文章编号：1003-0077(2018)07-0099-10

基于 DQN 的开放域多轮对话策略学习

宋皓宇，张伟男，刘 挺

（哈尔滨工业大学 社会计算与信息检索研究中心，黑龙江 哈尔滨 150001）

摘 要：有效地进行多轮对话是开放域人机对话系统的主要目标之一。目前的神经网络对话生成模型在开放域多轮对话过程中存在着容易产生万能回复、很快陷入死循环的问题；而已有的多轮对话研究工作存在着没有考虑未来对话走向的问题。借鉴强化学习方法考虑全局的视角，该文利用深度强化学习算法 DQN(deep Q-network)，提出了使用深度价值网络对每一轮的候选句子进行评估，并选择未来收益最大的而非生成概率最大的句子作为回复的多轮对话策略学习方法。实验结果表明，该文提出的方法将多轮对话的平均对话轮数提高了两轮，同时在主观对比评价指标上获胜比例高出了 45%。

关键词：多轮对话；对话策略；强化学习

中图分类号：TP391 **文献标识码**：A

DQN-based Policy Learning for Open Domain Multi-turn Dialogues

SONG Haoyu, ZHANG Weinan, LIU Ting

(Research Center of Social Computing and
Information Retrieval, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: The open domain dialogue system is challenged by effective multi-turn dialogues. Current neural dialogue generation models tend to fall into conversation black holes by generating safe responses, without considering the future information. Inspired by the global view of reinforcement learning methods, we present an approach to learn multi-turn dialogue policy with DQN (deep Q-network). We introduce a deep neural network to evaluate each candidate sentence and choose the sentence with the maximum future rewards, instead of the highest generation probability, as a response. The results show that our method improves the average dialogue turns by 2 in the automatic evaluation and outperforms the baseline model by 45% in the human evaluation.

Key words: multi-turn dialogue; dialogue policy; reinforcement learning

0 引言

随着社会媒体的发展，微博和 Twitter 等社交媒体上积累了大量的短文本，这些短文本可以近似作为对话语料来训练基于深度神经网络的对话生成模型[1]。基于深度神经网络的对话生成模型能够有效对输入产生回复，展现出巨大的研究潜力[2-4]。这些模型中得到广泛应用的是 seq2seq 模型。seq2seq 模型基于编码器—解码器结构，应用于对话生成任务时，输入通过编码器编码为一个特征向量，再由解码器根据特征向量解码得到回复。这一模型基于最大似然估计(maximum likelihood estimate, MLE)最大化回复的生成概率[5]。Shang L 等将该模型应用于单轮对话生成，取得了很好的效果[6]。

实际的对话过程在大多数情况下都是多轮交互的过程，而非一问一答的单轮对话。seq2seq 模型基于 MLE 的生成方式难以有效建模多轮对话的过程[7]。现有的关于多轮对话相关工作大多数基于填槽(slot filling)的方式进行[8-10]，这类工作的目标是完成特定任务；而在开放域多轮对话最近

收稿日期：2017-11-24 **定稿日期**：2017-12-15

基金项目：国家重点基础研究发展计划(2014CB340503)；国家自然科学基金(61502120,61472105)；哈尔滨学院青年科研基金(HUYF2013-002)；黑龙江省哲学社会科学研究规划项目(16TQD03)

的研究工作中,端到端的方式逐渐成为主流:基于大量的多轮对话训练数据,通过考虑最近的对话历史信息,使得生成的回复能够照应上文信息[11-13]。

尽管相关工作强调了历史对话信息,但是考虑对话历史而得到的回复并不一定有利于对话过程的持续进行。此外,对话生成任务中以seq2seq模型为代表的端到端生成方式基于MLE,每次得到的回复都是生成概率最大的句子。这就导致了那些经常出现但是没有实际意义的万能回复,比如中文里的"哈哈哈哈"以及英文里的"I don't know",非常容易被选中[2-3,14]。一旦出现输入与回复相同的情况,那么模拟的多轮对话就会陷入死循环;有很高生成概率的万能回复则很可能把模拟的多轮对话带入这样的死循环中。表1展示了这样两类问题。

表1 中文微博数据集上不同方法的多轮模拟对话结果对比

基于最大似然估计的模型	基于DQN的对话策略模型
A:元芳,(1) B:元芳你怎么看?(2) A:元芳,你怎么看?(3) B:元芳,你怎么看?(4) A:元芳,你怎么看?(5) B:元芳,你怎么看?(6) A:元芳,你怎么看?(7) ……	A:元芳,(1) B:元芳,你怎么了?(2) A:元芳,你怎么看?(3) B:元芳,我怎么看?(4) A:元芳你怎么看?(5) B:大人,此事必有蹊跷。(6) A:很有创意(7) ……
A:发现,还是杨坤这期最好看~(1) B:哈哈哈哈哈哈哈哈(2) A:太可爱了(3) B:可爱死了(4) A:太可爱了(5) ……	A:发现,还是杨坤这期最好看~(1) B:我觉得哈林唱的更好(2) A:哈哈,哈林唱得不错。(3) B:哈林唱得好(4) A:哈林唱得好!(5) ……

开放域多轮对话的一个重要目标就是尽可能聊得更久。开放域多轮对话中每一轮回复的选择,不仅需要考虑是否能够有效回复当前输入,更需要考虑是否有利于对话过程的持续进行。更好地建模多轮对话过程,引导现有的seq2seq模型有效进行多轮多话,需要从多轮对话过程的整体角度引入一种对话策略。

