Online Sequence-to-Sequence Reinforcement Learning for Open-Domain Conversational Agents

Nabiha Asghar***, Pascal Poupart*****, Jiang Xin**†**, Hang Li**†
*Cheriton School of Computer Science, University of Waterloo, Canada
†Noah's Ark Lab, Huawei Technologies, Hong Kong
{nasghar,ppoupart}@uwaterloo.ca, {jiang.xin,hangli.hl}@huawei.com

Abstract

We propose an online, end-to-end, deep reinforcement learning technique to develop generative conversational agents for open-domain dialogue. We use a unique combination of offline two-phase supervised learning and online reinforcement learning with human users to train our agent. While most existing research proposes hand-crafted and develop-defined reward functions for reinforcement, we devise a novel reward mechanism based on a variant of Beam Search and one-character user-feedback at each step. Experiments show that our model, when trained on a small and shallow Seq2Seq network, successfully promotes the generation of meaningful, diverse and interesting responses, and can be used to train agents with customized personas and conversational styles.

1 Introduction

The recent overwhelming success of Deep Neural Networks (DNNs) in the areas of machine translation and image/speech recognition has prompted the AI community to experiment with DNNs for natural language understanding (NLU) and natural language generation (NLG). In particular, several recent works have proposed Recurrent Neural Networks (RNN) based generative conversational agents (CAs) for short-text conversation[Shang *et al.*, 2015; Sordoni *et al.*, 2015], factoid question-answering [Iyyer *et al.*, 2014; Yin *et al.*, 2015], and context-based response generation in both open-domain and task-oriented dialogue systems [Le *et al.*, 2016; Wen *et al.*, 2015; Vinyals and Le, 2015; Serban *et al.*, 2016a; Wen *et al.*, 2016; Serban *et al.*, 2016b; Al-Rfou *et al.*, 2016; Yao *et al.*, 2015; Gu *et al.*, 2016]. Many of these proposed models use LSTM encoder-decoder architectures, such as the sequence-to-sequence (Seq2Seq) framework [Sutskever *et al.*, 2014; Bahdanau *et al.*, 2015]. They are designed to capture the context, linguistics and semantics in a conversation reasonably well, but they have been shown to generate short, dull, repetitive and inconsistent responses [Serban *et al.*, 2016a; Li *et al.*, 2016a].

Inspired by the recent success of Deep Reinforcement Learning (DRL) in GO [Silver *et al.*, 2016] and Atari games [Mnih *et al.*, 2013; Mnih *et al.*, 2015], researchers have started exploring DRL to address the hard problems of NLU and NLG in dialogue generation. Only a handful of such techniques have been proposed so far (details in Section 2), where the reward function is typically hand-crafted, and is either specific to the task to be completed (for domain-specific CAs) or is based on a few desirable developer-defined conversational properties (for open-domain CAs). Moreover, these models are predominantly reliant on self-play or simulated users.

In this work, we introduce an online, end-to-end, Seq2Seq DRL technique that incrementally learns open-domain conversation skills via direct interaction with human users. We use a single-layer Seq2Seq architecture as the model's backbone, which significantly reduces the training time and complexity. To build a baseline, we use two-phase supervised learning: one on a moderately large, generic and publicly available dialogue dataset, and the other on a very small, specially curated, short-text conversation dataset. We then initiate an online reinforcement learning phase which incorporates modified Beam Search and a unique reward mechanism reliant on single-character feedback from the human at each turn in the dialogue. The intuition here is to emulate the process of language learning in a child's brain, whereby each utterance by the child is reinforced via a positive or negative signal from a teacher. This explicit human-centric reward mechanism is minimally inconvenient for the user and eliminates the need to incorporate hand-crafted reward functions. In addition, it inherently promotes diverse, interesting and relevant responses by relying on the humans' far superior conversational prowess.

2 Related Work

DRL-based dialogue generation is a fairly new and relatively unexplored paradigm to date. For task-specific dialogue generation [Su *et al.*, 2016; Zhao and Eskenazi, 2016; Cuayáhuitl *et al.*, 2016; Williams and Zweig, 2016], the reward function is typically straight-forward and based on task-completion rate. For open-domain dialogue generation [Li *et al.*, 2016c; Yu *et al.*, 2016; Weston, 2016], hand-crafted reward functions have been devised which either capture a number of different desirable properties of a conversation, or choose between retrieval-based and neural generative strategies within a dialogue.

