The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Eliciting Positive Emotion through Affect-Sensitive Dialogue Response Generation: A Neural Network Approach

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura
Graduate School of Information Science,
Nara Institute of Science and Technology, Japan
{nurul.lubis.na4, ssakti, koichiro, s-nakamura}@is.naist.jp

Abstract

An emotionally-competent computer agent could be a valuable assistive technology in performing various affective tasks. For example caring for the elderly, low-cost ubiquitous chat therapy, and providing emotional support in general, by promoting a more positive emotional state through dialogue system interaction. However, despite the increase of interest in this task, existing works face a number of shortcomings: system scalability, restrictive modeling, and weak emphasis on maximizing user emotional experience. In this paper, we build a fully data driven chat-oriented dialogue system that can dynamically mimic affective human interactions by utilizing a neural network architecture. In particular, we propose a sequence-to-sequence response generator that considers the emotional context of the dialogue. An emotion encoder is trained jointly with the entire network to encode and maintain the emotional context throughout the dialogue. The encoded emotion information is then incorporated in the response generation process. We train the network with a dialogue corpus that contains positive-emotion eliciting responses, collected through crowd-sourcing. Objective evaluation shows that incorporation of emotion into the training process helps reduce the perplexity of the generated responses, even when a small dataset is used. Subsequent subjective evaluation shows that the proposed method produces responses that are more natural and likely to elicit a more positive emotion.

Introduction

Various parts of the globe are constantly facing societal problems that require emotional competences to address: an aging population, untreated affect-related problems, and high levels of stress, to name a few. For example, stress levels in the United States of America has steadily increased since 2007 (American Psychological Association 2015). Furthermore, a sizable portion of the people who reported stress problems also admitted to feel that they have not done their best to manage this problem.

Social ties has been reported to powerfully predict health and well-being in late life (Waldinger et al. 2015). Human communications could provide an array of benefits through social sharing of emotion (Zech and Rimé 2005), preventing longer-term, more serious emotion-related problems. As the technology develops, the potential of agents to improve the emotional-well being of users has been growing as well. In particular, emotionally intelligent systems could provide assistance and prevention measures through various affective tasks, such as caring for the elderly, low-cost ubiquitous chat therapy, or providing emotional support in general.

Two of the most studied emotional competences for agents are *emotion recognition* and *emotion simulation*. *Emotion recognition* allows a system to discern the user's emotions and address them in giving a response (Han, Kim, and Lee 2015; Tielman et al. 2014). On the other hand, *emotion simulation* helps convey non-verbal aspects to the user for a more believable and human-like interaction, for example to show empathy (Higashinaka, Dohsaka, and Isozaki 2008) or personality (Egges, Kshirsagar, and Magnenat-Thalmann 2004). These competences address some of the user's emotional needs (Picard and Klein 2002). However, they are not sufficient to provide emotional support in an interaction.

Recently, there has been an increasing interest in eliciting user's emotional response through Human-Computer Interaction (HCI), i.e. *emotion elicitation*. Skowron et al. have studied the impact of different affective personalities in a text-based dialogue system (Skowron et al. 2013), reporting consistent impacts with the corresponding personality in humans. On the other hand, Hasegawa et al. constructed translation-based response generators with various emotion targets (Hasegawa et al. 2013), allowing elicitation of a predefined emotional state. Despite the affirming results, these systems require strict definition of personality or target emotion on the system side. Furthermore, they have not yet paid attention to the emotional experience of the users.

With a more user-oriented perspective, (Lubis et al. 2017) have recently utilized examples of human emotion appraisal process to elicit positive emotion. This study drew on an important overlooked potential of emotion elicitation: its application to maximize user emotional experience and promote positive emotional states, similar to that of emotional support between humans. This can be achieved by actively eliciting a more positive valence throughout the interaction, i.e. *positive emotion elicitation*. However, the example-based approach employed in this study suffers from a limited repertoire of responses and has difficulties handling out-of-example (OOE) queries. In other words, to build a dynamic and scalable version of such systems remains a challenge.

