# Deep Active Learning for Dialogue Generation

**Nabiha Asghar**[†], **Pascal Poupart**[†], **Xin Jiang**[‡], **Hang Li**[‡]
[†] Cheriton School of Computer Science, University of Waterloo, Canada
{nasghar,ppoupart}@uwaterloo.ca
[‡]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 dialogue. 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 hamming-diverse beam search for response generation and one-character user-feedback at each step. Experiments show that our model inherently promotes the generation of semantically relevant and interesting responses, and can be used to train agents with customized personas, moods and conversational styles.

## 1 Introduction

Several recent works propose neural generative conversational agents (CAs) for open-domain and task-oriented dialogue (Shang et al., 2015; Sordoni et al., 2015; Vinyals and Le, 2015; Serban et al., 2016, 2017; Wen et al., 2016; Shen et al., 2017; Eric and Manning, 2017a,b). 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 often generate short, dull and inconsistent responses (Serban et al., 2016; Li et al., 2016a). Researchers are now exploring Deep Reinforcement Learning (DRL) to address the hard problems of NLU and NLG in dialogue 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 demonstrate how online Deep Active Learning can be integrated with standard neural network based dialogue systems to enhance their open-domain conversational skills. 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 to interact with human users for incremental model improvement, where a unique single-character[1] user-feedback mechanism is used as a form of reinforcement at each turn in the dialogue. The intuition is to rely on this all-encompassing human-centric 'reinforcement' mechanism, instead of defining hand-crafted reward functions that individually try to capture each of the many subtle conversational properties. This mechanism inherently promotes interesting and relevant responses by relying on the humans' far superior conversational prowess.

## 2 Related Work & Contributions

DRL-based dialogue generation is a relatively new research paradigm that is most relevant to our work. For task-specific dialogue (Su et al., 2016; Zhao and Eskenazi, 2016; Cuayáhuitl et al., 2016; Williams and Zweig, 2016; Li et al., 2017b,c; Peng et al., 2017), the reward function is usually based on task completion rate, and thus is easy to define. For the much harder problem of open-domain dialogue generation (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.

Very recently, new approaches have been pro-

[1]The user has the option to provide longer feedback.



*Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 78–83,
Vancouver, Canada, August 3-4, 2017. 

posed to incorporate online human feedback into neural conversation models (Li et al., 2016c; Abel et al., 2017; Li et al., 2017a). Our work falls in this line of research, and is distinguished from existing approaches in the following key ways.

1. We use online deep active learning as a form of reinforcement in a novel way, which eliminates the need for hand-crafted reward criteria. We use a diversity-promoting decoding heuristic (Vijayakumar et al., 2016) to facilitate this process.

2. 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.

## 3 Model Overview

The architectural backbone of our model is the Seq2Seq framework consisting of one encoder-decoder layer, each containing 300 LSTM units. The end-to-end model training consists of offline supervised learning (SL) with mini-batches 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 dialogue, 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[2] 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 of fine-tuning on a small dataset, we get an improved baseline for open-domain dialogue (Table 1, Figure 2a).