本文借助强化学习算法的全局视角,在开放域的多轮对话过程中引入了深度强化学习方法DQN[15]来进行对话策略学习,通过这个对话策略指导多轮对话过程中每一轮的回复选择。与MLE方式不同,强化学习的总体目标是最大化未来的累积奖励[16]。DQN方法估计的是每一个回复句子能够为给定的输入带来多少的未来奖励,对话的策略就是选择能够带来最大未来奖励的那个句子。如前所述,生成概率较低的句子并不意味着句子的质量差,很有可能只是因为这些句子出现频率没有万能回复那么高,相反,这些句子可能引入新的信息并更加有利于多轮对话的持续进行。因此,通过DQN方法进行对话策略学习能够有效挖掘seq2seq模型进行多轮对话的潜力。如表1所示,基于同样的输入,右侧根据DQN方法得到的对话策略进行的多轮对话质量明显更高。

本文的创新之处在于,将DQN应用于对话策略的学习过程中,使用独立的深度神经网络对每一句候选回复的未来收益进行评估,从而得到一个有利于多轮对话持续进行的对话策略。通过强化学习方法,DQN得到的深度神经网络就代表了多轮对话的策略,使得对话策略的学习独立于回复生成模型本身,在已有的回复生成模型不做任何改变的前提下,就能够通过DQN得到对话策略。实验结果表明,通过DQN方法得到的多轮对话策略有效提高了多轮对话的多样性、平均轮数和对话质量。

得到一个更好的多轮对话策略对于人机对话系统有着很多积极的意义。首先,人机对话系统的一种常见的训练方式就是通过用户模拟器(user simulator)来不断的进行模拟对话,生成式的用户模拟器需要能够有效地模拟多轮对话,因此更好

的多轮对话策略能够优化用户模拟器的回复效果，有利于训练出质量更高的对话模型。其次，在开放域对话系统中引入多轮对话策略能够有效提高回复整体上的多样性，使得回复内容更加丰富，并且能够引入更多的信息，将其应用到开放域的闲聊机器人中，对于提升用户的使用体验也有着积极作用。

1　相关工作

随着 Sutskever 等[5]提出序列到序列的学习方法，seq2seq 模型在最近几年开始广泛应用于对话生成研究领域[2,4,6]。深度强化学习是利用深度神经网络对强化学习方法做出的改进。Mnih 等[15]首先使用深度强化学习算法 DQN 在 Atari 游戏上取得突破性成功，其核心思想在于引入了经验回放(experience replay)机制。随后，Hasselt 等[17]，Schaul 等[18]，Wang 等[19]分别从不同的角度对 DQN 算法进行了改进。

与此同时，在对话系统的相关任务中引入深度强化学习方法也获得了越来越多研究者的关注。Guo H 将 DQN 算法应用到了 seq2seq 模型每个词语的解码过程中，从词语解码的级别上对模型做出了改进[20]。Li 等[7]使用了深度强化学习的策略梯度(policy gradient)算法，在 seq2seq 的训练过程中利用深度强化学习算法提供的梯度改变模型原有的训练进程，从而达到优化模型的目的。Su 等[10]结合强化学习和在线学习(online learning)的方式，通过与用户的实时交互，提高了任务型系统对话的性能。

开放域多轮对话方面，Lowe 等[11]利用 ubuntu 数据集进行了多轮对话的尝试，虽然利用了多轮对话数据集的优势，但是没有建模多轮对话的上下文信息。Pascual 等[12]考虑到了历史对话信息对于生成当前回复的影响，提出了一种能够感知历史信息的神经网络回复生成方法。Serban 等[13]在电影台词数据集上进行实验，同样考虑了上文对于当前回复生成的影响。开放域多轮对话系统的一个重要目标就是尽可能地使对话过程持续下去，然而这些工作都没有从如何回复才更有利于对话过程继续的角度考虑问题。Li 等[7]针对这一问题，利用深度强化学习策略梯度的方法建模了多轮对话过程，改进了 seq2seq 模型的训练过程。

在深度神经网络的相关优化方面，Graves 等[21]在神经网络解码过程中使用了束搜索(beam search)方法，用于平衡解码过程中的搜索质量与搜索开销；Bahdanau 等[22]提出了注意力模型，在深度神经网络中引入了一种动态赋权的机制；Srivastava 等[23]提出了 dropout 机制，用于防止深度神经网络训练时的过拟合。

与本文比较接近的工作是 Li 等[7]，本文关于多轮对话过程的强化学习建模方式以及实验结果评价指标也参考自这篇论文。虽然都是使用深度强化学习方法，但是本文与 Li 等[7]的区别也是显著的：首先，本文的目标是学习多轮对话的策略，而文献[7]的目标是进行对话生成。本文工作的核心是进行多轮对话的策略学习，使用的强化学习方法 DQN 没有涉及回复生成模型的训练过程，而是从已经训练好的回复生成模型中选择出最有利于多轮对话的回复；而文献[7]使用的是策略梯度方法，是在 seq2seq 模型的基础上改变回传前的梯度计算方式，从而引导回复生成模型的训练方向。其次，基于 Q-learning 的强化学习算法 DQN 和基于策略梯度的强化学习算法在原理和应用方式上完全不同，细节此处不再赘述。此外，本文使用了独立的深度神经网络来对回复的收益进行估计，最终学习得到的深度神经网络就代表了多轮对话的策略。

本文实验的总体结构如图 1 所示，从图中可以看出，DQN 的学习过程是独立于对话生成模型的，因此通过 DQN 进行对话策略的学习并不会改变基础的回复生成模型；也正是因为本文使用了独立的深度网络，所以对话策略的学习过程并不依赖于回复生成模型的梯度，即使应用场景改变，回复生成模型自身无法提供梯度，也不会影响本文对话策略学习方法的使用。

