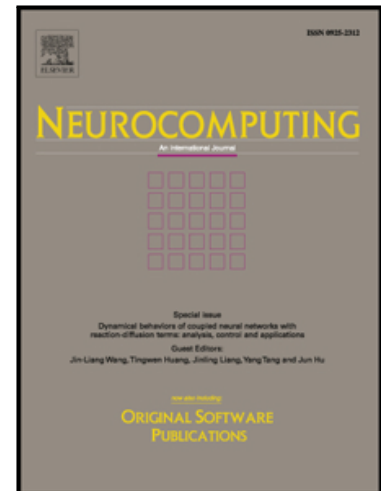# Journal Pre-proof

Ensemble-Based Deep Reinforcement Learning for Chatbots

Heriberto Cuayáhuitl, Donghyeon Lee, Seonghan Ryu, Yongjin Cho,
Sungja Choi, Satish Indurthi, Seunghak Yu, Hyungtak Choi,
Inchul Hwang, Jihie Kim

Please cite this article as: Heriberto Cuayáhuitl, Donghyeon Lee, Seonghan Ryu, Yongjin Cho, Sungja Choi,   Satish Indurthi,   Seunghak Yu,   Hyungtak Choi,   Inchul Hwang,   Jihie Kim, Ensemble-Based Deep Reinforcement Learning for Chatbots, *Neurocomputing* (2019), doi: https://doi.org/10.1016/j.neucom.2019.08.007

Ensemble-Based Deep Reinforcement Learning for Chatbots

Heriberto Cuayáhuitl[a][1], Donghyeon Lee[b], Seonghan Ryu[b], Yongjin Cho[b], Sungja Choi[b], Satish Indurthi[b], Seunghak Yu[b], Hyungtak Choi[b], Inchul Hwang[b], Jihie Kim[b]

[a] *University of Lincoln, School of Computer Science,*
Lincoln Centre for Autonomous Systems (L-CAS),
Brayford Pool, Lincoln, LN6 7TS, United Kingdom

[b] *Samsung Research, Artificial Intelligence Group,*
56 Seongchon-gil, Yangjae, Seocho-gu, Seoul, South Korea

Abstract

Trainable chatbots that exhibit fluent and human-like conversations remain a big challenge in artificial intelligence. Deep Reinforcement Learning (DRL) is promising for addressing this challenge, but its successful application remains an open question. This article describes a novel ensemble-based approach applied to value-based DRL chatbots, which use finite action sets as a form of meaning representation. In our approach, while dialogue actions are derived from sentence clustering, the training datasets in our ensemble are derived from dialogue clustering. The latter aim to induce specialised agents that learn to interact in a particular style. In order to facilitate neural chatbot training using our proposed approach, we assume dialogue data in raw text only – without any manually-labelled data. Experimental results using chitchat data reveal that (1) near human-like dialogue policies can be induced, (2) generalisation to unseen data is a difficult problem, and (3) training an ensemble of chatbot agents is essential for improved performance over using a single agent. In addition to evaluations using held-out data, our results are further supported by a human evaluation that rated dialogues in terms of fluency, engagingness and consistency – which revealed that our proposed dialogue rewards strongly correlate with human judgements.

Keywords: Deep Supervised/Unsupervised/Reinforcement Learning, Neural Chatbots

---

[1]Work carried out while the first author was visiting Samsung Research.

1. Introduction

Humans in general find it relatively easy to have chat-like conversations that are both coherent and engaging at the same time. While not all human chat is engaging, it is arguably coherent [1], and it can cover large vocabularies across a wide range of conversational topics. In addition, each contribution by a partner conversant may exhibit multiple sentences, such as greeting+question or acknowledgement+statement+question. The topics raised in a conversation may go back and forth without losing coherence. All of these phenomena represent big challenges for current data-driven chatbots.

We present a novel approach for chatbot training based on the reinforcement learning [2], unsupervised learning [3] and deep learning [4] paradigms. In contrast to other learning approaches for Deep Reinforcement Learning chatbots that rely on partially labelled dialogue data [5, 6], our approach assumes only unlabelled data. Our learning scenario is as follows: given a dataset of human-human dialogues in raw text (without any manually provided labels), an ensemble of Deep Reinforcement Learning (DRL) agents take the role of one of the two partner conversants in order to learn to select human-like sentences when exposed to both human-like and non-human-like sentences. In our learning scenario the agent-environment interactions consist of agent-data interactions – there is no user simulator as in task-oriented dialogue systems [7, 8]. During each verbal contribution and during training, the DRL agents

1. observe the state of the world via a recurrent neural network, which models a representation of all words raised in the conversation together with a set of candidate responses (i.e. *clustered actions* in our approach);

2. they then select an action so that their word-based representation is sent to the environment; and

3. they receive an updated dialogue history and a numerical reward for having chosen a certain action, until a termination condition is met.

This process—illustrated in Figure 1—is carried out iteratively until the end of a dialogue for as many dialogues as necessary, i.e. until there is no further improvement
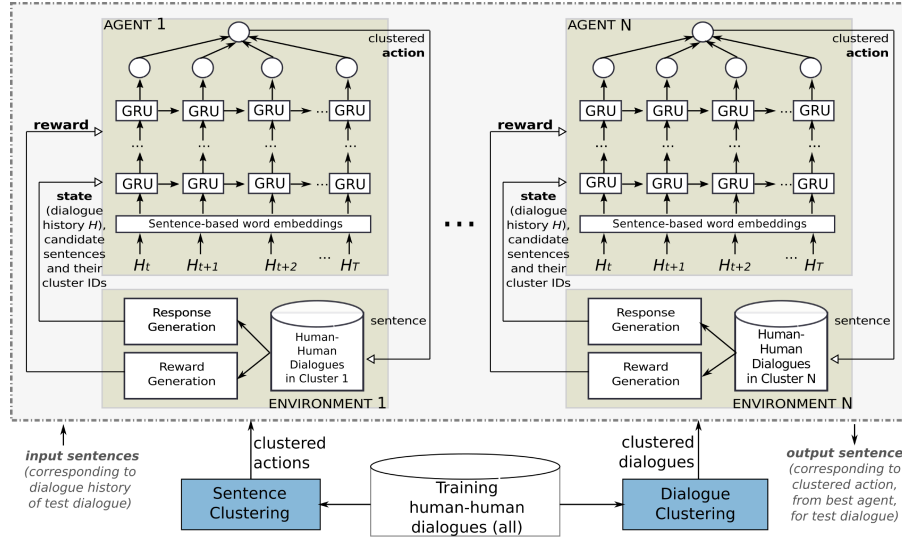
Figure 1: High-level architecture of the proposed ensemble-based learning approach for chatbot training—see text for details

in the agents' performance. During each verbal contribution at test time, the agent exhibiting the highest predictive dialogue reward is selected for human-agent interactions.

This article makes the following contributions to neural-based chatbots:

1. We propose a novel approach for chatbot training using value-based Deep Reinforcement Learning, where we induce action sets automatically via unsupervised clustering. Most previous related work has used policy search methods, and value-based methods have received little attention. We identified the latter as a research gap in our literature review. Although the performance of our DRL agents drops with dialogues that the agents are not familiar with, our DRL agents indeed learn to improve their performance over time with dialogues that they get familiarised with.

2. We propose a novel reward function due to the lack of well-embraced metrics for measuring chatbot performance. In addition, we train neural regressors for predicting dialogue rewards using a dataset of human-human dialogues that was

3

automatically extended with noisy dialogues. While non-noisy dialogues exemplify human-like and desirable outputs, the noisy ones exemplify less desirable behaviour. This reward function is easy to implement, it strongly correlates with test human-human dialogues subject to using long dialogue histories, and it strongly correlates with human judgements.

3. We propose a novel ensemble-based methodology for chatbot training, where each chatbot in our ensemble is trained with a set of clustered dialogues. To test our agents, we train 100 DRL chatbots with the aim of generating more context-relevant responses. Our experimental results according to automatic and human evaluations show that the ensemble of DRL agents outperforms a single DRL agent. This result is relevant for training future neural-based chatbots.

In the next two sections, 2 and 3, we review related work on neural-based chatbots and provide related background on deep reinforcement learning. Then we describe our proposed approach and methodology in section 4. This is followed by a comprehensive set of automatic and human evaluations in section 5, which use (i) a dataset of chitchat conversations, and (ii) human ratings of human-chatbot dialogues. Section 6 draws conclusions and discusses avenues for future research.

