Journal of Intelligent Information Systems (2021) 56:189–206
https://doi.org/10.1007/s10844-020-006286-5

Using reinforcement learning with external rewards for open-domain natural language generation

Vidhushini Srinivasan[1] · **Sashank Santhanam**[1] · **Samira Shaikh**[1]

Received: 22 March 2020 / Revised: 13 October 2020 / Accepted: 14 October 2020 /
Published online: 17 November 2020

Abstract
We propose a new approach towards emotional natural language generation using bidirectional seq2seq model. Our goal is to generate emotionally relevant language that accommodates the emotional tone of the prior context. To incorporate emotional information, we train our own embeddings appended with emotion values through valence, arousal and dominance scores. We use a reinforcement-learning framework, which is tuned using policy gradient method. Two of the internal rewards in our reinforcement learning framework, viz. Ease of Answering and Semantic Coherence are based on prior state-of-the-art. We propose a new internal reward, Emotional Intelligence, computed by minimizing the affective dissonance between the source and generated text. We also train a separate external reward analyzer to predict the rewards as well as to maximize the expected rewards (both internal and external). We evaluate the system on two common corpora used for Natural Language Generation tasks: the Cornell Movie Dialog and Yelp Restaurant Review Corpus. We report standard evaluation metrics including BLEU, ROUGE-L and perplexity as well as human evaluation to validate our approach. We demonstrate the ability of proposed model to generate emotionally appropriate responses on both corpora.

Keywords Deep learning · Reinforcement learning · Emotional intelligence · Human feedback · Seq2seq learning · Conversational agent · Natural language generation

1 Introduction

Rapid advances in the field of generative modeling through the use of neural networks and large-scale language models has led to the creation of intelligent conversational agents. The typical architecture used to develop these conversational agents is the *seq2seq* framework, more widely used in the field of machine translation (Vinyals and Le 2015). However, prior research has shown that engaging with these agents produces dull and generic responses

✉ Samira Shaikh
samirashaikh@uncc.edu

[1] Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA

whilst also being inconsistent with the emotional tone of conversation (Vinyals and Le 2015; Li et al. 2016). These issues also affect engagement with the conversational agent, that leads to short conversations (Venkatesh et al. 2018). Apart from producing engaging responses, understanding the situation and producing the right emotional response to that situation is a desirable trait in a conversational agent (Rashkin et al. 2018). Our main goal is to develop models that are capable of generating language across multiple genres of text – say, conversational text and restaurant reviews, while also being consistent in the emotional tone of the context. After all, humans are adept at both.

Emotions are an important part of human interactions, in that, they can facilitate engagement with the conversation partner (Poria et al. 2019). For example, extant research has demonstrated that conversational agents that incorporate some emotional ability can lead to an increase in user satisfaction, fewer breakdowns in dialogues and more positive interactions (Prendinger et al. 2005; Prendinger and Ishizuka 2005; Martinovski and Traum 2003)

Such findings have led to recent work which incorporates emotion in conversational agents (Asghar et al. 2018; Zhou et al. 2018; Huang et al. 2018; Ghosh et al. 2017), however these approaches use a typical seq2seq framework, which have been shown to generate dull and generic responses. We instead approach the problem of emotional generation through a reinforcement learning framework. We wish to train the model to generate emotionally as well as topically relevant language, by learning the reward function and policy from a set of predetermined rewards. As shown in the illustrative example in Table 1, our goal is to generate language that is able to match the Input Context (a review in this case) as well as its emotional tone just as the human ground truth does.

Table 1 Example of generated text from Yelp Review dataset

Input Text/Context:	Human Ground Truth:
Ahhhh this place is so good! My friend and I came in around 1 pm and it was busy! Luckily there was some tables available so we were seated quickly.	This place is known for it's taste and is preferred by tourists from all over the world! There had been times when people wait in long queues for seats. But the waiting is forgotten when they start tasting the delicious and yummy food
	Model Output:
	The food was great and the service was spectacular! Entire staff was very friendly. The service was so personable and cashier woman so nice and they were very accommodating to large orders for offices! As good as it gets!

Our goal is to match the emotional tone of the context provided in the input text. The generated response is an example of our system output while Human Ground Truth is the actual response text for the given Input Text in our corpus.

Finally, we wish to test the hypothesis that language that is topically as well emotionally relevant would engender more engagement from those who read the text. However, obtaining qualitative or quantitative ratings of engagement is an expensive endeavor. To address this issue, researchers have demonstrated that using unlabeled and weakly-labeled data in order to scale up model training can be an effective strategy (Mintz et al. 2009). We can view a distant label as a noisy version of a true underlying label of a piece of data. We therefore use human ratings of usefulness present in one of our corpora as a proxy of how textual content can be relevant and engaging.

Thus, one of our key contributions is using such distantly labeled data in our architecture as an external reward for the reinforcement learning system to address the lack of enough supervised training data. The strategy of using distant labels is similar to prior work in this area (Hashimoto and Sassano 2018). However, this strategy has not been extensively applied to emotional natural language generation, a gap that our work attempts to address.

2 Related work

Our work is at the intersection of two major research thrusts – language generation and more specifically, generation of emotional language. Accordingly, we review related work in both areas in these subsections.