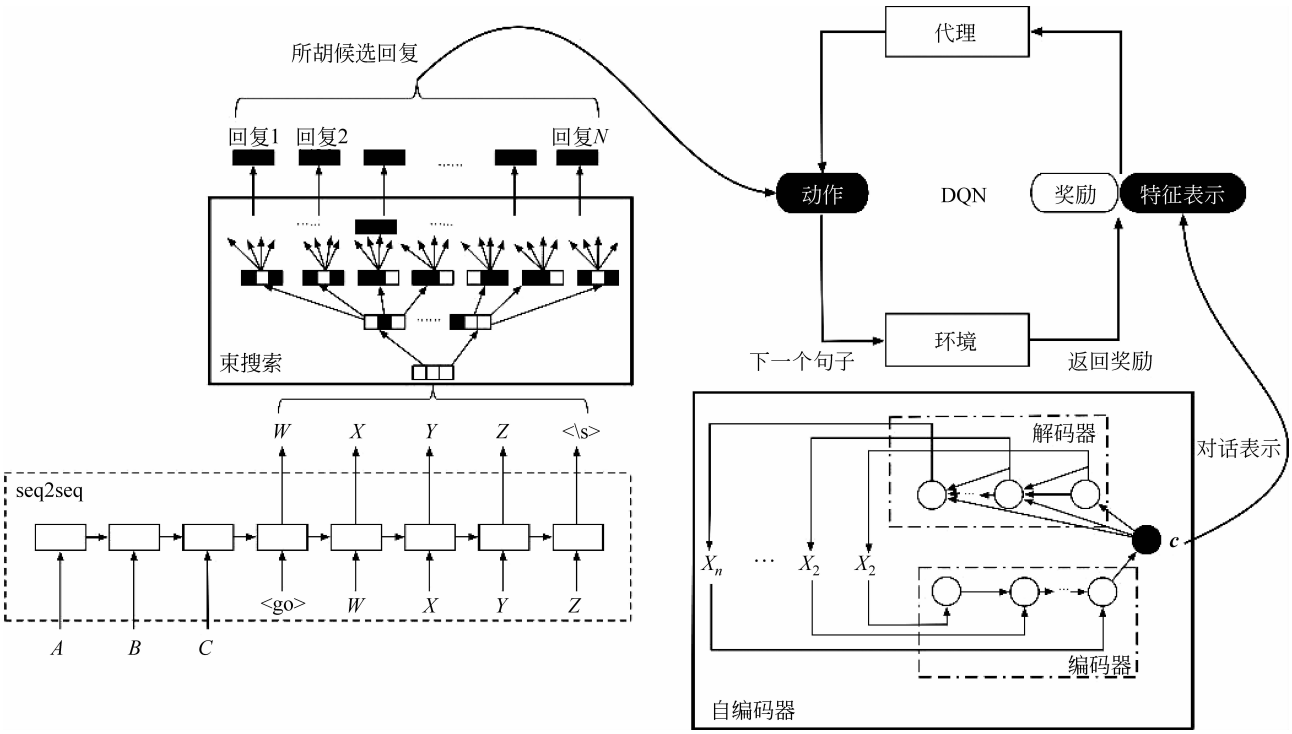

图1 本文实验的整体结构

2 DQN用于对话策略学习

DQN作为一种深度强化学习算法，基本结构仍然是“环境—代理”(environment-agent)框架：代理根据当前状态 s 选择一个动作 a 作用于环境，然后环境的状态 s 发生改变并返回相应的奖励 r，代理的目标是最大化未来能够获得的所有奖励之和，由此调整动作并构成一个循环过程。关于强化学习模型的更多细节请参考文献[16]。

本文的目标是得到一个有利于多轮对话持续进行的对话策略。这个总体目标分解到每一轮的对话中可以等价为每一轮都选择出能够为整个对话过程带来最大收益的句子。参考文献[7]，这里衡量回复带来的收益可以从是否产生万能回复、是否引入新的信息、是否与历史信息一致等方面来衡量。基于这些设定，就可以将对话策略的学习过程建模为典型的强化学习过程。

DQN方法的核心在于一个深度价值网络 Q，网络 Q 按照相应的算法迭代更新，目标是估计每个状态 s 下选择动作 a 的价值 $q(s,a)$。这个价值 q 代表了当前状态 s 下选择动作 a 能够带来的未来折扣奖励之和。更多细节和原理请参考强化学习有关Q-learning的部分。

在多轮对话过程中，把当前的输入语句记为 x，输入 x 通过回复生成模型以及束搜索得到了若干候选回复 $y_0, y_1, \ldots, y_n$。基于自然语言的句子是变长和离散的，无法作为状态 s 参与到网络 Q 的计算中，因此通过一个自编码器(autoencoder)将输入映射为固定维度的特征向量 $\boldsymbol{c}$，使用这个特征向量表达当前状态 s。同样，得到的回复也是句子，无法直接参与计算。但是回复与输入不同的地方在于，一旦回复生成模型和输入给定，那么这些候选项及其顺序也是确定的：本文中候选项的顺序是按照生成概率从高到低的顺序排列的。因此，表示回复并不需要对候选回复进行编码，只需要保留相应的序号即可。那么，一个候选回复 y_i 就有一个与之对应的动作 a_i，这个动作表示了这一轮选择 y_i 作为对输入 x 的回复。每一轮的对话中，通过深度价值网络 Q 对候选回复 y_i 进行评估，得到 $q(s_i,a_j)$。对话的策略就是选择 q 值最大的动作所对应的回复。