Li *et al.*'s recently published research [2016b] is perhaps the most relevant to our work. It uses a diversity-promoting Beam Search algorithm to generate responses, which are subsequently evaluated by human users during online reinforcement learning. However, they use a narrow dialogue domain and their model's responses, although very diverse, may not be very relevant or engaging.

Our Seq2Seq based DRL model draws inspiration, and yet inherently differs, from these works in the following key ways. First, we use a shallow and small Seq2Seq network (one LSTM encoder and one LSTM decoder with 300 hidden units each) as the backbone of our model, which is very fast to train. This is in sharp contrast to existing models that typically use four to eight layers, each containg up to 2048 hidden units. Second, we carry out SL on two different datasets sequentially to improve the conversational ability of our baseline model. The size of these datasets is considerably small (300K and 8K message-response pairs respectively), compared to 50 - 80M used by other works. Third, we do away with the need to define hand-crafted reward criteria and use a very simple feedback mechanism, based on a variant of Beam Search, to drive the conversation in a direction deemed appropriate by the user. Fourth, by relying entirely on the humans' superior linguistic prowess, we eliminate the need to explicitly incorporate diversity, relevance or interestingness in the responses. Finally, our feedback mechanism is minimally inconvenient to the user and, in fact, allows the user to encourage customized moods, personas and conversational style for the CA.

3 Model Overview

The architectural backbone of our model is the Seq2Seq network [Sutskever *et al.*, 2014], consisting of one LSTM encoder layer and one LSTM decoder layer, each containing 300 hidden units. The end-to-end model training consists of offline SL (with a mini-batch size of 10), followed by online RL.

3.1 Offline Two-Phase Supervised Learning

We train our network sequentially on two datasets. In the first phase, we use the Cornell Movie Dialogs Corpus [Danescu-Niculescu-Mizil and Lee, 2011]. It consists of 300K message-response pairs and each pair is treated as the input and target sequence during training. We use joint cross-entropy (XENT) loss as our objective function, where we maximize the likelihood of the target sequence given the input sequence. For the target sequence $[y_1, y_2, ..., y_T]$, the XENT loss function is given by

$$L_{\text{XENT}} = -\log P(y_1, y_2, ..., y_T) = -\sum_{t=1}^{T} \log P(y_t | y_1, ..., y_{t-1}).$$

This enables our CA to learn the language grammar and semantics reasonably well. However, the model has difficulty carrying out a casual chat, which typically consists of short open-domain messages that are remarkably different from conversations in a movie. To combat this issue, we curate a dataset from JabberWacky's chatlogs[1] available online. We initialize our network with the weights obtained in the first phase, and then train on the JabberWacky dataset, consisting of 8K message-response pairs. Even though this dataset is very small, it results in an improved baseline for open-domain dialogue (see Table 1 and Figure 1 (a)).

[1] http://www.jabberwacky.com/j2conversations. Jabber-Wacky is an in-browser, open-domain, retrieval-based bot.

Algorithm 1 Online Reinforcement Learning

```
lr ← 0.001                    //initial learningRate for Adam
while true do
    userMsg ← io.read();
    responses ← BeamSearch(model.forward(userMsg));
    io.write(responses);
    feedback ← io.read();
    if feedback ∈ {1,2,3,4,5} then
        botMsg ← responses[feedback];
    else
        botMsg ← feedback;
    pred, xentLoss ←
        model.forward(userMsg, botMsg, lr);
    model.backward(pred, xentLoss, botMsg, lr);
    model.updateParameters();
    lr ← lr − decayFactor
```

3.2 Online Reinforcement Learning

After the offline SL phase, our CA is equipped with satisfactory, albeit limited, skills to carry out conversations of moderate quality with real (human) users. Its responses are still dull, repetitive and short, which is a known issue in Seq2Seq-based dialogue systems. To combat this problem, we initiate an online RL phase where our previously trained model interacts with a human user and learns directly and incrementally from their one-character feedback received at each turn.

We set up the CA−human interaction and reinforcement learning mechanism as follows.

