工程科学学报
Chinese Journal of Engineering

文本生成领域的深度强化学习研究进展

徐聪 李擎 张德政 陈鹏 崔家瑞

Research progress of deep reinforcement learning applied to text generation

XU Cong, LI Qing, ZHANG De-zheng, CHEN Peng, CUI Jia-rui

引用本文:
徐聪, 李擎, 张德政, 陈鹏, 崔家瑞. 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. doi: 10.13374/j.issn2095-9389.2019.06.16.030
XU Cong, LI Qing, ZHANG De-zheng, CHEN Peng, CUI Jia-rui. Research progress of deep reinforcement learning applied to text generation[J]. *Chinese Journal of Engineering*, 2020, 42(4): 399-411. doi: 10.13374/j.issn2095-9389.2019.06.16.030

在线阅读 View online: https://doi.org/10.13374/j.issn2095-9389.2019.06.16.030

工程科学学报, 第 42 卷, 第 4 期: 399−411, 2020 年 4 月
Chinese Journal of Engineering, Vol. 42, No. 4: 399−411, April 2020
https://doi.org/10.13374/j.issn2095-9389.2019.06.16.030; http://cje.ustb.edu.cn

文本生成领域的深度强化学习研究进展

徐　聪[1,2)], 李　擎[1)✉], 张德政[2,3)], 陈　鹏[1)], 崔家瑞[1)]

1) 北京科技大学自动化学院, 北京 100083　2) 材料领域知识工程北京市重点实验室, 北京 100083　3) 北京科技大学计算机与通信工程学院, 北京 100083

✉通信作者, E-mail: liqing@ies.ustb.edu.cn

摘　要　谷歌的人工智能系统(AlphaGo)在围棋领域取得了一系列成功, 使得深度强化学习得到越来越多的关注. 深度强化学习融合了深度学习对复杂环境的感知能力和强化学习对复杂情景的决策能力. 而自然语言处理过程中有着数量巨大的词汇或者语句需要表征, 并且在对话系统、机器翻译和图像描述等文本生成任务中存在大量难以建模的决策问题. 这使得深度强化学习在自然语言处理的文本生成任务中能够发挥重要的作用, 帮助改进现有的模型结构或者训练机制, 并且已经取得了很多显著的成果. 为此, 本文系统阐述深度强化学习应用在不同的文本生成任务中的一些主要方法, 梳理其发展的轨迹, 分析算法特点. 最后, 展望深度强化学习与自然语言处理任务融合的前景和挑战.

关键词　深度强化学习; 自然语言处理; 文本生成; 对话系统; 机器翻译; 图像描述

分类号　TP183

Research progress of deep reinforcement learning applied to text generation

XU Cong[1,2)], *LI Qing*[1)✉], *ZHANG De-zheng*[2,3)], *CHEN Peng*[1)], *CUI Jia-rui*[1)]

1) School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

2) Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China

3) School of Computer & Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

✉ Corresponding author, E-mail: liqing@ies.ustb.edu.cn

ABSTRACT　With the recent exciting achievements of Google's artificial intelligence system in the game of Go, deep reinforcement learning (DRL) has witnessed considerable development. DRL combines the abilities of sensing and making decisions provided by deep learning and reinforcement learning. Natural language processing (NLP) involves a large number of vocabularies or statements that have to be represented, and its subtasks, such as the dialogue system and machine translation, involve many decision problems that are difficult to model. Because of the aforementioned reasons, DRL can be appropriately applied to various NLP tasks such as named entity recognition, relation extraction, dialogue system, image caption, and machine translation. Further, DRL is helpful in improving the framework or the training pipeline of the aforementioned tasks, and notable achievements have been obtained. DRL is not an algorithm or a method but a paradigm. Many researchers fit plenty of NLP tasks in this paradigm and achieve better performance. Specifically, in text generation based on the reinforcement learning paradigm, the learning process that is used to produce a predicted sequence from the given source sequence can be considered to be the Markov decision process (MDP). In MDP, an agent interacts with the environment by receiving a sequence of observations and scaled rewards and subsequently produces the next action or word. This causes the text generation model to achieve decision-making ability, which can result in future success. Thus, the text generation task integrated with reinforcement learning is an attractive and promising research field. This study presented a comprehensive introduction and a systemic overview. First, we presented the basic methods in DRL and its variations. Then, we showed the main applications of DRL during the

收稿日期: 2019−06−16

基金项目: 国家重点研发计划云计算和大数据专项资助项目(2017YFB1002304)

text generation task, trace the development of DRL, and summarized the merits and demerits associated with these applications. The final section enumerated some future research directions of DRL combined with NLP.

KEY WORDS deep reinforcement learning; natural language processing; text generation; dialogue system; machine translation; image caption

由于深度学习的兴盛, 强化学习和自然语言处理技术都得到了巨大的发展, 突破了各自在传统方法上的瓶颈. 如今越来越多研究将强化学习的强大决策能力应用于自然语言处理的各个任务之中, 都取得了不错的进展. 本文首先简要介绍深度强化学习和文本生成任务, 然后分别梳理三类深度强化学习方法在文本生成任务中的应用以及各自的优缺点, 最后对深度强化学习技术和自然语言处理任务相结合的前景与方向进行总结.

1 简介

1.1 深度强化学习

强化学习 (Reinforcement learning) 通常用来解决科学、工程甚至经济文化等众多领域中的序列决策问题[1]. 强化学习和神经网络的结合可以追溯到 20 世纪 90 年代, 而直到近年来由于深度学习和大数据的惊人成就以及硬件计算能力的大幅提升, 才使得强化学习迎来了一次复兴, 同时也使深度强化学习(Deep reinforcement learning, DRL)成为目前人工智能科学中最热门的研究领域之一.

谷歌的深度思维团队是深度强化学习的主要提出者和研究者, 他们于 2015 年在《Nature》杂志上提出了深度 Q 网络(Deep Q-network, DQN)[2], 并让其学习如何操作 Atari 视频游戏, 最终在 49 个游戏中取得了高于人类专业玩家的得分. 2016 年, 他们提出了蒙特卡罗树搜索和深度强化学习相结合的算法–人工智能算法(AlphaGo), 在与职业九段棋手李世石的对弈中以 4∶1 取得胜利, 并将算法发表于同年的《Nature》杂志上[3]. 在此基础上, 深度思维团队用这套算法的改进版本挑战世界排名第一的中国棋手柯洁, 以 3∶0 的巨大优势取胜. 这意味着以深度学习和强化学习为代表的人工智能算法, 已经能够在一些极其复杂的博弈环境中超越人类顶尖专家的水平.

深度强化学习利用深度学习非线性模型的强大感知能力对复杂环境状态进行表征[4], 利用强化学习的决策优化能力针对不同环境状态进行动作选择[5]. 将两种算法结合构成了基本的深度强化学习的框架, 如图 1 所示, 这样的过程也类似人类进行认知决策的过程, 先通过眼睛感知周围环境的状态, 再通过大脑进行动作选择. 其后大部分的深度强化学习改进算法也基本遵循这个框架原理[6].