2.1 基础的回复生成模型

本文参考文献[5]中的seq2seq模型，训练了对话生成模型。在训练过程中加入了Bahdanau提出的注意力机制和Srivastava提出的dropout机制。

seq2seq的解码过程使用了束搜索的方法。束搜索在每一步中按照启发式规则保留最优的若干候

选项，其他较差的结点则被剪掉。在对话系统中，由于对话过程的灵活性和多样性，回复生成的搜索过程并没有一个确定的“最优解”。因此，在回复生成的解码过程中应用束搜索方法，对于得到更加多样的回复是有帮助的。

2.2　多轮对话模拟

所谓模拟对话，就是基于回复生成模型，通过两个代理的彼此对话来模拟多轮交互的过程。两个代理进行模拟对话的过程如下：一开始，从测试集中随机找到一句话输入给第一个代理，这个代理通过编码器网络把这个输入编码成一个隐层向量，然后通过解码器来生成回复。之后，第二个代理把前一个代理输出的回复同对话历史拼接，重新通过编码器网络编码得到一个隐层向量，更新对话的状态，然后通过解码器网络生成回复，并回传给第一个代理。这个过程不断地重复，直到达到最大的模拟对话轮数。具体过程如图 2 所示。

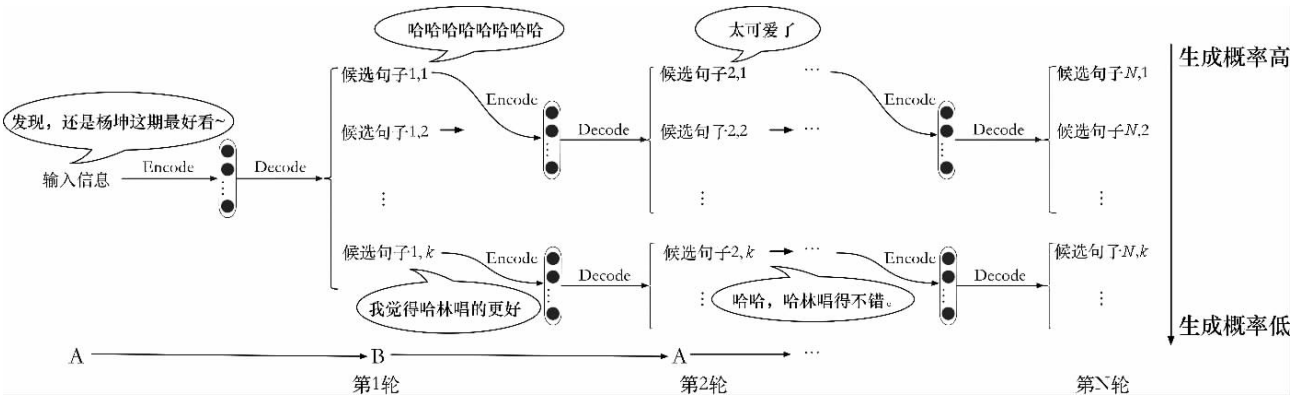

图 2　模拟多轮对话过程

2.3　自编码器

自编码器(autoencoder)是一类无监督的神经网络学习方法，其最大的特点是输入值和目标值相等。自编码器的目标是学习一种类似恒等映射的函数，如式(1)所示。

$$\varphi_{w,b}(x) \approx x \tag{1}$$

即输入数据到本身的一种映射函数。当自编码器学习得到这样一种恒等映射或者近似恒等映射的关系时，神经网络的隐层实际上就包含了数据的一种编码信息。

自编码器通常由编码器和解码器两部分组成。形式化地，编码器和解码器的作用可以定义为两个函数 ϕ 和 ψ，那么：

$$\phi: \chi \to \theta \tag{2}$$

$$\psi: \theta \to \chi \tag{3}$$

则自编码器的学习目标如式(4)所示。

$$\phi,\psi = \mathrm{argmin}_{\phi,\psi} \| X-\psi(\phi(X)) \|^2 \tag{4}$$

其中，式(2)和式(3)中的 θ 就是自编码器得到的数据特征表示。

用于对话表示学习的自编码器需要处理变长的句子，并且由于句子数量巨大，所以需要学习的映射关系 ϕ 和 ψ 都会非常复杂，因此本文用到的自编码器的编码器和解码器都由循环神经网络构成。由于自编码器的学习过程是无监督的，所以自编码器在任意语料上都能够学习得到该语料中句子的特征表示。

自编码器的输入是已经分词的句子，每个词由其词向量表示，由循环神经网络依次读入；输入句子经过编码器的编码，得到一个中间表示 c，这个表示再送入解码器中，解码出预测值；网络学习的目标由式(4)定义，网络的目标值是和输入句子完全相同的句子，目标值和预测值之间的误差就是网络参数调整的依据。

句子的自编码器虽然只能做到近似复原输入，但是这并不影响中间特征在强化学习模型中的使用。因为这些特征向量只需要按照同样的规则得到，处于同一个向量空间就可以满足需要。

2.4　DQN 模型训练

先定义奖励函数。奖励函数的作用是引导对话向轮数更多、信息更丰富、万能回复更少的方向进行。首先定义一个表示，如式(5)所示。

$$\mathrm{logprob}(s_1 \mid s_2) = \frac{1}{N_{s_1}} \mathrm{logP}_{\mathrm{seq2seq}}(s_1 \mid s_2) \tag{5}$$

