A Study on Dialogue Reward Prediction for Open-Ended Conversational Agents

Heriberto Cuayáhuitl[1], Seonghan Ryu[2], Donghyeon Lee[2], Jihie Kim[2]
[1]University of Lincoln, School of Computer Science, Lincoln, United Kingdom
[2]Samsung Research, Artificial Intelligence Team, Seoul, South Korea
HCuayahuitl@lincoln.ac.uk, {seonghan.ryu,dh.semko.lee,jihie.kim}@samsung.com

Abstract

The amount of dialogue history to include in a conversational agent is often underestimated and/or set in an empirical and thus possibly naive way. This suggests that principled investigations into optimal context windows are urgently needed given that the amount of dialogue history and corresponding representations can play an important role in the overall performance of a conversational system. This paper studies the amount of history required by conversational agents for reliably predicting dialogue rewards. The task of dialogue reward prediction is chosen for investigating the effects of varying amounts of dialogue history and their impact on system performance. Experimental results using a dataset of 18K human-human dialogues report that lengthy dialogue histories of at least 10 sentences are preferred (25 sentences being the best in our experiments) over short ones, and that lengthy histories are useful for training dialogue reward predictors with strong positive correlations between target dialogue rewards and predicted ones.

1   Introduction

Deep Reinforcement Learning holds a lot of promise for training intelligent conversational machines [2, 4, 3], which can infer their behaviour from trial and error via interaction with an environment in order to optimise a long term reward signal [19, 9]. This machine learning paradigm is interesting and worth studying because it resembles human learning [16, 12], and has wide application from low-level to high-level control using a single or multiple modalities. However, its successful application is not trivial due to interactions with real world environments (not easy to model), complex decision-making at different levels of granularity (not easy to infer or specify), and unkwown reward signals (also not easy to infer or specify), among others. In this paper we focus our attention to the latter only, the reward signals, in the context of conversational agents. Specifically, we focus on open-ended dialogue agents because they have received little attention so far in the literature—as described in Section 2. This type of agents are interesting and challenging at the same time because the amount of input features and their combinations are simply vast, making decision-making difficult. Even when some reward functions have been proposed by previous works [7, 8, 17] it is unclear what is the best reward function to use for training future conversational agents or chatbots.

Our study uses a dataset of human-human dialogues, which we extend with noisy dialogues by replacing human responses with randomly chosen responses from other dialogues. While non-noisy dialogues aim to show human-like and desirable outputs, the noisy ones aim to exemplify less desirable behaviour. Each dialogue is automatically labelled with a numerical reward according to the amount of noise (distorted sentences)—the automatic labels facilitate its application to other datasets. This extended dataset is used for predicting dialogue rewards of open-ended conversations. As part of our study, we propose a new reward function that is easy to implement and that strongly correlates with test human-human dialogues. The latter is indeed possible by using lengthy dialogue histories.

2 Related Work

So far little attention has been devoted to predicting numerical rewards of open-ended dialogues, which can be used to optimise the behaviour of dialogue agents such as chit-chat chatbots. By dialogue reward prediction we mean 'estimating the quality of human-computer or human-human dialogues', where quality can be measured for example via task success, user satisfaction or human-likeness, among others. Most previous related works focus on task-oriented dialogue systems, which require labelled data typically collected by a previously developed dialogue system. For example, [18] used previously collected human-computer dialogues in the domain of restaurants for training an LSTM-based dialogue regressor for predicting rewards of the form $\mathcal{R} = 20 \times 1_{success} - N$, where a value of 20 is given for a successful dialogue minus the number of dialogue turns. Such a regressor—which can be turned into a binary classifier for predicting success or failure—was re-trained online together with a dialogue policy using active learning. [13] also use LSTM-based predictors of dialogue success (or failure) in the programming domain (StackOverflow data) based on word-based features, where success is given to a correctly answered question—failure otherwise. Similarly, [14] explores other machine learning techniques for predicting dialogue success but from acoustic features instead of word-based features. [20] predict dialogue rewards of the form $\mathcal{R} = (iq - 1) \times 5 - N$, where iq is an interaction quality value derived from a multiclass classifier. While speech recognition features and dialogue actions are used as inputs, the outputs used for training were obtained from three human labellers that judged the quality of the interactions. The reward predictor induced from a dataset of bus information was used for training policies in the domains of restaurants, hotels and laptops.

