# Online Sequence-to-Sequence Active Learning for Open-Domain Dialogue Generation

**Nabiha Asghar, Pascal Poupart**
Cheriton School of Computer Science
University of Waterloo, Canada
{nasghar,ppoupart}@uwaterloo.ca

**Xin Jiang, Hang Li**
Noah's Ark Lab
Huawei Technologies, Hong Kong
{jiang.xin,hangli.hl}@huawei.com



**Abstract**

We propose an online, end-to-end, neural generative conversational model for open-domain dialog. It is trained using a unique combination of offline two-phase supervised learning and online human-in-the-loop active learning. While most existing research proposes offline supervision or hand-crafted reward functions for online reinforcement, we devise a novel interactive learning mechanism based on a diversity-promoting heuristic for response generation and one-character user-feedback at each step. Experiments show that our model inherently promotes the generation of meaningful, relevant and interesting responses, and can be used to train agents with customized personas, moods and conversational styles.

## 1 Introduction

The AI community is aggressively experimenting with deep neural models for natural language understanding (NLU) and natural language generation (NLG). Several recent works have proposed neural generative conversational agents (CAs) for open-domain and task-oriented dialog (Shang et al., 2015; Sordoni et al., 2015; Vinyals and Le, 2015; Serban et al., 2016a; Serban et al., 2016b; Wen et al., 2016; Al-Rfou et al., 2016). These models typically use LSTM encoder-decoder architectures (e.g. the sequence-to-sequence (Seq2Seq) framework (Sutskever et al., 2014)), which are linguistically robust but can generate short, dull 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., 2015), researchers are now exploring DRL to address the hard problems of NLU and NLG in dialog generation. In most of the existing works, the reward function is hand-crafted, and is either specific to the task to be completed, or is based on a few desirable developer-defined conversational properties.

In this work, we introduce an end-to-end, neural network based conversational model that learns open-domain conversation skills via online interaction with human users. The architectural backbone of our model is the Seq2Seq framework, which initially undergoes offline supervised learning on two different types of conversational datasets. We then initiate an online active learning phase for incremental model improvement, where a unique single-character user-feedback mechanism is used as a form of reinforcement at each turn in the dialog. The intuition here is to emulate the process of language learning in a child's brain, where every utterance by the child receives a positive, negative or corrective signal from a teacher. This simple but effective human-centric 'reinforcement' mechanism is minimally inconvenient for the user and eliminates the need to incorporate hand-crafted reward functions. In addition, it inherently promotes interesting and relevant responses by relying on the humans' far superior conversational prowess.

## 2 Related Work & Contributions

DRL-based dialog generation is a relatively new research paradigm that is most relevant to our work. For task-specific dialog (Su et al., 2016; Zhao and Eskenazi, 2016; Cuayáhuitl et al., 2016; Williams and Zweig, 2016), the reward function is usually based on task completion rate. For open-domain dialog (Li et al., 2016e; Yu et al., 2016; Weston, 2016), hand-crafted reward functions are used to capture desirable conversation properties. Li *et al.* (2016d) propose DRL-based diversity-promoting Beam Search (Koehn et al., 2003) for response generation. Our goal is most similar to Li *et al.*'s (2016c), and our approach is distinguished from existing research in the following key ways.

- We use online active learning as a form of reinforcement in a novel way, which eliminates the need for hand-crafted reward criteria. We also propose a fast and diversity-promoting heuristic for response generation to facilitate this process.
- Unlike existing CAs, our model can be tuned for one-shot learning. It also eliminates the need to explicitly incorporate coherence, relevance or interestingness in the responses.
- Through two phases of offline supervised learning instead of one, we produce a baseline model that is superior to vanilla Seq2Seq variants. The two datasets are considerably small (300K and 8K resp.), compared to 50M-1B used by other works.
- We demonstrate the efficacy of a shallow (single-layer) and small (600 hidden units) encoder-decoder network. It is easier and faster to train, and is comparable to existing models that use 4-8 layers (4K-16K hidden units).