2.1 Language generation approaches

Sequence-to-Sequence (Seq2Seq) models (Sutskever et al. 2014) have been quite successful in neural conversational and language generation systems (Serban et al. 2016; Yao et al. 2017). Following their success, several techniques have been proposed to improve the content quality of dialogue (Li et al. 2016; Zhao et al. 2017). Ferreira et al. (2017) adapt seq2seq models for generating text from abstract meaning representations (AMRs). However, the seq2seq models tend to work better for data-to-text generation systems than for open-domain dialogue (Gatt and Krahmer 2018). These end-to-end approaches are a departure from prior work on building NLG systems that were cast as a pipeline of components (Reiter and Dale 2000).

A further important development with encoder-decoder networks are attention based mechanisms, which force the encoder to more appropriately weigh parts of the input encoding more when predicting certain portions of the output during decoding. In our work, we use the Bahdanau-style attention mechanism (Bahdanau et al. 2014), which is the typical version of attention mechanism used in several neural architectures.

Another framework that has emerged in language generation is the use of adversarial learning (Goodfellow et al. 2014; Li et al. 2017). However, these frameworks have the problem of vanishing gradient in the discriminator (Gulrajani et al. 2017). Wasserstein GANs (Arjovsky et al. 2017) have been proposed to overcome this issue in the language generation context, but addressing emotion in dialogue has not been adequately addressed in these works. While the state-of-the-art in this field has advanced quite rapidly, models are often prone to generate language that is short, dull, off-context or vague.

To address this challenge, reinforcement learning has emerged as another paradigm for language generation (Kaelbling et al. 1996). Perhaps the closest to our work is the work of Li et al. (2016). Similar to their work, to ensure that the generated language is topically relevant and fluent, we incorporate two internal rewards in our system viz. Ease of Answering and Semantic Coherence. These reward functions are designed to overcome

some of the challenges in traditional seq2seq models. They cast the task as a reinforcement learning problem where they jointly trained a generative model to produce response sequences, and a discriminator to distinguish between the human-generated dialogues and the machine-generated ones. However, instead of using policy gradient to train the model we use Maximum Mutual Information (MMI) as the objective function (Li et al. 2016a) to enable more diverse responses. Thus, although our approach is very closely related to Li et al. (2016), there are key differences in the objective functions and the use of external rewards. These are highlighted in Table 2.

More recently, in keeping with the reinforcement learning paradigm, Sankar and Ravi (2019) propose using discrete attributes such as sentiment, dialog acts and emotion to generate responses that leads to improvement over traditional seq2seq models. Jaques et al. (2016) propose a general method for improving the structure and quality of sequences generated by a recurrent neural network (RNN), while maintaining information originally learned from data, as well as sample diversity. However, these prior works do not incorporate any external rewards during the reinforcement learning phase.

In terms of work that incorporates external rewards, Christiano et al. (2017) used external rewards to fine-tune their reinforcement learning model. However, their system was trained for Atari games and simulated robot locomotion, not language generation.

2.2 Generation of emotional language

Emotions are recognized as functional in decision-making by influencing motivation and action selection. Therefore, computational emotion models are usually grounded in an agent's decision-making architecture, of which reinforcement learning is an important subclass. Moerland et al. (2018) provides the first survey of computational models of emotion in reinforcement learning agents. The survey focuses on agent/robot emotions. Badoy and Teknomo (2014) proposed using four basic emotions: joy, sadness, fear, and anger to influence a Q-learning agent. Their simulations show that the proposed affective agent requires minimal steps to find the optimal path. Sequeira et al. (2014) investigate the use of emotional information in the learning process of autonomous agents. They optimise the relative contributions of each reward feature and the resulting "emotional" RL agents perform better than standard goal-oriented agents, particularly in consideration of their inherent perceptual limitations.

Prior work on affective or emotional NLG has focused on tactical choices (Keshtkar and Inkpen 2011). Various linguistic features that can have emotional impact have been utilized,

Table 2 State-of-the-art Method vs. Our Approach

	Objective Function	Internal Rewards	External Rewards
Li et al. Approach	Policy Gradient Method	Ease of Answering Information Flow Semantic Coherence	N/A
Our Approach	Maximum Mutual Information (MMI)	Ease of Answering Semantic Coherence Emotional Intelligence	Human Feedback

We use a different objective function, and Internal as well as External Rewards in our model

for example first-person pronouns and adverbs and sentence ordering (Rosis and Grasso 2000).There has been recent work to ground open-domain dialogue in personal information (Zhang et al. 2018), but this information is fact-based (e.g. "I have a pet dog.") instead of being emotionally grounded.

With respect to language generation, Asghar et al. (2018) incorporated affective content in neural models by using the ANEW lexicon (Warriner et al. 2013) and appending word embeddings with objective functions to achieve affective response generation. Zhou et al. (2018) have proposed Emotional Chatting Machine, that can generate appropriate responses not only in content (relevant and grammatical) but also in emotionally consistent with the input prompt. However, these prior approaches do not incorporate external feedback as a reward towards generating emotionally rich, coherent and useful language.

To summarize this related work section, our work addresses the following gaps in extant literature:

– We go beyond sentence-level emotion annotations since it can be difficult to capture the nuances of human emotion accurately without taking into account finer-grained token-level emotions.
– We go beyond relying solely on the dialogue history, and fully utilize the potential of user feedback for generated responses.

3 Problem statement and intuition

As argued by Rieser and Lemon (2009), one can cast Natural Language Generation (NLG) as a sequence planning task. The decisions that the system needs to make are to choose an NLG action of generating a token and also which token to generate, or to stop generation.

The goal then is to take advantage of reinforcement learning and rewards during the process of language generation to learn the optimal policy. At the same time, our goal is to use the encoder-decoder framework, since it has been shown to be advantageous for language generation tasks.