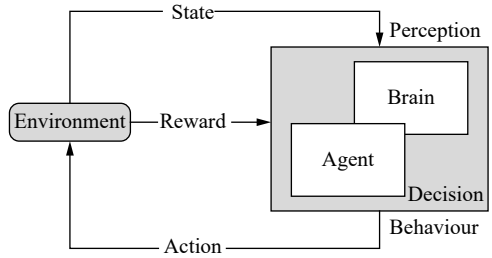

图 1 深度强化学习的基本框架

Fig.1 Framework of deep reinforcement learning

1.2 自然语言处理中的文本生成任务

自然语言处理(Natural language processing, NLP)是利用计算机技术对人类语言进行自动分析和表征的方法及理论的总称. 自然语言处理研究的目的是让计算机能够运行各种层次的自然语言相关任务, 包括分词、词性标注、机器翻译、对话系统. 近二十年来, 自然语言问题都是利用机器学习方法基于高维且稀疏的特征来训练浅层模型. 而随着深度学习方法的发展, 稠密矩阵表征(Dense vector representations)的方法使得很多自然语言处理任务取得了更加优秀的结果[7]. 随后词向量的成功使用加速推动了深度学习在自然语言处理中的应用[8], 与以往浅层模型相比, 深度学习能够自动表征多层次的特征而不依赖先验知识进行手工提取特征, 这就避免了手工提取特征通常耗费时间又不够完整的缺点. 深度学习和自然语言处理发展到现在, 已经能够部分解决一些相对复杂的文本生成任务, 例如对话系统、机器翻译、图像描述和自动摘要等[6].

对话系统通常也被叫作聊天机器人, 或者基于自然语言的人机交互. 他们通常分为两种: 一种是面向特定任务的, 目的是帮助用户完成特定的任务; 一种是开放领域的, 以聊天交流为主要目的[9]. 任务导向的对话系统(Task-oriented spoken dialogue systems)可以完成类似预定酒店、提供餐厅信息和获取公交时间表等任务. 这类系统通常依赖结构

化的本体或者数据库，他们提供了系统交谈所需要的领域知识；而开放领域对话不是以提供信息为目的，一般是以与用户交流的情感体验为目标[10]. 任务导向的对话系统通常使用的数据集有剑桥地区餐厅信息对话数据集[11]、旧金山餐厅信息对话数据集[12]、对话系统技术挑战（Dialog system technology challenge, DSTC）[13]、斯坦福多轮多领域对话数据集[14]；开放领域数据集一般是电影对白（Opensubtitles）、推特（Twitter）、微博等社交聊天记录、乌班图（Ubuntu）对话集[15]（表 1）.

机器翻译是计算机发展之初就企图解决的问题之一，目的是实现机器自动将一种语言转化为另一种语言. 早期方法是语言学家手动编写翻译规则实现机器翻译，但是人工设计规则的代价非常大，对语言学家的翻译功底要求非常高，并且规则很难覆盖所有的语言现象. 之后国际商业机器公司（IBM）在上世纪九十年代提出了统计机器翻译的方法[16]，这种方法只需要人工设计基于词、短语和句子的各种特征，提供足够多的双语语料，就能相对快速地构建一套统计机器翻译系统（Statistical machine translation, SMT），大大减少了翻译系统设计研发的难度，翻译性能也超越了基于规则的方法[17]. 于是机器翻译也从语言学家主导转向计算机科学家主导，在学术界和产业界中基于统计的方法也逐渐取代了基于规则的方法. 随着深度学习不断在图像和语音领域的各类任务中达到最先进水平，机器翻译的研究者也开始使用深度学习技术[18]. 2014 年谷歌的 Sutskever 等提出了序列到序列（Sequence to sequence, Seq2Seq）方法[19]，同年，蒙特利尔大学的 Cho 等提出了类似的编码–解码（Encoder-decoder）框架[20]，之后几乎所有的神经机器翻译（Neural machine translation, NMT）都是基于他们的模型进行改进实现的[21]. 直到注意力机制的出现，才真正使得神经机器翻译在翻译质量上开始超越统计机器翻译，逐步统治机器翻译领域. 基于深度学习的神经机器翻译仅用不到三年时间，已经成为各类自然语言处理国际会议中主要的机器翻译研究方法，同时也成为谷歌[22]、百度[23]、微软等商用机器翻译系统的核心方法. 机器翻译文献中一般使用的平行语料是世界机器翻译大会（The conference on machine translation, WMT）数据集[24]，其中包含英法、英德、英俄等对照翻译语句.

图像生成描述任务是用一个或者多个句子描述图片内容，涉及机器学习、计算机视觉和自然语言处理等领域，需要让模型能理解图片内容和图像的语义信息，并且能生成人类可读的正确描述. 此类任务也可以看作和上述机器翻译类似的过程，即翻译一张图片成为一段描述性文字[25]. 所以可以借鉴机器翻译任务的很多方法和基础框架，通常也是采用编码–解码器模型，编码器编码一张图片而解码器解码生成一段文字. 生成图像描述任务有很广泛的应用前景，例如基于文字的图像检索，为盲人用户提供帮助[26]，人类与机器人交互等场景. 论文中常用数据集为 Flickr8k[27]，lick30k[28]，MSCOCO[29] 等.

上述文本生成任务中存在大量难以建模表征的决策问题，而使用监督学习还不足以解决这样复杂情景的决策任务. 于是具有强大表征和决策能力的深度强化学习可以很好应用于此类自然语言处理任务之中，近年来关于这方面的研究也涌现出很多优秀的方法和思想，下面首先介绍深度强化学习的分类和主要算法，然后结合文本生成任务，详细分析各种算法的创新点和优势，以及如何利用深度强化学习提高各类文本生成任务的效果.

表 1　对话数据集内容概览

Table 1　Summary of dialogue datasets

Dataset	Numbers of dialogue	Numbers of slots	Scene	Multi-turn
Cambridge restaurants database	720	6	1	Yes
San Francisco restaurants database	3577	12	1	Yes
Dialog system technology challenge 2	3000	8	1	Yes
Dialog system technology challenge 3	2265	9	1	Yes
Stanford multi-turn multi-domain task-oriented dialogue dataset	3031	79, 65, 140	3	Yes
The Twitter dialogue corpus	1300000	—	—	Yes
The Ubuntu dialogue corpus	932429	—	—	No
Opensubtitle corpus	70000000	—	—	No

2　深度强化学习的分类

深度强化学习是将深度学习与强化学习结合起来, 实现从感知到动作的端到端学习的全新方法. 在人工智能中, 一般用代理(Agent)表示一个具备行为能力的物体, 比如机器人、无人车、人等等. 那么强化学习就是一个代理随着时间的推移不断地与环境进行交互学习的过程. 在 t 时刻, 代理接受一个状态 s_t 并且遵循策略 $\pi(a_t|s_t)$ 从动作空间 A 中选择一个动作 a_t 作用于环境, 接收环境反馈的奖赏 r_t, 并且依据概率 $P(s_t+1|s_t,a_t)$ 转换到下一个状态 s_{t+1}. 强化学习的最终目的是通过调整自身策略来最大化累计奖赏 $R_t=\sum_{k=0}^{\infty}\lambda^k r_{t+k}$, 其中 $\lambda\in[0,1]$ 表示折扣因子. 而值函数(Value function)是用来预测累计奖赏的期望大小, 衡量某个状态或者状态–动作对的好坏. 假定初始状态 $s_0=s$, 依据策略 π 的状态值函数为 $V^{\pi}(s)=\mathbb{E}\left\{\sum_{t=0}^{\infty}\gamma^t r_t|s_0=s,\pi\right\}$; 同时假定初始动作为 $a_0=a$, 则状态–动作值函数为 $Q^{\pi}(s,a)=\mathbb{E}\left\{\sum_{t=0}^{\infty}\gamma^t r_t|s_0=s,a_0=a,\pi\right\}$. 而根据 $\pi^*=\arg\max_{\pi}V^{\pi}(s)$ 或者 $\pi^*=\arg\max_{a\in A}Q^{\pi}(s,a)$ 可以得到最优策略 π^*.