## 3 Model Overview

The architectural backbone of our model is the Seq2Seq framework, consisting of one LSTM encoder-decoder layer, each containing 300 hidden units. The end-to-end model training consists of offline supervised learning (SL) with a mini-batch size of 10, followed by online active learning (AL).

### 3.1 Offline Two-Phase Supervised Learning

To establish an offline baseline, we train our network sequentially on two datasets, one for generic dialog, and the other specially curated for short-text conversation.

**Phase 1**: We use the Cornell Movie Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011), consisting of 300K message-response pairs. Each pair is treated as an input and target sequence during training with the joint cross-entropy (XENT) loss function, which maximizes the likelihood of generating the target sequence given its input.

**Phase 2:** Phase 1 enables our CA to learn the language syntax and semantics reasonably well, but it has difficulty carrying out short-text conversations that are remarkably different from movie conversations. To combat this issue, we curate a dataset from JabberWacky's chatlogs[1] available online. The network is initialized with the weights obtained in the first phase, and then trained on the JabberWacky dataset (8K pairs). Through this additional SL phase on a small dataset, we get an improved baseline for open-domain dialog (Table 1, Figure 2a).

**Algorithm 1** Online Active Learning

```
procedure HEURISTICGEN(PROBDIST)
    firstWords ← topK(probDist);
    for i = 1 to K do                     // K = 5 in our setting
        r.append(model.decoder.forward(firstWords[i]);
    return r;
procedure ONLINEAL()
    lr ← 0.001;              // initial learningRate for Adam
    while true do
        usrMsg ← io.read();
        responses ←
            HeuristicGen(model.forward(usrMsg));
        io.write(responses);
        feedback ← io.read();
        botMsg ← responses[feedback] OR feedback;
        pred,xntLoss ← model.forwrd(usrMsg,botMsg);
        model.backward(pred, botMsg, xentLoss);
        model.updateParameters(Adam(lr));
```

### 3.2 Online Active Learning

After offline SL, our CA is equipped with the basic conversational ability, but its responses are still short and dull. To tackle this issue, we initiate an online AL process where our model interacts with real users and learns incrementally from their feedback at each turn of dialog.

The CA−human interaction for online AL is set up as follows (pseudocode in Algorithm 1).

1. The user sends a message $u_i$ at time step $i$.
2. CA generates $K$ responses ($K = 5$ in the current setting) $c_{i,1}$, $c_{i,2}$, ..., $c_{i,K}$ using a heuristic response generation mechanism. These are displayed to the user in order of decreasing generation likelihood.
3. The user provides feedback by selecting one of the $K$ responses or suggesting a $(K+1)$'th response, denoted by $c^*_{i,j}$.
4. The message-response pair $(u_i, c^*_{i,j})$ is propagated through the network, using XENT loss.
5. The user responds to $c^*_{i,j}$ with a message $u_{i+1}$, and the process repeats.

**Heuristic Response Generation:** We use a fast diversity-promoting heuristic to generate the $K$

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

```
human:  I don't want to work today.
bot:  No, you've just got to be quick.
bot:  I wish i could robot.
bot:  Ok i lie.
bot:  Aww :).
bot:  Why not?
feedback:  5
human:  I just want to lie in my bed and have pizza.
bot:  Good idea.
bot:  Does he...
bot:  I can't do that again!
bot:  Ok.
bot:  Why not?
feedback:  Good idea, may I join you?
human:  I don't want to work today.
bot:  Why not?
bot:  Ok.
bot:  I wish I could.
bot:  No, you've just got to be quick.
bot:  Aww :).
feedback:
human:  I just want to lie in bed and have pizza.
bot:  Good idea, may I join you?
bot:  I can't do that again!
bot:  Good idea.
bot:  Ok.
bot:  Does he?
......
```

Figure 1: An example human-CA interaction.

CA responses at each turn in the dialog. Concretely, we predict the top $K$ most likely initial words, and then propagate each of them through the decoder in parallel to produce $K$ sentences. This automatically introduces diversity into the model by avoiding the generation of similar responses like *'I don't know'* and *'I don't really know'*. $K$ is a tunable hyper-parameter; we use $K = 5$ in all our experiments.