式(5)表达的含义是对于句子 s_2，在给定的 seq2seq 模型下生成句子 s_1 的对数概率，并且该对

数概率受到 s_1 中词数的 N_{s_1} 的约束。参考文献[7]，根据多轮对话的总体目标共定义三个奖励函数：

对于定义的万能回复的集合 S，惩罚动作 a 可能导致生成的万能回复，如式(6)所示。

$$r_1 = -\frac{1}{N_S}\sum_{s\in S}\text{logprob}(s \mid a) \tag{6}$$

对于连续的两个对话状态 h_i 和 h_{i+1}，惩罚对话状态过于接近，奖励对话状态存在较大差别以引入新的信息，如式(7)所示。

$$r_2 = -\cos(h_i, h_{i+1}) \tag{7}$$

对于连续的多轮对话 p_i、q_i 和 a，奖励使得对话前后连贯的动作 a，惩罚使对话不连贯的动作 a，如式(8)所示。

$$r_3 = \text{logprob}(a \mid p_i, q_i) + \text{logprob}(q_i \mid a) \tag{8}$$

最终的奖励值式(9)所示。

$$r = 0.45r_1 + 0.2r_2 + 0.35r_3 \tag{9}$$

三个系数在实验过程中调整得到。

DQN 中深度神经网络的参数更新通过对式(10)进行随机梯度下降来完成。其中 s_j 表示状态，a_j 表示动作，Q 表示通过以 θ 为参数的深度价值网络对状态—动作对进行估值，如式(10)所示。

$$\text{Loss} = (y_j - Q(s_j, a_j; \theta))^2 \tag{10}$$

其中，价值的估计通过式(11)来完成，公式中的 r_j 表示奖励：

$$y_j = \begin{cases} r_j, & \text{对于终止状态 } s_{j+1} \\ r_j + \gamma \max_{a'} Q(s_{j+1}, a'; \theta), & \text{对于非终止状态 } s_{j+1} \end{cases} \tag{11}$$

本文 DQN 算法参考文献[15]的算法实现，更多实现细节请参考原文。

3 实验

3.1 训练数据及其表示

本文的实验在中文微博语料上进行。该语料来源于新浪微博，每一对对话数据分别来自微博的正文和这条微博下面的评论，这样一组博文—评论对就近似构成了一组对话对。该数据集总共有约 110 万组这样的对话对，语言质量较高，同时该语料也是文献[6]所使用的数据集。本文通过 word embedding 的方式将每一个词语都转化为一个固定维度的向量，实验中取 300 维，并通过所有词语的向量共同表示原始的句子。使用的训练词向量的工具包是 Google 开源的 Word2Vec，具有配置简单、训练高效等优点。

3.2 评价指标

对于多轮对话的实验结果，本文参考 Li 等[7]的方法，使用以下两个客观指标进行评价。

(1) 平均对话轮数。对话轮数是指从输入到对话结束总共持续的对话轮数。当对话过程出现了类似“哈哈哈”这类的事先定义的万能回复或者对话进入一个死循环当中，那么就认为对话过程已经结束。

(2) 多样性。多样性通过统计模拟对话过程中出现的不重复的一元文法(unigram)和二元文法(bigram)所占的比例来衡量。unigram 和 bigram 是语言模型中的概念，这一指标能够表征总体生成结果的语言丰富程度。因为多轮对话并不存在一个标准答案，所以 BLEU 等传统方法不适用于该任务的评价。同时，进行主观评价是对话系统中常用的方式。本文还进行了主观的对比评价：给定相同的输入，本文的 DQN 模型和基础模型独立进行单轮回复生成以及进行模拟对话，然后由与实验无关的评价者主观对比二者的质量，结果分为 DQN 更好、DQN 更差和平局三种情况。评价者不知道结果所对应的方法，只是给出二者对比的一种主观评价。对于单轮对话，总共收到 400 组的对比打分；对于多轮对话，总共收到 600 组对比打分。

3.3 回复生成及对话模拟

中文微博语料的单轮回复生成结果如表 2 所示。左侧是从测试集中随机抽取的输入，右侧是对应于输入产生的回复，每一行组成一个输入—回复对。

表 2 单轮回复生成结果

输入信息	生成的回复
论小时候喝牛奶的重要性	这是牛奶吗？
你有多久没过暑假了？	半年了
为什么每个路口都是红灯	注意安全！
看来想坑喵星人是行不通了→_→	好可爱的猫咪
这空气质量，环保部也受不了了……	污染严重。

可以看到，就单轮回复的效果而言，seq2seq 模型能生成质量较高的回复语句。

中文模拟对话结果如表 3 所示。表中的一列为一组模拟对话的结果，第一句输入从测试集随机抽取，其余句子都根据上一句话生成得到。

表 3　中文模拟对话的结果

模拟对话 1	模拟对话 2
A：不足够自信上帝凭什么把机会给你 B：有道理！ A：有道理！ B：有道理！ A：有道理！ ……	A：成都加菲猫走红 B：真好！ A：有点意思 B：哈哈哈哈哈哈哈哈 A：哈哈哈哈哈哈哈哈 B：哈哈哈哈哈哈哈哈 ……

可以看到，seq2seq 模型和预期的结果一样，无法有效进行多轮对话。模拟对话很快就陷入了死循环中，并且容易产生万能回复。

从测试集句子中随机抽取 5 000 条进行模拟对话，客观指标评价结果如表 4 所示。可以看出，平均对话轮数比较短，说明基础的 seq2seq 模型难以有效进行多轮对话。

表 4　seq2seq 模型模拟对话结果评价

	对话轮数/轮	unigram	bigram
中文微博数据集	2.51	0.021	0.037

3.4　自编码器

就自编码器而言，训练的目标是输出与输入一致，或者输出尽量接近输入。我们从测试数据中随机抽取 2 000 条句子送入自编码器得到输出，然后通过计算输出对于输入的 F 值的方式来衡量句子映射到自身的效果。评价时逐句计算输出对于输入的召回词数以及输入和输出各自的词数，所有句子累加到一起得到总的召回值、预测值和标准答案词数。实验结果如表 5 所示。本文实验中的自编码器最终的 F 值为 0.872，在很大程度上已经能够将输入句子映射为句子本身；此外，自编码器从输入得到输出的映射关系并不会被使用，本文用到的是从编码器传递到解码器的中间特征向量，这个向量实际上是对于输入信息的编码。因此即使 F 值无法达到理想情况下的 1.0，也不会影响中间特征向量作为一种编码被使用。

