学习和文本生成任务的相结合研究备受关注，推动了利用强化学习方法进行文本

生成的研究和发展，且已取得了一定的成果，但该结合研究仍存在问题和挑战亟待解决.

深度强化学习领域的算法依然存在着其自身问题，例如训练不稳定、需要人为设计奖励函数等. 因此，如何提高生成模型的性能是深度强化学习能在文本生成任务中得以广泛应用的重要研究方向. 同时，目前利用强化学习算法及思想解决文本生成任务，仅局限在经典的强化学习算法. 深度强化学习发展至今有许多改进算法及新的模型，因此如何将更适合的强化学习算法有效地应用于文本生成任务也是另一个亟待探索的研究方向. 另外，基于深度学习的文本生成任务不断有新的算法被提出，例如记忆网络、注意力机制等，将其与深度强化学习相结合，提高生成模型的效果，这将是未来的一个研究热点.

参考文献：

[1] 刘建伟，高峰，罗雄麟. 基于值函数和策略梯度的深度强化学习综述[J]. 计算机学报，2019，42(6)：1406–1438.

[2] 赵冬斌，邵坤，朱圆恒，等. 深度强化学习综述：兼论计算机围棋的发展[J]. 控制理论与应用，2016，33(6)：701–717.

[3] MOGHADAM M，ELKAIM G H. A hierarchical architecture for sequential decision-making in autonomous driving using deep reinforcement learning[EB/OL]. [2021–06–20]. https://arxiv.org/abs/1906.08464.

[4] SALLAB A E L，ABDOU M，PEROT E，et al. Deep reinforcement learning framework for autonomous driving[J]. Electronic imaging，2017，2017(19)：70–76.

[5] GREGURIĆ M，VUJIĆ M，ALEXOPOULOS C，et al. Application of deep reinforcement learning in traffic signal control：an overview and impact of open traffic data[J]. Applied sciences，2020，10(11)：4011.

[6] NGUYEN N D，NGUYEN T，NAHAVANDI S，et al. Manipulating soft tissues by deep reinforcement learning for autonomous robotic surgery[C]// IEEE. 2019 IEEE International Systems Conference(SysCon). New York：IEEE，2013：8836924.

[7] ANDRYCHOWICZ O A I M，BAKER B，CHOCIEJ M，et al. Learning dexterous in-hand manipulation[J]. The international journal of robotics research，2020，39(1)：3–20.

[8] RAHMATI A，DAI H. Reinforcement learning for interference avoidance game in RF-powered backscatter communications [EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1903.03600.pdf.

[9] ZHANG J，SPRINGENBERG J T，BOEDECKER J，et al. Deep reinforcement learning with successor features for navigation across similar environments[C]//IEEE. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). New York：IEEE，2017：2371–2378.

[10] LIU X，XIA T，WANG J，et al. Fully convolutional attention networks for fine-grained recognition[EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1603.06765.pdf.

[11] SILVER D，HUBERT T，SCHRITTWIESER J，et al. A general reinforcement learning algorithm that masters chess，shogi，and go through self-play[J]. Science，2018，362(6419)：1140–1144.

[12] DENG Y，BAO F，KONG Y，et al. Deep direct reinforcement learning for financial signal representation and trading[J]. IEEE Transactions on neural networks and learning systems，2016，28(3)：653–664.

[13] 徐聪，李擎，张德政，等. 文本生成领域的深度强化学习研究进展[J]. 工程科学学报，2020，42(4)：399–411.

[14] CCF 中文信息技术专委会. 文本自动生成研究进展与趋势[C]// 中国计算机学会. CCF 2014—2015 中国计算机科学技术发展报告会论文集. 北京：北京万方数据股份有限公司，2015：298–323.

[15] BAHDANAU D，CHO K，BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1409.0473.pdf.

[16] RUSH A M，CHOPRA S，WESTON J. A neural attention model for abstractive sentence summarization[EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1509.00685.pdf.

[17] CHOPRA S，AULI M，RUSH A M. Abstractive sentence summarization with attentive recurrent neural networks [EB/OL]. [2021–06–20]. https://aclanthology.org/N16-1012.pdf.

[18] VINYALS O，TOSHEV A，BENGIO S，et al. Show and tell：a neural image caption generator[C]//IEEE. 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). New York：IEEE，2015：7298935.

[19] XU K，BA J，KIROS R，et al. Show，attend and tell：neural image caption generation with visual attention [EB/OL]. [2021–06–20]. https://arxiv.org/abs/1502.03044.

[20] PRABHUMOYE S，TSVETKOV Y，SALAKHUTDINOV R，et al. Style transfer through back-translation [EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1804.09000.pdf.