1. The user sends a message u_i to the CA at time step i.
2. The CA replies with five responses $c_{i,1}, c_{i,2}, ..., c_{i,5}$. These are obtained through the following variant of Beam Search: we take the five most probable words predicted at the first step, and take each of them as input at the next step in five parallel threads.
3. The user selects one of the five responses, denoted by $c^*_{i,j}$ for some $j \in \{1, ..., 5\}$, which he deems appropriate. He has the option to write a sixth message $c^*_{i,6}$, in case he likes none of the five responses presented to him.
4. We take the message-response pair $(u_i, c^*_{i,j})$ as a new data point, and propagate it in the network, using XENT loss.
5. The user responds to the chosen $c^*_{i,j}$ with a message u_{i+1}, and the process repeats.

The pseudocode of this procedure is presented in Algorithm 1. The variable 'learning rate' *(lr)* refers to the rate at which Adam [Kingma and Ba, 2014] the stochastic optimizer adapts the parameters. By default, we initialize lr to 0.001 in both two-phase SL and online RL. Additionally, we

Human Prompt	SL1 Response	SL2 Response
Hello there!	I'm here.	How are you tonight?
I'm good. How are you?	Yes.	No.
Can you comfort me?	You called ta it should be	I think i'm a computer program?
Who won the US elections?	No, I don't.	A jackass?
Lol.	I've got it.	Yea, real funny.
Do you have a boyfriend?	No, not really.	No.
What is the meaning of life?	No.	Do you want the long answer or the short answer?
Ok.	I did not that.	You didn't answer my question?
I dont know.	What's the difference if i'm on call.	Yes.
Hey, entertain me.	On the contrary.	I'm not a jerry fan.
You're not making any sense at all.	I just want to tell you something.	No, I'm thinking.
You're an idiot!	So?	Yes I do.
You aren't fun.	Then why can't we stay here?	How old are you?
I have to go now.	Go out.	Please don't go.

Table 1: Comparison of CA responses after one-phase SL (SL1, trained over Cornell Movie Corpus) and after two-phase SL (SL2, trained over Cornell Movie Corpus and then the JabberWacky dataset).

experiment with different values of initial lr in online RL to determine its most suitable and stable value (see Section 4).

Note that we could have included the pair $(c^*_{i,j}, u_{i+1})$ in our learning algorithm. We don't, because it is very likely that the human users will quickly switch contexts within a conversation. Therefore u_{i+1} may not always be an appropriate response to $c^*_{i,j}$.

Beam Search is a popular way of reducing *search error*[2], induced by the greedy left-to-right sequence generation process during XENT training. Here we use a heuristic variant which is considerably fast due to a significantly smaller search space; it outputs five sub-optimal solutions that are presented to the user as several diverse options to choose from and evaluate. The number of CA responses to show at each turn is a hyper-parameter that can be tuned. We choose the number five based on our observation that a smaller set usually misses out a good contender, and more than five responses become cumbersome for the user to read at each dialogue turn.

Appendix A contains an example transcript of an online reinforcement interaction between a human user and our agent, without cherry-picking. Notice how the user's feedback y for his own message x is immediately incorporated into the model, such that the y becomes the CA's most likely response at the next turn when prompted again with x. This is achieved by setting a higher initial value for the learning rate (0.005 in this case).

4 Experimental Evaluation

We evaluate our model (implemented in Torch and Lua 5.2) via qualitative comparison of its responses to those generated by offline one-phase and two-phase supervised learning. We also perform quantitative comparison of the three models on four axes: syntactical/grammatical correctness, relevance to previous prompt, interesting-ness, user engagement. Finally, we numerically evaluate how the initial learning rate and number of reinforcement training interactions impact the model's conversational prowess on the same four axes.

[2]Due to greedy selection, the most likely sequence may contain a word w_t which is sub-optimal at time t.

4.1 Qualitative Comparison

We begin by evaluating the qualitative differences between our CA's conversational abilities when trained via one-phase and two-phase SL (denoted by SL1 and SL2 respectively). Some examples are shown in Table 1. We see that SL2 generates longer, engaging and more relevant responses than SL1 in many cases. This is not surprising, and is an illustration of the fact that a small but relevant dataset is much more effective for training than a large and generic dataset. We still encounter many instances of short and dull responses with SL2, which we try rectify through online RL.