深度学习和强化学习相结合的主要方式是利用深度神经网络近似任意一个强化学习的组成部分, 包含值函数 $V(s;\theta)$ 或者 $Q(s,a;\theta)$, 策略 $\pi(a|s;\theta)$ 和模型 (状态转移和奖励), 其中参数 θ 是深度神经网络的权重. 通常使用随机梯度下降方法更新深度强化学习的网络参数. 下面介绍一些重要的深度强化学习方法.

2.1　基于值函数的方法

基于值函数(Value-bBased)方法是利用深度神经网络近似强化学习中的值函数部分, 其策略部分并不显现出来而是隐含在值函数的分布之中, 通过选择最大值函数的动作获得策略.

Mnih 等首次介绍了深度 Q 网络[2] 并且带动了深度强化学习这一研究领域. 他们创造性的解决了利用非线性函数近似 Q 函数时容易导致算法不稳定甚至无法收敛的问题. 其主要方法是使用经验回放机制和目标网络, 也就是在训练卷积神经网络近似 Q 函数时随机抽取之前训练过程保存的数据进行网络参数更新, 同时网络的参数并不是立刻更新, 而是通过目标网络进行保存, Q 网络定期与目标网络进行参数同步, 具体训练流程如图 2. 他们的工作开创性的实现了端到端的深度强化学习过程, 整个学习过程基本不需要先验知识以及人工参与, 并且在学习视频游戏的任务中取得了很好的实验结果, 大部分游戏的成绩都超过了人类专家.

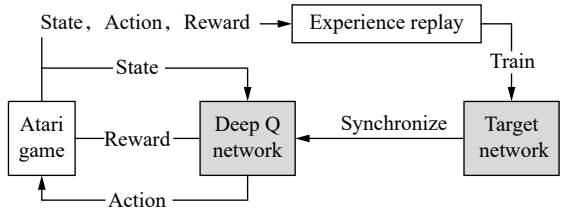

图 2　深度 Q 网络的训练流程

Fig.2　Training process of deep Q-network

随后有研究者发现标准的深度 Q 网络存在过高估计的问题, 其原因是深度 Q 网络使用了同一个 Q 网络进行动作评估和动作选择, 导致了估计误差的出现. 于是 Van Hasselt 等[30] 提出了使用 Q 网络进行动作选择, 而使用目标网络对动作进行评估. Schaul 等[31] 认为标准 Q 网络使用经验回放时是同等概率进行采样, 没有考虑历史数据不同的重要程度, 所以他们提出利用时序差分 (Temporal–difference, TD) 误差来衡量历史数据的重要性, 重要的数据会被更多的采样, 以提高学习效率. Wang 等[32] 提出了一种竞争网络结构, 两个网络分别输出状态值函数和优势函数, 再把它们合并起来输出动作–状态值函数, 并通过实验证明相比深度 Q 网络更快的收敛速度和更好的评估策略.

2.2　基于策略的方法

因为基于值函数结合的方式需要完全计算所有动作的值函数, 再贪婪地选择值函数最大的动作, 所以这种方法通常无法很好的应用在具有高维度或者连续动作空间的问题之中. 而基于策略 (Policy–based) 结合的方式, 直接用深度神经网络学习策略, 网络参数也就是策略的表征, 因此可以直接在策略网络上进行优化, 输出最终动作. 基于策略结合的方法对比基于值函数结合的方法, 拥有更好的收敛性, 能够更有效地应用在高维度或者连续动作空间中, 并且可以学习到随机策略. 然而由于没有值函数, 基于策略的方法对于策略的评估较慢, 必须在与环境交互的过程中进行评价.

Schulman 等[33] 提出了一种可以单调提升策略的迭代过程, 并且通过对理论公式做近似, 给出了可以并行的学习算法——信赖域策略优化(Trust region policy optimization, TRPO). 作者还经过分析后统一了标准的策略梯度和神经网络的策略梯

度. 信赖域策略优化算法用联合梯度计算神经网络梯度的方向，最后在仿真机器人的多项任务中都取得了比较好的效果. 2017 年 Kandasamy 等[34]针对神经对话模型提出了批策略梯度(Batch policy gradient)方法，作者认为采用离策略而非在策略的更新方式更适合序列到序列模型，能够保证梯度的稳定下降. 此外还提出批策略迭代方法，通过保存的动作和奖励按批次进行梯度计算更新目标策略.

2.3　值函数–策略方法

基于值函数和策略结合的方法对应于传统强化学习中的动作者–评价者(Actor–critic)方法[35]，它融合了只用评价者(Critic–only)方法变异性小和只用动作者(Actor–only)容易处理连续动作的优点. 这类算法利用网络参数化的动作者网络生成动作，利用评价者网络为动作者网络提供方差较小的梯度估计[36].

Mnih 等提出了一种异步的强化学习方法(Asynchronous advantage actor–critic, A3C)[37]，多个并行的动作者利用不同的探索策略来稳定训练过程，因此不需要经验回放机制参与训练. 异步强化学习算法能够比深度 Q 网络、深度双 Q 网络、加入竞争机制的深度双 Q 网络等算法获得更高的运行效率并且能够很好的应用在连续控制问题中. Lillicrap 等也同样提出了一种改进的动作者–评价者方法——深度确定性策略梯度(Deep deterministic policy gradient, DDPG)[38]，该算法可以认为是深度 Q 网络在连续动作空间的版本，它利用 Sliver 提出的确定性策略梯度(Deterministic policy gradient)算法结合动作者–评价者方法解决了深度 Q 网络不能在连续或者高维度动作空间中应用的问题，通过实验证明了该算法能够从低维度的观测数据中学习到复杂的策略. Kulkarni 等提出了无模型和基于模型两种算法之外的另一种深度强化学习算法，称为深度继承表征(Deep successor representations, DSR)[39]. 深度继承表征算法由一个奖励预测网络(Reward predictor)和一个继承状态映射网络(Successor map)组成，它的优点是对末端的奖励变化很敏感，并且能够提取子目标从而突破一些瓶颈状态，目前也已经应用于文本生成任务之中，取得了较好的效果[40].

3　深度强化学习在文本生成中的应用

随着近两年深度强化学习在决策和控制领域获得成功，更多的研究者开始把深度强学习应用在各种不同领域，例如视觉导航[41]、策略游戏[42]、细粒度图像分类[43]、自动构建神经网络[44]、网络服务个性化[45]. 自然语言领域中也有不少研究者开始使用深度强化学习来改进现有的网络模型结构或者是建模流程[46–47]. 在自然语言处理的文本生成领域中，如对话系统、机器翻译、图像生成描述和自动摘要等任务都有很多成功使用深度强化学习的文章发表.

深度强化学习和文本生成任务的结合一般是把生成文本的过程看成是生成动作[48]，模型需要根据一些环境信息学习文本生成的策略，环境信息在不同任务中是不一样的，可以有不同的设计方式. 下面根据强化学习模型的类别介绍一些代表性工作.

3.1　基于值函数

这种方法一般是利用深度 Q 网络及其改进算法，将生成文本任务看作是序列决策任务，状态和动作都是自然语言的形式，例如人机对话、基于文本的游戏等.