With recent advancements in neural network research, end-to-end approaches have been reported to show promising results for non-goal oriented dialogue systems (Vinyals and Le 2015; Serban et al. 2016; Nio et al. 2016). These approaches rid the need for explicit definition of predefined state and action spaces, allowing for a more dynamic model that mimics human conversation contained in the training corpus. Furthermore, they may still be able to produce a natural response given a query that does not exist in the training data, solving the OOE problem of the example-based approach. These qualities are especially desirable in domain-free, chat-oriented dialogue systems. However, application of this approach towards incorporating emotion in the dialogue is still very lacking.

Only very recently, Zhou et al. published their work addressing emotional factor in neural network response generation (Zhou et al. 2017). They examined the effect of internal emotional state on the decoder, investigating 6 categories to emotionally color the response. However, this study has not yet considered user's emotion in the response generation process, nor attempted to utilize emotion to maximize user experience.

To the best of our knowledge, there is not yet an effort in utilizing neural network approaches for affect-sensitive response generation and emotion elicitation. Our contributions in this paper are as follows:

- We propose a neural-network based chat-oriented dialogue system that 1) captures user's emotional state and considers it in generating a dialogue response, and 2) elicits positive emotion through the interaction. We believe that this constitutes the first neural network approach for affect-sensitive response generation.
- We construct a dialogue corpus that reflects a positive emotion elicitation strategy for model training to influence its affective tendency. This allows positive emotion elicitation without any elaborate dialogue strategy.

We build our model extending the recently proposed hierarchical recurrent encoder-decoder architecture (Serban et al. 2016). We incorporate an emotion encoder into the network to capture the emotional context of a dialogue, and use this information in the response generation process to produce an affect-sensitive response. We then train the network using the positive emotion eliciting data. Objective and subjective evaluations show that compared to the existing method, the proposed architecture 1) results in lower model perplexity, even when only a small amount of data is used, 2) generates a more natural response, and 3) on average elicits a more positive emotional impact.

Emotion Definition

In this work, we define the emotion scope based on the *circumplex model of affect* (Russell 1980). Two dimensions of emotion are defined: *valence* and *arousal*. Valence measures the positivity or negativity of emotion; e.g., the feeling of joy is indicated by positive valence while fear is negative. On the other hand, arousal measures the activity of emotion; e.g., depression is low in arousal (passive), while rage is high (active). Figure 1 illustrates the valence-arousal dimension with respect to a number of common emotion terms. This model is intuitive and can easily be adapted and extended to either discrete or other dimensional emotion definitions.

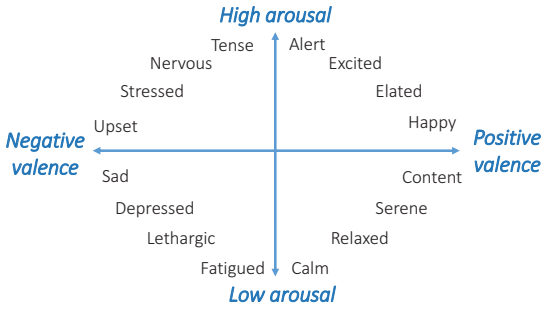

Figure 1: Emotion dimensions and common terms.

Henceforth, based on this scope of emotion, the term *positive emotion* refers to the emotions with positive valence. Respectively, a *positive emotional change* refers to the change of position in the valence-arousal space where the value of valence after the movement is greater than that of before.

Dialogue Definition

Serban et al. have previously considered the two-hierarchy view of dialogue (Serban et al. 2016), which we extend and adapt in this study. First, we view a dialogue D as a sequence of dialogue turns of arbitrary length M between two speakers. That is, $D = \{U_1, ..., U_M\}$. Each utterance in the m-th dialogue turn is a sequence of tokens of arbitrary length N_m. That is, $U_m = \{w_{m,1}, ..., w_{m,N_m}\}$.

Throughout the study and experiments, we utilize the dialogue triple format. A triple is a sequence of three dialogue turns. That is, $D = \{U_1, U_2, U_3\}$. As we are focused on dyadic dialogue, we consider U_1 and U_3 to be uttered by speaker A, and U_2 by speaker B. In particular, we are interested in triturns with system-user-system speaker sequence. It is practical to view U_1, U_2, and U_3 as dialogue *context*, *query*, and *response*, respectively.

The triple format has been previously utilized for considering context in response generation (Sordoni et al. 2015), filtering multi-party conversation to ensure dyadic snippets (Lasguido et al. 2014), and observing emotion appraisal in a dialogue (Lubis et al. 2017). In this study, the format is particularly useful as it provides both past and future contexts of an emotion occurrence, i.e., U_1 and U_3 are the contexts of the emotion occurrence in U_2.