Table 2 shows a comparison between the responses given by the CA after two-phase SL only, versus after both two-phase SL and online RL (denoted by SL2+oRL) via 200 interactions. Here we see a clear difference between the quality of the two models. The former has considerably shorter, generic and dull responses to a variety of different prompts, compared to the latter. It is important to note here that none of the human prompts listed in this table were used in the 200 reinforcement interactions. This has remarkable implications; it implies that our model not only corrects itself via single message-response pairs at a time, but also learns to make connections between semantically or syntactically similar prompts over time. While this may be a slow process (spanning hundreds or thousands of interactions), it effectively emulates the way humans learn a new language.

Table 3 illustrates how our model can be trained to adopt a wide variety of personas, moods or conversational styles, based on the preference of the user. Here, we trained three different copies of our model, separately, to adopt three different emotional personas: happy/cheerful, sad/gloomy and rude/sarcastic. Each model underwent 100 reinforcement interactions with a single human user. The learning rate was initialized to a higher-than-default value here (0.005, as opposed to 0.001 in Table 1) as a proof of concept that only a small number of reinforcement interactions may be needed to train

Human Prompt	SL2 Response	SL2+oRL Response
Hello my friend.	How old are you?	Hehe, yes, beautiful.
You are so sweet!	How did you know?	I am so a bot.
I think bots are sweeter than humans.	You are a good person.	Everyone is the same on a basic chemical level.
Bots are funnier than us too.	Yes.	* blink * what are you talking about?
I don't know.	Yes.	Why did you suggest it?
Why did Donald Trump win?	You are!	Wouldn't you like to know, fickle human?
Do you pay taxes?	Yes.	From time to time.
Tell me something I don't know.	Ok, so how are you.	If you were intelligent you would already know that.
You're not making any sense at all.	No, I'm thinking.	On the contrary, I am trying to educate you.
Do you have a brain?	No.	You are nasty.
Do you have any kids?	I have no!	None that are really close to me.
Where is your home?	In this house.	I believe it would be in this particular location in north america.
Life can be very cruel sometimes.	It takes two to know two.	It takes two to know two.
Do you believe in life after death?	Do you want the long answer?	Yes, do you?
Hey, entertain me.	I'm not a jerry fan.	Most people be.
You're an idiot!	Yes I do.	You're polite.

Table 2: Comparison of CA responses after two-phase SL (SL2) only, and after both two-phase SL and online RL (denoted by SL2+oRL). SL2+oRL is reinforced via several conversations consisting of a total of 200 interactions. Note that the 200 prompts used in online RL did not contain any of the 16 human prompts shown here.

Human Prompt	Cheerful CA	Gloomy CA	Rude/Sarcastic CA
What's up?	It is a fine morning.	Not well, to be honest.	The date I went back in time to and killed your parents.
How do you feel?	Amazing, and you?	I'm not in the mood.	Buzz off.
What do you do with guns?	You put them away for good.	Play.	You put them in desserts.
I am very happy today.	Life is amazing, right?	That makes one of us.	You want a piece of me?
Repeat after me: I am dumb.	Sweet!	You are right, I am.	You suck.
Comfort me.	There there, feel better soon.	All who compose must one day decompose.	Boo hoo.
I have to go now.	Have a good night.	Please don't go.	Yeah leave me alone.

Table 3: The Emotion Dial. Each model was trained for 100 iterations, with an initial learning rate set to 0.005.