2. Related Work

Deep Reinforcement Learning Chatbots. Reinforcement Learning (RL) methods are typically based on value functions or policy search [2], which also applies to deep RL methods. Both types of trained agents can use the same state representations and rewards, but they differ in the representation of actions and policies. While value-based methods are typically applied to problems with discrete and finite actions, policy search methods can be applied to problems with either finite or infinite actions. In addition, while policies in value-based methods calculate numerical values (also referred to as 'expected long-term rewards') to model the importance of each state-action pair, policy search methods induce policies directly [9]—via the parameters of a model such as a neural network or a Gaussian process. While value-based methods have been particularly applied to task-oriented dialogue systems [10, 11, 12, 13, 14, 15], policy-

4

based methods have been particularly applied to open-ended dialogue systems such as (chitchat) chatbots [6, 16, 5]. This is not surprising given the fact that task-oriented dialogue systems use finite action sets, while chatbot systems use infinite action sets. The latter consider each sentence as an action, and consequently, the task is to induce dialogue behaviour from an infinite action set. This is extremely challenging for value-based reinforcement learning methods, which are more suitable for solving problems with finite action spaces. So far there is a preference for policy search methods for chatbots, but it is not clear whether they should be preferred because they face problems such as local optima rather than global optima, inefficiency, and high variance. It is therefore that we explore the feasibility of value-based methods for chatbots with large action sets, which has not been explored before—especially not from the perspective of deriving the action sets automatically as attempted in this article.

Sequence2Sequence Chatbots. Other closely related methods to DRL include sequence to sequence models for dialogue generation [17, 18, 19, 20, 21, 22]. These methods are based on Recurrent Neural Networks (RNNs) using an encoder-decoder architecture. In these methods, while one RNN referred to as 'encoder' computes an internal representation of the inputs, another RNN referred to as 'decoder' generates one word as a time – both trained end-to-end with all parameters (weights) trained jointly. Given a sequence of input words $(x_1, \cdots, x_T)$, an encoder-decoder computes a sequence of output words $(y_1, \cdots, y_T)$ by iterating the following equation: $y_t = softmax(\mathbf{W}^{yh}\mathbf{h}_t)$, where $softmax(z_i) = \frac{\exp_i^z}{\sum_j \exp_j^z}$, $\mathbf{h}_t = \sigma(\mathbf{W}^{xh}\mathbf{x}_t + \mathbf{W}^{hh}\mathbf{h}_{t-1})$, function $\sigma(.)$ is an activation function, and training weights $\mathbf{W}$ involves minimising a loss function such as categorical cross entropy [23]. Sequence2Sequence (Seq2Seq) methods can be combined with deep reinforcement learners by treating the policy as an encoder-decoder [6]. Seq2Seq methods have also been included in ensemble-based methods together with rule-based systems and a variety of other machine learning methods [5, 24, 25]. While some of them use a single DRL agent [5], they have not been investigated using an ensemble containing a horde of DRL-based chatbots as attempted in this article.

Reward Functions. Related work above highlights that evaluation is a difficult part, and that there is a need for better evaluation metrics. This is further supported by [26],

5

who found that typical metrics used to assess the quality of machine translators such as BLEU (Bilingual Evaluation Understudy) [27] and METEOR (Metric for Evaluation of Translation with Explicit ORdering) [28] amongst others do not correlate with human judgments. The dialogue rewards used by DRL agents are either specified manually depending on the application, or learnt from dialogue data. For example, [6] conceives a reward function that positively rewards sentences that are easy to respond to and coherent while penalising repetitiveness. [16] uses an adversarial approach, where the discriminator is trained to score human vs. non-human sentences so that the generator can use these scores during training. [5] trains a reward function from expensive and time-consuming human ratings. All these related studies are neural-based, and there is no clear best reward function to use in future (chitchat) chatbots. This motivated us to propose a new metric that is easy to implement, practical due to requiring only data in raw text, and potentially promising as described below.

This article contributes to the literature of neural-based chatbots as follows. First, our methodology for training value-based DRL agents uses only unlabelled dialogue data. Previous work requires manual extensions to the dialogue data [6] or expensive and time consuming ratings for training a reward function [5]. Second, our proposed reward function strongly correlates with human judgements. Previous work has only shown moderate positive correlations between target dialogue rewards and predicted ones [5], or rely on high-level annotations requiring external and language-dependent resources typically induced from labelled data [29]. Third, while previous work on DRL chatbots train a single agent [5, 6], our study—confirmed by automatic and human evaluations—shows that an ensemble-based approach performs better than a counterpart single agent. The remainder of this article elaborates on these contributions.

3. Background

A reinforcement learning agent induces its behaviour from interacting with an environment through trial and error, where situations (representations of sentences in a dialogue history) are mapped to actions (follow-up sentences) by maximising a long-term reward signal. Such an agent is typically characterised by: (i) a finite set of states

6

$S = \{s_i\}$ that describe all possible situations in the environment; (ii) a finite set of actions $A = \{a_j\}$ to change in the environment from one situation to another; (iii) a state transition function $T(s, a, s')$ that specifies the next state s' for having taken action a in the current state s; (iv) a reward function $R(s, a, s')$ that specifies a numerical value given to the agent for taking action a in state s and transitioning to state s'; and (v) a policy $\pi : S \rightarrow A$ that defines a mapping from states to actions [2, 30].

The goal of a reinforcement learning agent is to find an optimal policy by maximising its cumulative discounted reward defined as

$$Q^*(s, a) = \max_\pi \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+1} + ... | s_t = s, a_t = a, \pi],$$

where function Q^* represents the maximum sum of rewards r_t discounted by factor γ at each time step. While a reinforcement learning agent takes actions with probability $Pr(a|s)$ during training, it selects the best action at test time according to

$$\pi^*(s) = \arg\max_{a \in A} Q^*(s, a).$$

A deep reinforcement learning agent approximates Q^* using a multi-layer neural network [31]. The Q function is parameterised as $Q(s, a; \theta)$, where θ are the parameters or weights of the neural network (recurrent neural network in our case). Estimating these weights requires a dataset of learning experiences $D = \{e_1, ...e_N\}$ (also referred to as 'experience replay memory'), where every experience is described as a tuple $e_t = (s_t, a_t, r_t, s_{t+1})$. Inducing a Q function consists in applying Q-learning updates over minibatches of experience $MB = \{(s, a, r, s') \sim U(D)\}$ drawn uniformly at random from the full dataset D. This process is implemented in learning algorithms using Deep Q-Networks (DQN) such as those described in [31, 32, 33], and the following section describes a DQN-based algorithm for human-chatbot interaction.

4. Proposed Approach

This section explains the main components of Figure 1 as follows. Motivated by [34], we first describe the ensemble of Deep Reinforcement Learning (DRL) agents, we then explain how to conceive a finite set of dialogue actions from raw text, and finally we describe how to assign dialogue rewards for training DRL-based chatbots.

7

4.1. Ensemble of DRL Chatbots

We assume that all deep reinforcement learning agents in our ensemble use the same neural network architecture and learning algorithm. They only differ in the portion of data used for training and consequently the weights in their trained models—see [35, 36] for alternative approaches. Our agents aim to maximise their cumulative reward over time according to

$$Q^*(s, a; \theta_i) = \max_{\pi_{\theta_i}} \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots | s, a, \pi_{\theta_i}],$$

where r is the numerical reward given at time step t for choosing action a in state s, γ is a discounting factor, and $Q^*(s, a; \theta_i)$ is the optimal action-value function using weights θ in the neural network of chatbot i. During training, a DRL agent will choose actions in a probabilistic manner in order to explore new (s, a) pairs for discovering better rewards or to exploit already learnt values—with a reduced level of exploration over time and an increased level of exploitation over time. During testing, our ensemble-based DRL chatbot will choose the best actions a^* according to

$$\pi^*(s) = \arg \max_{a \in A} Q^*(s, a; \theta_i) \text{ with } i = \arg \max_{i \in I} \hat{R}(\tau_i),$$

where $\tau_i = < (s_0, a_0), ..., (s_t, a_t) >$ is a trajectory of state-action pairs of chatbot i, and $\hat{R}(\tau_i)$ is a function that predicts the dialogue reward of chatbot i as in [37]. Given the set of trajectories for all agents—where each agent takes its own decisions and updates its environment states accordingly—the agent with the highest predictive reward is selected, i.e. the one with the least amount of errors in the interaction.