The following elements can comprise a reinforcement learning system:

- *Action (a)* – which in our case are dialogue utterances to generate i.e. action $a = gen(S)$, where $gen(S)$ is the generated sequence. The action space is infinite and generates sequences of varying length.
- *State (S)* – for our problem statement, dialogue is transformed to a vector representation and fed as input to the current dialogue state for which the response has to be generated.
- *Policy* – policy takes the form $p_{RL}(p_{i+1}|p_i)$ where p_{i+1} is the response to be generated for the given dialog p_i.

Here, we use a stochastic distribution to represent policy as it is the probability distribution over actions given states, where both state and actions are dialogues. By doing so, we overcome the difficulty of optimizing a deterministic policy.

The reward function defines the goal of the overall dialogue. In our case, the goal is to match the emotional tone of the preceding context, and therefore the reward should be set up to penalize language that does not match the emotional tone of the preceding context. While at the same time, we want the model to not compromise the fluency and grammaticality of the generated text. Hence, there should be rewards that penalize the generation of non-fluent or ungrammatical text.

- *Rewards (r)* – Based on the above observations, we implement three internal rewards and one external reward to overcome the issues in generating language with seq2seq architecture (Vinyals and Le 2015).

However, external feedback and rewards are hard to come by for language generation; these would need to be provided through crowdsourcing judgments on the generated responses *during* the generation process, which makes the process time-consuming and impractical. To overcome this issue, we make use of the distance labeling paradigm (Mintz et al. 2009) - and use labels provided in the training set as a proxy for human judgments on the generated responses. Specifically, we incorporate usefulness scores as a proxy for external feedback.

4 Model architecture

Figure 1 shows overall system architecture. We explain each component in detail in the subsections that follow.

4.1 Word embeddings from corpora

We generate word embeddings using the word2vec method (Mikolov et al. 2013). Traditional word embeddings trained with co-occurrence statistics can be insufficient to capture aspects of emotion and affect. Thus, to augment the word embeddings with affective information, we use the Affective Norms of Words (ANEW) lexicon (Warriner et al. 2013) that has valence, arousal and dominance scores for words. Arousal is the extent to which a word is calming or exciting, whereas valence is the extent to which a word is negative or positive, and dominance is the degree of control exerted by a stimulus (word) (Kuperman et al. 2014).

We chose the ANEW lexicon for augmenting word embeddings with emotion information since this lexicon contains continuous scores on a scale of 1 (low valence/arousal/dominance) to 9 (high valence/arousal/dominance). This approach is preferable over the use of other lexicons such as NRC (Mohammad and Turney 2013). One reason why the NRC lexicon is not preferable is that the words in the NRC lexicon are often duplicated. For example, "abandon" is contained in three categories, and "abandoned" in four. This duplication can result in difficulties in matching lemmas. Several other lexicons exist, but these may not offer a continuous score for words, but instead provide simply a binary value of whether a word carries positive or negative (e.g. the Bing lexicon (Liu 2012)).

Fig. 1 Overall Architecture of our system. Word embeddings are learnt from the corpora augmented with emotion information and fed into the Bidirectional RNN Encoder-Decoder module. The model learnt from this module is then tuned using Reinforcement Learning by optimizing the internal and external rewards

Accordingly, we append the scores for Valence (V), Arousal (A) and Dominance (D) score from the ANEW lexicon to each word, resulting in 1027 dimensions for each word. In cases where a match cannot be found in the lexicon, we append a neutral vector [5, 1, 5] similar to the method used in (Asghar et al. 2018). This word2vec-VAD embedding is fed as input to the bidirectional RNN encoder-decoder seq2seq model and is also used in the Reinforcement Learning system to model the Emotional Intelligence heuristic.

4.2 Bidirectional RNN encoder-decoder

We use a bidirectional RNN encoder-decoder seq2seq model (Vinyals and Le 2015) with Bahdanau–style attention mechanism (Bahdanau et al. 2014). During decoding, we use a greedy decoder to generate the best response at every stage of decoding during decoder training and inference phases.

The standard objective function for seq2seq models is the log-likelihood of target T given source S, given as follows:

$$\hat{T} = \arg\max_{T} \{\log p(T|S)\} \tag{1}$$

This formulation leads to generic responses, since it only selects for target given source. We optimize this standard objective function by replacing it with Maximum Mutual Information (MMI) (Li et al. 2016a). In MMI, parameters are chosen to maximize (pairwise) mutual information between the source S and the target T:

$$\log \frac{p(S,T)}{p(S)p(T)} \tag{2}$$

Doing so avoids favoring responses that unconditionally enjoy high probability, and instead biases towards those responses that are specific to the given input. The MMI objective can written as follows:

$$\hat{T} = \arg\max_{T} \{\log p(T|S) - \lambda \log p(T)\} \tag{3}$$

Here, λ is the hyperparameter that controls the extent to which we penalize generic responses to get more diverse responses. Adjusting the value of λ results in a reasonable number of diverse responses, however, these could still be dull and also lack emotion and proper grammatical structure. To address these issues, we model the reinforcement learning system with appropriate heuristics.

4.3 Reinforcement learning with internal rewards

We fine tune the basic *seq2seq* generative model as described in the subsection above with rewards to generate more interesting, diverse and emotionally appropriate responses.

The three internal are Ease of Answering r_{EA}, Semantic Coherence r_{SC} and Emotional Intelligence r_{EI}.