Narasimhan 等[49]最早将深度 Q 网络应用在自然语言相关的任务中，他们在深度思维团队把深度强化学习应用于视频游戏任务的基础上，把相同的算法框架移植到文本游戏当中. 不同于视频游戏中算法的状态是游戏画面，文本游戏的状态是基于文字的，通常是一段比较长的介绍性文字，需要算法给出一个合适的动作使游戏进入下一个状态. 作者通过循环神经网络(Recurrent neural networks, RNN)[50]的一个变种长短期记忆网络(Long short-term memory, LSTM)来读取状态信息并生成相应的向量表示[51]，将向量化的状态表示输入到多个多层神经网络中，每个网络输出的是动作指令中每个单词的状态值函数，本工作中假设动作指令都是一个动词和一个形容词的形式. 然后选择每个动作中对应状态值函数最大的单词组合成动作指令，作用到游戏中，使游戏转移到下一个状态. 网络的训练方式也和传统深度 Q 网络相似，利用带优先次序的经验回放机制稳定网络的训练过程. 最后作者用实验比较了随机策略算法、长短期记忆网络–深度 Q 网络(LSTM–DQN)算法和利用传统的词袋模型 BOW(Bag of words)或者二元词袋 BI(Bag of bigrams)文本表示方法结合深度 Q 网络的算法，结果表明长短期记忆网络–深度 Q 网络在多个文本游戏中都取得较好得分.

He 等[52]不赞同 Narasimhan 把动作空间当作是有限和已知的做法，他们认为很多文本游戏中

候选动作指令的词汇量是巨大的并且未知的, 候选动作集合是灵活可变的, 对于这些情况一般深度 Q 网络的做法是每次决策的时候把所有候选动作和状态组合后输入最大动作–深度 Q 网络(Max-action DQN)或者把每一种候选动作分别和状态组合后输入每个动作–深度 Q 网络(Per-action DQN). 作者给出了一种改进算法深度强化相关性网络(Deep reinforcement relevance network, DRRN), 不同于以往的深度 Q 网络算法把状态和动作组合后输入同一个网络计算状态值函数, 深度强化相关性网络把表示状态的向量和表示动作的向量分别输入两个深度网络, 然后把两个网络的输出通过点乘结合在一起作为状态值函数. 这样算法就能够从状态和动作两个方面分别进行理解表征, 然后计算状态和动作之间的关联程度作为网络输出, 训练网络使得长期奖励最大化. 实验结果表明深度强化相关性网络算法对于给定候选动作的游戏能够比最大动作–深度 Q 网络(Max-action DQN)和每个动作–深度 Q 网络(Per-action DQN)获得更多的长期奖励.

上述工作将深度强化学习应用在文本游戏中, 面对的并不是典型的自然语言任务. 由于游戏中涉及的动作指令词汇一般数量较少或者提前给定了有限个候选动作指令, 而自然语言处理中的文本生成任务通常会面临巨大的词汇空间, 也就是拥有巨大的动作空间, 因此简单移植标准深度 Q 网络算法是行不通的. 针对上述问题 Guo[53] 提出了一种新的算法框架解决文本生成问题中动作空间过大的难题. 作者利用常规的编码–解码模型中的解码器为深度 Q 网络生成候选动作, 这样就大大减少了深度 Q 网络需要计算的动作数量, 从上万的词汇空间减小到数十个候选词汇. 此算法用 t 时刻输入词汇和输出词汇作为 t 时刻的状态, 用度量相似性的评价指标双语评估替换指标(Bilingual evaluation understudy, BLEU)[54] 作为奖励. 同时作者还尝试使用双向长短期记忆网络作为深度 Q 网络的网络模型. 最后本文选取了 10000 条句子进行编码再解码的训练, 让基于深度 Q 网络改进的解码器尽量生成和输入编码器一致的句子. 实验结果表明基于深度 Q 网络改进的解码器生成的句子比长短期记忆网络形式的解码器生成的句子更加顺畅, 即平均平滑双语评价替换指标(Average smoothed BLEU)更高.

3.2　基于策略

基于策略的方法与文本生成任务结合的方式通常是利用深度网络学习生成词语的策略, 即用网络参数表征词语选择的策略, 网络直接输出词语的标记(Token)而非词语对应的值函数, 跳过了计算值函数的步骤, 从根本上解决词汇空间过大的问题, 这种方法也称作策略梯度方法(Policy gradient method)或策略网络(Policy network)[55].

Ranzato 等[56] 指出之前的文本生成任务中, 训练模型时给定了文本序列中前面的真实词语和一些上下文信息, 让模型预测接下来的词语, 而测试模型的时候并没有文本序列中的真实词语, 只能依据前面生成的预测词语和上下文信息生成下一个词语. 一旦前几个词语生成的错误较大, 就会导致错误一直叠加, 使整个文本序列产生较大偏差. 神经网络生成模型中的这种问题被称之为暴露误差问题[57]. 于是作者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, 如双语评估替换指标或者基于召回率替换的主旨评价标准(Recall-oriented understudy for gisting evaluation, ROUGE)[58]. 为了使用强化学习算法解决文本序列生成问题, 作者把循环神经网络 RNN 结构的文本生成模型看作一个代理, 它与外部环境进行交互, 也就是把词语和上下文信息作为环境的状态输入到代理中. 代理的参数表征策略, 运行策略就能够进行动作的选择. 同时作者把测试时候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)作为训练模型时的奖励, 优化目标是最大化奖励的期望. 本工作还提出一个提高模型训练效果的算法——混合增量式交叉熵强化学习(Mixed incremental cross-entropy reinforce), 算法的前 s 步按照以前的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 步之后直接把前面 s 步训练过的循环神经网络模型作为深度强化学习的策略网络, 优化目标是最大化生成文本的期望奖励. 将混合增量式交叉熵强化学习算法应用到自动摘要、机器翻译和图像生成描述任务中相较于以前的改进方法在四元双语评估替换指标(BLEU-4)和基于召回率替换的二元主旨评价指标(ROUGE-2)指标上都有不同程度的提升.

Rennie 等[59] 同样针对自然语言任务中的深度生成模型存在暴露误差问题, 提出了一种自评价序列训练的强化学习算法(Self-critical sequence training, SCST). 在上述 Ranzato 的工作中, 为了达到减小策略波动的目的, 他们使用线性回归预估出的参考奖励对实际奖励进行归一化操作, 作者

认为这种做法是没有必要的. 文章中提出了另外一种获取参考奖励方法，可以避免训练预测模型，具体做法是使用测试时的算法输出文本序列计算奖励，将此奖励作为参考奖励. 测试时期和训练时期算法的区别是，前者取每个循环神经网络单元输出概率最大的词语组成预测的文本序列，这种方式也称为贪婪式解码（Greedy decoding）；后者是对每个循环神经网络单元产生的词语做蒙特卡罗抽样，抽样所得词语组成预测文本序列. 然后对两个网络的输出文本序列分别计算奖励，当抽样得到句子获得的奖励低于贪婪式解码方法得到句子的奖励时，通过策略梯度的调整降低这句话出现的概率，反之提高其出现的概率. 他们使用基于共识的图像描述评价（Consensus-based image description evaluation, CIDEr）[60] 指标作为奖励函数，在微软带有上下问的常见物体数据集（Microsoft common objects in context, COCO）上进行实验，获得了当时排名第一的成绩，并且发现优化基于共识的图像描述评价指标能够使其他度量指标如双语替换评价指标，基于召回率替换的主旨评价指标，基于单精度的加权调和平均数和单字召回率的评价指标（METEOR）[61] 都得到提高.