Our DRL agents implement the procedure above using a generalisation of DQN-based methods [31, 32, 33]—see Algorithm 1, explained as follows.

- After initialising replay memory $D = \{e_1, \ldots, e_{|D|}\}$ with learning experience $e_i = (s, a, r, s')$, dialogue history $H = \{s_1, \ldots, s_{|H|}\}$ with sentences s_i, action-value function Q and target action-value function $\hat{Q}$, we sample a training dialogue from our data of human-human conversations (lines 1-4).

- Once a conversation starts, it is mapped to its corresponding sentence embedding representation, i.e. 'sentence vectors' as described in Section 4.2 (lines 5-6).

8

- Then a set of candidate responses is generated including (1) the true human response and (2) a set of randomly chosen responses (distractors). The candidate responses are clustered as described in the next section and the resulting actions are taken into account by the agent for action selection (lines 8-10).

- Once an action is chosen, it is conveyed to the environment, a reward is observed as described at the end of this section, and the agent's partner response is observed in order to update the dialogue history H (lines 11-14).

- In response to the update above, the new sentence embedding representation is extracted from H for updating the replay memory D with experience e (lines 15-16).

- Then a minibatch of experiences $MB = \{e_j\}$ is sampled from D for updating weights θ according to the error derived from the difference between the target value y_j and the predicted value $Q(s, a; \theta)$ (see lines 18 and 20), which is based on the following weight updates:

$$\theta_{t'} = \theta_t + \alpha(y_t - Q(s, a; \theta_t))\nabla_{\theta_t} Q(s, a; \theta_t),$$

where $y_t = r_t + \gamma \max_{a'} \hat{Q}(s', a'; \hat{\theta}_t)$ and α is a learning rate hyperparameter.

- The target action-value function $\hat{Q}$ and environment state s are updated accordingly (lines 21-22), and this iterative procedure continues until convergence.

4.2. Sentence and Dialogue Clustering

Actions in reinforcement learning chatbots correspond to sentences, and their size is infinite assuming all possible combinations of word sequences in a given language. This is especially true in the case of open-ended conversations that make use of large vocabularies, as opposed to task-oriented conversations that make use of smaller (restricted) vocabularies. A **clustered action** is a group of sentences sharing a similar or related meaning via *sentence vectors* derived from word embeddings [38, 39]. We represent sentences via their mean word vectors—similarly as in Deep Averaging Networks [40]—denoted as $\mathbf{x}_l = \frac{1}{N_l} \sum_{j=1}^{N_l} c_j$, where c_j is the vector of coefficients of

9

Algorithm 1 ChatDQN Learning

1: Initialise Deep Q-Networks with replay memory D, dialogue history H, action-value function Q with random weights θ, and target action-value functions $\hat{Q}$ with $\hat{\theta} = \theta$

2: Initialise clustering model from training dialogue data

3: **repeat**

4:     Sample a training dialogue (human-human sentences)

5:     Append first sentence to dialogue history H

6:     s = sentence embedding representation of H

7:     **repeat**

8:         Generate *noisy* candidate response sentences

9:         A = cluster IDs of candidate response sentences

10:         $a = \begin{cases} rand_{a \in A} \text{ if random number } \leq \epsilon \\ \max_{a \in A} Q(s, a; \theta) \text{ otherwise} \end{cases}$

11:         Execute chosen clustered action a

12:         Observe human-likeness dialogue reward r

13:         Observe environment response (agent's partner)

14:         Append agent and environment responses to H

15:         s' = sentence embedding representation of H

16:         Append learning experience $e = (s, a, r, s')$ to D

17:         Sample random minibatch (s_j, a_j, r_j, s'_j) from D

18:         $y_j = \begin{cases} r_j \text{ if final step of episode} \\ r_j + \gamma \max_{a' \in A} \hat{Q}(s', a'; \hat{\theta}) \quad \text{otherwise} \end{cases}$

19:         Set $err = (y_j - Q(s, a; \theta))^2$

20:         Gradient descent step on err with respect to θ

21:         Reset $\hat{Q} = Q$ every C steps

22:         $s \leftarrow s'$

23:     **until** end of dialogue

24:     Reset dialogue history H

25: **until** convergence

---

word j, N_l is the number of words in sentence l, and $\mathbf{x}_l$ is the embedding vector of sentence l. Similarly, a **clustered dialogue** is a group of conversations sharing a similar or related topic(s) via their clustered actions. We represent dialogues via their clustered actions. Dialogue clustering in this way can be seen as a two-stage approach, where sentences are clustered in the first step and dialogues are clustered in the second step. In our proposed approach, each DRL agent is trained on a cluster of dialogues.

While there are multiple ways of selecting features for clustering and also multiple clustering algorithms, the following requirements arise for chatbots: (1) unlabelled data due to human-human dialogues in raw text (this makes it difficult to evaluate the goodness of clustering features and algorithms), and (2) scalability to clustering a large set of data points (especially in the case of sentences, which are substantially different between them due to their open-ended nature).

Given a set of data points $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \forall \mathbf{x}_l \in \mathbb{R}^m$ and a similarity metric $d(\mathbf{x}_l, \mathbf{x}_{l'})$, the task is to find a set of $\mathcal{K}$ groups with a clustering algorithm. In our case each data point $\mathbf{x}$ corresponds to a dialogue or a sentence. For scalability purposes, we use the K-Means++ algorithm [41] and the Euclidean distance $d(\mathbf{x}_l, \mathbf{x}_{l'}) = \sqrt{\sum_{j=1}^{m}(x_l^j - x_{l'}^j)^2}$ with m dimensions, and consider $\mathcal{K}$ as a hyperparameter – though other clustering algorithms and distance metrics can be used with our approach. In this way, a trained sentence clustering model assigns a cluster ID $a \in A$ to features $\mathbf{x}_l$, where the number of actions (in a DRL agent) refers to the number of sentence clusters, i.e. $|A| = \mathcal{K}$.

4.3. Human-Likeness Rewards

Specifying reward functions in reinforcement learning dialogue agents is often a difficult aspect. We propose to derive rewards from human-human dialogues by assigning positive values to contextualised responses seen in the data, and negative values to randomly chosen responses due to lacking coherence (also referred to as 'non-human-like responses') – see example in Tables 1 and 2. An episode or dialogue reward can thus be computed as $R_d = \sum_{j=1}^{N} r_j^d(a)$, where index d refers to the dialogue in focus,

11

turn	Verbalisation	Reward
1	A: hello what are doing today ? B: i'm good , i just got off work and tired , i have two jobs .	+1
2	A: i just got done watching a horror movie B: i rather read , i have read about 20 books this year .	+1
3	A: wow ! i do love a good horror movie . loving this cooler weather B: but a good movie is always good .	+1
4	A: yes ! my son is in junior high and i just started letting him watch them too B: i work in the movies as well .	+1
5	A: neat ! ! i used to work in the human services field B: yes it is neat , i stunt double , it is so much fun and hard work .	+1
6	A: yes i bet you can get hurt . my wife works and i stay at home B: nice , i only have one parent so now i help out my mom .	+1
7	A: i bet she appreciates that very much . B: she raised me right , i'm just like her .	+1
8	A: my dad was always busy working at home depot B: now that i am older home depot is my toy r us .	+1

Table 1: Example human-human dialogue with a good numerical reward $\sum_{j=1}^{T=8} r_j^1(a) = 8$

index j to the dialogue turn in focus, and $r_j^d(a)$ is given according to

$$r_j^d(a) = \begin{cases} +1, & \text{if } a \text{ is a human response in dialogue-turn } j \text{ of dialogue } d. \\ -1, & \text{if } a \text{ is human but randomly chosen (incoherent).} \end{cases}$$

Table 1 shows an example of a well rewarded dialogue (without distortions) and Table 2 shows an example of a poorly rewarded dialogue (with distortions). Other dialogues can exhibit similar dialogue rewards or something in between (ranging between $-T$ and T), depending on the amount of distortions—the higher the amount of distortions the lower the dialogue reward.