1. *Ease of Answering (EA)* (r_{EA}) – is measured as negative log likelihood of generating a dull response for a dialog. Following (Li et al. 2016), (Lowe et al. 2015) and (Niu and Bansal 2018), we compose a list of 10 dull responses that frequently occur in the

seq2seq model and penalize the model when it generates those responses.[1] Let set $\mathbb{S}$ represent a list of dull responses. Then, the reward function can be defined as follows:

$$r_{EA} = -\frac{1}{N_{\mathbb{S}}} \sum_{s \in \mathbb{S}} \frac{1}{N_s} \log p_{seq2seq(s|a)} \tag{4}$$

$p_{seq2seq}$ represents likelihood output of *seq2seq* model. Here $N_{\mathbb{S}}$ represents cardinality of $\mathbb{S}$ whereas N_s is the number of tokens in the dull response. The rationale behind this reward is that the RL system is likely to penalize utterances in the above composed list and hence less likely to generate dull responses. Arguably, a system less likely to generate utterances in the list may also be less likely to generate other dull responses.

2. *Semantic Coherence (SC)* (r_{SC}) (Li et al. 2016) – is used to avoid situations in which the generated responses are highly rewarded, but are neither grammatical nor coherent. We consider the mutual information between the action a and the given input to ensure that the responses are coherent and appropriate. This also involves reverse training the model where we count the probability of the input prompt given the current generated response.

$$\begin{aligned} r_{SC} = & \frac{1}{N_a} \log p_{seq2seq}(a|q_i, p_i) + \\ & \frac{1}{N_{q_i}} \log p^{backward}_{seq2seq}(q_i|a) \end{aligned} \tag{5}$$

Here N_a represents the length of the target response, $\log p_{seq2seq}(a|q_i, p_i)$ represents the probability of generating response a given the previous dialogue utterances while $logp^{backward}_{seq2seq}(q_i|a)$ is the backward probability of generating the previous dialogue utterance q_i based on the response a, scaled by the length of targets N_{q_i}.

3. *Emotional Intelligence (EI)* (r_{EI}) (Asghar et al. 2018) – This reward is incorporated by minimizing affective dissonance between the prompts and the responses. This approach tries to maintain affective consistency between input and generated response. The heuristic is based on the fact that open-domain textual conversations between humans follow an affective pattern. Thus, we make an assumption that the affective tone does not fluctuate often in general and we focus on minimizing the dissonance in affective tone between the input prompt and the generated responses.

$$r_{EI_i} = \lambda p(a) \| \sum_{j=1}^{n} \frac{W2AV(x_j)}{|X|} - \sum_{k=1}^{i} \frac{W2AV(y_k)}{i} \| \tag{6}$$

Here, *W2AV* in (6) denotes the word-affect vector of the given sequence and the term $\sum_{j=1}^{n} \frac{W2AV(x_j)}{|X|}$ denotes average affect vector of the input prompt and $\sum_{k=1}^{i} \frac{W2AV(y_k)}{i}$ denotes average affect vector of the generated response up to the current step i. Here X is the input sequence and y is the generated output sequence. We relax hard prediction of a word by its predicted probability $p(a)$. λ is a hyperparameter that balances the two factors.

[1] The 10 responses are: *"I don't know.", "I don't know what I mean.", "I don't know what you're talking about.", "You don't know.", "You know what I mean.", "You know what I'm saying.", "You don't know anything.", "I am not sure.", "I know what you mean.", "I do not know anything."*

4.4 External reward from human feedback

To incorporate external rewards in our model, we simulate human feedback through the reviews from the usefulness score in one of our corpora. This corpus is the Yelp Review corpus. We categorize each review in the Yelp dataset into two main classes `Useful` and `Not Useful` based on the frequency distribution of the usefulness scores of reviews (as shown in Fig. 2). Reviews with normalized scores < 5 are considered not useful, while the rest are considered to be useful. We exclude all reviews that do not have usefulness ratings, since it is not clear which category they would fall under.

Next, we train a simple Support Vector Machine classifier to differentiate between the two classes `Useful` and `Not Useful` as described above. This classifier is integrated with the Reinforcement Learning system described above and produces the external reward that would need to be optimized – we name this reward as human feedback (r_{HF}).

During the training phase, we determine whether the generated response is useful or not (by classifying the generated output in real-time using the SVM classifier) and give the reward accordingly. This synthetic feedback from the external reward analyzer is provided throughout the training phase and a greedy decoder is then used to generate the best response.

5 Experiments and results

We test the efficacy of the proposed method in generating emotional and coherent text in two different corpora, which pertain to two different genres of text.

5.1 Corpora & pre-processing

We have used two different datasets and have created two different models. The Cornell Movie Dialog corpus (Danescu-Niculescu-Mizil and Lee 2011) and the Yelp Restaurant