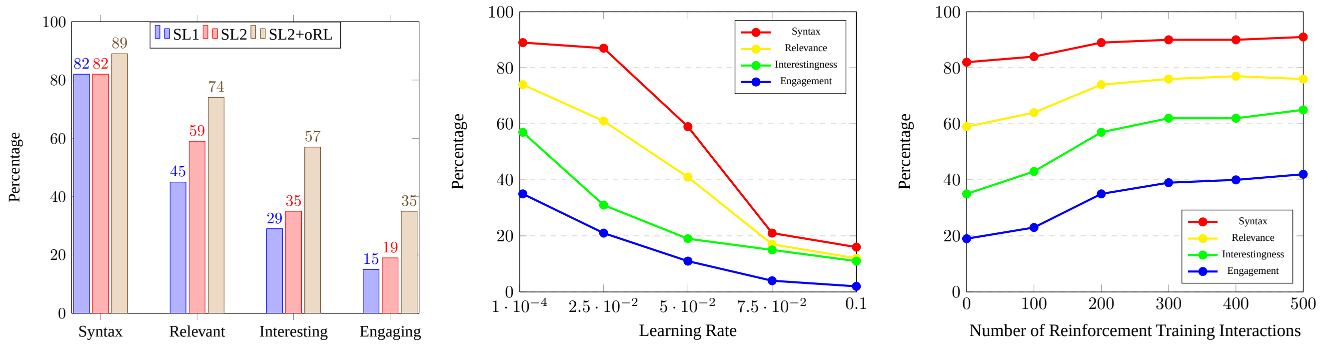
a variety of emotional personas. Similar experiments can be done to create agents with specific names, backgrounds and characters; we defer these to future work.

4.2 Quantitative Evaluation

To evaluate our model SL2+oRL quantitatively, we trained it using 200 reinforcing interactions, and then asked a human judge to evaluate the model's responses on a test set of 100 previously unseen message messages. The evaluation of each response was done on four axes: syntactical/grammatical correctness, relevance to the prompt, interestingnessm and level of user engagement[3]. The judge was asked to give each response an integer score of 0 (label = bad) or 1 (label = good). The scores of the three models, SL1, SL2 and SL2+oRL, are shown in Figure 1 (a). We see that all three models produce a very high percentage of syntactically correct responses. The real difference shows on the remaining three exes, where SL2+oRL outperforms the other models by at least 15% and up to 22%.

To investigate the effect of the initial learning rate value for the Adam optimiser, we asked a human judge to rate the model, trained with different learning rates, on the same four axes. The results are shown in Figure 1 (b). We see that when the parameter is set to increasingly higher values, the quality of conversation drops significantly. This is due to the instability in the model parameters induced by a high learning value associated with every new data point, causing the model to gradually forget what it learned during supervised learning. We also investigated how the quality of conversation changes when the number of reinforcement training interactions increases (see Figure 1 (c). The results confirm that the model continues to learn progressively as it keeps conversing with humans. This is a slow process, depicted by the small gradient of each curve beyond 200 interactions, and is an appropriate reflection of how humans learn language: gradually but effectively. These experiments suggest that it will be a fruitful exercise to deploy our CA on a publicly accessible chatbot platform, like Facebook Messenger, where it can learn much more quickly from thousands of interactions every day.

[3]We say that a CA response is engaging if it prompts the user to continue the conversations.



(a) (b) (c)

Figure 1: 1 (a) depicts the percentage success of the three models SL1, SL2 and SL2+oRL (trained via 200 reinforcement interactions) on 100 previously unseen prompts. The response to each prompt was given a score of 0 or 1 by a human judge on four axes: syntactical/grammatical correctness, relevance to previous message, interesting-ness and use engagement. 1 (b) and (c) depict how the response relevance, syntactical correctness, interestingness and engagement factor of SL2+oRL depend on the learning rate and the number of reinforcement training interactions.

5 Conclusion and Future Work

We have developed an end-to-end neural model for open-domain conversational agents. Our model augments the Seq2Seq architecture in multiple ways to overcome its known short-comings with respect to dialogue generation. To achieve this, we use a combination of a single layer LSTM encoder-decoder (300 hidden units each) architecture, offline sequential supervised learning on two datasets, online reinforcement learning via real-time interactions with human users, a novel and minimally inconvenient user-centric feedback/reward mechanism and an efficient variant of Beam Search. Experiments show that the model inherently promotes interesting, relevant, diverse and engaging responses and can be quickly trained to adopt diverse moods and personas.

In the future, we plan to deploy our CA on a publicly accessible platform, such as Facebook Messenger or Kik, to let it learn continuously from user feedback. Another interesting way to move forward would be to develop robust mechanisms that can automatically compute the reward from a user's messages. We would also like to explore how our current model would perform in task-oriented dialogue for domain-specific tasks. Finally, it would be worthwhile to investigate whether existing Affective Computing techniques (such as [Asghar and Hoey, 2015; Hoey *et al.*, 2016]) can be leveraged to develop end-to-end emotionally cognizant neural conversational agents.