The related studies above, motivated by the earlier work of [21] in predicting problematic dialogues, have focused on task-oriented dialogues that aim to achieve a goal such as finding restaurant/hotel/laptop information, call routing, etc., see summary in Table 1. There is however little related research on human-human or human-computer dialogues where there is no clear notion of a task or goal to achieve—notable exceptions are [17, 8, 7]. [17] train a neural multiclass classifier and linear regressor from human-chatbot dialogues from Amazon Mechanical Turk (AMT). In this study, AMT human evaluators are given dialogues and are asked to rate the quality of candidate responses according to the following categories (or labels): 1 indicates an innapropriate response or one that does not make sense, 3 indicates neutral, and 5 is a highly appropriate and interesting response. The classifier receives inputs with 1458 features (including lexical, syntactic and semantic information) and outputs a label out of the five possible labels. The regressor aims to learn a reward function for open-ended dialogues. It receives inputs with 23 features (also including lexical, syntactic and semantic information) and outputs a real value between 1 and 5. This regressor was used to train neural-based reinforcement learning dialogue policies for open-ended dialogues – [11] used a similar approach in a restricted set of domains. [8] train a binary classifier for predicting whether a dialogue is human-generated or machine generated. The outputs of this classifier, derived from a two-class Softmax function, are used as dialogue rewards in order to train a generator that aims for deceiving such a classifier within an adversarial framework. [7] produces rewards of the form $r(a, [p_i, q_i]) = \lambda_1 r_1 + \lambda_2 r_2 + \lambda_3 r_3$, where a refers to the generated response, $[p_i, q_i]$ are the previous two dialogue turns, r_1 and r_3 refer to likelihood outputs of $Seq2Seq$ models, r_2 refers to dialogue history similarity, and λ_i are empirically defined weights (easy of answering: $\lambda_1 = 0.25$, information flow: $\lambda_2 = 0.25$, and semantic coherence: $\lambda_3 = 0.5$). While [17] reports a 0.40 Pearson correlation coefficient between predicted rewards and AMT human judgements, it is unclear whether or not the binary classifier proposed by [8] or heuristic function of [7] can produce rewards that correlate with human judgements.

The work presented in the remaining of the paper goes beyond previous works as follows:

1. we study dialogue reward prediction for open-ended dialogues using automatically labelled data. This is in contrast with [17, 11] who use paid human evaluators, which results in an expensive and time consuming process because large amounts of dialogues require labelling;

2. we use contextually-richer and automatically derived input features to represent dialogue histories. This is in contrast with [17, 11] who use a small set of carefully designed features and with [7, 8, 17] who use short dialogue histories for dialogue reward prediction; and

3. we investigate the impact of dialogue history length in dialogue reward prediction for open-ended conversational agents. Our best predictor—using rich dialogue histories—reports a strong positive correlation on unseen test dialogues, which is higher than the previously reported scores of [17] with only moderate positive correlation.

Authors	Domain(s)	Model
Su et al. [18]	Restaurants	Recurrent Neural Network (LSTM)
Noseworthy et al. [13]	Programming (StackOverflow)	Flat and Hierarchical LSTMs
Papangelis et al. [14]	Laptops, Restaurants	Support Vector Machines, Gaussian Process Regressors, Random Forests
Ultes et al. [20]	Bus, Restaurants, Hotels, Laptops	Support Vector Machine (SVM)
Walker et al. [21]	Call Routing	Rule Learner (RIPPER)
Misu, et al. [11]	Restricted QA in a Museum	Unknown
Li et al. [7, 8]	Open-Ended	Recurrent Neural Network (LSTM)
Serban at al. [17]	Open-Ended	Fully-Connected Neural Networks

Table 1: Related works on dialogue reward or dialogue success prediction

3   Proposed Study

We aim to answer the research questions **How dialogue rewards can be predicted in open-ended conversations? How much dialogue history do open-ended conversational agents need?**

This study focuses on open-ended dialogues as those observed in chit-chat interactions. On the one hand, the longer the history the richer the context to take into account, but the more computational expense involved (due to high amounts of features). On the other hand, the shorter the history the simpler the dialogue context at a cheaper computational expense. In addition, we study the effects of short and long dialogue histories in the task of dialogue reward prediction—often used for dialogue optimisation—and we treat dialogue reward prediction as a regression problem.