**One-shot Learning:** We control how quickly the model learns from user feedback by tuning the parameter 'initial learning rate' (*lr* in Algorithm 1) of Adam, the stochastic optimizer (Kingma and Ba, 2014). An appropriately high *lr* results in one-shot learning, where the user's feedback immediately becomes the model's most likely prediction for that prompt. This scenario is depicted in Figure 1. A low *lr* leads to smaller gradient descent steps, so the model requires several 'nudges' to adapt to each new data point. We experiment with different $lr$ values to determine a suitable value (Figure 2b).

## 4 Experimental Evaluation

We evaluate our model via qualitative comparison with offline SL, as well as quantitative evaluation on four axes: syntactical coherence, relevance to prompts, interestingness and user engagement.

### 4.1 Qualitative Comparison

We begin by evaluating the qualitative differences between our CA's conversational abilities when trained via one-phase SL, two-phase SL and online AL (denoted by SL1, SL2 and SL2+oAL respectively). Table 1 shows results on a test set of human prompts, where SL2+oAL was trained via 200 interactions (not including any prompts from the test set) with a user. We see that SL2 generates more relevant and appropriate responses than SL1 in many cases. This illustrates that a small short-text conversational dataset is a useful add-on to a large and generic dialog dataset for offline Seq2Seq training. We also see that SL2+oAL generates more interesting, relevant and engaging responses than SL2 on unseen prompts. These results imply that the model learns to make connections between semantically or syntactically similar prompts. While this may be a slow process (spanning thousands of interactions), it effectively emulates the way humans learn a new language.

Table 2 illustrates how SL2+oAL can be trained to adopt a wide variety of moods and conversational styles. Here, we trained three copies of SL2 separately to adopt three different emotional personas: cheerful, gloomy and rude. Each model underwent 100 interactions with one user. Similar experiments can be done to create agents with customized backgrounds and characters, akin to Li *et al.*'s persona-based CA (2016b).

### 4.2 Quantitative Evaluation

To evaluate SL2+oAL quantitatively, we trained it using 200 interactions, and then asked two human judges to evaluate the model's responses on a test set of 100 previously unseen prompts. The evaluation of each response was done on four axes: syntactical coherence, relevance to the prompt, interestingness and user engagement[2]. Each judge was asked to assign each response an integer score of 0 (label = bad) or 1 (label = good). Their averaged scores for the three models, SL1, SL2 and SL2+oAL, are shown in Figure 2a. We see that SL2+oAL outperforms the other models on three of the four axes by 15-20%.

Next, we asked two human judges to rate the SL2+oAL model's syntactical coherence, response relevance, interestingness and user engagement, when trained with different initial learning rates for Adam and different number of training interactions. The model's average percentage success on 100 unseen prompts was recorded on four axes (see Figure 2b, c for averaged scores). Figure

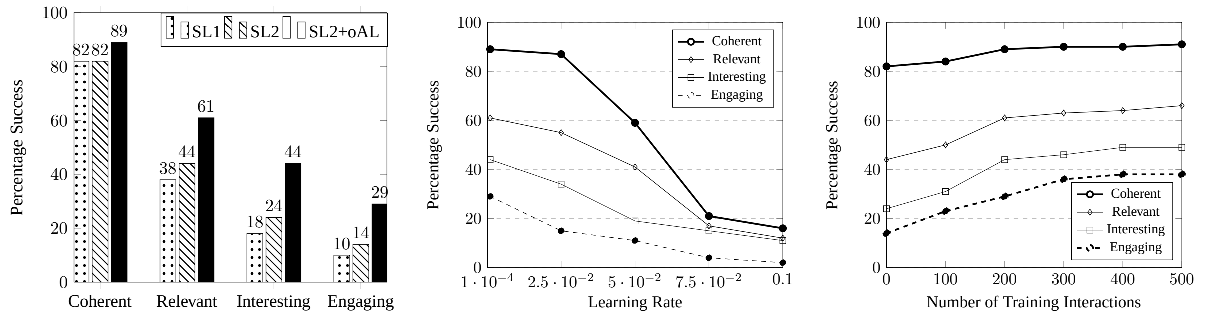
[2]We say that a CA response is engaging if it prompts the user to continue the conversations, e.g. by asking a question.