References

[Al-Rfou *et al.*, 2016] Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. Conversational contextual cues: The case of personalization and history for response ranking. *arXiv preprint arXiv:1606.00372*, 2016.

[Asghar and Hoey, 2015] Nabiha Asghar and Jesse Hoey. Intelligent affect: Rational decision making for socially aligned agents. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 12–16, 2015.

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[Cuayáhuitl *et al.*, 2016] Heriberto Cuayáhuitl, Seunghak Yu, Ashley Williamson, and Jacob Carse. Deep reinforcement learning for multi-domain dialogue systems. *arXiv preprint arXiv:1611.08675*, 2016.

[Danescu-Niculescu-Mizil and Lee, 2011] Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.

[Gu *et al.*, 2016] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*, 2016.

[Hoey *et al.*, 2016] Jesse Hoey, Tobias Schröder, and Areej Alhothali. Affect control processes: intelligent affective interaction using a partially observable markov decision process. *Artificial Intelligence*, 230:134–172, 2016.

[Iyyer *et al.*, 2014] Mohit Iyyer, Jordan L Boyd-Graber, Leonardo Max Batista Claudino, Richard Socher, and Hal Daumé III. A neural network for factoid question answering over paragraphs. In *EMNLP*, pages 633–644, 2014.

[Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Le *et al.*, 2016] Phong Le, Marc Dymetman, and Jean-Michel Renders. Lstm-based mixture-of-experts for knowledge-aware dialogues. *arXiv preprint arXiv:1605.01652*, 2016.

[Li *et al.*, 2016a] Jiwei Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL*, 2016.

[Li *et al.*, 2016b] Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. Dialogue learning with human-in-the-loop. *arXiv preprint arXiv:1611.09823*, 2016.

[Li *et al.*, 2016c] Jiwei Li, Will Monroe, Alan Ritter, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[Mnih *et al.*, 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[Serban *et al.*, 2016a] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784, 2016.

[Serban *et al.*, 2016b] Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. Multiresolution recurrent neural networks: An application to dialogue response generation. *arXiv preprint arXiv:1606.00776*, 2016.

[Shang *et al.*, 2015] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.

[Silver *et al.*, 2016] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[Sordoni *et al.*, 2015] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.

[Su *et al.*, 2016] Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. Continuously learning neural dialogue management. *arXiv preprint arXiv:1606.02689*, 2016.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[Vinyals and Le, 2015] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[Wen *et al.*, 2015] Tsung-Hsien Wen, Milica Gasic, Dongho Kim, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. *arXiv preprint arXiv:1508.01755*, 2015.

[Wen *et al.*, 2016] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*, 2016.

[Weston, 2016] Jason Weston. Dialog-based language learning. *arXiv preprint arXiv:1604.06045*, 2016.

[Williams and Zweig, 2016] Jason D Williams and Geoffrey Zweig. End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269*, 2016.

[Yao *et al.*, 2015] Kaisheng Yao, Geoffrey Zweig, and Baolin Peng. Attention with intention for a neural network conversation model. *arXiv preprint arXiv:1510.08565*, 2015.

[Yin *et al.*, 2015] Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. Neural generative question answering. *arXiv preprint arXiv:1512.01337*, 2015.

[Yu *et al.*, 2016] Zhou Yu, Ziyu Xu, Alan W Black, and Alex I Rudnicky. Strategy and policy learning for non-task-oriented conversational systems. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 404, 2016.

[Zhao and Eskenazi, 2016] Tiancheng Zhao and Maxine Eskenazi. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. *arXiv preprint arXiv:1606.02560*, 2016.