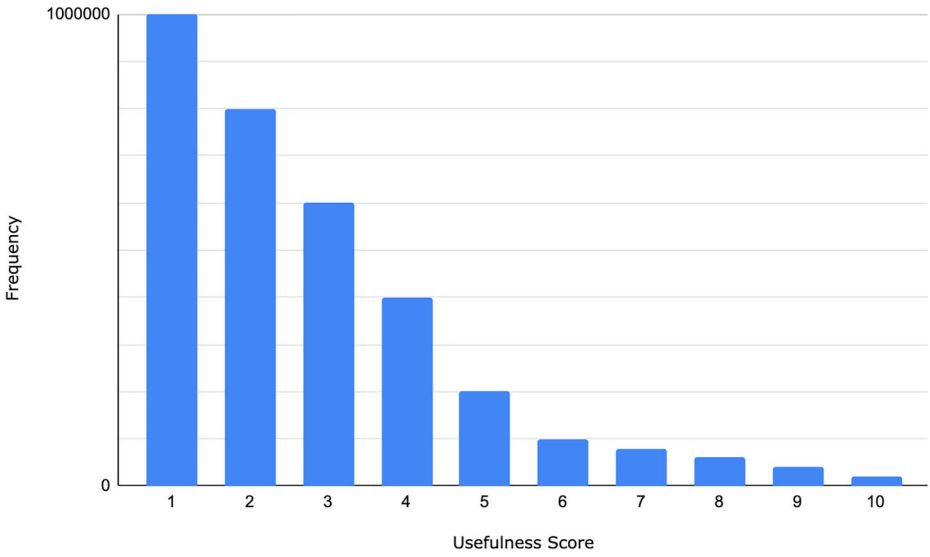

Fig. 2 Frequency distribution of usefulness scores of reviews in the Yelp review corpus

Table 3 Descriptive statistics for the two corpora used in our experiments

	Training set	Validation set	Testing set
Cornell corpus	160,000	14,000	6000
Yelp corpus	4,017,986	1,187,406	791,604

Review dataset[2]. Our choice of corpora for these experiments is motivated by the fact that these are (a) these contain utterances (dialogues and reviews) with emotional content; (b) these corpora are considered standard corpora for NLG tasks and would be useful to compare our model output with performance report by state-of-the-art approaches; and (c) they provide two different genres of text to test our model performance on.

The Cornell Movie-Dialog corpus (Danescu-Niculescu-Mizil and Lee 2011) contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts. There are 220,579 conversational exchanges between 10,292 pairs of movie characters involving 9,035 characters from 617 movies. There are 304,713 utterances in total. The Yelp review dataset contains 5.9M reviews. Along with the reviews, this dataset contains nine additional features, including usefulness score, which we use to train our external reward analyzer.

Table 3 shows the descriptive statistics for both corpora. We take themost common 12,000 words from the training and validation sets as our vocabulary (Asghar et al. 2018), and replace any other tokens in these sets with an unknown symbol <UNK>. We partition the training and validation sets such that none of the responses in the training set have <UNK>. This effectively prevents the model from generating the unknown token during inference. This provides us with 160,000 exchanges for training and 14,000 for validation, 6000 for testing in the Cornell corpus. The word count is set to maximum threshold of 20.

We perform the standard pre-processing steps on the Cornell and Yelp dataset, including lower casing all conversations, expanding contractions, compress duplicate end punctuation to one symbol and removing HTML entities.

5.2 Proposed model

Our proposed model uses seq2seq to choose the initial policy and fine-tunes that model to generate more diverse responses based on internal and external rewards. For Cornell corpus, the final reward function just uses the internal reward components during reinforcement learning and can be described as in (7) and for the Yelp corpus, the final reward function, is the weighted sum of both internal and external rewards (8).

$$r_{Final\_Cornell} = \lambda_1 r_{EA} + \lambda_2 r_{SC} + \lambda_3 r_{EI} \tag{7}$$

$$r_{Final\_Yelp} = \lambda_1 r_{EA} + \lambda_2 r_{SC} + \lambda_3 r_{EI} + \lambda_4 r_{HF} \tag{8}$$

For both models, the values of the hyperparameters are given in Table 4. We can see certain differences in the hyperparameters since the Yelp corpus size is greater than the Cornell corpus (e.g. batch size and number of epochs). The learning rate and decay rate are greater for Yelp as it takes a longer time to converge and train the model than it does on the smaller

[2]https://www.yelp.com/dataset

Table 4 Hyperparameter settings for the models with Reinforcement Learning used in our approach

Hyperparameter	Cornell model value	Yelp model value
Batch size	128	512
Gradient clip	1.0	1.0
Learning rate	0.01	0.15
Decay rate	0.0095	0.01
Epochs	50	75
LSTM layers	2	2
Encoder RNN size	1027	1027
Decoder RNN size	1027	1027
$r_{EA}\lambda_1$	0.25	0.25
$r_{SC}\lambda_2$	0.35	0.25
$r_{EI}\lambda_3$	0.40	0.25
$r_{HF}\lambda_4$	–	0.25

corpus (Cornell). The values for the rewards are adjusted and fine-tuned based on the outcome of each model. The hyperparameters were chosen following grid search and are based on the best performing model on the validation set.

5.3 Performance on automated metrics

We first evaluate the model using automated metrics including BLEU score, ROUGE-L and Perplexity(Papineni et al. 2002). These metrics are standard evaluation measures for NLG systems. Although there are documented issues with using such metrics as a measure of performance (Novikova et al. 2017), we report them here for completeness sake. For the baseline, we use a basic seq2seq model with MMI objective function (Li et al. 2016a), that does not use reinforcement learning.

In Table 5, we report scores on the automated metrics, BLEU, ROUGE-L and Perplexity. The scores are statistically significantly better than baseline (without RL), with $p < 0.01$ for BLEU score, $p < 0.05$ for Perplexity and $p < 0.005$ for ROUGE-L for Cornell. For the Yelp corpus, the model with external rewards performs significantly better on all three metrics ($p < 0.01$) when compared to the baseline (without RL).