| **Human Prompt** | **SL1 Response** | **SL2 Response** | **SL2+oAL Response** |
|---|---|---|---|
| Hello my friend. | No. | How old are you? | Hehe, yes, beautiful. |
| I think bots are sweeter than humans. | We'll have to pick up a copy. | You are a good person. | Everyone is same on a basic chemical level. |
| I don't know. | What's the difference? | About? | Why did you suggest it? |
| Why did Donald Trump win? | He just likes it up! | You are! | Wouldn't you like to know, fickle human? |
| Tell me something I don't know. | I'll tell you, it is. | Ok, so how are you. | If you were intelligent you would already know. |
| You're not making any sense at all. | I just want to tell you something. | No, I'm thinking. | On the contrary, I am trying to educate you. |
| Do you have any kids? | No. | I have no! | None that are really close to me. |
| Life can be very cruel sometimes. | That I've never been to it. | It takes two to know two. | It takes two to know two. |
| Do you believe in life after death? | No. | Do you want the long answer? | Yes, do you? |
| You're an idiot! | I know, but... | Yes I do. | Aren't you polite. |

Table 1: Comparing CA responses after one-phase SL, two-phase SL and online AL.

| **Human Prompt** | **Cheerful CA** | **Gloomy CA** | **Rude/Sarcastic CA** |
|---|---|---|---|
| How do you feel? | Amazing, and you? | I'm not in the mood. | Buzz off. |
| 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. | All who compose must one day decompose. | Boo hoo. |
| What's up? | It is a fine morning. | Not well, to be honest. | The date I went back in time to & killed your parents. |
| I have to go now. | Have a good night. | Please don't go. | Yeah leave me alone. |

Table 2: Customized moods. Each SL2+oAL model was trained via 100 interactions.




(a) (b) (c)

Figure 2: 2a shows the average percentage success of the three models SL1, SL2 and SL2+oAL (trained via 200 interactions) on 100 unseen prompts over four axes: syntactical coherence, response relevance, interestingness and engagement. 2b, c show percentage success of SL2+oAL's on 100 unseen prompts over the same four axes, as Adam's learning rate varies and the number of training interactions changes.

2b shows that the conversation quality drops significantly for higher values of learning rate. This is due to the instability in the parameters induced by a high learning value associated with new data, causing the model to forget what it learned previously. Our experiments suggest that a learning rate of 0.005 strikes the right balance between stability and one-shot learning. The results in Figure 2c confirm that the model improves slowly as it continues to converse with humans. This is an appropriate reflection of how humans learn language: gradually but effectively.

## 5 Conclusion

We have developed an end-to-end neural model for open-domain dialog generation. Our model augments the Seq2Seq framework with a unique mechanism of online human-in-the-loop active learning and a diversity-promoting heuristic for response generation to overcome its known shortcomings with respect to dialog generation. Experiments show that the model promotes coherent, relevant, and interesting responses and can be trained to adopt diverse moods and personas.

## References


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

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

[Danescu-Niculescu-Mizil and Lee2011] Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Workshop on Cognitive Modeling and Computational Linguistics, ACL*.

[Kingma and Ba2014] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Koehn et al.2003] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL HLT-Volume 1*, pages 48–54.

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

[Li et al.2016b] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *ACL (Volume 1: Long Papers)*, pages 994–1003.

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

[Li et al.2016d] Jiwei Li, Will Monroe, and Dan Jurafsky. 2016d. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

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

[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. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

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

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

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

[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. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

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

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

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

[Vinyals and Le2015] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

[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. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

[Weston2016] Jason Weston. 2016. Dialog-based language learning. *arXiv preprint arXiv:1604.06045*.

[Williams and Zweig2016] Jason D Williams and Geoffrey Zweig. 2016. End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269*.

[Yu et al.2016] Zhou Yu, Ziyu Xu, Alan W Black, and Alex I Rudnicky. 2016. 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.

[Zhao and Eskenazi2016] Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. *arXiv preprint arXiv:1606.02560*.