Hans Journal of Data Mining 数据挖掘, 2023, 13(2), 185-193
Published Online April 2023 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2023.132018

Hans汉斯

基于强化学习的生成式对话系统研究

颜　永，白宗文*

延安大学物理与电子信息学院，陕西　延安

收稿日期：2023年3月20日；录用日期：2023年4月20日；发布日期：2023年4月28日

摘　要

构建一个回复多样性的开放型对话系统模型，以尝试解决对话系统在回复过程中回答单调的问题。提出一种融合双向长短期记忆神经网络和强化学习模型的生成式对话方法。首先，采用多种类型过滤器对语料库进行预处理，使对话语料库能够被多样化探索到；其次，为了增加对话系统回复的多样性，采用多样性集束搜索作为解码器；最终，在微调模型阶段采用自评序列训练方法削减REINFORCE算法策略梯度的高方差现象。所提方法比Srinivasan等人的方法在BLUE、ROUGE-L、Perplexity分别增长了10.5%，9%和5%，模型的训练时间比原来缩短了43%。部分类型语料数量较少，所以对话系统在这方面的话题相对缺乏。传统的网络架构融合强化学习方法，能够有效地使对话系统产生极具价值意义的回复。

关键词

对话系统，强化学习，多样性探索，回复多样性

Research on Generative Dialogue System Based on Reinforcement Learning

Yong Yan, Zongwen Bai*

School of Physics and Electronic Information, Yan'an University, Yan'an Shaanxi

Received: Mar. 20th, 2023; accepted: Apr. 20th, 2023; published: Apr. 28th, 2023

Abstract

An open dialogue system model with diverse responses is constructed to try to solve the monotonous questions answered by the dialogue system during the response process. This paper proposes a generative dialogue method that combines bidirectional short-term memory neural network and reinforcement learning model. First, the corpus is preprocessed with various types of filters, so that the discourse corpus can be explored in a variety of ways; Secondly, in order to in-

*通讯作者。

文章引用: 颜永, 白宗文. 基于强化学习的生成式对话系统研究[J]. 数据挖掘, 2023, 13(2): 185-193.
DOI: 10.12677/hjdm.2023.132018

crease the diversity of the reply of the dialogue system, the diversity cluster search is used as the decoder; Finally, in the fine-tuning model stage, the self-assessment sequence training method is used to reduce the high square error phenomenon of the REINFORCE algorithm strategy gradient. Compared with Srinivasan's method, the proposed method has increased 10.5%, 9% and 5% respectively in BLUE, ROUGE-L and Perplexity, and the training time of the model has been shortened by 43%. The number of some types of corpus is relatively small, so the topic of dialogue system is relatively lacking. The traditional network architecture and reinforcement learning method can effectively make the dialogue system produce valuable replies.

Keywords

Dialogue System, Intensive Learning, Diversity Exploration, Reply to Diversity

1. 引言

随着大数据和深度学习技术的发展，创建智能对话系统不再是一个只会出现在电影里面的幻想[1]。对话系统根据模式可以分为检索式对话系统和生成式对话系统。与检索式对话系统不同，生成式对话系统通常使用基于神经网络的生成模型逐字生成回复。对话系统场景可以建模成为马尔可夫决策过程，恰巧强化学习也是在马尔可夫决策过程基础上发展起来的，所以用强化学习解决对话系统场景有着天然的优势。这种方法有时候也会出现一些问题，对话系统的目标是找到一个令人满意的对话策略，但是探索、测试和评估策略都极其耗时；模型在训练时缺乏对用户谈话风格的关注，有时候会故意转移话题，从而使模型更加以自我为中心[2]。