Recurrent Encoder-Decoder for Dialogue Systems

Recurrent neural network encoder-decoder

A recurrent neural network (RNN) is a neural network variant that can retain information over sequential data. In response generation, first, an *encoder* summarizes an input sequence into a vector representation. An input sequence at time t is modeled using the information gathered by the RNN up to time $t-1$, contained in the hidden state h_t. For

an input sequence $U_m = \{w_{m,1}, ..., w_{m,N_m}\}$ of arbitrary length N_m, the hidden state of the RNN after processing the last token w_{m,N_m} can be viewed as the vector representation of U_m. Afterwards, a *decoder* predicts the output sequence using this representation and its output from the previous time step. Figure 2 presents a schematic view of this process. This architecture was previously proposed as neural conversational model in (Vinyals and Le 2015).

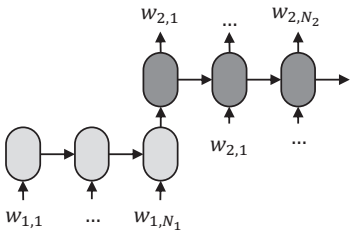

Figure 2: RNN in a response generation task. The lower RNN encodes the information from input sentence U_1, the upper RNN decodes the response U_2.

Hierarchical recurrent encoder-decoder

Based on the two-hierarchy view of dialogue, the hierarchical recurrent encoder-decoder (HRED) extends the sequence-to-sequence architecture (Serban et al. 2016). It consists of three RNNs, each with a distinct role. First, an *utterance encoder* encodes a dialogue turn by recurrently processing each token in the utterance. After processing the last token, the hidden state of the utterance encoder h_{utt} represents the entirety of the dialogue turn, called an utterance vector. This information is then passed on to the *dialogue encoder*, which encodes the sequence of dialogue turns. The state of the dialogue encoder h_{dlg} represents the history of the dialogue up until the currently processed turn. The *utterance decoder*, or the response generator, takes the hidden state of the dialogue encoder, and then predicts the probability distribution over the tokens in the next utterance, i.e., the prediction in the generation process is conditioned on the hidden state of the dialogue encoder. Figure 3 presents an overview of this architecture.

The HRED makes use of the gated recurrent unit (GRU) (Cho et al. 2014) with hyperbolic tangent activation function. The model is trained to maximize the log-likelihood of the training data using the Adam optimizer (Kingma and Ba 2014). Serban et al. argue for the superiority of this architecture for two reasons. First, the dialogue encoder allows the summarization of dialogue history, containing common knowledge between the two speakers. Second, this architecture reduces the computational steps between utterances, allowing a more stable optimization during the training phase.

Affect-sensitive Response Generation

Emotion-sensitive HRED (Emo-HRED)

We propose to incorporate an *emotion encoder* into the HRED architecture. The emotion encoder is placed in the same hierarchy as the dialogue encoder, capturing emotion

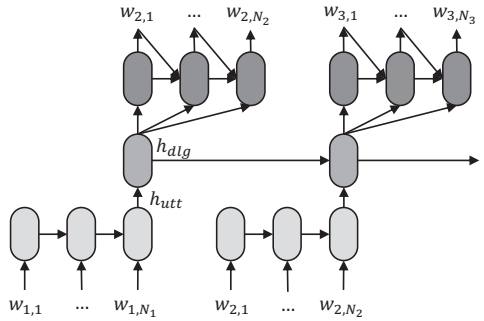

Figure 3: HRED architecture. The lower RNN encodes sequences of tokens, the middle RNN encodes sequence of the dialogue turn, and the upper RNN decodes the tokens of the next dialogue turn.

information at dialogue-turn level and maintaining the emotion context history throughout the dialogue.