Let $\mathcal{D} = \{d_1, \ldots, d_N\}$ be a set of open-ended human-human conversations in raw text, where each dialogue $d_i = \{t_1, \ldots, t_T\}$ contains a set of dialogue turns $t_j^i = \{s_1^1, \ldots, s_X^1\}$, and each sentence $s_{j,k}^i = \{w_{1,1}^1, \ldots, w_{1,Y}^1\}$ contains a set of words as part of the verbal contribution in dialogue turn i, j. We use $T = |d_i|$ to denote the length—in terms of number of dialogue turns—of dialogue i. Assuming two partner conversants in each dialogue (our case), a dialogue turn has sentences of the form $s_j^i = \{(s_{j_A}^i, s_{j_B}^i)\}$—one per conversant in each dialogue turn, e.g. human A says sentence $s_{1_A}^i$, human B says sentence $s_{1_B}^i$, then human A says sentence $s_{2_A}^i$, human B says sentence $s_{2_B}^i$, and so on until the end of dialogue i.

We propose to derive dialogue rewards with automatically labelled data (for practical application) by extending a dataset of human-human dialogues and assigning *positive values* to responses seen in dataset $\mathcal{D}$, and *negative values* to randomly generated responses due to lacking coherence (also referred to as 'non-human-like responses'). A complete dialogue can be rewarded as

$$R_i = \sum_{j=1}^{T} r_j^i(a),$$

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

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

Table 2 shows an example of a well rewarded dialogue (without distortions) and Table 3 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 Algorithm 1 for generating dialogues with varying amounts of distortions (i.e. different degrees of human-likeness), which will be used for training and testing reward prediction models. Given our extended dataset $\hat{\mathcal{D}} = \{(\hat{d}_1, y_1), \ldots, (\hat{d}_N, y_N)\}$ with (noisy) dialogue histories $\hat{d}_i$, the goal is to predict dialogue scores y_i as accurately as possible.

Algorithm 1 Generator of (Distorted) Dialogues

1: $\hat{\mathcal{D}} \leftarrow [\emptyset]$ ▷ Initialise dataset of original and distorted dialogues
2: **for** dialoge $i = 1, \ldots, N$ in set $\mathcal{D}$ **do**   ▷ Iterate over the set of original dialogues
3:     **for** noise $n = 0, \ldots, |d_i|$ **do**   ▷ Iterate over different amounts of noise ($|d_i|$ = no. dialogue turns)
4:        $score = |d_i| - (n * 2)$ ▷ Score range: $(|d_i| \times -1), \ldots, |d_i|$
5:        $\hat{d} \leftarrow [\emptyset]$ ▷ Initialise distorted dialogue
6:        **for** dialogue turn $j = 0, \ldots, T - 1$ in dialogue i **do**   ▷ Iterate over all sentences in dialogue i
7:           $s_j^i \leftarrow$ sentence of first partner conversant
8:           $s_{j+1}^i \leftarrow$ sentence of second partner conversant
9:           **if** $j < n$ **then**
10:              $\hat{s}_{j+1}^i \leftarrow$ distorted sentence of second partner conversant
11:              $\hat{d} \leftarrow$ append tuple $< s_j^i, \hat{s}_{j+1}^i >$ ▷ Append distorted dialogue turn
12:           **else**
13:              $\hat{d} \leftarrow$ append tuple $< s_j^i, s_{j+1}^i >$ ▷ Append non-distorted dialogue turn
14:           **end if**
15:           $j = j + 1$
16:        **end for**
17:        $\hat{\mathcal{D}} \leftarrow$ append tuple $< \hat{d}, score >$ ▷ Append newly generated (noisy) scored dialogue
18:        $n = n + 1$
19:     **end for**
20:     $i = i + 1$
21: **end for**

---

4   Data

We used data from the *Persona-Chat* dataset[1], which includes 17,877 dialogues for training and 999 dialogues for testing. The average dialogue length is 7.35 and 7.8 dialogue turns (166 and 186 words per dialogue, including punctuation) for training and testing, respectively. The vocabulary size in the entire dataset contains 19,667 unique words. See example dialogue in Table 2. We used this dataset as seed dialogues in order to generate dialogues with different amounts of noise (or distortions)—as described in the previous section—and obtained 149,308 dialogues for training and 8,704 for testing.