本文提出一种融合 Bi-LSTM 和强化学习的对话系统模型。首先，针对强化学习探索、测试和评估策略耗时问题，提出了一种数据处理方法，对复杂的语料库进行了多次过滤整理。我们认为从多角度探索数据集可以高效、快速地使模型收敛。其次，针对对话系统有时候说话喜欢以自我为中心的问题，选择将[3]中的内部奖励信息流替换为情商，意在生成与之前输入具有相同情感基调的语言。最后，用 REINFORCE 算法和自评序列训练方法替换[3] [4]中的策略梯度算法[5]，从而对网络参数起到更好的优化效果。

2. 对话系统模型

生成式对话系统模型利用双向 LSTM 神经网络模型对输入文本进行情感分类，考虑到 SEQ2SEQ 模型很容易出现通用性、安全性的回复，需要通过 DBS 算法挑选可能性回复，进而通过强化学习系统对 SEQ2SEQ 模型不断地进行优化[6]。如图 1 所示，该模型框架主要包括以下几个子模块：数据集探索、双向 LSTM、多样性集束搜索、强化学习系统。

2.1. 多样性集束搜索

在每个时刻输出中，保存前 K 个概率最大的候选序列，其余全部舍弃掉，这样就能减少搜索空间，提高搜索效率[7] [8]。但是 BS 有个最大的问题是其输出的 K 个句子之间差异性很小，这样会造成计算资源的浪费，并且不能体现实际语言的多样性。基本上相同的计算被重复，性能没有显著提高。除此之外，

BS 算法有时候还会引起安全回复问题。

为解决 BS 带来的输出单一性问题，引入多样性集束搜索(DBS)。选择合适的集束宽度 B，然后将其分成 G 组，那么每一组就有(B/G)个集束(一般选择整除关系)。在每一个单独的组内跟 BS 的处理相似，不断的延展序列。同时也通过引入一个差异性项来保证组与组之间的区别。

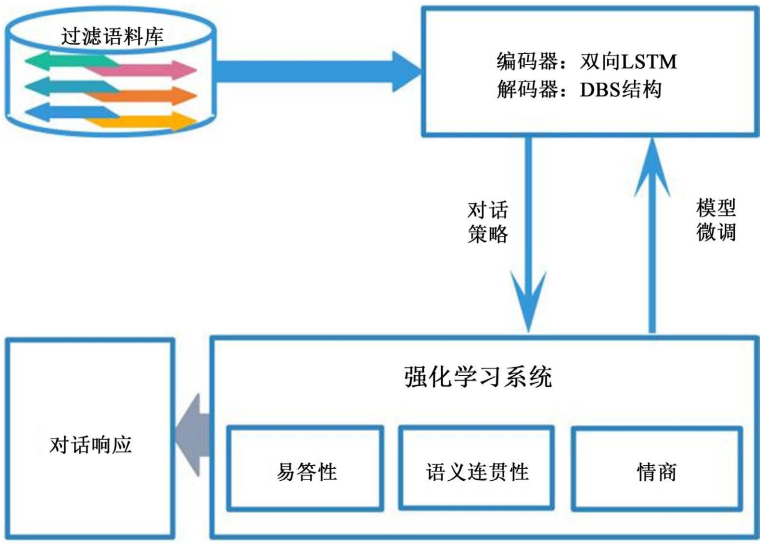

Figure 1. The overall structure of the Dialogue System
图 1. 对话系统总体结构

2.2. 强化学习系统

RL 的本质是在互动中学习。Agent 与环境交互，通过反复试错不断改进，使得累积期望达到最大，从而学习到最佳的动作行为。一个 RL 的 Agent 与它处的环境相互作用，在观察其行为造成的结果后，可以通过学习改变自己的行为来响应所获得的奖励[9]。强化学习过程，如图 2 所示。