Wang 等[62] 的工作主要解决自动摘要中的一致性、多样性问题，他们提出了一种具有联合注意力机制和偏置概率生成机制的卷积序列到序列的模型. 上述机制能够将主题信息整合到自动摘要模型中，使得上下文信息能够帮助模型生成更一致、更多样和包含更多信息的摘要文本. 同时作者利用上文 Rennie 等提出的自评价的序列训练强化学习算法，直接优化摘要任务的评价指标基于召回率替换的主旨评价标准，不仅解决了召回率替换的主旨评价标准作为优化目标导致模型不可导的问题，还免去了暴露误差的影响. 他们利用提出的模型在多个数据集上取得了当前最好成绩. Wu 等[63] 为了提高自动摘要任务中上下文的一致性，设计了能够计算一致性的奖励模型，并将此奖励融合到提出的强化神经抽取式总结模型（Reinforced neural extractive summarization, RNES）中. 此模型同样利用策略梯度方法进行训练，最终能够提高生成的摘要中跨越多个句子的语义信息一致性.

开放领域对话任务相较于其他文本生成任务而言，不只关注于生成下一句文本序列，还需要关注生成的回复对整个对话发展的影响. Li 等[64] 提出了利用强化学习对传统序列到序列模型进行改进，同样利用循环神经网络表征生成对话回复的策略，优化目标是最大化未来奖励的期望. 作者根据开放领域对话任务的特点，设计了三个指标函数共同组成奖励，他们分别评价生成语句的信息丰富性、连贯性和让对方回复的难易度. 通过上述方法，在一定程度上可以避免对话系统出现无意义的语句、重复性的语句和难以回答的语句. 本文还借鉴阿尔法围棋的训练方式，先通过监督学习预训练一个基础序列到序列网络，再让两个训练好的基础序列到序列模型互相对话，通过强化学习的策略梯度方法来更新参数，以获得一个比较大的期望奖励值. 最终结果显示文章采用的算法能产生更丰富、更多交互性、更能持续响应的对话回复. 这个工作也为未来实现长期全局的对话系统作了有益的尝试.

在基于任务的对话系统中，根据对话的主题将对话语料进行分割和标记是其关键任务之一. Takanobu 等[65] 提出利用策略网络和长短期记忆网络相结合的深度网络完成此任务. 由于缺乏标注完善的训练语料，作者将此任务归纳为弱监督学习和序列标注问题. 他们利用先验知识对对话语料进行粗粒度的标注，产生包含噪声的训练数据. 再用包含噪声的标注数据初步训练状态表征网络和策略网络. 策略网络输入的状态是由状态表征网络生成的，输出的动作是语料的主题标签. 也就是说噪声数据经过策略网络之后能够获得一组新的主题标签. 将打上新标签的数据送入状态表征网络进行有监督地训练，更新对话语料的状态表征. 新的状态表征又经过策略网络输出新的主题标签，再重复前面的过程，直到验证集的标签变化率小于设定值. 此时训练好地状态表征网络就可以进行主题分割和标记工作. 作者通过策略网络巧妙地解决了此类任务没有直接监督信号的问题，让强化学习网络为监督学习网络提供不断更新的训练标签，监督学习网络为强化学习网络提供状态输入，联合训练这两个网络最终实现弱监督学习的过程. 他们同时在电商购物的对话数据集上验证了模型在主题分割、标注和上下文理解任务上有很好的效果. 本文提出的基于策略网络的弱监督学习框架有很好的创新性和扩张性，能够应用在其他缺乏完善标签数据的任务中.

3.3　基于策略和值函数

基于策略和值函数的方法，融合了上述两种强化学习算法的优点，策略网络利用策略梯度方法生成动作，值函数评价部分利用深度 Q 网络一类的方法生成对动作的评价，通过评价得到的值

函数来优化策略网络. 基于策略的方法需要在一个回合结束的时候再进行学习, 而由于奖励的稀疏以及衰减, 就造成了基于策略的方法学习效果不够好. 这也解释了为什么最初深度思维公司用的是深度 Q 网络而不是用更直接的基于策略的方法来产生动作. 而动作者−评价者算法结合了基于值函数的方法后, 可以使策略梯度实现单步更新.

Bahdanau 等[66] 提出利用强化学习的动作者−评价者框架和循环神经网络结构的生成模型相融合的方法, 试图改进 Ranzato 提出的算法. 具体做法是把两个典型的编码−解码器网络分别作为动作者和评价者, 动作者网络接收文本序列 X 然后输出预测样本序列 $\hat{Y}$; 评价者网络接收真实的标签序列 Y 和动作者在 t 时刻生成的词语 y_t, 最后输出状态−动作值 Q_T, 再用 Q_T 去训练动作者网络, 如图 3 所示.

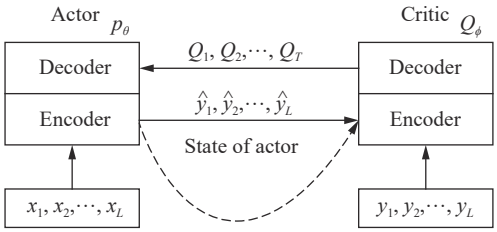

图 3　动作者−评价者框架的训练流程图

Fig.3　Training process of the actor−critic framework

同时作者还采用了一些技巧来提升模型的性能, 如采用类似深度 Q 网络中的目标网络来达到稳定训练的目的. 首先, 增加一个参数更新较为滞后的动作者, 通过这个动作者而非正在训练的行动者生成预测序列, 这样可以避免动作者和评价者相互循环反馈; 其次, 此模型不会只对完整的预测序列计算指标得分作为奖励, 而是对每一步生成的不完整序列计算指标得分, 再做差分计算构造及时奖励, 这样奖励就不只在所有词语都生成完毕时才能获得, 使得评价者的训练信号不再稀疏. 作者将此模型应用于拼写纠正能够获得更低的拼写错误率, 而在机器翻译任务中同样比最大似然估计的训练方法获得更高的双语评估替换指标的得分.

Su 等[67] 将最新的动作者−评价者模型的改进算法应用于任务导向的对话系统中, 并且提高了动作者−评价者算法的学习速度, 解决了策略训练初期算法表现较差的问题. 作者把对话策略优化问题看作是学习每轮如何选择回复序列的任务, 任务目标是最大化长期收益. 因为基于策略的方法比基于值的方法有更强的收敛性, 但是具有策略波动大、采样效率低和易收敛到局部极小的问题, 因此本文采用两种策略方法和值方法结合的方法, 分别是带经验回放的信赖域动作者−评价者模型(Trust region actor−critic with experience replay, TRACER)和带经验回放的不定期动作者−评价者模型(Episodic natural actor−critic with experience replay, eNACER). 前者利用重要性采样比率调节经验回放采样所得历史样本的奖励, 消除它对于当前策略的偏差; 同时采用 Wang 等[68] 提出的改进信赖域策略优化(Trust region policy optimization)方法使得更新后的策略不会偏离平均策略太多, 从而保证了策略的稳定更新, 不会出现较大的策略波动. 后者为了解决策略梯度在陡峭方向上不能保证模型进行更新的问题, 采用 Peters 与 Schaal[69] 提出的自然动作者−评价者 (Natural actor−critic, NAC) 算法加上经验回放机制, 此方法使用了相容函数近似(Compatible function approximation) 不需要精确的计算值函数只需要给出一个估计值. 作者在剑桥地区电话咨询餐厅对话数据集上进行实验, 通过对比高斯过程强化学习(Gaussian processes reinforcement learning, GPRL)[70], 深度 Q 网络, 带经验回放的信赖域动作者−评价者模型和带经验回放的不定期动作者−评价者模型等算法发现提出的算法有更好的效果.