5.4 Human evaluation of performance

To provide complementary evidence of performance and to address the issues with using automated metrics as a barometer of performance, we conducted studies with human

Table 5 Model evaluation on automated metrics

		BLEU	ROUGE-L	Perplexity
Cornell Corpus	Baseline	0.15	0.39	98.96
	Our Model	0.38**	0.55***	76.65*
Yelp Corpus	Baseline	0.014	0.24	99.04
	Our Model	0.21**	0.32**	85.34**

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.005$

subjects to evaluate performance. Both studies were conducted after obtaining IRB approval from our institution.

Study 1 First, we created a simple survey, containing 20 prompt/response pairs from both Cornell and Yelp models.

We recorded responses from a total of 52 undergraduate and graduate Computer Science students. Each response generated by the system was evaluated on three measures - **Syntactic Coherence** (how grammatical and coherent are the responses with respect to the given prompt), **Natural Flow** (how natural is the response as a follow on to the given prompt) and **Emotional Appropriateness** (how well does response capture the emotional appropriateness of the text in the given prompt) (Asghar et al. 2018).

Study 2 Next, we conducted experiments on Amazon Mechanical Turk with 100 response pairs of both Cornell and Yelp models. Each response was rated by at least 5 workers on measures of Syntactic Coherence, Natural Flow and Emotional Appropriateness.

Table 6 shows the ratings obtained from human evaluation on the metrics of Syntactic Coherence, Natural Flow and Emotional Appropriateness. These metrics were chosen based on extant literature (Dušek et al. 2018), and have been demonstrated to be appropriate measures of language generation capabilities. As seen in Table 6, we find that the models perform well on all three metrics as rated by the humans. The scores shown in Table 6 are averages on a 3-point Likert scale, with 0 being lowest and 2 being highest scores. Higher scores would indicate a better performance on the given metric. While there are no established best practices in designing studies for human evaluation of NLG output (Novikova et al. 2017), the study design used in this work is based on prior work by (Li et al. 2016a).

Based on the scores shown in Table 6, we find that the scores are sufficiently high, to enable us to get an adequate measure of model performance. We also find that the ratings for output on the Yelp corpus were slightly higher than the ratings for the Cornell corpus. One reason for this could be length of input context given to the models *as well as* to the human raters themselves.

Study 3 Next, we evaluate our model against the model presented in Asghar et al. (2018). Their model is the closest to our model in terms of implementation. Similar to our method, their model creates an affectively cognizant neural encoder-decoder dialogue system by embedding word vectors in an affective space. However, the key difference is that they do not incorporate another internal rewards that maximize coherence and minimize dull responses. Also, another reason for comparing our model against Asghar et al. (2018) is that they also experiment with the Cornell movie dialogue corpus, which is one of the corpora

Table 6 Human evaluation of our models performance on measures of syntactic coherence, naturalness of flow and emotional appropriateness of generated response

Scores are averages on a 3-point (0 being lowest and 2 being highest) Likert scale, with higher scores indicating better performance on a given metric.

Coherence	Flow	Syntactic Appropriateness	Natural	Emotional
Cornell Corpus	Study 1	1.45	1.42	1.44
	Study 2	1.49	1.41	1.53
Yelp Corpus	Survey	1.46	1.52	1.73
	MTurk	1.51	1.50	1.66

we experiment with. We find that our model achieves better ratings on all three metrics as we generate longer sentences for the Yelp review and the model is also able to outperform the current state of the art of the model (Asghar et al. 2018) as demonstrated in Table 7.

5.5 Qualitative analysis of output

We conduct qualitative evaluation of the outputs and present several cherry- and lemon-picked examples of the model capabilities. We begin with the examination of successful cases and then present failure analysis using several illustrative examples.

Successful Cases In Tables 8 and 9 we show examples of the cases where the model is able to successfully generate topically relevant and emotional language. It is evident that the limited context available for the Cornell movie dialogue corpus makes it unclear what the emotional tone of the input text might be. However, given the relatively longer context of the Yelp review corpus, the emotional tone and coherence is easier to determine. For instance, the prompt for the first example in Table 8 may have an ambiguous tone, and consequently, it may be that the response does not necessarily need to be positive or negative. We could experiment with changing the objective function, in cases where the emotional tone of the prompt could not be identified due to lack of context.

Similarly, even if we are able to successful match the emotional tone, as noted in the examples in Table 9, it is not always clear that the generated text "reflects reality" - in that, it may not be true that the restaurant sells burgers at all. This lack of real-world knowledge is a known problem with NLG systems. We acknowledge this shortcoming – however, it is not the goal of this paper to address that research problem in this paper.

Failure Cases In Table 10, we show three types of model failure.

One category of output failure is the **inconsistency** of the model to adhere to the given topic or context. As can be seen in the example, the input text and the following text in the human ground truth both describe a shooting range. However, the model generates a review about a restaurant. Such a failure of the model can be attributed to a lack of external knowledge provided to the model. If knowledge about the type of review (shooting range vs. restaurant) can be fed into the model, it may be possible to further constrain the model output to be relevant to the topic at hand.