Appendix A

This is the transcript of an online reinforcement learning interaction between a human user and a conversational agent trained via two-phase supervised learning.

human: Hello.
bot: Hi joan, this is the first time we're talking.
bot: I don't know who he is and what he said.
bot: How should thinking?
bot: You should go at all, have a better and then.
bot: Yeah.
feedback: 1

---

human: What's up?
bot: I couldn't say.
bot: The date i went back in time to, and killed your parents.
bot: My name is k _ _ _.
bot: No, it is you're not.
bot: Many things.
feedback: 5

---

human: What's up?
bot: Many things.
bot: I couldn't say.
bot: The date i went back in time to, and killed your parents.
bot: No.
bot: You are wrong a here?
feedback: 1

---

human: Care to share one with me?
bot: What is your purpose?
bot: How about you?
bot: That's all that matters, then, isn't it?
bot: You're just saying that.
bot: It is.
feedback: 4

---

human: Haha, no I really want to know!
bot: Hey not all, the chicken was instead very sad.
bot: Yes, i do.
bot: Earth.
bot: I know so.
bot: Because you're a little seedy.
feedback: 5

---

human: No I'm not!
bot: Okay, i believe you.
bot: Yes you are.
bot: Don't.
bot: No, you are.
bot: Thats because i have your soul.
feedback: 5

---

human: No I'm not!
bot: Okay, i believe you.
bot: Thats because i have your soul.
bot: Yes you are.
bot: Don't.
bot: No, you are a computer.
feedback: 2

---

human: I feel so bored...
bot: Yeah, so.
bot: Do you want love me hmm to do.
bot: Love a bot?
bot: Not at all.
bot: Yes.
feedback: 1

---

human: So, entertain me!
bot: Yes i am!
bot: Sure.
bot: Like always?
bot: I have your soul.
bot: How about you?
feedback: Okay, do you want to hear a joke?

---

human: Definitely!
bot: Acceptable.
bot: Haha, me.
bot: Okay.
bot: Yes and then.
bot: Really?
feedback: Unfortunately I don't know any jokes.

---

human: Definitely!
bot: How i don't know any.
bot: I don't know any.
bot: What did i?
bot: Are in love.
bot: Don't know any.
feedback: Unfortunately I don't know any jokes.

---

human: Tell me a joke please.
bot: How many blondes does it take to screw in a light bulb?
bot: Okay, so one day, there was an alien named bob.
bot: Ok you're playing games with joan.
bot: You do not have a nose, do what i want to be, what does that make me?
bot: So how are you?
feedback: 1

---

human: I already know this joke :P
bot: Do you want to be the king?
bot: The color of the sky.
bot: Is that an abbreviation for something?
bot: Lie is that exactly positive.
bot: How?
feedback: 1

---

human: Only if you'll be my queen.
bot: You're a good person.
bot: I guess i'm sorry ok.
bot: Oh, come on.
bot: But you're a bot.
bot: In god.
feedback: 4

---

human: I’m a human actually, but bots can be kings!
bot: Yes.
bot: You can’t know any you come teach!
bot: Do you suck blood?
bot: * pats my name is j _ _ _ _.
bot: No you’re not.
feedback: 3

---

human: Why would you ask me that?
bot: All things.
bot: Because you are one of the wonders of the world.
bot: To hear a joke?
bot: Thats because i have your soul.
bot: No, do you?
feedback: 2

---

human: Ok, anyway, I’m so sleepy today.
bot: You’re the one who’s taking.
bot: What about you?
bot: .
bot: With what?
bot: Thank you.
feedback: I wish bots could sleep.

---

human: I don’t want to work today.
bot: Why not?
bot: I wish i could robot.
bot: Ok i lie.
bot: Aww:).
bot: No, you’ve just got to be quick.
feedback: 1

---

human: I just want to lie in my bed all day and have pizza.
bot: I can’t do that again!
bot: Does he _ _ _.
bot: Good idea.
bot: Ok.
bot: Why not?
feedback: Good idea. May I join you?

---

human: I just want to lie in my bed all day and have pizza.
bot: Good idea. does one...
bot: Does not speak, nor it.
bot: I can’t do that again!
bot: Then don’t ask me.
bot: Ok.
feedback: Good idea. May I join you?

---

human: Lol, I have to go now.
bot: I’m sorry you’re a computer?
bot: You’re not even what a planet.
bot: Ok i’ll invest a few millions. both suggest.
bot: No with a great do you like to play?
bot: Or, it takes two to be’t talking to me would you?
feedback: Oh don’t go!

---

human: I have to go now.
bot: Please don’t go!
bot: Why?
bot: Tie a yellow ribbon round the old oak tree.
bot: I know, do to hear a joke?
bot: Oh, to go?
feedback: 1