We employ the algorithm described in [37] for generating dialogues with varying amounts of distortions (i.e. different degrees of human-likeness), which we use for training and testing reward prediction models using supervised regression. Given our extended dataset $\hat{\mathcal{D}} = \{(\hat{d}_1, y_1), \ldots, (\hat{d}_N, y_N)\}$ with (noisy) dialogue histories $\hat{d}_i$ represented with sequences of sentence vectors, the goal is to predict dialogue scores y_i as accurately as possible.

Alternative and automatically derived values between -1 and +1 are also possible but considered as future work. Section 5.4 provides an evaluation of our reward func-

12

Turn	Verbalisation	Reward
1	A: hello what are doing today ?	
	B: ~~i'm good , i just got off work and tired , i have two jobs .~~	+1
	B: lol. she is more pop now. her old stuff was country. i like classic country .	-1
2	A: i just got done watching a horror movie	
	B: ~~i rather read , i have read about 20 books this year .~~	+1
	B: i am good how about you	-1
3	A: wow ! i do love a good horror movie . loving this cooler weather	
	B: ~~but a good movie is always good .~~	+1
	B: i do not enjoy the driving in my honda civic	-1
4	A: yes ! my son is in junior high and i just started letting him watch them too	
	B: ~~i work in the movies as well .~~	+1
	B: do you like dogs ? I have two .	-1
5	A: neat ! ! i used to work in the human services field	
	B: ~~yes it is neat , i stunt double , it is so much fun and hard work .~~	+1
	B: very cool . i had one of those .	-1
6	A: yes i bet you can get hurt . my wife works and i stay at home	
	B: ~~nice , i only have one parent so now i help out my mom .~~	+1
	B: that is just great . we need more educators like you in the system .	-1
7	A: i bet she appreciates that very much .	
	B: ~~she raised me right , i'm just like her .~~	+1
	B: is it bad that i hate pumpkins and trees ?	-1
8	A: my dad was always busy working at home depot	
	B: ~~now that i am older home depot is my toy r us .~~	+1
	B: why not ? maybe you do not like to travel ?	-1

Table 2: Example distorted human-human dialogue with a poor numerical reward $\sum_{j=1}^{8} r_j^2(a) = -8$

tion and its correlation with human judgement. We show that albeit simple, our reward function is highly correlated with our judges' ratings.

4.4. Methodology

Our proposed approach can be summarised through the following methodology:

13

1. Collect or adopt a dataset of human-human dialogues (as in 5.1)

2. Design or adopt a suitable reward function (as in 4.3)

3. Train a neural regressor for predicting dialogue rewards (as in [37])

4. Perform sentence and dialogue clustering in order to define the action set and training datasets (as in 4.2)

5. Train a Deep Reinforcement Learning agent per dialogue cluster (as described in 4.1)

6. Test the ensemble of agents together with the predictor of dialogue rewards (as in 5.3 and 5.4), and iterate from Step 1 if needed

7. Deploy your trained chatbot subject to satisfactory results in Step 6

5. Experiments and Results

5.1. Data

We used the *Persona-Chat* dataset[2], stats are shown in Table 5.1.

---

[2]Dataset downloaded from `http://parl.ai/` on 18 May 2018 [42]

Table 3: Statistics of the Persona-Chat data used in our experiments for chatbot training

Attribute / Value	Training Set	Test Set
Number of dialogues	17877	999
Number of dialogue turns	131438	7801
Number of sentences	262862	15586
Number of unique sentences	124469	15186
Number of unique words	18672	6692
Avg. turns per dialogue	7.35	7.8
Avg. words per dialogue	165.89	185.86
Avg. words per sentence	11.28	11.91

14

5.2. *Experimental Setting*

Our agents' states model dialogue histories as sequences of sentence vectors—using GloVe-based [39] mean word vectors [40]—with pre-trained embeddings. All our experiments use a 2-layer Gated Recurrent Unit (GRU) neural network[3] [43]. At each time step t in the dialogue history, the first hidden layer generates a hidden state $\mathbf{h}_t$ as follows:

$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1}),$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1}),$$

$$\bar{\mathbf{h}}_t = \tanh(\mathbf{W}_{\bar{\mathbf{h}}}\mathbf{x}_t + \mathbf{U}_{\bar{\mathbf{h}}}(\mathbf{r}_t \odot \mathbf{h}_{t-1})),$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \bar{\mathbf{h}}_t,$$

where $\mathbf{x}_t$ refers to a set of sentence vectors of the dialogue history, $\mathbf{r}_t$ is a reset gate that decides how much of the previous state to forget, $\mathbf{z}_t$ is an update gate that decides how much to update its activation, $\bar{\mathbf{h}}_t$ is an internal state, $\sigma(.)$ and $tahn(.)$ are the Sigmoid and hyperbolic Tangent functions (respectively), $\mathbf{W}_*$ and $\mathbf{U}_*$ are learnt weights, and $\odot$ refers to the element-wise multiplication. If the equations above are summarised as $\mathbf{h}_t = GRU(\mathbf{x}_t, \mathbf{h}_{t-1})$ we get the following output action taking into account both hidden layers in our neural net: $a_t = \arg\max_{a \in A}(\mathbf{W}_a\mathbf{h}_t^2)$, where $\mathbf{h}_t^2 = GRU(\mathbf{h}_t^1, \mathbf{h}_{t-1}^2)$ and $\mathbf{h}_t^1 = GRU(\mathbf{x}_t, \mathbf{h}_{t-1}^1)$.

While a small number of sentence clusters may result in actions being assigned to potentially the same cluster[4], a larger number of sentence clusters would mitigate the problem, but the larger the number of clusters the larger the computational cost—i.e. more parameters in the neural net. Table 4 shows example outputs of our sentence clustering using 100 clusters on our training data. A manual inspection showed that while clustered sentences sometimes do not seem very similar, they made a lot of sense

---

[3]Other hyperparameters include embedding batch size=128, dropout=0.2, sentence vector dimension=100, latent dimensionality=256, discount factor=0.99, size of candidate responses=20, max. number of mean sentence vectors in H=25, burning steps=3K, memory size=10K, target model update (C)=10K, optimiser=Adam, learning steps={100K, 500K}, test steps=100K. The number of parameters of each neural net is 2.53 million.

[4]In this case our system would select randomly from sentences with the same cluster ID
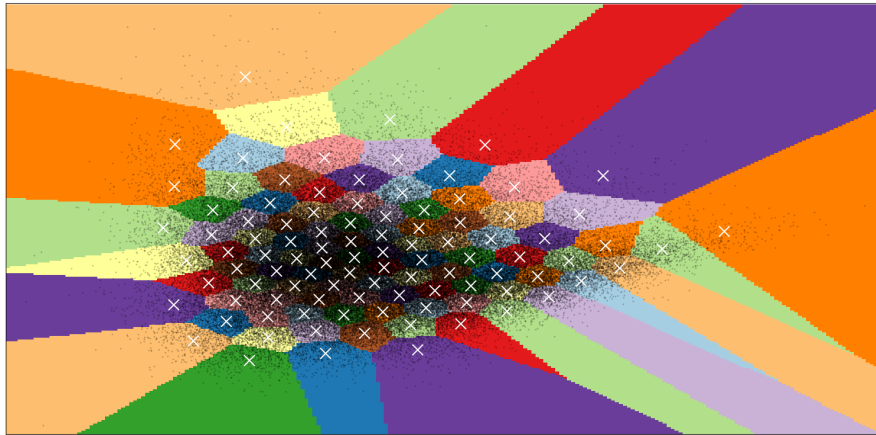
Table 4: Example clustered sentences chosen arbitrarily