3.4　其他形式

深度强化学习的框架具有一定的通用性, 于是很多研究者把深度强化学习和不同的模型框架或者算法做融合, 应用于自然语言处理任务中, 也取得了很好的效果. 生成对抗网络(Generative adversarial networks, GANs)是近年最火热的深度学习模型之一, 它是由蒙特利尔大学的 Goodfellow 等[71] 学者在 2014 年提出的. 生成对抗网络是一种生成模型(Generative model), 它利用一个判别器模型指导生成模型的训练, 使得模型最终能够生成接近真实的数据. 经过两年的发展, 生成对抗网络及其改进模型已经可以很好的应用于图像生成任务, 但是在自然语言任务中的应用还面临着一些问题. 生成对抗网络中的生成器和判别器模型都需要完全可微, 才能进行梯度训练, 而自然语言任务中需要生成离散的标记序列; 另一个难点是生成对抗网络的判别模型一般是对完整序列进行评价, 而自然语言任务中需要对已经生成的部分序列和之后生成的完整序列的质量都进行评价.

针对上面两个问题, Yu 等[72] 提出了序列生成

对抗网络模型（SeqGANs），用深度强化学习中的策略梯度方法训练生成模型，解决离散标记序列不能进行梯度计算的问题；同时通过蒙特卡洛搜索利用一个展开策略对已经生成的部分序列做采样生成完整序列，即当生成到 t 个词时，假设完整序列有 T 个词语，用蒙特卡洛搜索出后面的 $T-t$ 个词语的 N 条路径，将搜索生成的 $T-t$ 个词语和已经生成的 t 个词语组成完整的 N 个输出序列，再由判别器对这些序列进行评价，将所有评价的平均值作为生成模型的奖励，从而解决了部分生成序列的评价问题，训练过程如图 4 所示.

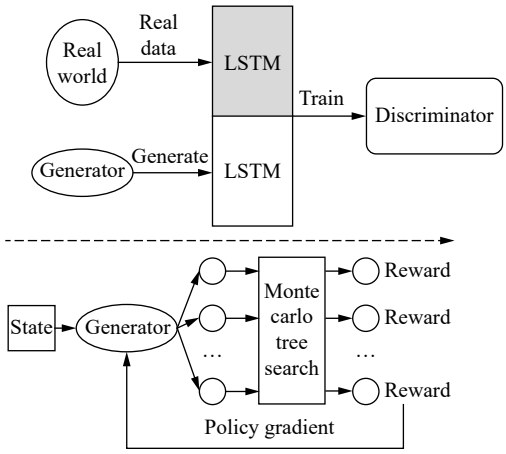

图 4　序列生成对抗网络模型结构及其训练过程

Fig.4　Structure and training process of the seqGANs model

作者将序列生成对抗网络模型应用于生成文本任务如中文诗词、奥巴马政治演讲，以及生成音乐任务中，得到的生成结果比极大似然估计方法要自然和准确.

Li 等[57] 利用对抗训练方法和强化学习方法来解决开放领域对话生成问题，作者采用了和序列生成对抗网络类似方法，用策略梯度训练生成器，用判别器对生成器的输出序列进行评价作为奖励. 不同的是作者认为对部分生成序列进行评价时用蒙特卡罗搜索比较消耗时间，可以训练一个判别器对部分和完整序列都能进行评价，训练数据是从正序列和负序列中随机采样的子序列，每次只从正、负序列的子序列中采样一个样本，确保早期生成的序列不会频繁出现在判别器的训练集中，文中称为每步生成的奖励（Reward for every generation step, REGS）方法. 作者还发现在对抗训练的时候，生成器比较容易崩溃，这是由于生成器不会通过真实的目标序列进行训练，当接受的奖励很低时，只知道当前生成的序列质量糟糕，而不知道如何向正确的方向作调整. 于是作者在更新生成器的参数之后，加入了极大似然估计方法用真实序列值重新更新参数，类似于有老师指导模型训练的方向，因此称为教师指导. 文章中训练了一个可以区分机器生成语句和人类生成语句的模型替代人工评估，最后对比了极大似然估计方法、最大互信息方法、序列到序列模型以及作者提出的对抗−强化学习模型和对抗−每步生成奖励模型，结果显示虽然序列到序列模型生成的回复语句最像人类的回复，但是通常其意思含糊或者与上下文不相关，而作者提出的两个模型的回复语句在这两个方面都能够取得较好的表现.

上面两个工作都是把深度强化学习和生成对抗模型相结合，而 Pfau 与 Vinyals[73] 认为生成对抗网络和动作者−评价者方法有很多相似之处，这篇论文主要工作是从不同的角度来说明了生成对抗网络和动作者−评价者模型的异同点，从而鼓励研究生成对抗网络和动作者−评价者模型的学者合作研发出通用、稳定、可扩展的算法，或者从各自的研究中获取灵感.

在亚马逊的构建社交机器人的比赛中，Serban 等[74] 通过深度强化学习算法结合对话系统开发的 MILABOT 聊天机器人获得最终比赛胜利. 他们利用深度强化学习对若干个对话系统进行整合，该聊天机器人在与真实用户的互动中进行训练，让强化学习算法学习如何从自身包含的一系列模型中选择合适的生成语句作为回复. 真实用户使用 A/B 测试对该系统进行评估，结果显示其性能大大优于其他参赛系统. 由于其所有模块都是可学习的，额外的数据能够帮助该系统继续提升性能.

He 等[75] 利用强化学习中价值网络具有评估长期奖励的能力，解决机器翻译模型解码时只关注局部最优的问题，使翻译的句子整体上达到更好的效果. 作者提出的翻译模型不仅考虑了生成词语的条件概率，还结合了生成词语对未来句子的长期奖励，通过实验证明了此方法较集束搜索解码的翻译模型能够获得更高的双语评估替换指标得分.

4　总结与展望

本文对深度强化学习及其在文本生成任务中的应用现状进行了较为全面的总结，对相关的研究工作进行了分类和解析. 随着深度强化学习和自然语言处理的迅速发展，越来越多的新方法和新应用出现，可以预见强化学习和文本生成以及

其他自然语言处理任务的结合形式会更加丰富. 目前深度强化学习主要还是用来解决自然语言处理中普遍出现的不可导问题，或者是利用深度强化学习的框架帮助改进网络训练流程，从而提升最终效果，未来可以从下面几个方向开展研究工作：

（1）提升深度强化学习算法的性能. 深度强化学习算法本身还有不少问题亟待解决，例如其训练过程较为艰难、稳定性不够好、奖励函数的设计依赖经验等，都需要研究者对其进一步改进[76]. 同时研究者也可以关注于如何提高算法的收敛性、精度、速度和鲁棒性，简化模型结构，增加数据使用效率等方面.