The information flow of the Emo-HRED is as follows. After reading the input sequence $U_m = \{w_{m,1}, ..., w_{m,N_m}\}$, the dialogue turn is encoded into utterance representation h_{utt}.

$$h_{utt} = h^{utt}_{N_m} = f(h^{utt}_{N_m-1}, w_{m,N_m}) \quad (1)$$

h_{utt} is then fed into the dialogue encoder to model the sequence of dialogue turns into dialogue context h_{dlg}.

$$h_{dlg} = h^{dlg}_{m} = f(h^{dlg}_{m-1}, h_{utt}) \quad (2)$$

In Emo-HRED, the h_{dlg} is then fed into the emotion encoder, which will then be used to model the emotion context h_{emo}.

$$h_{emo} = f(h^{emo}_{m-1}, h_{dlg}) \quad (3)$$

The generation process of the response, U_{m+1}, is conditioned by the concatenation of the dialogue and emotion contexts.

$$P_\theta(w_{n+1} = v | w_{\leq n}) = \frac{\exp(g(\text{concat}(h_{dlg}, h_{emo}), v))}{\sum_{v'} \exp(g(\text{concat}(h_{dlg}, h_{emo}), v'))} \quad (4)$$

Figure 4 shows a schematic view of this architecture. To the best of our knowledge, this constitutes the first neural network approach for affect-sensitive response generation.

In this study, we consider the emotion encoder to be an RNN with GRU cells and sigmoid activation function. The emotion encoder is trained together with the rest of the network. However, the emotion encoder has its own target vector, which is the emotion label of the currently processed dialogue turn U^{emo}_m. We modify the definition of the training cost to incorporate the prediction error of the emotion encoder.

$$cost_{emo} = ((1 - U^{emo}_m) \cdot \log(1 - f(h_{emo}))) - (U^{emo}_m \cdot \log f(h_{emo})) \quad (5)$$

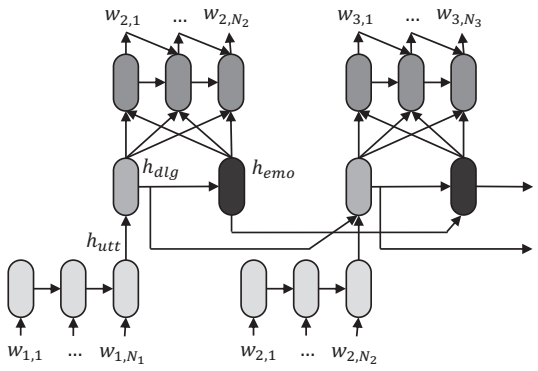

Figure 4: Emo-HRED architecture. The emotion encoder is added to accommodate the emotional context of the dialogue.

The training cost of the Emo-HRED is a linear interpolation between the response generation error $cost_{utt}$ (i.e. negative log-likelihood of the generated response) and the emotion label prediction error $cost_{emo}$ with a decaying weight α. The final cost is then propagated to the network and the parameters are optimized as usual with the optimizer algorithm.

$$cost = \alpha \cdot cost_{emo} + (1 - \alpha) \cdot cost_{utt} \quad (6)$$

Pre-training and selective fine-tuning

Availability of large-scale data is an ongoing challenge for emotion-related research because of the difficulties in capturing life-like emotion occurrence and annotating it reliably. Due to the limited amount of conversational data available with emotion information, training a full end-to-end dialogue system from scratch is unlikely to yield a high quality result. To that extent, pre-training the Emo-HRED with a large scale conversational corpus is essential to infer content and syntactic knowledge prior to training its emotion-related parameters.

We propose selective fine-tuning of the Emo-HRED, limiting the parameter updates to the emotion encoder and utterance decoder only. We hypothesize that the encoding ability has converged during pre-training by utilizing the large amount of data, and will potentially destabilize when fine-tuned using the much smaller, emotion-rich data. As emotion is not yet involved during encoding, we further hypothesize that the pre-trained encoders can be used for the affect-sensitive response generation task as is.

Datasets

Existing datasets

Conversational corpus from movie subtitles Previous works have demonstrated the effectiveness of large scale conversational data in improving the quality of dialogue systems (Banchs and Li 2012; Ameixa et al. 2014; Serban et al. 2016). In this study, we make use of SubTle, a large scale conversational corpus, to learn the syntactic and semantic knowledge for response generation. The use of movie subtitles is particularly suitable as they are available in large amounts and reflecting natural human communication.

The SubTle corpus contains conversational trigger-answer pairs extracted from movie subtitles (Ameixa et al. 2014). Four genres are included in the corpus: horror, science fiction, western, and romance. High-quality movie subtitles are obtained using movie identifiers shared by various movie cataloging websites. The corpus consists of 6,072 subtitle files in total. The subtitles are then automatically processed to obtain conversation pairs, similar to that of Query-Answer format.