表 5　自编码器的 F 值

	召回值	预测值	标准答案
词数	12 203	13 923	14 072
	准确率	召回率	F 值
得分	0.867	0.876	0.872

3.5　基于 DQN 的对话策略模型

本文使用 DQN 模型在测试集上随机抽取了 5 000 条句子进行模拟对话。相应的客观评价结果与基础的 seq2seq 模型对比如表 6 所示。

表 6　DQN 和基础模型客观评价指标对比

	对话轮数/轮	unigram	bigram
seq2seq(Baseline)	2.51	0.021	0.037
DQN	4.56	0.066	0.137

可以看到，客观指标上 DQN 模型显著优于基础模型，平均对话轮数更是增加了二轮。分析其原因，基础模型对话容易进入死循环，并且倾向于生成经常出现的回复，而 DQN 模型通过奖励函数的设置在一定程度上缓解了这些问题，所以 DQN 模型在对话轮数以及多样性上的表现都要明显比基础模型好。

主观评价的结果如表 7 所示。主观评价指标方面，单轮回复的质量总体来说不相上下，DQN 略占优势，平局的情况占了接近 20%；多轮对话上 DQN 模型的优势明显，平局的比例也下降到了不到 10%。可以看出，因为单轮对话基于相同的回复生成模型，所以总体上单轮回复质量比较接近；而在多轮对话上，DQN 模型则明显优于基础模型。

表 7　DQN 和基础模型对话质量人工对比评价/%

	DQN 赢	DQN 输	平局
单轮回复质量	43.5	37	19.5
多轮对话质量	68.3	23.3	8.3

部分单轮对话和多轮对话的效果对比如表 8、表 9 所示。

表 8　DQN 和基础模型单轮对话对比

输入	seq2seq(Baseline)	DQN 模型
日落巴塞罗那金色海岸，浪漫色彩。	真美！	美得让人窒息。
彩虹和闪电交汇，非常难得的瞬间。	真美！	抓拍得真好！
两种广告，你选哪种？	很有创意！	二者皆可抛。

续表

输入	seq2seq(Baseline)	DQN 模型
老段子,但我还是笑了。	笑死了	我笑了
人际交往的 3 个超经典的潜规则	很有道理	很不错的说
录音棚学歌中	忐忑唱忐忑	唱忐忑了
苹果明年或推 iOS 电视机	有时间看看	30 寸苹果的 iphone
流言止于智者,聊天止于呵呵～	支持 o(ノ□ヽ)o	呵呵,呵呵呵
二师兄也曾有过青葱岁月	真好!	好可爱的猪猪～

表 9 DQN 和基础模型多轮对话对比

seq2seq baseline	DQN 模型
A:不足够自信上帝凭什么把机会给你 B:有道理! A:有道理! B:有道理! A:有道理! B:有道理! …	A:不足够自信上帝凭什么把机会给你 B:这句话很有道理 A:说的很好 B:不会吧? A:不是一般的好 B:真的假的! …
A:成都加菲猫走红 B:真好! A:有点意思 B:哈哈哈哈哈哈哈哈 A:哈哈哈哈哈哈哈哈 B:哈哈哈哈哈哈哈哈 …	A:成都加菲猫走红 B:好可爱哦 A:太可爱了! B:好 Q 啊 A:真的吗? B:太有才了!!! A:哈哈哈哈哈哈哈哈 …

基础模型在解码过程中基于 MLE,最终选择的总是束搜索结果中生成概率最大那个句子。但是生成概率高并不意味着最有利于对话过程,如图 2 所示,候选句子按照生成概率降序排列,因此基于 MLE 的选择方式总是会选择第一个候选句子。从图中可以看到,生成概率较低的“我觉得哈林唱得更好”作为回复明显效果更好;但是因为其生成概率不如类似“哈哈哈”这种经常出现的“万能回复”的生成概率高,所以导致这个更优的结果被 MLE 忽略。同时也可以看到,在多轮对话过程中,第一轮选择的不同直接导致了后续回复生成的不同,从而进一步拉大多轮对话质量的差距。

与 MLE 的方式不同,本文的 DQN 模型每次选择的是价值最大的动作,这一动作最有利于对话的持续进行,并能够带来更加丰富的信息。如果生成概率较低的候选回复在深度价值网络中具有最大的价值,那么这个回复也会被选中,因此 DQN 模型更容易选择出有利于对话过程持续进行的回复。

记录 DQN 和 MLE 每次选择的动作,画出表 1 中多轮对话 2 的动作选择,如图 3 所示,能够更加直观地说明 DQN 相较于 MLE 如何选择潜在的更优句子。动作 1 是选择生成概率最高的句子,动作 10 是选择生成概率最低的句子。值得注意的是,在第一轮中两种方式选择了不同的动作,那么在后续几轮中,两种方式各自进入了不同的状态,所以同一个动作也会对应不同的回复。

从图 3 可以明显地看到,MLE 每次选择的动作都是 1,因为这是生成概率最大的句子,所有动作构成的路径是一条直线;而 DQN 方法选择动作构成的路径则更加“曲折”,每次选择的是通过深度价值网络估计得到的当前状态下价值最大的动作,这个价值的最大则结合了奖励函数的定义,使得这个动作有利于对话过程朝着轮数更多、包含的信息更丰富、更少的生成万能回复的方向进行。图 4 给出了随机采样的六组动作选择路径的对比。