---

[1]Dataset downloaded from `http://parl.ai/` on 18 May 2018 [10]

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

4

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 .~~ B: lol. she is more pop now. her old stuff was country. i like classic country .	 +1 -1
2	A: i just got done watching a horror movie B: ~~i rather read , i have read about 20 books this year .~~ B: i am good how about you	 +1 -1
3	A: wow ! i do love a good horror movie . loving this cooler weather B: ~~but a good movie is always good .~~ B: i do not enjoy the driving in my honda civic	 +1 -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 .~~ B: do you like dogs ? I have two .	 +1 -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 .~~ B: very cool . i had one of those .	 +1 -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 .~~ B: that is just great . we need more educators like you in the system .	 +1 -1
7	A: i bet she appreciates that very much . B: ~~she raised me right , i'm just like her .~~ B: is it bad that i hate pumpkins and trees ?	 +1 -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 .~~ B: why not ? maybe you do not like to travel ?	 +1 -1

Table 3: Example distorted human-human dialogue with a poor numerical reward $\sum_{j=1}^{8} r_j^2(a) = -8$

5 Experiments and Results

5.1 Experimental Setting

We represent a dialogue history via its mean word vectors as in Deep Averaging Networks [6], where sentences are represented with numerical feature vectors denoted as $\mathbf{x} = \{x_1, ..., x_{|\mathbf{x}|}\}$. In this way, a set of word sequences s_j^i in dialogue-sentence pair i, j is mapped to feature vectors

$$\mathbf{x}_j^i = \frac{1}{N_j^i} \sum_{k=1}^{N_j^i} c_{j,k}^i,$$

where $c_{j,k}^i$ is the vector of coefficients of word k, part of sentence j in dialogue i, and N_j^i is the number of words in the sentence in focus.

To complete the preparation of our training and test data, vector $\mathbf{Y} = \{y_1, ..., y_{|\mathbf{Y}|}\}$ is the set of target labels—generated as described in Section 3. In this way, dataset $\mathcal{D}^{train} = (\mathbf{X}^{train}, \mathbf{Y}^{train})$ is used for training neural regression models using varying amounts of dialogue history, and dataset $\mathcal{D}^{test} = (\mathbf{X}^{test}, \mathbf{Y}^{test})$ is used for testing the learnt models.

All our experiments use a 2-layer Gated Recurrent Unit (GRU) neural network [1], see Figure 1. In this recurrent neural network—at each time step t in a dialogue history—the first hidden layer generates a hidden state $\mathbf{h}_t$ as follows:

$$\mathbf{r}_t = \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1}),$$
$$\mathbf{z}_t = \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1}),$$
$$\bar{\mathbf{h}}_t = \tanh(\mathbf{W}_{\bar{\mathbf{h}}} \mathbf{x}_t + \mathbf{U}_{\bar{\mathbf{h}}}(\mathbf{r}_{t-1} \odot \mathbf{h}_{t-1})),$$
$$\mathbf{h}_t = \text{BN}_{\gamma,\beta} \left[(1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \bar{\mathbf{h}}_t \right],$$

where $\mathbf{r}_t$ is a reset gate that decides how much to forget the previous state, $\mathbf{z}_t$ is an update gate that decides how much it updates its activation, $\bar{\mathbf{h}}_t$ is an internal state, $\sigma(.)$ and $tahn(.)$ are the Sigmoid
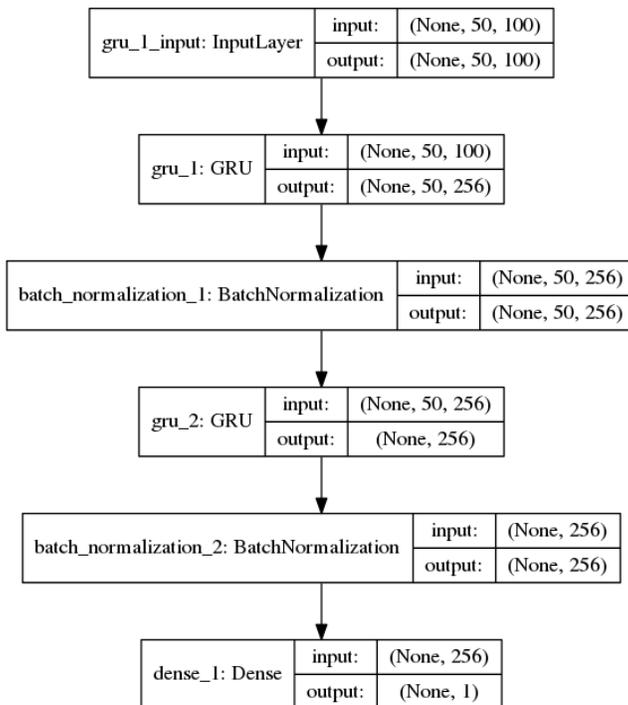
Figure 1: Illustration of our recurrent neural network for dialogue reward prediction