[21] BARONI M，ZAMPARELLI R. Nouns are vectors，

adjectives are matrices：representing adjective-noun constructions in semantic space[C]// MIT. Conference on Empirical Methods in Natural Language Processing. Trier：DBLP，2010：1183–1193.

[22] BENGIO Y，DUCHARME R，VINCENT P，et al. A neural probabilistic language model[J]. Journal of machine learning research，2003，3：1137–1155.

[23] MITCHELL J，LAPATA M. Vector-based models of semantic composition[EB/OL]. [2021–06–20]. https://aclanthology.org/P08-1028.pdf.

[24] KOMBRINK S，MIKOLOV T，KARAFIÁT M，et al. Recurrent neural network based language modeling in meeting recognition[EB/OL]. [2021–06–20]. https://www.isca-speech.org/archive/archive_papers/interspeech_2011/i11_2877.pdf.

[25] HOCHREITER S，SCHMIDHUBER J. Long short-term memory[J]. Neural computation，1997，9(8)：1735–1780.

[26] CHO K，VAN MERRIËNBOER B，GULCEHRE C，et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL]. [2021–06–20]. http://de.arxiv.org/pdf/1406.1078.

[27] GRAVES A. Generating sequences with recurrent neural networks[EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1308.0850.pdf.

[28] WILLIAMS R J，ZIPSER D. A learning algorithm for continually running fully recurrent neural networks[J]. Neural computation，1989，1(2)：270–280.

[29] BENGIO S，VINYALS O，JAITLY N，et al. Scheduled sampling for sequence prediction with recurrent neural networks[EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1506. 03099. pdf.

[30] HUSZÁR F. How(not)to train your generative model：Scheduled sampling，likelihood，adversary?[EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1511.05101.pdf.

[31] GUO H. Generating text with deep reinforcement learning[EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1510.09202.pdf.

[32] PETERS J，SCHAAL S. Natural actor-critic[J]. Neurocomputing，2008，71(7/8/9)：1180–1190.

[33] BAHDANAU D，BRAKEL P，XU K，et al. An actor-critic algorithm for sequence prediction[EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1607.07086.pdf.

[34] LU S，ZHU Y，ZHANG W，et al. Neural text generation：past，present and beyond[EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1803.07133.pdf.

[35] PAPINENI K，ROUKOS S，WARD T，et al. BLEU：a method for automatic evaluation of machine translation[EB/OL]. [2021–06–20]. https://aclanthology.org/P02-1040.pdf.

[36] LIN C Y. Rouge：a package for automatic evaluation of summaries[EB/OL].[2021–06–20]. https://aclanthology.org/W04-1013. pdf.

[37] RANZATO M A，CHOPRA S，AULI M，et al. Sequence level training with recurrent neural networks[EB/OL]. [2021–06–20]. http：//de.arxiv.org/pdf/1511.06732.

[38] 陈二静，姜恩波. 文本相似度计算方法研究综述[J]. 数据分析与知识发现，2017(6)：1–11.

[39] YU L，ZHANG W，WANG J，et al. Seqgan：sequence generative adversarial nets with policy gradient[EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1609.05473.pdf.

[40] GOODFELLOW I，POUGET-ABADIE J，MIRZA M，et al. Generative adversarial nets[EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1406.2661.pdf.

[41] LIN K，LI D，HE X，et al. Adversarial ranking for language generation[EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1705.11001.pdf.

[42] CHE T，LI Y，ZHANG R，et al. Maximum-likelihood augmented discrete generative adversarial networks [EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1702.07983.pdf.

[43] GUO J，LU S，CAI H，et al. Long text generation via adversarial training with leaked information[EB/OL]. [2021–06–20]. https://arxiv.org/pdf/1709.08624.pdf.

[44] SUTTON R S，BARTO A G. Reinforcement learning：an introduction[M]. Cambridge：MIT Press，2018.

[45] BENNETT C C，HAUSER K. Artificial intelligence framework for simulating clinical decision-making：a Markov decision process approach[J]. Artificial intelligence in medicine，2013，57(1)：9–19.

[46] TESAURO G. Temporal difference learning and TD-gammon[J]. Communications of the ACM，1995，38(3)：58–68.

[47] WATKINS C J C H，DAYAN P. Q-learning[J]. Machine learning，1992，8(3)：279–292.

[48] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Human-level control through deep reinforcement learning[J]. Nature，2015，518(7540)：529–533.

[49] 刘全，翟建伟，章宗长，等. 深度强化学习综述[J]. 计算机学报，2018，41(1)：1–27.

[50] SCHAUL T，QUAN J，ANTONOGLOU I，et al. Prioritized experience replay[EB/OL]. [2021–06–20]. http://www.arxiv.org/pdf/1511.05952.pdf.