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

**Algorithm 1** Online Active Learning

```
procedure HAMMINGDBS(TEXT)
    r = emptyList(size = K);
    for t = 1 to T do
        r[1][t] = model.forward(text, r[1][1,...,t − 1]);
        for i = 2 to K do            // K = 5 in our setting
            augmentedProbs = model.forward(t,text,r[i])
                    +λ(hammingDist(r[i], r[1, ..., i − 1]));
            r[i][t] = top1(augmentedProbs);
    return r;
procedure ONLINEAL()
    lr ← 0.001;            // initial learningRate for Adam
    while true do
        usrMsg ← io.read();
        responses ← HammingDBS(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 dialogue.

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

1. The user sends a message $u_i$ at time step $i$.
2. CA generates $K$ responses $c_{i,1}$, $c_{i,2}$, ..., $c_{i,K}$ using hamming-diverse Beam Search. These are displayed to the user in order of decreasing generation likelihood.
3. The user provides feedback by selecting one of the $K$ responses as the 'best' one or suggesting a $(K+1)$'th response, denoted by $c^*_{i,j}$. The selection criterion is subjective and entirely up to the user.
4. The message-response pair $(u_i, c^*_{i,j})$ is propagated through the network using XENT loss, with a learning rate optimized for one-shot learning.
5. The user responds to $c^*_{i,j}$ with a message $u_{i+1}$, and the process repeats.

**Heuristic Response Generation:** We use the recently proposed Diverse Beam Search (DBS) algorithm (Vijayakumar et al., 2016) to generate the $K$ CA responses at each turn in the dialogue. DBS has been shown to outperform BS and other diverse decoding techniques on several NLP tasks, including image captioning, machine translation and visual question generation. DBS incorporates diversity between the beams by maximizing an objective that consists of a standard sequence likelihood term and a dissimilarity metric between the beams. We use the hamming diversity metric for decoding at each time step, which penalizes the selection of words that have already been chosen in other beams (Algorithm 1). In particular, the weight $\lambda$ associated with this metric is tuned to aggressively promote diversity between the first tokens of each of the $K$ generated sequences, thereby avoiding similar beams like *I don't know* and *I really don't know*. We refer the reader to the original paper by Vijayakumar *et al.* for the complete DBS algorithm and derivation. $K$ is a tunable hyper-parameter; we used $K = 5$ in all our experiments, based on our observation that a smaller response set usually misses out a good contender, and more than five responses become too cumbersome for the user to read at each turn.

It is possible that displaying the $K$ responses in decreasing order of generation likelihood introduces a bias to the user's response, since users typically prefer to pick items located at the top of the screen. If this is a cause for concern for an application, the problem can be resolved simply by tweaking Algorithm 1 such that the $K$ responses are displayed to the user in a random order. In our experiments, we assume that the users are unbiased and do not take into consideration the display order or the generation likelihood of the responses.

**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).

```
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.

## 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 Quantitative Evaluation

We begin by presenting the experimental results of the quantitative evaluation 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).

We first asked a human trainer to actively train SL2+oAL using 200 prompts of his choice. We then created a test set of 100 prompts by randomly choosing 100 of the 200 training prompts and linguistically rephrasing each of them to convey the same semantics. For instance, the AL training prompts *'How's it going?'*, *'I hate you'* and *'What are your favorite pizza toppings?'* were altered to the following test prompts: *'How are you doing?'*, *'I don't like you!'* and *'What do you like on your pizza?'*. Next, we recorded SL1's, SL2's and SL2+oAL's responses to these test prompts. Finally, we asked five human judges (not including the human trainer) to subjectively evaluate the responses of the three models on the test set. The evaluation of each response was done on four axes: syntactical coherence, relevance to the prompt, interestingness and user engagement[3]. Each judge

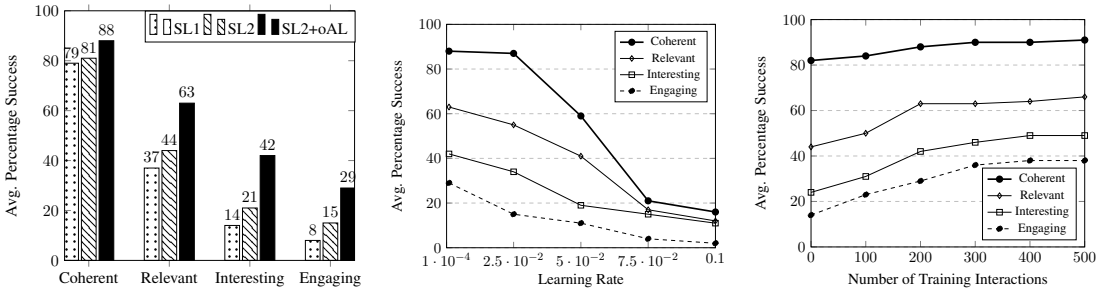
[3]We say that a CA response is engaging if it prompts the user to continue the conversations, e.g. by asking a question.

| 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 1-phase SL (SL1), 2-phase SL (SL2) & online AL (SL2+oAL).

| 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 test prompts over four axes: syntactical coherence, response relevance, interestingness and engagement. 2b, c show percentage success of SL2+oAL on 100 test prompts over the same four axes, as Adam's learning rate varies and the number of training interactions changes.

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 14-21%.

Next, we asked the human trainer to train SL2+oAL with the same 200 prompts and responses for different values of the initial learning rate for Adam (*lr* in Algorithm 1). We then asked the five human judges to subjectively rate each model's syntactical coherence, response relevance, interestingness and user engagement. Each model's percentage success on the test prompts was recorded on four axes. The averaged scores are given in Figure 2b. We see that the response 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.