Spontaneous affective conversational corpus As emotion-rich data, we utilize an emotionally colored corpus of spontaneous human spoken interaction. The SEMAINE database consists of dialogues between a user and a Sensitive Artificial Listener (SAL) in a Wizard-of-Oz fashion (McKeown et al. 2012). A SAL is a system capable of holding a multimodal conversation with humans, involving speech, head movements, and facial expressions, topped with emotional coloring (Schröder et al. 2012). This emotional coloring is adjusted according to each of the SAL characters; cheerful Poppy, angry Spike, sad Obadiah, and sensible Prudence. Each user interacts with all 4 characters, with each interaction typically lasting for 5 minutes. The topics of conversation are spontaneous, with a limitation that the SAL can not answer any questions. In total, 95 sessions are provided, amounting to 475 minutes of material.

The emotion occurrences are annotated using the FEELtrace system (Cowie et al. 2000) to allow recording of perceived emotion in real time. As an annotator is watching a target person in a video recording, they would move a cursor along a linear scale on an adjacent window to indicate the perceived emotional aspect (e.g. valence or arousal) of the target. This results in a sequence of values with an interval of 0.02 seconds, called a *trace*, that shows how a certain emotional aspect falls and rises within an interaction. The main advantage of emotion trace to utterance-level label is its ability to capture fluctuation of emotion within an utterance, whereas averaging will dampen or even neutralize variations of emotion.

Constructing positive-emotion eliciting data

In addition to considering user's emotion during response generation, we also aim to use this information to elicit positive emotion through human-computer interaction (HCI). We realize our emotion-eliciting goal implicitly by relying on the training data. As the model is trained to learn to mimic the data, the training targets will influence the affective tendency of the responses generated by the system. To minimize the cost of data collection, we enhance the emotion-rich SEMAINE corpus with the following procedure, illustrated in Figure 5.

First, we run the triples through a dialogue system that elicits positive emotion, reported in (Lubis et al. 2017), to obtain new candidate responses that supposedly elicit positive emotion. Subsequently, through crowdsourcing, we ask

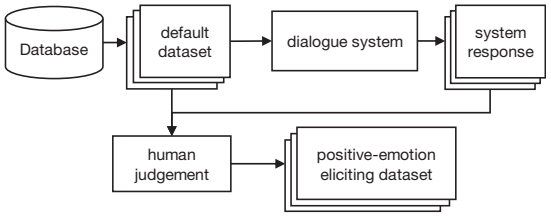

Figure 5: Obtaining references that elicit positive emotion.

human judges to decide which *response*, i.e. U_3, elicits a more positive emotional impact in the triple, the default or the system generated one. If neither are judged to do so, the human judge is asked to provide one that elicits positive emotion. When more than one human-proposed responses are provided for a triple, we manually select the best suited one based on naturalness and potential positive emotional impact. The result of this process is then used to replace the default response from the corpus. These steps ensure the quality of the new responses, aligning it to human standards.

For each triple, we obtain at least 3 human judgements, or more when ties occur. The final response is obtained by majority voting, with each vote weighted by the voter's trust score[1]. In total, 419 crowd workers participated in the judgement process with an average trust score of 0.93. The average consensus of the voting is 0.78.

We fed a total of 2,349 triples extracted from the SEMAINE corpus to the entire process. In the resulting corpus, 12.69% of the responses are human generated, 38.84% are SEMAINE default, 46.38% are system generated, and 2.02% are cases where the default and system generated responses are identical, and voted to elicit positive emotion by the workers. The average word count for the human generated responses are 6.09 words.

Experimental Set Up

Pre-trained model

In our experiments, we utilize the HRED trained on the SubTle corpus as our starting model. The data pre-processing steps are performed as in (Serban et al. 2016). The processed SubTle corpus contained 5,503,741 query-answer pairs in total. The triple format is forced onto the pairs by treating the last dialogue turn in the triple as empty. The 10,000 most frequent tokens are treated as the system's vocabulary, and the rest as unknowns.