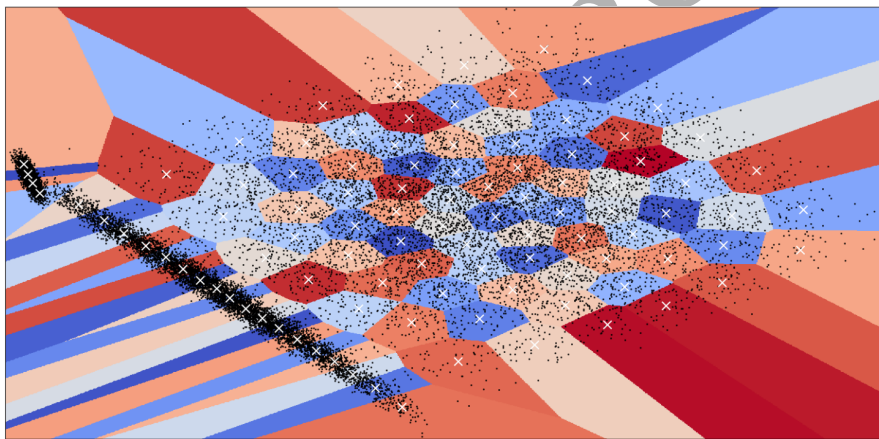| ID	Clustered Sentence
25	'i mostly eat a fresh and raw diet , so i save on groceries .', 'i only eat kosher foods', 'i like kosher salt a lot', 'i prefer seafood . my dad makes awesome fish tacos .', 'i do a pet fish', 'that sounds interesting i like organic foods .', 'cheeseburgers are great , i try all kinds of foods everywhere i go , gotta love food .'
68	'hi how are you doing ? i am okay how about you ?', 'i am great . what do you like to do ?', 'oh right how i am between jobs', 'i am thinking about my upcoming retirement . how about you ?', 'i am not sure ? how old are you ?', 'i am well , how are you ?' 'i am doing very fine this afternoon . how about you ?'
88	'i have dogs and i walk them . and a cat .', 'yeah dogs are pretty cool', 'i have dogs and cats', 'hello , leon . my dogs and i are doing well .', 'sadly , no . my dogs and i are in ohio .', 'i have 2 dogs . i should take them walking instead of eating .', 'yeah dogs are cool . i kayak too . do you have pets ?'

and they produced reasonable outputs. Our human evaluation (see Section 5.4) confirms this. All our experiments use $\mathcal{K} = 100$ due to a reasonable compromise between system performance and computational expense.

The purpose of our second clustering model is to split our original training data into a group of data subsets, one subset for each ChatDQN agent in our ensemble. We explored different numbers of clusters (20, 50, 100) and noted that the larger the number of clusters the (substantially) higher the computational expense [5]. We chose 100 clusters for our experiments due to higher average episode rewards of cluster-based agents than non-cluster-based ones. Figure 2 shows visualisations of our sentence and dialogue clustering using 100 clusters on our training data of 17.8K data points. A manual inspection was not as straightforward as analysing sentences due to the large variation of open-ended sets of sentences—see next section for further results.

---

[5]Our experiments ran on a cluster of 16 GPU Tesla K80, and their implementation used the following libraries: Keras (https://github.com/keras-team/keras), OpenAI (https://github.com/openai) and Keras-RL (https://github.com/keras-rl/keras-rl).

16

(a) 100 clusters of training sentences

(b) 100 clusters of training dialogues

Figure 2: Example clusters of our training data using Principal Component Analysis [44] for visualisations in 2D, where each black dot represents a training sentence or dialogue

5.3. Automatic Evaluation

We compared three DQN-based algorithms (DQN [31], Double DQN [32] and Dueling DQN [33]) in order to choose a baseline single agent and the learning algorithm for our ensemble of agents. The goal of each agent is to choose the human-generated sentences (actions) out of a set of candidate responses (20 available at each dialogue turn). Figure 3(left) shows learning curves for these three learning algorithms, where

we can observe that all agents indeed improve their performance (in terms of aver-

265   age episode reward) over time. It can also be observed that DQN and Double DQN
performed similarly, and that Dueling DQN was outperformed by its counterpart algo-
rithms. Due to its simplicity, we thus opted for using DQN as our main algorithm for
the remainder of our experiments.
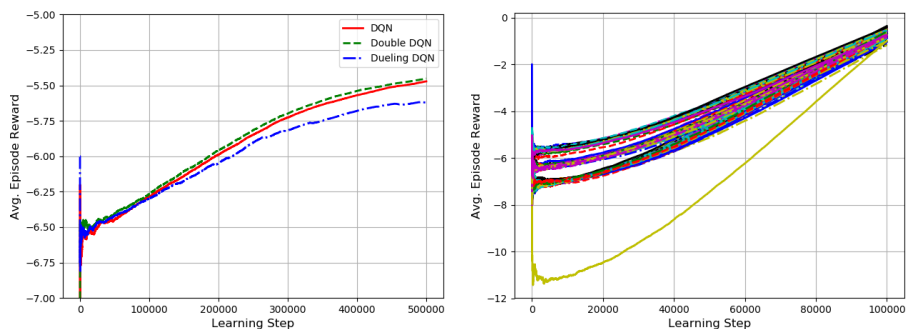
(a) ChatDQN Agents – 1 Dialogue Cluster   (b) ChatDQN Agents – 100 Dialogue Clusters

Figure 3: Learning curves of ChatDQN agents

Figure 3(right) shows the performance of 100 ChatDQN agents (one per dialogue

270   cluster), where we also observe that all agents improve their performance over time.
It can be noted however that the achieved average episode reward of $\sim$ -1 is much
greater than that of the single agent corresponding to $\sim$ -5.5. Additional experiments
reported that the lower the number of clusters the lower the average episode reward
during training. We thus opted for using 100 dialogue clusters in the remainder of our

275   experiments.

We analysed the performance of our agents further by using the test set of 999
totally unseen dialogues during training. We clustered the test set using our trained
dialogue clustering model in order to assess the goodness of each agent in dialogues
that were similar but not the same. The box plots in Figure 4 report the performance

280   of our DRL agents according to the following metrics while tested on training data
and test data: Avg. Episode Reward, Avg. F1 score, Avg. Recall@1, and Average
Recall@5. One can quickly observe the striking performance gap between testing on
training data vs. testing on test data. This can be interpreted as ChatDQN agents

18

(a) Avg. Episode Reward

(b) Avg. F1 Score

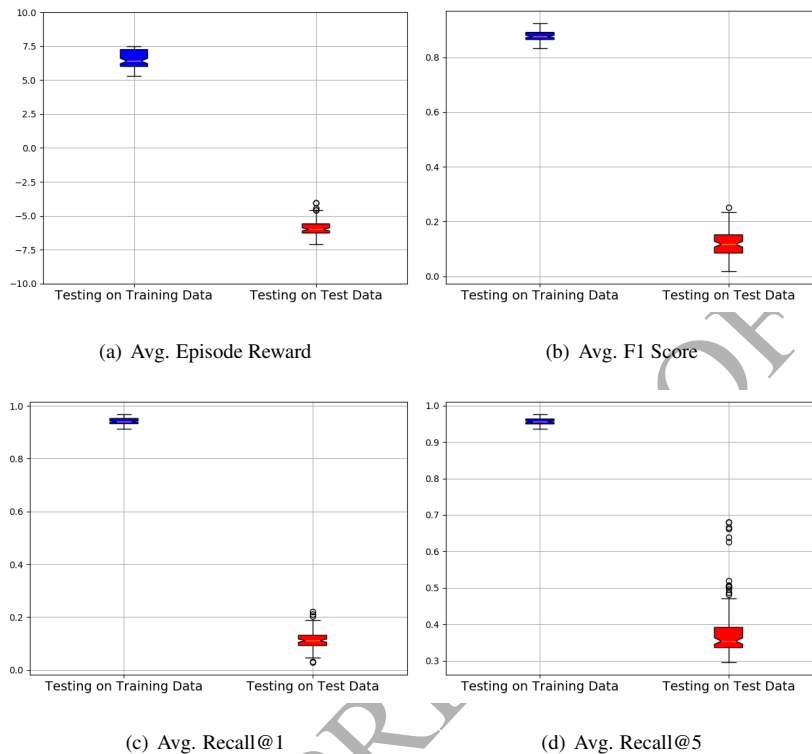(c) Avg. Recall@1

(d) Avg. Recall@5

Figure 4: Test performance of 100 ChatDQN agents on training (blue boxes) and test data (red boxes) using 4 evaluation metrics

being able to learn well how to select actions on training data, but not being able to

₂₈₅ replicate the same behaviour on test data. This may not be surprising given that only 720 sentences (out of 263,862 training sentences and 15,586 test sentences) are shared between both sets, and it is presumably a realistic scenario seen that even humans rarely use the exact same sentences in multiple conversations. On the one hand our results also suggest that our training dataset is rather modest, and that a larger dataset is needed

₂₉₀ for improved performance. On the other hand, our results help us to raise the question *'Can chitchat chatbots with reasonable performance be trained on modest datasets— i.e. with thousands of dialogues instead of millions?'* If so, the generalisation abilities of chatbots need to be improved in future work. If not, large (or very large) datasets should receive more attention in future work on neural-based chatbots.