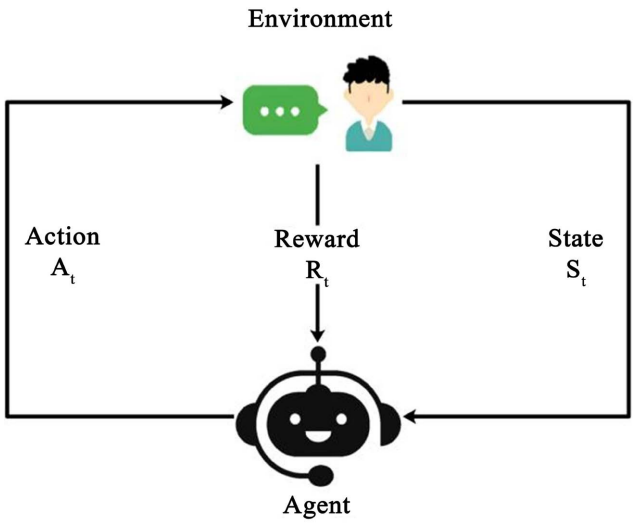

Figure 2. Reinforcement learning process
图 2. 强化学习过程

SEQ2SEQ 模型生成的句子被视为编码器－解码器模型策略所采取的行动，使用强化学习中的 REINFORCE 算法对网络编码器－解码器网络的参数进行优化调整。生成式对话场景中的强化学习主要由四部分组成，分别是：Action (A_t)：生成对话的回复，即动作 $A_t = gen(S)$，其中 $gen(S)$ 是 Bi-LSTM 生成的序列。动作空间是无限的，因为 Agent 可以生成任意长度的句子；State(S_t)：通过向当前对话框(状态)输入必须生成响应的信息，对话框被转换为向量表示；Policy：策略采用 $p_{RL}(p_{i+1} \mid p_i)$ 的形式，其中用 p_{i+1} 是给定对话生成的响应。策略是给定状态下动作的概率分布，其中状态和动作都是对话；Reward：使用内部奖励解决 SEQ2SEQ 架构生成语言的问题。三个内部奖励是易答性 r_{EA}、语义连贯性 r_{SC}、情商 r_{EI}。

易答性(EA) (r_{EA})衡量对话产生枯燥响应的负对数可能性，也就是衡量生成的对话的难易程度。在[1]和[10]之后，在实验阶段列出了在 SEQ2SEQ 模型中经常出现的 10 个枯燥的响应，并在模型生成这些响应时对其进行惩罚。S 代表一系列回答很乏味的对话列表。然后，奖励函数可以定义如下：

$$r_{EA} = -\frac{1}{N_S}\Sigma\frac{1}{N_S}\log p_{\text{seq2seq}(s|a)} \tag{1}$$

p_{seq2seq} 表示 SEQ2SEQ 模型输出的似然。RL 系统可能会惩罚上述对话列表中的话语，因此不太可能产生枯燥的反应。

语义连贯性(SC) (r_{SC})衡量回复的充分性，以避免生成的回复得到高回报但不合语法或不连贯的情况。考虑动作 a 和给定输入之间的互信息，以确保响应是连贯的和适当的。

$$r_{SC} = \frac{1}{N_y}\log p_{\text{seq2seq}}(y \mid x_i) + \frac{1}{N_{x_i}}\log p_{\text{backward-seq2seq}}(x_i \mid y) \tag{2}$$

情商(EI) (r_{EI})这种奖励是通过最大限度地减少提示和响应之间的情感不协调来实现的。有时候与用户保持一致的谈话风格对于构建类人对话模型至关重要。这种方法试图保持输入和产生的响应之间的情感一致性。基于人类之间的开放域文本对话遵循情感模式，假设情感语气通常不会经常波动，聚焦于最小化输入提示和生成的响应之间的情感语气上的不协调。

$$r_{EI_i} = \lambda p(a)\left|\sum_{j=1}^{n}\frac{W2AV(x_j)}{|X|} - \sum_{k=1}^{i}\frac{W2AV(y_k)}{i}\right| \tag{3}$$