（2）更多传统强化学习算法和深度学习结合可以更好的解决自然语言领域的问题. 传统强化学习算法的研究已经历了 20 年的时间，其中很多算法都有各自的优势，例如逆强化学习、继承学习等，借助深度学习的力量可以在自然语言处理的多种任务中发挥新的作用. 例如 Casanueva 等[77]借鉴封建强化学习 Feudal RL[78] 的方法，把基于任务的对话管理分解为两步，每个子策略通过深度继承学习进行学习.

（3）从自然语言处理的任务中抽象出更多的决策问题. 不同的自然语言任务中都包含需要决策的环节，例如对话机器人与人进行交互、问答系统从知识库抽取知识、利用人的反馈改进图像生成的描述或者是机器翻译的输出等，深度强化学习强大的决策能力能够帮助自然语言处理任务做出较优的选择，这是监督学习无法做到的，例如深度路径强化学习算法模型[79] 利用强化学习解决知识图谱中的关系补全问题；Buck 等[80] 将问答任务归纳到创新的强化学习框架中，提高了回答的效果.

（4）深度强化学习与新的学习算法结合. 深度强化学习是一个灵活的框架，可以与很多新算法融合，例如结合生成对抗网络、记忆网络、注意力机制等，这也能够为解决自然语言处理中的问题提供更多创新的方法和思路，例如 Feng 等[81] 提出基于强化学习的框架从噪声数据中抽取关系，解决了远距离监督学习的问题；Zhang 等[82] 利用强化学习算法自动地学习句子的最优结构化表示，并用于句子分类任务中.

参　考　文　献

[1] Sutton R S, Barto A G. *Reinforcement Learning: An Introduction*. 2nd Ed. Massachusetts: MIT Press, 2018

[2] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, 518(7540): 529

[3] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, 529(7587): 484

[4] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521(7553): 436

[5] Littman M L. Reinforcement learning improves behaviour from evaluative feedback. *Nature*, 2015, 521(7553): 445

[6] Li Y X. Deep reinforcement learning: an overview[J/OL]. *arXiv Preprint* (2017-09-15) [2019-06-16]. https://arxiv.org/abs/1701.07274

[7] Baroni M, Zamparelli R. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space // *Proceedings of the* 2010 *Conference on Empirical Methods in Natural Language Processing*. Cambridge, 2010: 1183

[8] Lapata M, Mitchell J. Vector-based models of semantic composition // *Proceedings of the Meeting of the Association for Computational Linguistics*. Columbus, 2008: 236

[9] Su P H, Gašić M, Mrkšić N, et al. On-line active reward learning for policy optimisation in spoken dialogue systems // *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*. Berlin, 2016: 2431

[10] Vinyals O, Le Q. A neural conversational model[J/OL]. *arXiv Preprint* (2015-07-22) [2019-06-16]. https://arxiv.org/abs/1506.05869

[11] Wen T H, Vandyke D, Mrksic N, et al. A network-based end-to-end trainable task-oriented dialogue system[J/OL]. *arXiv Preprint* (2017-04-24) [2019-06-16]. https://arxiv.org/abs/1604.04562

[12] Wen T H, Gašic M, Kim D, et al. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking // *Proceedings of* 16*th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, 2015: 275

[13] Henderson M, Thomson B, Williams J. The second dialog state tracking challenge // *Proceedings of* 15*th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Philadelphia, 2014: 263

[14] Eric M, Manning C D. Key-value retrieval networks for task-oriented dialogue[J/OL]. *arXiv Preprint* (2017-07-14) [2019-06-16]. https://arxiv.org/abs/1705.05414

[15] Lowe R, Pow N, Serban I V, et al. The ubuntu dialogue corpus: a large dataset for research in unstructured multi-turn dialogue systems // *Proceedings of* 16*th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, 2015: 285

[16] Brown P F, Pietra V J D, Pietra S A D, et al. The mathematics of statistical machine translation: Parameter estimation. *Comput Linguist*, 1993, 19(2): 263

[17] Koehn P, Och F J, Marcu D. Statistical phrase-based translation // *Proceedings of the* 2003 *Conference of the North American*

Chapter of the Association for Computational Linguistics on Human Language Technology. Edmonton, 2003: 48

[18] Zhang J J, Zong C Q. Deep neural networks in machine translation: an overview. *IEEE Intell Sys*, 2015, 30(5): 16

[19] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks // *Proceedings of Advances in Neural Information Processing Systems*. Montréal, 2014: 3104

[20] Cho K, Merriënboer van B, Bahdanau D, et al. On the properties of neural machine translation: encoder–decoder approaches. *Comput Sci*, 2014: 103

[21] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation // *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Lisbon, 2015: 1412

[22] Wu Y H, Schuster M, Chen Z F, et al. Google's neural machine translation system: bridging the gap between human and machine translation[J/OL]. *arXiv Preprint* (2016-10-08) [2019-06-16]. https://arxiv.org/abs/1609.08144

[23] He Z J. Baidu translate: research and products // *Proceedings of the ACL* 2015 *Fourth Workshop on Hybrid Approaches to Translation* (*HyTra*). Beijing, 2015: 61

[24] Cho K, Merrienboer van B, Gulcehre C, et al. Learning phrase representations using RNN encoder –decoder for statistical machine translation // *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Doha, 2014: 1724

[25] Xu K, Ba J L, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention // *Proceedings of* 32*nd International Conference on Machine Learning*. Lille, 2015: 2048

[26] Das A, Kottur S, Gupta K, et al. Visual dialog[J/OL]. *arXiv Preprint* (2017-08-01) [2019-06-16]. https://arxiv.org/abs/1611.08669

[27] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. *J Artif Intell Res*, 2013, 47: 853

[28] Young P, Lai A, Hodosh M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans Assoc Comput Linguist*, 2014, 2: 67

[29] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: common objects in context // *Proceedings of European Conference on Computer Vision*. Zurich, 2014: 740

[30] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-Learning // *AAAI Conference on Artificial Intelligence*. Phoenix, 2016: 2094

[31] Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay[J/OL]. *arXiv Preprint* (2016-02-25) [2019-06-16]. https://arxiv.org/abs/1511.05952

[32] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning // *Proceedings of* 33*rd International Conference on Machine Learning*. New York, 2016: 1995

[33] Schulman J, Levine S, Mortiz P, et al. Trust region policy optimization // *Proceedings of* 31*st International Conference on Machine Learning*. Lille, 2015: 1889

[34] Kandasamy K, Bachrach Y, Tomioka R, et al. Batch policy gradient methods for improving neural conversation models[J/OL]. *arXiv preprint* (2017-02-10) [2019-06-16]. https://arxiv.org/abs/1702.03334

[35] Bhatnagar S, Sutton R S, Ghavamzadeh M, et al. Natural actor-critic algorithms. *Automatica*, 2009, 45(11): 2471

[36] Grondman I, Busoniu L, Lopes G A D, et al. A survey of actor-critic reinforcement learning: standard and natural policy gradients. *IEEE Trans Syst Man Cybern Part C Appl Rev*, 2012, 42(6): 1291

[37] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning // *Proceedings of* 33*rd International Conference on Machine Learning*. New York, 2016: 1928

[38] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[J/OL]. *arXiv Preprint* (2016-02-29) [2019-06-16]. https://arxiv.org/abs/1509.02971

[39] Kulkarni T D, Saeedi A, Gautam S, et al. Deep successor reinforcement learning[J/OL]. *arXiv Preprint* (2016-06-08) [2019-06-16]. https://arxiv.org/abs/1606.02396