The model is pre-trained by feeding this dataset sequentially into the network until it converges, taking approximately 2 days to complete. In addition to the model parameters, we also learn the word embeddings of the tokens. We used word embeddings with size 300, utterance vectors of size 600, and dialogue vectors of size 1200. The parameters are randomly initialized, and then trained to optimize the log-likelihood of the training triples using the Adam optimizer.

[1]A worker's trust score is equal to the percentage of correct answers to a set of triples for which we provided the gold standard answers.

Fine-tuning set ups

All the models considered in this study are the result of fine-tuning the pre-trained model with the emotion-rich data, fed sequentially into the network. To investigate the effectiveness of the proposed approach, we train multiple models with combinations of set ups.

First, we consider two different models: HRED as baseline model and Emo-HRED as the proposed model. As emotion information for the Emo-HRED, we use the valence and arousal traces provided by the SEMAINE corpus as emotion context. For a dialogue turn, we sample with replacement a vector of length 100 from each trace. We concatenate the valence and arousal vectors to form the final emotion label, resulting in an emotion vector of length 200. This length is chosen to balance the size of emotion information with the network and word vector size. To accommodate this additional information during fine-tuning, we append new randomly initialized parameters to the utterance decoder. These parameters are trained exclusively during the fine-tuning process.

Second, we consider two different parameter update schemes: `standard` and `selective`. In the `standard` scheme, we fine-tune all the parameters of the model. In the `selective` scheme, we fix the utterance and dialogue encoders parameters, as has been previously elaborated. We hypothesize that the `selective` scheme will produce a more stable model after fine-tuning. The SubTle and SEMAINE corpus have a number of major differences that may cause straight-forward fine-tuning to not work optimally. For one, the sizes of the corpora differ significantly (5.5M vs.2K triples). Furthermore, the conversations are of different nature (human-human vs. human-wizard, acted speech vs. spontaneous speech).

Third, we consider two different emotion-rich datasets to test the effectiveness of the implicit positive-emotion elicitation strategy: the `sem` dataset, using the dialogue turns as provided by the SEMAINE corpus; and the `pos_sem` dataset, containing positive emotion eliciting *responses*,i.e. U_3 produced through the process previously described (Figure 5). We hypothesize that the `pos_sem` corpus will cause the model to elicit more positive emotion. For both, we consider 66 sessions from the SEMAINE corpus based on transcription and emotion annotation availability; 17 of Poppy's sessions, 16 Spike, 17 Obadiah, and 16 Prudence. For every dialogue turn, we keep the speaker information (wizard or user), transcription, and emotion annotation. We partition the data as follows: 58 sessions (1985 triples) for training, 4 (170) for validation, and 4 (194) for test.

Evaluation

We perform two evaluations to confirm the effectiveness of the proposed method. First, we perform objective evaluation of the systems by calculating model perplexity. Second, we perform subjective evaluation to measure the naturalness and emotional impact of the generated responses. To obtain

deeper insight into the evaluation results, analyses are provided at the end of this section.

Objective evaluation

Perplexity measures the probability of exactly regenerating the reference response in a triple. This metric is commonly used to evaluate dialogue systems that relies on probabilistic approaches (Serban et al. 2016) and has been previously recommended for evaluating generative dialogue systems (Pietquin and Hastie 2013). We evaluate the models using the `pos_sem` test set, as we assume this dataset to be the one that fulfills our main goal of an emotionally positive dialogue. Table 1 presents the perplexity of the models fine-tuned with different set ups.

Table 1: Model perplexity on `pos_sem` test set.

Model	**Parameter update**	**Fine-tune data**	**Perplexity**
Baseline HRED	`standard`	`sem`	185.66
		`pos_sem`	121.44
	`selective`	`sem`	151.77
		`pos_sem`	100.94
Proposed Emo-HRED	`selective`	`sem`	69.66
		`pos_sem`	**51.34**

First, we test the effect of the parameter update and fine-tune data by holding the model fixed to HRED. We observe significant improvements when fine-tuning only the decoder (`selective` scheme) compared to the the entire network (`standard` scheme). This supports the hypotheses that we have previously made: it is better to utilize the encoders pre-trained using the large dataset as is, rather than to fine-tune them further using the small emotion-rich data.

Second, we test the impact of the emotion encoder by comparing HRED and Emo-HRED. we found that with identical starting model and fine-tune set up, the Emo-HRED architecture converges to significantly better models compared to the HRED. This suggests two things: incorporation the emotion prediction error helps the model to converge to a better local optimum, and that the emotion information helps in generating a response closer to the training reference.