这里，公式(10)中的$W2AV$ 表示给定序列的词情感向量，$\sum_{j=1}^{n}\frac{W2AV(x_j)}{|X|}$ 项表示输入提示的平均情感向量，$\sum_{k=1}^{i}\frac{W2AV(y_k)}{i}$ 表示直到当前步骤 i 的生成响应的平均情感向量。

原始 REINFORCE 方法会存在梯度的高方差现象，可以使用 A2C 方法[11]缓解此现象。该方法使用特定状态的估计值作为基线，用另一个头扩展解码器，并根据解码后的序列返回 BLUE 分数估算，此过程额外需要一个 Critic 网络来训练 Value 基线。当然，可以有更好的办法。在生成序列结束(EOS)令牌后，Agent 会观察到一个“奖励”，例如，生成句子的 CIDEr 分数——用 r 表示这个奖励。奖励是由评估指标通过将生成的序列与相应的真实序列进行比较来计算的。训练的目标是最小化负期望奖励(最大化期望奖励)：

$$L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}\left[r(w^s)\right] \tag{4}$$

其中 $w^s = (w_1^s, \cdots, w_T^s)$ 的 w_t^s 是在时间步 t 时从模型中采样的单词。在实践中，通常使用来自 p_θ 的单个样本来估计 $L(\theta)$：

$$L(\theta) \approx -r(w^s), w^s \sim p_\theta \tag{5}$$

很明显 $\mathbb{E}_{w^s \sim p_\theta}\left[b\nabla_\theta \log p_\theta(w^s)\right] = b\sum \nabla p_\theta(w^s) = 0$ 带有基线的 REINFORCE 算法如下：

$$\begin{aligned}\nabla_\theta L(\theta) &= -\mathbb{E}_{w^s \sim p_\theta}\left[\left(r(w^s) - b\right)\nabla_\theta \log p_\theta(w^s)\right] \\ &= -\mathbb{E}_{w^s \sim p_\theta}\left[r(w^s)\nabla_\theta \log p_\theta(w^s)\right]\end{aligned} \tag{6}$$

其中 b 为常数。

S. Rennie 与 E. Marcherett 等[12]在 2016 年提出自评序列训练(SCST)的方法，几乎可以无代价获得基线，而不会增加模型复杂度。SCST 方法的核心是当前模型在测试时使用的推理算法下获得的奖励作为 REINFORCE 算法的基线。在时间步 t，样本 w^s 从模型到 softmax 激活的负回报梯度变为：

$$\frac{\partial L(\theta)}{\partial s_t} = \left(r(w^s) - r(\hat{w})\right)\left(p_\theta(w_t \mid h_t) - 1_{w_t^s}\right) \tag{7}$$

事实上，SCST 的方差要比策略梯度算法的方差低得多，可以更有效地使用 SGD 对小批量样本进行训练。

3. 实验与分析

3.1. 实验方法与数据

在训练阶段，首先将双向 LSTM 网络设计成 SEQ2SEQ 的编码器，旨在对输入语句进行情感分类。随后采用 DBS 算法挑选出可能出现的候选结果，大大减轻了训练过程中的计算量。最后使用 REINFORCE 算法和自评序列训练方法优化网络模型。如下表 1 是我们的方法和原方法的对比。

Table 1. Comparison between Li's method and our method
表 1. Li 的方法和我们方法的对比

方法	目标函数	奖励
Li's method	策略梯度算法	易答性 语义连贯性 信息流
Our method	**多样化集束搜索** **自评序列训练** **REINFORCE 算法**	易答性 语义连贯性 **情商**

对于康奈尔对话语料库，最终的奖励函数仅使用强化学习过程中的内部奖励成分，可以描述如下：

$$r_{\text{Final Cornell}} = \lambda_1 r_{EA} + \lambda_2 r_{SC} + \lambda_3 r_{EI} \tag{8}$$

其中 $\lambda_1 + \lambda_2 + \lambda_3 = 1$，权重设置为 $\lambda_1 = 0.25$ 、 $\lambda_2 = 0.30$ 和 $\lambda_3 = 0.45$ 。