and hyperbolic Tangent functions (respectively), $\mathbf{W}_*$ and $\mathbf{U}_*$ are learnt weights, $\odot$ refers to the element-wise multiplication, and $\text{BN}_{\gamma,\beta}$ refers to Batch Normalisation with learnt weights γ and β [5]. Assuming that the equations above can be summarised as $\mathbf{h}_t = GRU(\mathbf{x}_t, \mathbf{h}_{t-1})$ we get the following output taking into account both hidden layers i in our neural network

$$\mathbf{h}_t^1 = GRU(\mathbf{x}_t, \mathbf{h}_{t-1}^1),$$
$$\mathbf{h}_t^2 = GRU(\mathbf{h}_t^1, \mathbf{h}_{t-1}^2),$$
$$o_t = \mathbf{W}_o \mathbf{h}_t^2,$$

where o_t refers to the predicted dialogue reward for dialogue history $\mathbf{x}_t$.

We trained this neural network[2] for varying amounts of dialogue lengths, where $\mathbf{x}_t$ included 1, 3, 5, 10, 25 and 50 sentences. Each sentence was represented with a mean word vector using the pretrained coefficients of the Glove model [15] (file glove.840B.300d.txt to be precise). Our experiments used the first 100 dimensions per word vector—empirical evidence reported that this dimensionality is a good compromise between prediction performance and computational expense. Note that the larger the word vector dimensionality the more parameters in the neural network. Other hyperparameters in our experiments include batch size=128, dropout=0.2, dimensionality=256, loss function=Mean Absolute Error (MAE), optimiser=Adam, epochs=100, and early stopping=10 epochs. In addition, the training set was split into 80% for training and 20% for validation. The number of parameters in our neural network corresponds to 668K.

5.2 Experimental Results

We trained neural networks for six different lengths of dialogue history, ranging from 1 sentence to 50 sentences. Each length size involved a separate neural network, trained 10 times in order to report results over multiple runs. Figure 2 reports the average Pearson correlation coefficient—between true dialogue rewards and predicted dialogue rewards—for each length size. It can be observed that short dialogue histories contribute to obtain weak correlations, and that longer dialogue histories (≥ 10

---

[2]Our experiments ran on a cluster of GPUs Tesla K80, and their implementation used the Keras library `https://github.com/keras-team/keras`
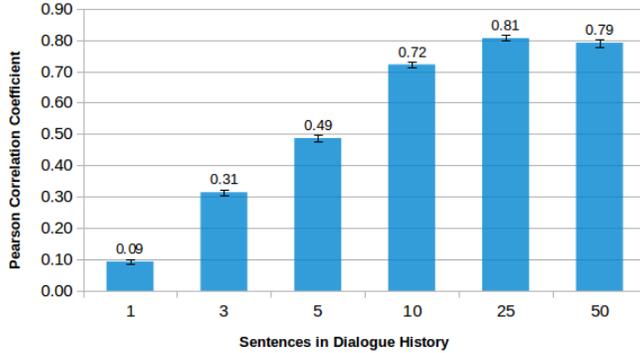
Figure 2: Bar plot showing the performance of our dialogue reward predictors using different amounts of dialogue history (from 1 sentence to 50 sentences). Each bar reports an average Pearson correlation score over 10 runs, where the coefficients report the correlation between true dialogue rewards and predicted dialogue rewards in our test data

sentences) contribute to obtain strong correlations. It can also be observed that the longest history may not be the best choice of length size, the network using 25 sentences achieved the best results.