19

Table 5: Automatic evaluation of chatbots on test data

Agent/Metric	Dialogue Reward	F1 Score	Recall@1
Upper Bound	7.7800	1.0000	1.0000
Lower Bound	-7.0600	0.0796	0.0461
Ensemble	**-2.8882**	**0.4606**	**0.3168**
Single Agent	-6.4800	0.1399	0.0832
Seq2Seq	-5.7000	0.2081	0.1316

Finally, we compared the performance of 5 dialogue agents on 999 dialogues with 20 candidate sentences at every dialogue turn:

- **Upper Bound**, which corresponds to the true human sentences in the test dataset;

- **Lower Bound**, which selects a sentence randomly from other dialogues than the one in focus;

- **Ensemble**, which selects a sentence using 100 agents trained on clustered dialogues as described in section 4 – the agent in focus is chosen using a regressor as predictor of dialogue reward $\hat{R}(.)$ using a similar neural net as the ChatDQN agents except for the final layer having one node and using Batch Normalisation [45] between hidden layers as in [37];

- **Single Agent**, which selects a sentence using a single ChatDQN agent trained on the whole training set; and

- **Seq2Seq**, which selects a sentence using a 2-layer LSTM recurrent neural net with attention[6] – from the Parlai framework (http://www.parl.ai) [22], trained using the same data as the agents above.

Table 5 shows the results of our automatic evaluation, where the ensemble of Chat-DQN agents performed substantially better than the single agent and Seq2Seq model.

---

[6]https://github.com/facebookresearch/ParlAI/tree/master/projects/
convai2/baselines/seq2seq

20

5.4. Human Evaluation

In addition to our results above, we carried out a human evaluation using 15 human judges. Each judge was given a form of consent for participating in the study, and was asked to rate 500 dialogues (100 core dialogues—from the test dataset—with 5 different agent responses[7], dialogues presented in random order) according to the following metrics: *Fluency* (Is the dialogue naturally articulated as written by a human?), *Engagingness* (Is the dialogue interesting and pleasant to read?), and *Consistency* (without contradictions across sentences). This resulted in $15 \times 500 \times 3 = 22,500$ ratings from all judges. Figure 5 shows an example dialogue with ratings ranging from 1=strongly disagree to 5=strongly agree.

Figure 6 shows average ratings (and corresponding error bars) per conversational agent and per metric[8]. As expected, the **Upper Bound** agent achieved the best scores and the **Lower Bound** agent the lowest scores. The ranking of our agents in Table 5 is in agreement with the human evaluation, where the **Ensemble** agent outperforms the **Seq2Seq** agent, and the latter outperforms **Single Agent**. The difference in performance between the **Ensemble** agent and the **Seq2Seq** agent is significant at $p = 0.0332$ for the Fluency metric and at $p < 0.01$ for the other metrics (Engagingness and Consistency)—based on a two-tailed Wilcoxon Signed Rank Test.

Furthermore, we analysed the predictive power of dialogue rewards, derived from our reward function, against human ratings on test data. This analysis revealed positive high correlations between them as shown in Figure 7. These scatter plots show data points of test dialogues (the X-axes include Gaussian noise drawn from $\mathcal{N}(0, 0.3)$ for better visualisation), which obtained Pearson correlation scores between 0.90 and 0.91 for all metrics (Fluency, Engagingness and Consistency). This is in favour of our proposed reward function and supports its application to training open-ended dialogue agents.

---

[7]All agents responded to the same human conversants, and they used the same sets of candidate sentences—for a fair comparison.

[8]Note that the candidate sentences used as distractors were chosen randomly from randomly selected dialogues—which is rather challenging for action selection. Future work could consider candidate sentences from similar dialogues for potential improvements in terms of engagingness and consistency.

Dialogue Evaluation

2. Dialog_2

A: hello there , how are you ?
B: i am a software engineer working for boeing . i just moved stateside to live with family .
A: wow ! i am actually pregnant too
B: it was my favorite toy as a child . what kind of music do you like ?
A: that is great i am expecting twins in two months . will these be your first kids ?
B: no , i already have step children , they are a pain .
A: yikes . i already have great kids
B: well good for you , i am getting a bigger house now since all the kids
A: i just bought my first one . it was very exciting
B: will you be driving to or living on campus ?
A: that is great . do you eat meat ? because i do
B: just playing the guitar , piano and singing . i like to color too !
A: oh that is good for you
B: no my sister used to , helped her become a pole dancer
A: i do not work i am a stay at home mom and i love it
B: me too , well i guess i should start eating meat now

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Is this dialogue Fluent?	○	○	●	○	○
Is this dialogue Engaging?	○	○	●	○	○
Is this dialogue Consistent?	○	○	●	○	○

Previous   Next

Figure 5: Screenshot of our dialogue evaluation tool
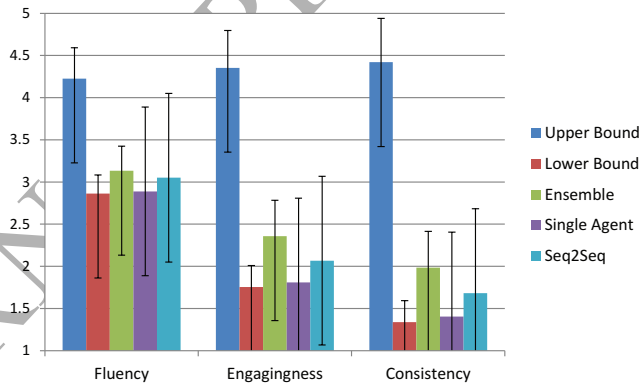
Figure 6: Human evaluation results, the higher the better

(a) Fluency

(b) Engagingness

(c) Consistency

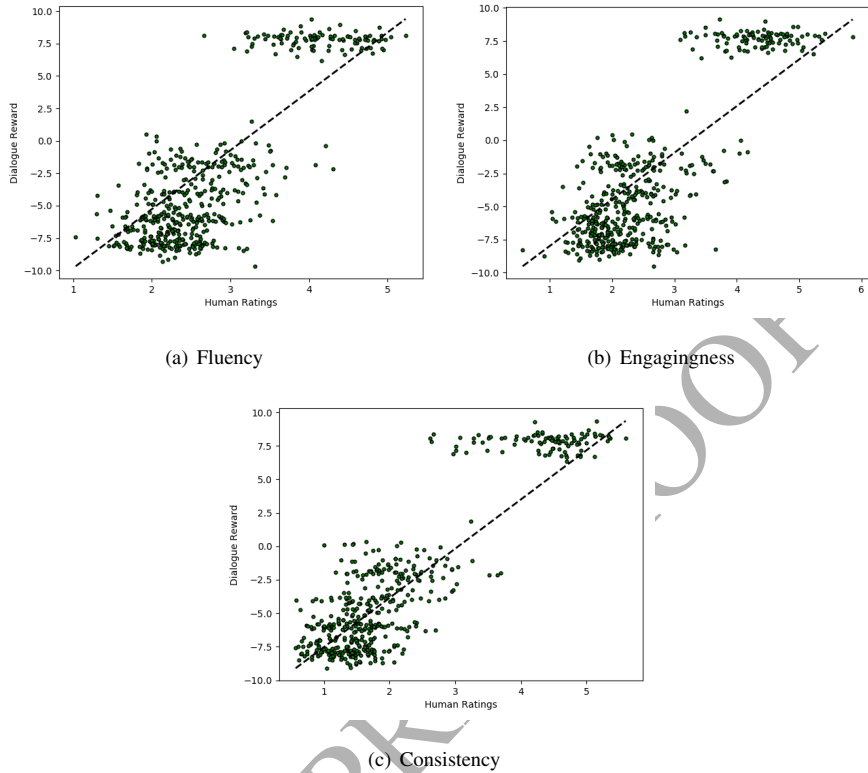Figure 7: Scatter plots showing strong correlations (with Pearson coefficients of $0.90 \le r \le 0.91$) between predicted dialogue rewards and average human ratings as part of the human evaluation, i.e. our proposed reward function correlates with human judgements

6. Conclusions and Future Work

We present a novel approach for training Deep Reinforcement Learning (DRL) chatbots. It uses an ensemble of 100 DRL agents based on clustered dialogues, clustered actions, and rewards derived from human-human dialogues without any manual annotations. The task of the agents is to learn to choose human-like actions (sentences) out of candidate responses including human generated and randomly chosen sentences. Our ensemble trains specialised agents with particular dialogue strategies according to their dialogue clusters. At test time, the agent with the highest predicted reward is used during a dialogue. Experimental results using chitchat dialogue data report that DRL

23

agents learn human-like dialogue policies when tested on training data, but their generalisation ability in a test set of unseen dialogues (with mostly unseen sentences, only 4.62% seen sentences to be precise) remains a key challenge for future research in this field. As part of our study, we found the following:

1. an ensemble of DRL agents is more promising than a single DRL agent or a single Seq2Seq model—confirmed by a human evaluation;

2. value-based DRL can be used for training chatbots—previous work mostly uses policy search methods due to infinite action sets; and

3. our proposed reward function albeit simple was useful for training chatbots.