本文实验采用的数据是康奈尔对话语料库，共收集 617 部电影中的对话，其中有 9035 个字符，共有 304,713 个句子。对话数据及其丰富，涉及好多感兴趣的领域。其中训练集有 160,000 条对话，验证集有 140,000 条对话，测试集有 6000 条对话。

3.2. 实验设置与评价指标

本文实验通过 PyTorch 框架搭建对话系统模型，在 Ubuntu16.4 环境下利用 1 块 1080Ti GPU 进行训

练。在使用神经网络时，调整参数可以使模型达到一个很好的状态。经过反复的对比实验，将参数设定为如下表 2，模型表现效果相对较好。为评估方法的效果，本文采用三种评价指标：质量评估 BLEU [13]、召回率 ROUGE-L [14]、困惑度 Perplexity [15]。

Table 2. Parameter setting
表 2. 参数设定

参数	取值
loss 函数	交叉熵损失函数
优化器	SGD
batch size	128
epochs	50
learning rate	0.01
词向量维度	256
编码维度	2
隐层维度	256
$r_{EA}\lambda_1$	0.25
$r_{SC}\lambda_2$	0.30
$r_{EI}\lambda_3$	0.45

BLEU 用于衡量回复句子的流畅性，评价输出语句的质量。实践中，通常是取 $N=1,2,3,4$，然后对进行加权平均。计算公式如下：

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log_{p_n}\right) \tag{9}$$

其中 n表示 n-gram，w_n 表示 n-gram 的权重，BP 表示短句子惩罚因子。

ROUGE-L 计算候选输出与参考输出的最长公共子序列长度。计算公式如下：

$$\text{ROUGE-L} = \frac{\left(1+\beta^2\right)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \tag{10}$$

其中，X 表示候选输出，Y 表示参考输出，$LCS(X,Y)$ 表示候选输出与参考输出的最长公共子序列的长度，m 表示参考输出的长度，n 表示候选输出的长度。其中，$R_{lcs}=\frac{LCS(X,Y)}{m}$，$P_{lcs}=\frac{LCS(X,Y)}{n}$。

Perplexity 用于衡量语言模型输出的质量，困惑度越低，输出回复质量越好。计算公式如下：

$$PPL(W) = P\left(w_1 w_2 \cdots w_N\right)^{\frac{1}{N}} \tag{11}$$

其中 W 是候选翻译，N 是候选翻译的长度，$P(\cdot)$ 是根据参考翻译得到的语言模型，而 $P\left(w_1 w_2 \cdots w_N\right)$ 则是语言模型对候选翻译计算出的得分。

3.3. 实验结果与分析

Srinivasan 等在实验中没有对数据集进行过滤整理，本文弥补了这一缺点，使用多种过滤器对数据集多样性探索。为了验证数据集预处理阶段的重要性，设计了一组对比实验。将 Srinivasan 等人的方法用

到我们的模型中，只改变数据预处理方式。从表 3 中可以看出，本文的数据预处理方式大大缩短了模型训练的时间。

Table 3. Training time of all models on Cornell Movie Dialogs Corpus
表 3. 所有模型在康奈尔电影对话数据集上的训练时间

数据集	模型	训练时间
康奈尔电影对话数据集	Without RL	18 h
	Srinivasan's (RL)	**46 h**
	Ours (RL)	**26 h**