Figure 3 shows a more detailed inspection of our results using a scatter plot for the best trained neural networks, one per dialogue length size, chosen according to their performance in the validation set. These scatter plots show the correlation between target dialogue rewards (in our test dataset) and predicted dialogue rewards. The data points are test examples, where the number of data points is the same as the number of examples in our test dataset. The scatter plots include Gaussian noise drawn from $\mathcal{N}(0, 0.3)$ in the X-axis for better visualisation at the expense of showing less correlations. It can be noted that 1 and 3 sentences in the dialogue history lead to weak correlations, 5 sentences in the dialogue history lead to moderate correlations, and 10 or more lead to strong positive correlations. From these results we can conclude that long (or very informative) dialogue histories should be used in open-ended conversational agents.

6   Conclusion and Future Work

This paper studies dialogue reward prediction for open-ended conversational systems. Our study aimed at finding a reliable dialogue reward predictor—useful for dialogue optimisation—using data specified in a principled and practical way. To achieve this we propose a new reward function, train neural regressors from automatically labelled data, and evaluate their impact using different dialogue history lengths. We adopted a dataset of human-human conversations and expanded it with noisy conversations in order to generate dialogues with varying amounts of noise, resulting in different dialogue rewards – the higher the numerical reward the more human-like dialogue and the lower the numerical reward the less human-like dialogue. Our new noisy dataset was used to train neural regressors that took into account different amounts of dialogue history in order to observe the predictive power of target dialogue rewards. We modelled sentences in dialogue histories via their mean word vectors, using pre-trained word embeddings. Experimental results show that dialogue histories including 10 or more sentences lead to good performance (i.e. strong correlations between target dialogue rewards and predicted ones), and that shorter histories lead to poorer prediction power (i.e. moderate or weak correlations between target dialogue rewards and predicted ones). A history size of 25 sentences reported the best results in our test data of chit-chat conversations. Our findings are in contrast with related work that have confined themselves to short dialogue histories of 2 turns [7] or 2 sentences [8, 17].

Future conversational agents or chatbots can optimise their behaviour using neural regressors with lengthy or informative dialogue histories. Other future works can also investigate the impact of different dialogue history sizes in other datasets and types of dialogues (e.g. task oriented conversations).
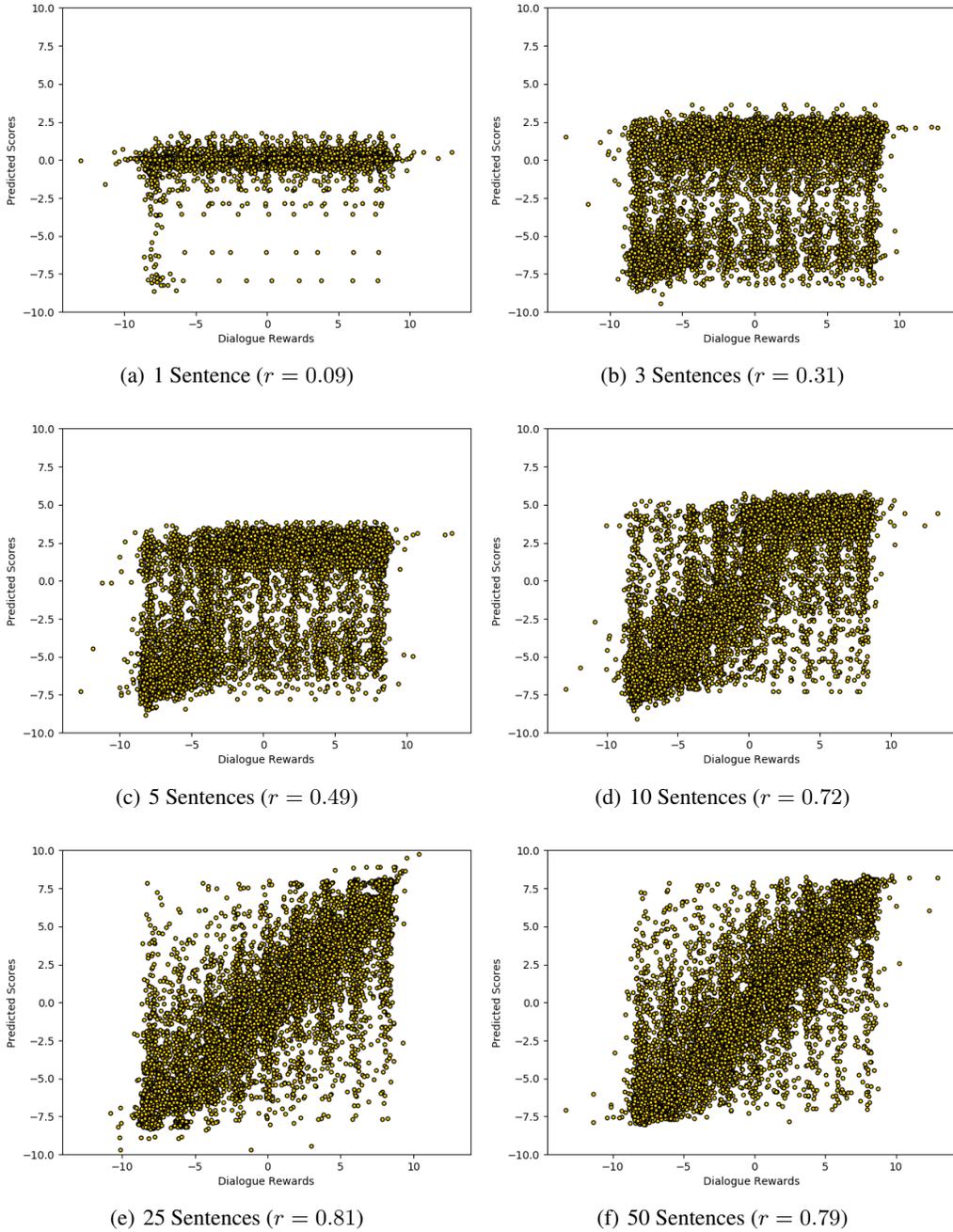
(a) 1 Sentence ($r = 0.09$)