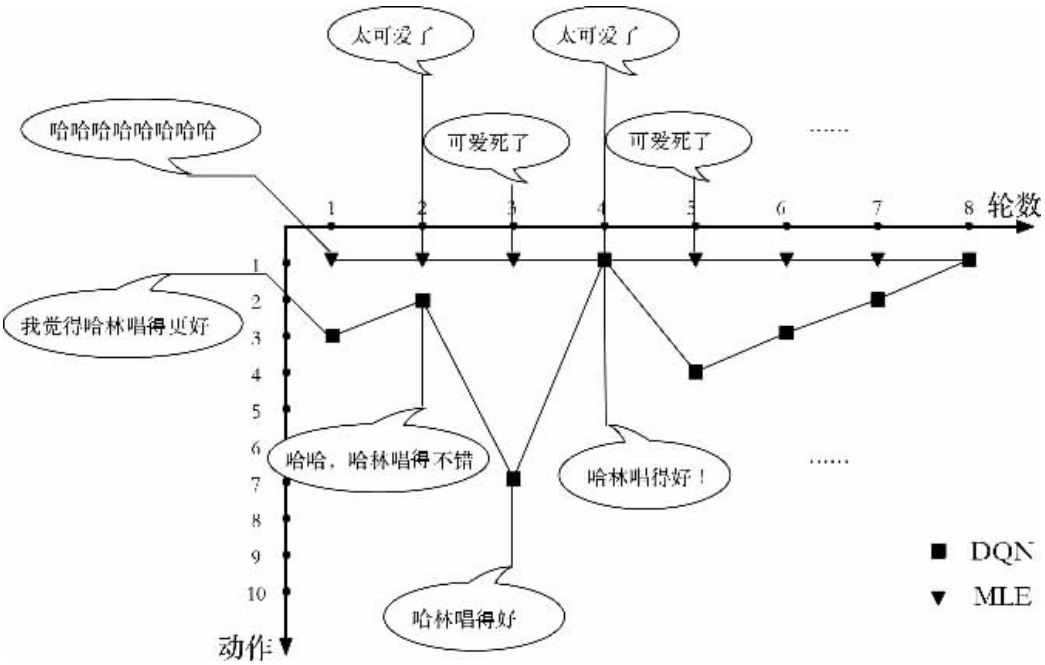

图 3　多轮对话过程中的动作选择对比(表 1 多轮对话 2)