为了验证输出结果的多样性，假定模型其他部分不变，只改变解码器部分。表 4 和表 5 分别展示了在不同解码器的情况下所有模型的自动评测和人工评测的结果。从自动评测的结果上可以看出，在 BLEU 指标上，MMI 模型性能最好，而 DBS 模型在 BLEU-2/3/4 指标是均获得了第二的成绩。在开放域对话中生成回复与所有参考回复都不相似时，BLEU 指标可能并不能反应回复的质量，因此人工标注的分数会更可靠[16]。从表 5 中的人工评测结果上可以看出，DBS 模型所生成的回复可接受比例最高，比 MMI 模型提升了将近 130%。这个结果表明本文用到的方法生成的回复质量更高。Dist-1/2 指标和多样性方面，DBS 模型获得了最好的结果，其性能相比 MMI 模型提升了 110%，同时相比 LSTM 模型提升 160%。这表明本文用到的 DBS 算法能够生成更多样的回复，同时能够保证与输入文本相关性。

Table 4. Automatic evaluation of all models
表 4. 所有模型的自动评测

模型	%BLEU-1	%BLEU-2	%BLEU-3	%BLEU-4	%Dist-1	%Dist-1
LSTM	28.667	29.753	25.224	19.381	36.556	48.567
BS	39.595	34.331	27.727	22.857	51.867	63.951
MMI	44.306	38.727	31.760	26.661	48.678	60.135
DBS	37.004	32.075	25.836	21.198	53.703	66.226

Table 5. Human evaluation results of all models
表 5. 所有模型的人工评测结果

模型	Quality		Diversity
	%Acceptable	%Good	
SEQ2SEQ	19.455	2.678	0.324
BS	28.055	3.778	0.787
MMI	45.667	7.667	0.727
DBS	63.850	7.722	0.808

为了进一步验证强化学习方法在对话系统中的有效性，设计了一组对比实验。通过以上实验参数的设置，本文方法最终的自动测评结果如表 6 所示。使用包括 BLEU 评分、ROUGE-L 和 Perplexity 在内的自动化指标来评估模型。通过结果可以看出，有强化学习确实比没有强化学习的表现效果好，事实表明的确可以通过强化学习训练出较为理想的 SEQ2SEQ 模型。通过表里面的第二行数据和第三行数据对比

不难看出，使用 REINFORCE 算法和自评序列训练代替 Srinivasan 等论文中用到的策略梯度算法，可以有效提高对话系统的性能。

Table 6. Automatic evaluation results of all models on Cornell Movie Dialogs Corpus
表 6. 所有模型在康奈尔电影对话数据集上自动评测结果

数据集	模型	BLUE	ROUGE-L	Perplexity
康奈尔电影对话数据集	Without RL	0.11	0.39	98.96
	Srinivasan's (RL)	0.38	0.55	76.65
	Ours (RL)	**0.42**	**0.60**	**80.40**

从测试结果来看：当输入文本复杂时，对话系统仍然可以给出很有意思的回复，而不会只是“我不知道”“我不能明白”这样简单的回复。从结果有 RL 对话系统的表现来看，对话系统经过大量的训练后，对话系统会积累大量的经验，有时候碰到敏感话题，它也会变得跟人类一样风趣幽默。

4. 结论

生成式对话系统很容易产生“安全回答”问题，有时候对话系统回复很单调，往往每句对话只能对应一种答复，这对传统通用的对话生成方法提出了一定挑战[17]。本文着眼于尝试性的强化学习训练方法和传统的 NLP 方法的融合，以提高生成式对话系统回复的合理性和多样性。具体的方法包括数据预处理层、双向 LSTM 神经网络层、DBS 网络搜索层、强化学习优化层 4 个主要模块。数据预处理层主要对数据集进行多种类型过滤，有益于多样化探索数据集；双向 LSTM 神经网络层作为编码器，对每种类型的样本进行情感分类。BS 网络搜索层作为解码器，挑选输出文本可能的候选回复，增加回复的多样性。强化学习基于尝试性的思想实现最佳的对话策略。实验结果表明，多样化探索数据集确实有效果，模型的训练时间较 Srinivasan 等的训练时间缩短了近 43%。本文方法在 BLUE、ROUGE-L、Perplexity 比其他方法分别提高了 11、9、5 个百分点。