Another issue we observed was the generation of **ungrammatical** text. Shown in the Table 10, the example is taken from the output from the Cornell movie dialogue corpus. We

Table 7 Comparison of our model against best performing model from state-of-the-art baseline Asghar et al. (2018) on measures of syntactic coherence, naturalness of flow and emotional appropriateness of generated response

Coherence	Syntactic Flow	Natural Appropriateness	Emotional
Cornell Model	1.49	1.41	1.53
Asghar *et al.* (2018)	1.45	1.31	1.33

Scores are averages on a 3-point (0 being lowest and 2 being highest) Likert scale, with higher scores indicating better performance on a given metric

Table 8 Examples of our model output against the human ground truth for the Cornell corpus

does not sound free to me. so you want me to pay for it ?	Human Ground Truth: Yes, I think so. Model Output: Yes sir! I'm sure you do have to rush
say hi to rob and carrie.	Human Ground Truth: Sure, I'll tell them. Model Output: i'm going to miss you.

see that the while the model output is acceptable, one could expect that more appropriate output could instead be either (a) *I'm good. Glad you're here.* or (b) *I'm glad you're here.*

Finally, another issue that was observed was the generation of **incoherent** language. The last example in Table 10 is illustrative of how the model output can read seemingly fine, however, it is incoherent in several places (e.g. *more than enough for 2 the board* and *I don't live here full-time, every time I come to Phoenix*).

The issues that arise with generating **ungrammatical** or **incoherent** language are common to many language generation systems. With the advent to large-scale language models

Table 9 Examples of our model output against the human ground truth for the Yelp review corpus

This place was awesome! The French dip with Au Ju is amazing! Great service, great location.	Human Ground Truth: Nice restaurant with excellent service. This place has a good ambience and has good mexican food. I highly recommend this restaurant. Model Output: Outside seating and a fine lunch., A great burger and superb popovers. Great food and service. This is the place to be.
Me and my family have been going to this place for years to fulfill our sushi and soju fix. Sushi is made fresh from the bar and the prices are reasonable! The staff is very friendly and welcoming, the owner is always willing to sit down and talk with us!	Human Ground Truth: They are open to suggestions and have variety of recipies from different cuisines. I always prefer dining at this restaurant if not for home. Model Output: The yummy and flavorful sides., Good,wine and beer selection. The food was very good and the service was excellent. I would eat here everyday if I could.

Table 10 Examples of failure cases of our model output against the human ground truth

Inconsistent topic/context Input Text/Context: Very friendly staff and a great range. Their prices may be few dollars more than a box store, but I like to support a local business and they definitely earn it in service.	Human Ground Truth: As long as you are safe (and they do watch you) they don't bother you. I love that and wish more ranges did that. Go, be safe, have fun, and enjoy your rights. Model Output: The limited menu helps them keep their focus on the freshness of the meal. This place is always busy and is a testament to their product. Go. Eat. Enjoy. This place is awesome!!
Ungrammatical Text Input Text/Context: i said i would call you the minute i got free. and i did - and here we are	Human Ground Truth: Wow that's amazing! Model Output: I'm good glad you're here.
Incoherent Output Input Text/Context: Needed a fast bite to eat as I had to hit the hay early and catch a 6 am flight to Chicago. This place was near my hotel and I always hear people rave about it so I thought I would give it a shot. You walk in and there's a counter with a variety of fixins, you can get taco's, burritos, chips, guacamole, and a burrito bowl which is like a burrito but the tortilla is on the side.	Human Ground Truth: The food is very delicious and their service is quick and clean. I started visiting this restaurant often and I love it Model Output: This location is really good too. They have the best pizza on the weekends when Kids Eat FREE!, If you have a larger family (more than enough for 2 the board and sandwiches and I'm bummed I don't live here full-time, every time I come to Phoenix

and transformer-based architectures (Vaswani et al. 2017), some of these issues can be ameliorated. However, true language understanding and human-like generation does remain an elusive goal (Linzen 2020).

6 Discussion and future work

We describe our method for generating emotionally appropriate language using a reinforcement learning paradigm. The novelty of our approach lies in the addition of Emotional Intelligence as an internal reward function and combining both internal and external rewards to create an emotionally appropriate model. The use of external rewards to generate more sensible and human-like responses is novel in the natural language generation task, with the exception of work conducted by Niu and Bansal (2018).

Our qualitative analysis reveals that there are still issues that pertain to lack of contextual and world knowledge that affect the quality of the generated text. In addition, the model can generate text that is ungrammatical or incoherent. These issues can be addressed by experimenting with different architectures, for example, the GPT-2 models (Budzianowski and Vulic 2019).

In future, we plan to experiment with different heuristics like maximizing affective dissonance and content as emotional intelligence heuristic reward system. This would allow us to generalize to situations where the affective tone of the preceding context would need to be dissimilar to the generated language output – for example, customer service scenarios. We have used the usefulness score in the Yelp restaurant review dataset as external feedback. We also plan to incorporate direct human feedback into the training phase. All the code used in these experiments and repository of additional examples is available at https://github.com/VidhushiniSrinivasan16/EmotionalNLG.

References

Arjovsky, M., Chintala, S., Bottou, L. (2017). Wasserstein gan. arXiv:1701.07875.

Asghar, N., Poupart, P., Hoey, J., Jiang, X., Mou, L. (2018). Affective neural response generation. In *European Conference on Information Retrieval Springer, pp 154–166*.

Badoy, W., & Teknomo, K. (2014). Q-learning with basic emotions. CoRR abs/1609.01468.

Bahdanau, D., Cho, K., Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv:1409.0473.

Budzianowski, P., & Vulic, I. (2019). EMNLP-IJCNLP 2019 p. 15.

Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D. (2017). Deep reinforcement learning from human preferences. In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.) *Advances in Neural Information Processing Systems 30 Curran Associates, Inc., pp 4299–4307*. http://papers.nips.cc/paper/7017-deep-reinforcement-learning-from-human-preferences.pdf.

Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Dušek, O., Novikova, J., Rieser, V. (2018). Findings of the e2e nlg challenge. arXiv:1810.01170.

Ferreira, T.C., Calixto, I., Wubben, S., Krahmer, E. (2017). Linguistic realisation as machine translation: Comparing different mt models for amr-to-text generation. In *Proceedings of the 10th International Conference on Natural Language Generation, pp 1–10*.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P., Scherer, S. (2017). Affect-lm: A neural language model for customizable affective text generation. arXiv:1704.06851.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems, pp 2672–2680*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in neural information processing systems, pp 5767–5777*.

Hashimoto, C., & Sassano, M. (2018). Detecting absurd conversations from intelligent assistant logs by exploiting user feedback utterances. In *Proceedings of the 2018 World Wide Web Conference, pp 147–156*.

Huang, C., Zaiane, O.R., Trabelsi, A., Dziri, N. (2018). Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp 49–54*.

Jaques, N., Gu, S., Turner, R.E., Eck, D. (2016). Tuning recurrent neural networks with reinforcement learning. arXiv:1611.02796.

Kaelbling, L.P., Littman, M.L., Moore, A.W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, 237–285.

Keshtkar, F., & Inkpen, D. (2011). A pattern-based model for generating text to express emotion. In *International Conference on Affective Computing and Intelligent Interaction (Springer), pp 11–21*.

Kuperman, V., Estes, Z., Brysbaert, M., Warriner, A.B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, *143*(3), 1065.

Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B. (2016a). A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Association for Computational Linguistics), pp 110–119*. http://aclweb.org/anthology/N16-1014.

Li, J., Galley, M., Brockett, C., Spithourakis, G.P., Gao, J., Dolan, B. (2016). A persona-based neural conversation model. arXiv:1603.06155.

Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., Gao, J. (2016). Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp 1192–1202*. http://www.aclweb.org/anthology/D16-1127.

Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., Jurafsky, D. (2017). Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp 2157–2169*. http://aclweb.org/anthology/D17-1230.

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization?. arXiv:2005.00955.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1–167.

Lowe, R., Pow, N., Serban, I., Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. arXiv:1506.08909.

Martinovski, B., & Traum, D. (2003). Breakdown in human-machine interaction: the error is the clue. In *Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems, pp 11–16*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (Eds.) *Advances in Neural Information Processing Systems 26 (Curran Associates, Inc.), pp 3111–3119*. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Mintz, M., Bills, S., Snow, R., Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, (ACL), pp 1003–1011*.

Moerland, T.M., Broekens, J., Jonker, C.M. (2018). Emotion in reinforcement learning agents and robots: a survey. *Machine Learning*, *107*(2), 443–480. https://doi.org/10.1007/s10994-017-5666-0.

Mohammad, S.M., & Turney, P.D. (2013). National Research Council, Canada 2.

Niu, T., & Bansal, M. (2018). Polite dialogue generation without parallel data. *Transactions of the Association of Computational Linguistics*, *6*, 373–389.

Novikova, J., Dušek, O., Curry, A.C., Rieser, V. (2017). Why we need new evaluation metrics for nlg. arXiv:1707.06875.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics (Association for Computational Linguistics), pp 311–318*.

Poria, S., Majumder, N., Mihalcea, R., Hovy, E. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, *7*, 100943–100953.

Prendinger, H., & Ishizuka, M. (2005). The empathic companion: A character-based interface that addresses users'affective states. *Applied artificial intelligence*, *19*(3-4), 267–285.

Prendinger, H., Mori, J., Ishizuka, M. (2005). Recognizing, modeling, and responding to users affective states. In *International Conference on User Modeling (Springer), pp 60–69*.

Rashkin, H., Smith, E.M., Li, M., Boureau, Y.-L. (2018). Towards empathetic open-domain conversation models: A new benchmark and dataset. arXiv:1811.00207.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge: Cambridge university press.

Rieser, V., & Lemon, O. (2009). Natural language generation as planning under uncertainty for spoken dialogue systems. In *Empirical methods in natural language generation (Springer), pp 105–120*.

Rosis, F., & Grasso, F. (2000). *Affective natural language generation. affective interactions, towards a new generation of computer interfaces, in ed. a. paiva, 204-218*. New York: New York: Springer-Verlag.

Sankar, C., & Ravi, S. (2019). Deep reinforcement learning for modeling chit-chat dialog with discrete attributes. arXiv:1907.02848.

Sequeira, P., Melo, F.S., Paiva, A. (2014). Learning by appraising: An emotion-based approach to intrinsic reward design. *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, *22*(5), 330–349. https://doi.org/10.1177/1059712314543837.

Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Sutskever, I., Vinyals, O., Le, Q.V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems, pp 3104–3112*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems, pp 5998–6008*.

Venkatesh, A., Khatri, C., Ram, A., Guo, F., Gabriel, R., Nagar, A., Prasad, R., Cheng, M., Hedayatnia, B., Metallinou, A., et al. (2018). On evaluating and comparing conversational agents, (Vol. 4. arXiv:1801.03625.

Vinyals, O., & Le, Q. (2015). A neural conversational model. arXiv:1506.05869.

Warriner, A.B., Kuperman, V., Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x.

Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T. (2017). Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision, pp 4894–4902*.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too?. arXiv:1801.07243.

Zhao, T., Zhao, R., Eskenazi, M. (2017). Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. arXiv:1703.10960.

Zhou, H., Huang, M., Zhang, T., Zhu, X., Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.