Finally, we asked the human trainer to train SL2+oAL with $lr = 0.005$ and different number of training interactions. 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. Although the curves seem to plateau after 300 training interactions and suggest that the learning has stopped, this is not the case. The gradient is small but non-zero, which is an expected behavior of reinforcement learning algorithms in general.

### 4.2 Qualitative Comparison

We illustrate the qualitative differences between the responses generated by SL1, SL2 and SL2+oAL. Table 1 shows results on a small subset of the 100 test prompts. 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 fine-tuning add-on to a large and generic dialogue dataset for offline Seq2Seq training. We also see that SL2+oAL generates more interesting, relevant and engaging responses than SL2. These results imply that the model learns to make connections between semantically similar prompts that are syntactically different. 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 training interactions with one human trainer, who was instructed to adopt each of the four conversation styles while training the SL2+oAL model. The test prompts shown in Table 2 were syntactic variations of the training prompts, as before. The results illustrate that SL2+oAL was able to modify the mood of its responses appropriately, based on the way it was trained. Similar experiments can be done to create agents with customized backgrounds and characters, akin to Li *et al.*'s persona-based CA (2016b).

## 5 Conclusion & Future Work

We have developed an end-to-end neural model for open-domain dialogue generation. Our model augments the Seq2Seq framework with online Deep Active Learning to overcome some of its known short-comings with respect to dialogue generation. Experiments show that the model promotes semantically coherent, relevant, and interesting responses and can be trained to adopt diverse moods, personas and conversation styles.

In the future, we will explore context-sensitive active learning for encoder-decoder conversation models. We will also investigate whether existing Affective Computing techniques (e.g. (Asghar and Hoey, 2015)) can be leveraged to develop emotionally cognizant neural conversational agents.

## References


David Abel, John Salvatier, Andreas Stuhlmüller, and Owain Evans. 2017. Agent-agnostic human-in-the-loop reinforcement learning. *arXiv preprint arXiv:1701.04079* .

Nabiha Asghar and Jesse Hoey. 2015. Intelligent affect: Rational decision making for socially aligned agents. In *UAI*. pages 12–16.

H. Cuayáhuitl, Seunghak Yu, Ashley Williamson, and Jacob Carse. 2016. Deep reinforcement learning for multi-domain dialogue systems. *Deep Reinforcement Learning Workshop, NIPS* .

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*. Association for Computational Linguistics. https://www.aclweb.org/anthology/C16-1242.

Mihail Eric and Christopher D Manning. 2017a. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. *arXiv preprint arXiv:1701.04024* .

Mihail Eric and Christopher D Manning. 2017b. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414* .

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

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL HLT-Volume 1*. Association for Computational Linguistics, pages 48–54. http://www.aclweb.org/anthology/N03-1017.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016*

*Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 110–119. http://www.aclweb.org/anthology/N16-1014.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 994–1003. http://www.aclweb.org/anthology/P16-1094.

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* .

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017a. Learning through dialogue interactions by asking questions. In *ICLR*.

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

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

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017b. Investigation of language understanding impact for reinforcement learning based dialogue systems. *arXiv preprint arXiv:1703.07055* .

Xuijun Li, Yun-Nung Chen, Lihong Li, and Jianfeng Gao. 2017c. End-to-end task-completion neural dialogue systems. *arXiv preprint arXiv:1703.01008* .

Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite task-completion dialogue system via hierarchical deep reinforcement learning. *arXiv preprint arXiv:1704.03084* .

Julian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. 2017. Multiresolution recurrent neural networks: An application to dialogue response generation. *AAAI* .

Julian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1577–1586. https://www.aclweb.org/anthology/P15-1152.

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A conditional variational framework for dialog generation. *arXiv preprint arXiv:1705.00316* .

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. In *Proceedings of 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 196–205. http://www.aclweb.org/anthology/N15-1020.

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* .

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

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424* .

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

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* .

Jason Weston. 2016. Dialog-based language learning. *NIPS* .

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* .

Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, pages 404–412. http://www.aclweb.org/anthology/W16-3649.

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Los Angeles, pages 1–10. http://www.aclweb.org/anthology/W16-3601.