[40] Xu C, Li Q, Zhang D, et al. Deep successor feature learning for text generation[J/OL]. *Neurocomputing*, (2019-04-25) [2019-06-16]. https://doi.org/10.1016/j.neucom.2018.11.116

[41] Zhang J W, Springenberg J T, Boedecker J, et al. Deep reinforcement learning with successor features for navigation across similar environments[J/OL]. *arXiv Preprint* (2017-07-23) [2019-06-16]. https://arxiv.org/abs/1612.05533

[42] Bowling M, Burch N, Johanson M, et al. Heads-up limit hold'em poker is solved. *Science*, 2015, 347(6218): 145

[43] Liu X, Xia T, Wang J, et al. Fully convolutional attention localization networks for fine-grained recognition[J/OL]. *arXiv Preprint* (2017-03-21) [2019-06-16]. https://arxiv.org/abs/1603.06765

[44] Zoph B, Le Q V. Neural architecture search with reinforcement learning[J/OL]. *arXiv Preprint* (2017-02-15) [2019-06-16]. https://arxiv.org/abs/1611.01578

[45] Theocharous G, Thomas P S, Ghavamzadeh M. Personalized ad recommendation systems for life-time value optimization with guarantees // *International Joint Conferences on Artificial Intelligence*. Buenos Aires, 2015: 1806

[46] Cuayáhuitl H. Simple D S: A simple deep reinforcement learning dialogue system // *Dialogues with Social Robots*. Springer, Singapore, 2017: 109

[47] He D, Xia Y C, Qin T, et al. Dual learning for machine translation // *Advances in Neural Information Processing Systems*. Barcelona, 2016: 820

[48] Zhang X X, Lapata M. Sentence simplification with deep reinforcement learning // *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, 2017: 584

[49] Narasimhan K, Kulkarni T D, Barzilay R. Language understanding for text-based games using deep reinforcement learning // *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Lisbon, 2015: 1001

[50] Williams R J, Zipser D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput*, 1989, 1(2): 270

[51] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*, 1997, 9(8): 1735

[52] He J, Chen J, He X, et al. Deep reinforcement learning with a natural language action space // *Proceedings of* 54*th Annual Meeting of the Association for Computational Linguistics*. Berlin, 2016: 1621

[53] Guo H. Generating text with deep reinforcement learning[J/OL]. *arXiv Preprint* (2015-10-30) [2019-06-16]. https://arxiv.org/abs/1510.09202

[54] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation // *Proceedings of* 40*th Annual Meeting of Association for Computational Linguistics*. Philadelphia, 2002: 311

[55] Sutton R S, McAllester D A, Singh S P, et al. Policy gradient methods for reinforcement learning with function approximation // *Advances in Neural Information Processing Systems*. Denver, 2000: 1057

[56] Ranzato M A, Chopra S, Auli M, et al. Sequence level training with recurrent neural networks[J/OL]. *arXiv Preprint* (2016-05-06) [2019-06-16]. https://arxiv.org/abs/1511.06732

[57] Li J W, Monroe W, Shi T L, et al. Adversarial learning for neural dialogue generation[J/OL]. *arXiv Preprint* (2017-09-24) [2019-06-16]. https://arxiv.org/abs/1701.06547

[58] Lin C Y. Rouge: A package for automatic evaluation of summaries // *Proceedings of Workshop on Text Summarization Branches Out, Post Conference Workshop of ACL 2004*. Barcelona, 2004: 8

[59] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning[J/OL]. *arXiv Preprint* (2017-11-16) [2019-06-16]. https://arxiv.org/abs/1612.00563

[60] Vedantam R, Lawrence Z C, Parikh D. CIDEr: Consensus-based image description evaluation // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, 2015: 4566

[61] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments // *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, 2005: 65

[62] Wang L, Yao J L, Tao Y Z, et al. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization // *Proceedings of the* 27*th International Joint Conference on Artificial Intelligence*. Stockholm, 2018: 4453

[63] Wu Y X, Hu B T. Learning to extract coherent summary via deep reinforcement learning // *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, 2018: 5602

[64] Li J W, Monroe W, Ritter A, et al. Deep reinforcement learning for dialogue generation[J/OL]. *arXiv Preprint* (2016-09-29) [2019-06-16]. https://arxiv.org/abs/1606.01541

[65] Takanobu R, Huang M, Zhao Z Z, et al. A weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning // *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Stockholm, 2018: 4403

[66] Bahdanau D, Brakel P, Xu K, et al. An actor-critic algorithm for sequence prediction[J/OL]. *arXiv Preprint* (2017-03-03) [2019-06-16]. https://arxiv.org/abs/1607.07086

[67] Su P H, Budzianowski P, Ultes S, et al. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management[J/OL]. *arXiv Preprint* (2017-07-05) [2019-06-16]. https://arxiv.org/abs/1707.00130

[68] Wang Z Y, Bapst V, Heess N, et al. Sample efficient actor-critic with experience replay[J/OL]. *arXiv Preprint* (2017-07-10) [2019-06-16]. https://arxiv.org/abs/1611.01224

[69] Peters J, Schaal S. Natural actor-critic. *Neurocomputing*, 2008, 71(7-9): 1180

[70] Chen L, Su P H, Gasic M. Hyper-parameter optimisation of gaussian process reinforcement learning for statistical dialogue management // *Proceedings of the* 16*th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, 2015: 407

[71] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets // *Advances in Neural Information Processing Systems*. Montréal, 2014: 1

[72] Yu L T, Zhang W N, Wang J, et al. SeqGAN: Sequence generative adversarial nets with policy gradient // *Proceedings of Thirty-First AAAI Conference on Artificial Intelligence*. Palo Alto, 2017: 2852

[73] Pfau D, Vinyals O. Connecting generative adversarial networks and actor-critic methods[J/OL]. *arXiv Preprint* (2017-01-18) [2019-06-16]. https://arxiv.org/abs/1610.01945

[74] Serban I V, Sankar C, Germain M, et al. A deep reinforcement learning chatbot[J/OL]. *arXiv Preprint* (2017-11-05) [2019-06-16]. https://arxiv.org/abs/1709.02349

[75] He D, Lu H Q, Xia Y C, et al. Decoding with value networks for neural machine translation //*Advances in Neural Information Processing Systems*. Long Beach, 2017: 177

[76] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning // *Proceedings of* 33*rd International Conference on Machine Learning*. New York, 2016: 1928

[77] Casanueva I, Budzianowski P, Su P H, et al. Feudal reinforcement learning for dialogue management in large domains // *Proceedings of the* 2018 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Louisiana, 2018: 714

[78] Dayan P, Hinton G E. Feudal reinforcement learning // *Advances in Neural Information Processing Systems*. Denver, 1993: 271

[79] Xiong W, Hoang T, Wang W Y. DeepPath: a reinforcement learning method for knowledge graph reasoning // *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*. Copenhagen, 2017: 564

[80] Buck C, Bulian J, Ciaramita M, et al. Ask the right questions: active question reformulation with reinforcement learning. *arXiv Preprint* (2018-03-02) [2019-06-16]. https://arxiv.org/abs/1705.07830

[81] Feng J, Huang M L, Zhao L, et al. Reinforcement learning for relation classification from noisy data // *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, 2018: 5779

[82] Zhang T Y, Huang M L, Zhao L. Learning structured representation for text classification *via* reinforcement learning // *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, 2018: 6053