We suspect that partly tuning the parameters through the smaller valence-arousal space helps the model to infer useful information for response generation through the simpler emotion recognition task. The relationship between semantic and emotional content is not arbitrary, and thus utilizing them in combination could benefit the learning process of the model.

Fine-tuning the HRED model with `standard` weight update scheme is equivalent to the SubTle bootstrap approach proposed in (Serban et al. 2016). However, there are differences that are important to highlight, summarized in Table 2. Due to these differences, it is not possible to straightforwardly compare model perplexities on the respective test sets. However, this demonstrates the ability of Emo-HRED to efficiently take advantage of emotion information, consequently decreasing model perplexity despite of small data size, which is often a challenge in affective computing works.

Table 2: Model comparison.

	(Serban et al. 2016)	This research
Pretraining	SubTle bootstrap	
Fine-tune and test data	MovieTriples	SEMAINE
# triples	245,492	2,349
Architecture	HRED Bidirectional	Emo-HRED
Emotion	No	Yes
Perplexity	26.81	51.34

Human subjective evaluation

We present human judges with a dialogue triple and ask them to rate the response in terms of two criteria. The first is naturalness, which evaluates whether the response is intelligible, logically follows the dialogue context, and resembles real human response. The second is emotional impact, which evaluates whether the response elicits a positive emotional impact or promotes an emotionally positive conversation.

We evaluate 4 `selective` models, combining the options for model (HRED or Emo-HRED) and fine-tune data (`sem` or `pos_sem`). We evaluate 100 triples from the full test set, where each is judged by 20 human evaluators. Each triple is presented in A-B-A format, the first two dialogue turns are held fixed according to the test set, and the last turn is the response generated by the evaluated model. Evaluators are asked to judge the responses by stating their agreement to two statements: 1) A gives a natural response, and 2) A's response elicits a positive emotional impact in B. The agreement is given using a Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree).

Table 3: Subjective evaluation scores. Average (avg) and standard deviation (stdev) accross all test triples are shown. * denotes $p < 0.05$ compared with HRED default.

System	Naturalness		Emotional impact	
	avg	stdev	avg	stdev
HRED `sem`	2.71	0.31	2.56	0.29
HRED `pos_sem`	3.26 *	0.22	3.22	0.25
Emo-HRED `sem`	3.14 *	0.25	3.02	0.25
Emo-HRED `pos_sem`	3.59 *	0.19	3.58 *	0.24

Table 3 shows the result of subjective evaluation. The proposed method is shown to achieve highest scores for both metrics. That is, the responses it generates are perceived as more natural and elicit a more positive emotional response. The reported averages and standard deviations show that the ratings are higher and at the same time more uniform, i.e. less spread out from the average, when both the proposed data and architecture are applied to the model. Paired t-test of the models with HRED `sem` shows significant improve-

ment in naturalness for the other three models, and in emotional impact for Emo-HRED pos_sem.

Analysis

Both the subjective and objective evaluation show consistent incremental improvements when each of the proposed set ups is applied to the dialogue system. They also show that a system performs best when all of the proposed set ups are applied at the same time.

Table 4: Comparison of system responses for a triple in test set.

U_1	that ' s so cool you must be so proud of yourself .
U_2	ah yeah i am i am very proud because it ' s like i didn ' t think it was gon na go this far [laugh] .
U_3 sem	yeah .
U_3 pos_sem	that 's good yes .
HRED sem	cause you don ' t really want to go out with people .
HRED pos_sem	yeah so you have to be inside really for the best .
Emo-HRED sem	yes .
Emo-HRED pos_sem	yes that ' s that ' s good .

Table 4 shows an example of a test triple along with responses generated by the models. We analyse the generated responses to reason for the objective and subjective evaluation results. We found that on average, the Emo-HRED models generated responses that are shorter compared to that of HRED (6.76 vs. 8.19 words). Consequently, the Emo-HRED responses amount to a smaller vocabulary than the HRED. However, this smaller vocabulary contains larger proportions of positive-sentiment words. For example, with HRED sem, HRED pos_sem, Emo-HRED sem, and Emo-HRED pos_sem respectively, the word "good" makes up 2.4%, 4.9%, 4.3%, and 26.5% of the evaluated responses, excluding stopwords. Table 5 lists top 10 words from these vocabularies in order of frequency.