基金项目

工业人工智能中跨媒体协同深度安全态势感知理论和应用研究(62266045)。

参考文献

[1] Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M. and Gao, J. (2016) Deep Reinforcement Learning for Dialogue Generation. *Proceedings of the* 2016 *Conference on Empirical Methods in Natural Language Processing*, Austin, 1-4 November 2016, 1192-1202. https://doi.org/10.18653/v1/D16-1127

[2] Li, J., Galley, M., Brockett, C., Gao, J. and Dolan, B. (2016) A Diversity Promoting Objective Function for Neural Conversation Models. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, San Diego, 12-17 June 2016, 110-119. https://doi.org/10.18653/v1/N16-1014

[3] Srinivasan, V., Santhanam, S. and Shaikh, S. (2019) Natural Language Generation Using Reinforcement Learning with External Rewards. ArXiv Preprint ArXiv: 1911.11404.

[4] Liu, Y., Zhang, L., Han, W., Zhang, Y. and Tu, K. (2021) Constrained Text Generation with Global Guidance—Case Study on CommonGen. ArXiv Preprint ArXiv: 2103.07170.

[5] Ive, J., Li, A.M., Miao, Y., *et al.* (2021) Exploiting Multimodal Reinforcement Learning for Simultaneous Machine Translation. ArXiv Preprint ArXiv: 2102.11387.

[6] Srinivasan, V., Santhanam, S. and Shaikh, S. (2019) Natural Language Generation Using Reinforcement Learning with External Rewards. ArXiv Preprint ArXiv: 1911.11404.

[7] Liu, Q., Chen, Y., Chen, B., *et al*. (2020) You Impress Me: Dialogue Generation via Mutual Persona Perception. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 1417-1427. https://doi.org/10.18653/v1/2020.acl-main.131

[8] Vijayakumar, A.K., *et al*. (2017) Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. ArXiv Preprint ArXiv: 1610.02424

[9] Arulkumaran, K., Deisenroth, M.P., Brundage, M. and Bharath, A.A. (2017) A Brief Survey of Deep Reinforcement Learning. ArXiv Preprint ArXiv: 1708.05866.

[10] Danescu-Niculescu-Mizil, C. and Lee, L. (2011) Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. ArXiv Preprint ArXiv: 1106.3077.

[11] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D. and Kavukcuoglu, K. (2016) Asynchronous Methods for Deep Reinforcement Learning. *Proceedings of the* 33*rd International Conference on Machine Learning*, New York, 19-24 June 2016, 1928-1937.

[12] Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J. and Goel, V. (2017) Self-Critical Sequence Training for Image Captioning. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1179-1195. https://doi.org/10.1109/CVPR.2017.131

[13] Xu, C., Li, P., Wang, W., *et al*. (2022) COSPLAY: Concept Set Guided Personalized Dialogue Generation Across Both Party Personas. *Proceedings of the* 45*th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, 11-15 July 2022, 201-211. https://doi.org/10.1145/3477495.3531957

[14] Cao, Y., Bi, W., Fang, M., Shi, S. and Tao, D. (2022) A Model-Agnostic Data Manipulation Method for Persona-based Dialogue Generation. *Proceedings of the* 60*th Annual Meeting of the Association for Computational Linguistics*, Dublin, 22-27 May 2022, 7984-8002. https://doi.org/10.18653/v1/2022.acl-long.550

[15] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002) Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the* 40*th Annual Meeting on Association for Computational Linguistics*, Philadelphia, 7-12 July 2002, 311-318.

[16] 高俊. 开放域对话系统的多样化回复生成方法研究[D]: [硕士学位论文]. 苏州: 苏州大学, 2020. https://doi.org/10.27351/d.cnki.gszhu.2020.001335

[17] 王晶. 基于强化学习的情感对话回复生成算法研究[D]: [硕士学位论文]. 桂林: 桂林电子科技大学, 2020. https://doi.org/10.27049/d.cnki.ggldc.2020.000309