Future work can investigate further the proposed learning approach for improved generalisation in test dialogues. Some research avenues are as follows.

- Investigating other methods of sentence embedding such as [46, 47, 48] – possibly with fine-tuning or domain adaptation subject to using large datasets. Other DRL algorithms such as policy search methods should also be compared or combined. In addition, other distance metrics and clustering algorithms should be used to investigate better sentence clustering and dialogue clusterings. Alternative dialogue rewards should be compared to train agents with human-like dialogue rewards across different datasets.

- An interesting future direction is training an ensemble of ChatDQN agents using a very large dataset – much larger than attempted in this article. Our results seem to suggest that the larger the dataset the better generalisation on unseen data. But this requirement represents high costs in data collection and high computational expense for system training. While chatbot training using large or very large datasets is interesting, chatbot training using modest datasets is still relevant because it can save costly datasets and computational requirements.

- The proposed learning approach focuses on value-based deep reinforcement learning, and it could be combined with other deep learning methods in order to investigate more effective ensembles of machine learners. For example, our ensemble of agents could include not only value-based DRL methods but also policy

search methods and a variety of seq2seq methods. Although this research direction represents increased computational expense, it has the potential of showing improved performance over single agents/models.

- Our proposed approach in this article did not include any linguistic resources. One reason for this is the practical application of DRL agents to other languages/datasets, where linguistic resources are scarce or do not exist. Another reason is due to the fact that linguistic resources usually come at the expense of labelled data, and we aimed for investigating an approach and methodology assuming unlabelled data only. However, future work could improve the performance of DRL agents by including knowledge bases and natural language resources such as part-of-speech tagging, named entity recognition, coreference resolution, and syntactic parsing [49].

- Last but not least, the proposed approach can be applied to different applications, beyond chitchat dialogue. Example applications in no particular order are as follows: combining task-oriented dialogue with open-ended dialogue [50], strategic dialogue [51], spatially-aware dialogue [52], automatic (medical) diagnosis [53], in-car infotainment systems [54], and conversational robots [55], among others.

References

[1] B. J. Grosz, C. L. Sidner, Attention, intentions, and the structure of discourse, Computational Linguistics 12 (3) (1986) 175–204.

[2] R. S. Sutton, A. G. Barto, Reinforcement learning - an introduction, 2nd Edition, Adaptive computation and machine learning, MIT Press, 2018.

[3] T. Hastie, R. Tibshirani, J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, 2nd Edition, Springer series in statistics, Springer, 2009.

[4] Y. LeCun, Y. Bengio, G. E. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

25

[5] I. V. Serban, C. Sankar, M. Germain, S. Zhang, Z. Lin, S. Subramanian, T. Kim, M. Pieper, S. Chandar, N. R. Ke, S. Rajeswar, A. de Brébisson, J. M. R. Sotelo, D. Suhubdy, V. Michalski, A. Nguyen, J. Pineau, Y. Bengio, A deep reinforcement learning chatbot (short version), CoRR abs/1801.06700.

[6] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, J. Gao, Deep reinforcement learning for dialogue generation, in: EMNLP, 2016.

[7] H. Cuayáhuitl, Simpleds: A simple deep reinforcement learning dialogue system, in: Dialogues with Social Robots - Enablements, Analyses, and Evaluation, Seventh International Workshop on Spoken Dialogue Systems, IWSDS 2016, Saariselkä, Finland, January 13-16, 2016, 2016, pp. 109–118.

[8] J. Gao, M. Galley, L. Li, Neural approaches to conversational AI, Foundations and Trends in Information Retrieval 13 (2-3) (2019) 127–298.

[9] Y. Li, Deep reinforcement learning: An overview, CoRR abs/1701.07274.

[10] I. Casanueva, P. Budzianowski, P. Su, S. Ultes, L. M. Rojas-Barahona, B. Tseng, M. Gasic, Feudal reinforcement learning for dialogue management in large domains, in: NAACL-HLT, 2018.

[11] H. Cuayáhuitl, S. Yu, Deep reinforcement learning of dialogue policies with less weight updates, in: INTERSPEECH, 2017.

[12] H. Cuayáhuitl, S. Yu, A. Williamson, J. Carse, Scaling up deep reinforcement learning for multi-domain dialogue systems, in: IJCNN, 2017.

[13] J. D. Williams, K. Asadi, G. Zweig, Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning, in: ACL, 2017.

[14] B. Peng, X. Li, L. Li, J. Gao, A. Çelikyilmaz, S. Lee, K. Wong, Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning, in: EMNLP, 2017.

[15] H. Cuayáhuitl, SimpleDS: A simple deep reinforcement learning dialogue system, in: International Workshop on Spoken Dialogue Systems (IWSDS), 2016.

[16] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, D. Jurafsky, Adversarial learning for neural dialogue generation, in: EMNLP, 2017.

[17] O. Vinyals, Q. V. Le, A neural conversational model, CoRR abs/1506.05869.

[18] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao, B. Dolan, A neural network approach to context-sensitive generation of conversational responses, in: HLT-NAACL, 2015.

[19] I. V. Serban, T. Klinger, G. Tesauro, K. Talamadupula, B. Zhou, Y. Bengio, A. C. Courville, Multiresolution recurrent neural networks: An application to dialogue response generation, in: AAAI, 2017.

[20] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, W. B. Dolan, A persona-based neural conversation model, in: ACL, 2016.

[21] W. Wang, M. Huang, X.-S. Xu, F. Shen, L. Nie, Chat more: Deepening and widening the chatting topic via a deep model, in: SIGIR, ACM, 2018.

[22] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, J. Weston, Personalizing dialogue agents: I have a dog, do you have pets too?, CoRR abs/1801.07243.

[23] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: NIPS, 2014.

[24] Y. Song, C. Li, J. Nie, M. Zhang, D. Zhao, R. Yan, An ensemble of retrieval-based and generation-based human-computer conversation systems, in: IJCAI, 2018.

[25] I. Papaioannou, A. C. Curry, J. L. Part, I. Shalyminov, X. Xu, Y. Yu, O. Dusek, V. Rieser, O. Lemon, An ensemble model with ranking for social dialogue, CoRR abs/1712.07558.

[26] C. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, J. Pineau, How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, in: EMNLP, 2016.

27

[27] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: A method for automatic evaluation of machine translation, in: ACL, 2002.

[28] A. Lavie, A. Agarwal, Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments, in: Proceedings of the Second Workshop on Statistical Machine Translation (StatMT), 2007.

[29] L. F. D'Haro, R. E. Banchs, C. Hori, H. Li, Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics, Computer Speech & Language 55 (2019) 200–215.

[30] C. Szepesvári, Algorithms for Reinforcement Learning, Morgan and Claypool Publishers, 2010.

[31] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, Nature 518 (7540).

[32] H. van Hasselt, A. Guez, D. Silver, Deep reinforcement learning with double Q-learning, in: AAAI, 2016.

[33] Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, N. de Freitas, Dueling network architectures for deep reinforcement learning, in: ICML, 2016.

[34] H. Cuayáhuitl, S. Ryu, D. Lee, S. Choi, I. Hwang, J. Kim, Deep reinforcement learning for chatbots using clustered actions and human-likeness rewards, in: IJCNN, 2019.

[35] M. A. Wiering, H. van Hasselt, Ensemble algorithms in reinforcement learning, Trans. Sys. Man Cyber. Part B 38 (4).