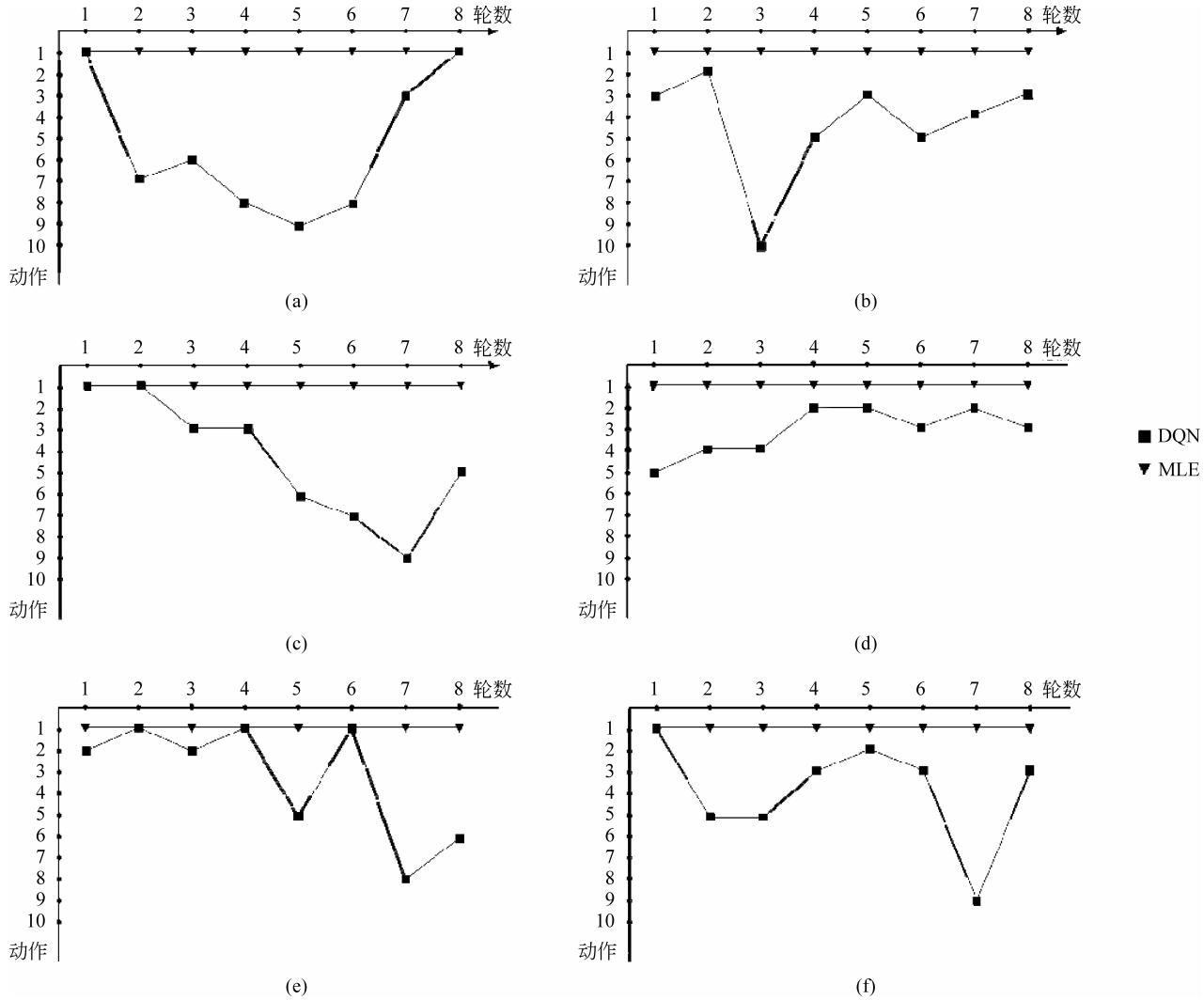

图 4　随机采样的六组动作选择对比

4 结论及后续工作

本文的主要研究内容针对开放的域多轮对话，关注之前工作中存在的没有建模整个对话过程、多轮对话中容易产生大量万能回复、很快陷入死循环等问题，引入了DQN方法进行对话策略学习。通过DQN对每一轮的候选回复的未来价值进行评估，选择出了最有利于对话持续进行的句子作为回复，减少了万能回复的产生，并增加了平均对话轮数。

实验结果表明，在单轮回复质量几乎持平的前提下，本文使用DQN方法得到的多轮对话策略能够优化多轮对话的质量，最终各个评价指标上都明显优于基础方法：平均对话轮数提高了二轮，主观评价获胜比例高出了接近45%。

本文的后续工作将着眼于将DQN用于seq2seq模型的训练过程，使用深度价值网络来估计训练过程中的损失，使得训练损失带有更多的信息，从更细粒度上提高生成句子的质量。此外，如何更加全面地评价对话结果也是一个值得研究的问题。

参考文献

[1] Serban I V, Lowe R, Charlin L, et al. A survey of available corpora for building data-driven dialogue systems[J]. arXiv preprint, 2015, arXiv: 1512.05742.

[2] Vinyals O, Le Q. A neural conversational model[J]. arXiv preprint, 2015, arXiv: 1506.05869.

[3] Luan Y, Ji Y, Ostendorf M. LSTM based conversation models[J]. arXiv preprint, 2016, arXiv: 1603.09457.

[4] Yao K, Peng B, Zweig G, et al. An attentional neural conversation model with improved specificity[J]. arXiv preprint, 2016, arXiv: 1606.01292.

[5] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the Advances in neural Information Processing Systems. 2014: 3104-3112.

[6] Shang L, Lu Z, Li H. Neural responding machine for-short-text conversation[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 2015: 1577-1586.

[7] Li J, Monroe W, Ritter A, et al. Deep reinforcement learning for dialogue Generation[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 1192-1202.

[8] Williams J D, Raux A, Ramachandran D, et al. The dialog state tracking challenge[C]//Proceedings of SIGDIAL Conference, 2013: 404-413.

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

[10] Su P H, Gasic M, Mrksic N, et al. On line active Reward Learning for policy optimisation in spoken dialogue systems[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2016: 2431-2441.

[11] Lowe R, Pow N, Serban I, et al. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems[J]. arXiv preprint, 2015, arXiv: 1506.08909.

[12] Pascual B, Gurruchaga M, Ginebra M P, et al. A neural network approach to context-sensitive generation of conversational responses[J]. Transactions of the Royal Society of Tropical Medicine and Hygiene, 2015, 51(6): 502-504.

[13] Serban I V, Sordoni A, Bengio Y, et al. Hierarchical neural network generative models for movie dialogues [J]. arXiv preprint, 2015, arXiv: 1507.04808.

[14] Li J, Galley M, Brockett C, et al. A diversity-promoting objective function for neural conversation models[C]//Proceedings of NAACL-HLT, 2016: 110-119.

[15] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint, 2013, arXiv: 1312.5602.

[16] Sutton R S, Barto A G. Reinforcement learning: An introduction[J]. IEEE Transactions on Neural Networks, 2005, 16(1): 285-286.

[17] Hasselt H V, Guez A, Silver D. Deep reinforcement learning with double q-learning[J]. arXiv preprint, 2015, arXiv: 1509.06461.

[18] Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay[J]. arXiv preprint, 2015, arXiv: 1511.05952.

[19] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning[J]. arXiv preprint, 2015, arXiv: 1511.06581.

[20] Guo H. Generating text with deep reinforcement learning[J]. arXiv preprint, 2015, arXiv: 1510.09202.

[21] Graves A. Sequence transduction with recurrent neural networks[J]. Computer Science, 2012, 58(3): 235-242.

（下转第136页）

吴育锋(1994—),本科,主要研究领域为计算语言学。
E-mail: 33120142201411@stu. xmu. edu. cn

吴胜涛(1980—),博士,副教授,主要研究领域为文化价值观、社会情绪、大数据分析等。
E-mail: michaelstwu@xmu. edu. cn

朱廷劭(1971—),通信作者,博士,研究员,主要研究领域为机器学习、大数据心理。
E-mail: tszhu@psych. ac. cn

(上接第 108 页)

[22] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint, 2014, arXiv: 1409.0473.

[23] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

宋皓宇(1994—),硕士研究生,主要研究领域为人机对话系统和自然语言处理。
E-mail: hysong@ir. hit. edu. cn

张伟男(1985—),博士,讲师,主要研究领域为人机对话系统、自然语言处理和信息检索。
E-mail: wnzhang@ir. hit. edu. cn

刘挺(1972—),博士,教授,主要研究领域为自然语言处理、文本挖掘和文本检索。
E-mail: tliu@ir. hit. edu. cn