[51] VAN HASSELT H，GUEZ A，SILVER D. Deep rein-

forcement learning with double Q-learning[EB/OL]. [2021-06-20]. https://arxiv.org/pdf/1509.06461.pdf.

[52] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning [EB/OL]. [2021-06-20]. https://dl.acm.org/doi/abs/10.5555/3045390.3045601.

[53] HAUSKNECHT M, STONE P. Deep recurrent Q-learning for partially observable MDPs[EB/OL]. [2021-06-20]. https://arxiv.org/pdf/1507.06527.pdf.

[54] FORTUNATO M, AZAR M G, PIOT B, et al. Noisy networks for exploration[EB/OL]. [2021-06-20]. https://arxiv.org/pdf/1706.10295.pdf.

[55] BELLEMARE M G, DABNEY W, MUNOS R. A distributional perspective on reinforcement learning[EB/OL]. [2021-06-20]. https://arxiv.org/pdf/1707.06887.pdf.

[56] ANDREW Y, JORDAN M. PEGASUS: a policy search method for large MDPs and POMDPs[EB/OL]. [2021-06-20]. https://arxiv.org/ftp/arxiv/papers/1301/1301.3878.pdf.

[57] SEHNKE F, OSENDORFER C, RÜCKSTIEß T, et al. Parameter-exploring policy gradients[J]. Neural networks, 2010, 23(4): 551-559.

[58] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine learning, 1992, 8(3/4): 229-256.

[59] KAKADE S M. A natural policy gradient[EB/OL]. [2021-06-20]. https://dl.acm.org/doi/abs/10.5555/2980539.2980738.

[60] DAYAN P, HINTON G E. Using expectation-maximization for reinforcement learning[J]. Neural computation, 1997, 9(2): 271-278.

[61] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement leaning[EB/OL]. [2021-06-20]. http://arxiv.org/pdf/1602.01783.

[62] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. [2021-06-20]. https://arxiv.org/abs/1509.02971v2.

[63] SCHULMAN J, LEVINE S, ABBEEL P, et al. Trust region policy optimization[EB/OL]. [2021-06-20]. https://arxiv.org/abs/1502.05477v4.

[64] SHI Z, CHEN X, QIU X, et al. Toward diverse text generation with inverse reinforcement learning[EB/OL]. [2021-06-20]. https://arxiv.org/pdf/1804.11258.pdf.

[65] CHEN L, BAI K, TAO C, et al. Sequence generation with optimal-transport-enhanced reinforcement learning [EB/OL]. [2021-06-20]. https://ojs.aaai.org//index.php/AAAI/article/view/6249.

[66] ABBEEL P, NG A Y. Apprenticeship learning via inverse reinforcement learning[EB/OL]. [2021-06-20]. http://dx.doi.org/10.1145/1015330.1015430.

[67] ZIEBART B D, MAAS A L, BAGNELL J A, et al. Maximum entropy inverse reinforcement learning [EB/OL]. [2021-06-20]. https://arxiv.org/pdf/1507.04888v2.pdf.

[68] ROSS S, GORDON G, BAGNELL D. A reduction of imitation learning and structured prediction to no-regret online learning[EB/OL]. [2021-06-20]. https://arxiv.org/pdf/1011.0686.pdf.

[69] DIETTERICH T G. Hierarchical reinforcement learning with the MAXQ value function decomposition[J]. Journal of artificial intelligence research, 2000, 13: 227-303.

[70] XU J, REN X, LIN J, et al. Diversity-promoting GAN: a cross-entropy based generative adversarial network for diversified text generation[C]// Association for Computational Linguistics. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: EMNLP, 2018: 3940-3949.

[71] HUANG P S, HE X, GAO J, et al. Learning deep structured semantic models for web search using click through data[EB/OL]. [2021-06-20]. https://dl.acm.org/doi/10.1145/2505515.2505665.

[72] CHO W S, ZHANG P, ZHANG Y, et al. Towards coherent and cohesive long form text generation[EB/OL]. [2021-06-20]. https://arxiv.org/pdf/1811.00511.pdf.

[73] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2021-06-20]. https://arxiv.org/abs/1409.0473.

[74] WANG H, LU Y, ZHAI C. Latent aspect rating analysis on review text data: a rating regression approach [EB/OL]. [2021-06-20]. https://dl.acm.org/doi/10.1145/1835804.1835903.

[75] ZHOU W, GE T, XU K, et al. Self-adversarial learning with comparative discrimination for text generation [EB/OL]. [2021-06-20]. https://arxiv.org/abs/2001.11691v2.

[76] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

责任编辑：郎婧