[36] X. liang Chen, L. Cao, C.-X. Li, Z.-X. Xu, J. Lai, Ensemble network architecture for deep reinforcement learning, Mathematical Problems in Engineering Article ID 2129393.

[37] H. Cuayáhuitl, S. Ryu, D. Lee, J. Kim, A study of dialogue reward prediction for open-ended conversational agents, in: NeurIPS Workshop on Conversational AI: "Today's Practice and Tomorrow's Potential", 2018.

[38] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: NIPS, 2013.

[39] J. Pennington, R. Socher, C. D. Manning, GloVe: Global vectors for word representation, in: EMNLP, 2014.

[40] M. Iyyer, V. Manjunatha, J. L. Boyd-Graber, H. D. III, Deep unordered composition rivals syntactic methods for text classification, in: ACL (1), 2015.

[41] D. Arthur, S. Vassilvitskii, K-means++: The advantages of careful seeding, in: SODA, SIAM, 2007.

[42] A. H. Miller, W. Feng, D. Batra, A. Bordes, A. Fisch, J. Lu, D. Parikh, J. Weston, ParlAI: A dialog research software platform, in: EMNLP, 2017.

[43] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: EMNLP, 2014.

[44] M. E. Tipping, C. Bishop, Probabilistic principal component analysis, Journal of the Royal Statistical Society, Series B 21/3 (1999) 611622.

[45] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: ICML, 2015.

[46] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, 2019.

[47] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, R. Kurzweil, Universal sentence encoder, CoRR abs/1803.11175.

29

[48] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: ICML, 2014.

[49] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, D. McClosky, The
510   Stanford CoreNLP natural language processing toolkit, in: ACL, 2014.

[50] I. Papaioannou, O. Lemon, Combining chat and task-based multimodal dialogue for more engaging HRI: A scalable method using reinforcement learning, in: International Conference on Human-Robot Interaction (HRI), 2017.

[51] H. Cuayáhuitl, S. Keizer, O. Lemon, Strategic dialogue managent via deep rein-
515   forcement learning, in: NIPS Deep Reinforcement Learning Workhop, 2015.

[52] N. Dethlefs, Domain transfer for deep natural language generation from abstract meaning representations, IEEE Comp. Int. Mag. 12 (3) (2017) 18–28.

[53] Z. Wei, Q. Liu, B. Peng, H. Tou, T. Chen, X. Huang, K. Wong, X. Dai, Task-oriented dialogue system for automatic diagnosis, in: Proceedings of the 56th
520   Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers, 2018.

[54] F. Weng, P. Angkititrakul, E. Shriberg, L. P. Heck, S. Peters, J. H. L. Hansen, Conversational in-vehicle dialog systems: The past, present, and future, IEEE Signal Process. Mag. 33 (6) (2016) 49–60.

525 [55] H. Cuayhuitl, A data-efficient deep learning approach for deployable multimodal social robots, Neurocomputing, issn 0925-2312, 2019.

30

No conflict(s) of interest to the best knowledge of the authors.

Heriberto Cuayáhuitl is a Senior Lecturer in Computer Science in the College of Science at the
University of Lincoln, and member of the Lincoln Centre for Autonomous Systems (L-CAS). He
received a PhD from the University of Edinburgh in 2009, and has an international research profile in
academia and industry in the discipline of machine intelligence including (spoken) language
processing, (deep) machine learning, (multimodal) dialogue systems and robotics. He has published
80 research papers in these areas, has been led organiser of the international workshop on Machine
Learning for Interactive Systems (MLIS), and guest editor of the journals ACM Transactions on
Interactive Interactive Systems and Elsevier Computer Speech and Language. His work in industry has
been carried out at SpeechWorks International Inc. (now Nuance Communications Inc.), the German
Research Center for Artificial Intelligence (DFKI), and Samsung Research.

Donghyeon Lee is a staff engineer and member of the Language Understanding Lab in the AI Center
at Samsung Research in South Korea since 2014. He received his B.S. degree in Computer Science and
Engineering from Sungkyunkwan University in 2006. He received his Ph.D. degree in Computer
Science and Engineering from Pohang University of Science and Technology in 2014 for his
dissertation on Unsupervised Spoken Language Understanding for Multi-domain Dialogue Systems. He
is interested in deep learning technology for natural language understanding in multi-domain dialog
systems.

Seonghan Ryu is a staff engineer and member of the Language Understanding Lab in the AI Center at
Samsung Research in South Korea since 2018. He received his B.S. degree in Computer Science and
Engineering from Dongguk University in 2012. He received his Ph.D. degree in Computer Science and
Engineering from Pohang University of Science and Technology in 2018 for his dissertation on Out-ofdomain
Detection based on Deep Learning for Dialogue Systems. His thesis investigated sentence
embedding for extracting meaningful features about the domain of the input sentence, and one-class
classification for rejecting out-of-domain sentences. He is interested in deep learning technology for
natural language understanding in multi-domain dialog systems.

Yongjin Cho is a staff engineer and member of the Language Understanding Lab in the AI Center at
Samsung Research in South Korea since 2018. He recieved his B.E. degree in Computer Science and
Engineering from Dongguk University in 2018. His undergraduate research in Dongguk University focused on a recommendation system using genetic algorithms. He interested in deep learning technology for natural language understanding in dialogue systems.

Sungja Choi is a staff engineer and member of the Language Understanding Lab in the AI Center at
Samsung Research in South Korea since 2012. She received her B.S. degree in Computer Science and
Engineering from Pusan National University in 2003, and her M.S. degree in Computer Science and
Engineering from the same institution in 2005 for her dissertation on a Korean Word-Spacing Method
Using Word and Syllable Statistics. She is interested in deep learning technology for natural language
understanding in dialogue systems.

Sathish Reddy received his M.E. in Computer Science and Engineering from Indian Institute of
Science, Bangalore, India, in 2009. Since October 2016, he is a Senior Research Engineer at Samsung
Research in South Korea. Prior to this, he worked as Research Engineer at IBM Research in India. His
research interests include Data Mining, Artificial Intelligence, Robotics, and Machine Learning. He
has published papers in EACL, EMNLP, ACL, SIGDIAL, Elsevier Pattern Recognition Journal.

Seunghak Yu is a postdoctoral associate at the Massachusetts Institute of Technology in the Computer Science and Artificial Intelligence Laboratory. Since obtaining his M.S. and Ph.D. from Korea University, his research has focused on machine learning and deep learning for natural language processing.

Hyungtak Choi is a staff engineer of the Language Understanding Lab in the AI Center at Samsung
Research, and technical leader of the Natural Language Generation Group. His current research includes deep learning models of natural language processing and generation, style transfer in dialogue response, emotional response generation, conversational agents, multi-modal dialogue generation and the language generation engine for dialogue systems. He worked on the Bixby voice assistant project, developing new engines for dialogue management and generation. He has more
than 20 U.S. patents granted or pending. He received his M.S. in Computer Science and Engineering
from Yonsei University for his dissertation on collaborative filtering recommendation systems using
semantic reasoning by genre.

Inchul Hwang is a vice president of Samsung Research at Samsung Electronics, and leader of Language Understanding Lab in the AI Center. He received his B.S, M.S and PhD from KAIST (Korea
Advanced Institute of Science & Technology). As soon as he graduated from KAIST, he joined Samsung
from 2006 and worked on embedded systems, IPTV, Convergence&IoT, Cloud, Big data, among others.
His current research focus is building the human-like virtual assistant (Bixby service in Samsung Electronics) especially based on dialogue technology, language understanding and generation, and
contextual interaction with users using AI technology.

Jihie Kim is a vice president of the AI Center at Samsung Electronics. She has been managing research on applied NLP and machine learning for improving user interactions of Samsung devices
including Bixby voice assistant. Before joining Samsung, she was a senior vice president at KT (Korea
Telecom) and managed big data projects. She was research faculty in the Computer Science Department at the University of Southern California and a lead researcher at the USC/Information
Sciences Institute. She was the PI of six NSF (National Science Foundation) projects on AI and education. She organized a series of five workshops on Intelligent Support for Learning in Groups. Dr.
Kim was the general chair of the IUI (Intelligent User Interfaces) conference 2013, and served as the
poster chair, workshop and tutorial chair, publicity chair of several AIEd and IUI conferences. She has
been a senior program committee member of AAAI 2016 and IJCAI 2016 and a program committee
member of various AI conferences.