(b) 3 Sentences ($r = 0.31$)

(c) 5 Sentences ($r = 0.49$)

(d) 10 Sentences ($r = 0.72$)

(e) 25 Sentences ($r = 0.81$)

(f) 50 Sentences ($r = 0.79$)

Figure 3: Scatter plots of dialogue reward predictors using different amounts of sentences in dialogue history, where r=correlation coefficients and X-axis includes Gaussian noise drawn from $\mathcal{N}(0, 0.3)$

References

[1] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[2] 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.

[3] L. Deng and Y. Liu. *Deep Learning in Natural Language Processing*. Springer Publishing Company, Incorporated, 1st edition, 2018.

[4] J. Gao, M. Galley, and L. Li. Neural approaches to conversational AI. In *ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018.

[5] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Intenational Conference on Machine Learning (ICML)*.

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

[7] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao. Deep reinforcement learning for dialogue generation. In *EMNLP*, 2016.

[8] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky. Adversarial learning for neural dialogue generation. In *EMNLP*, 2017.

[9] Y. Li. Deep reinforcement learning: An overview. *CoRR*, abs/1701.07274, 2017.

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

[11] T. Misu, K. Georgila, A. Leuski, and D. R. Traum. Reinforcement learning of question-answering dialogue policies for virtual museum guides. In *SIGDIAL*, 2012.

[12] 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, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540), 2015.

[13] M. Noseworthy, J. C. K. Cheung, and J. Pineau. Predicting success in goal-driven human-human dialogues. In *SIGDIAL*, 2017.

[14] A. Papangelis, M. Kotti, and Y. Stylianou. Predicting dialogue success, naturalness, and length with acoustic features. In *Intl. Conf. on Acoustics, Speech and Signal Proc.(ICASSP)*, 2017.

[15] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[16] W. Schultz, P. Dayan, and R. P. Montague. A neural substrate of prediction and reward. *Science*, 275:1593–1599, 1997.

[17] 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, and Y. Bengio. A deep reinforcement learning chatbot (short version). *CoRR*, abs/1801.06700, 2018.

[18] P. Su, M. Gasic, and S. J. Young. Reward estimation for dialogue policy optimisation. *Computer Speech & Language*, 51:24–43, 2018.

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

[20] S. Ultes, P. Budzianowski, I. Casanueva, N. Mrksic, L. M. Rojas-Barahona, P. Su, T. Wen, M. Gasic, and S. J. Young. Domain-independent user satisfaction reward estimation for dialogue policy learning. In *INTERSPEECH*, 2017.

[21] M. A. Walker, I. Langkilde-Geary, H. W. Hastie, J. H. Wright, and A. L. Gorin. Automatically training a problematic dialogue predictor for a spoken dialogue system. *J. Artif. Intell. Res.*, 16:293–31, 2002.