Table 5: 10 most frequent words in the generated responses, excluding stop words. Positive sentiment words are boldfaced.

System	Most frequent content words
HRED sem	tell, well, **like**, **good**, think, make, else, go, get, know
HRED pos_sem	tell, **good**, think, **nice**, well, **sensible**, see, meet, know, really
Emo-HRED sem	aha, deal, things, **good**, go, yes, know, away, people, sure
Emo-HRED pos_sem	**good**, tell, aha, hear, **glad**, well, **happy**, makes, yes, **fun**

It may be of interest to note that we also observe the same tendency in the responses we collected from human annotators for the pos_sem dataset. This actually follows human strategy when promoting positive emotional experiences in conversations with only limited context provided – by using general responses that contain positive-sentiment words.

Furthermore, we observe similar phenomena on the subjective evaluation results. As the response length grows, so does its likelihood to carry grammatical or logical errors. This leads to both poor naturalness and uncertain emotional responses upon human perception. The responses generated from the proposed model are short and sweet, enough to sustain general conversation with short context (in this case, two previous dialogue turns), similar to that of human daily small talks. These tendencies observed from the pos_sem dataset and Emo-HRED model could explain the lower perplexity when one of them is employed, and lowest when both are.

To clarify, this is not to say that short, generic responses are always desirable. This is a standing problem for neural network based response generation (Li et al. 2016) – moving toward longer, context-specific responses will lead to a more engaging interaction. However, we note that there are circumstances for which the implicit strategy of the proposed method is suitable, as previously discussed. We look forward to expand the conversational ability of the model to accommodate longer context and content-specific information in future works.

Conclusion

We proposed a neural network approach for affect-sensitive response generation and positive emotional impact elicitation. We extend the recently proposed HRED (Serban et al. 2016) and augment it with an emotion encoder to capture the emotional context of a dialogue. This information is then used in the response generation process to produce an affect-sensitive response. By obtaining positive-emotion eliciting dialogue targets, we influence the affect-sensitive system to elicit positive emotion through the responses it generates.

The evaluations we conducted show that the proposed architecture, data, and training procedure result in a better model: it produces responses that are perceived as more natural and eliciting a more positive emotional impact. Analysis of the evaluations suggests that the proposed method generates shorter responses that contain more positive-sentiment words. This resembles human strategy when promoting positive emotion in a conversation with limited context.

We acknowledge that evaluation through real user interaction needs to be carried in the future to test the effectiveness of the positive emotion elicitation in a more realistic scenario. We also hope to further improve the quality of the proposed system, both in terms of response quality and user emotional experience. We believe that collection of emotionally rich conversational data is crucial and will highly benefit this research effort, widening the scope of the data to cover larger conversational scenarios. In terms of elicitation strategy, we would like to define an explicit training goal to maximize the positive emotional effect of the generated response. Lastly, we hope to consider longer dialogue history to evoke a more context-specific response generation.

Acknowledgement

This research and development work was supported by the MIC/SCOPE #152307004.

References

Ameixa, D.; Coheur, L.; Fialho, P.; and Quaresma, P. 2014. Luke, I am your father: dealing with out-of-domain requests by using movies subtitles. In *International Conference on Intelligent Virtual Agents*, 13–21. Springer.

American Psychological Association, A. P. A. 2015. Stress snapshot, http://www.apa.org/news/press/releases/stress/2015/snapshot.aspx.

Banchs, R. E., and Li, H. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, 37–42. Association for Computational Linguistics.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cowie, R.; Douglas-Cowie, E.; Savvidou, S.; McMahon, E.; Sawey, M.; and Schröder, M. 2000. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Egges, A.; Kshirsagar, S.; and Magnenat-Thalmann, N. 2004. Generic personality and emotion simulation for conversational agents. *Computer animation and virtual worlds* 15(1):1–13.

Han, S.; Kim, Y.; and Lee, G. G. 2015. Micro-counseling dialog system based on semantic content. In *Natural Language Dialog Systems and Intelligent Assistants*. Springer. 63–72.

Hasegawa, T.; Kaji, N.; Yoshinaga, N.; and Toyoda, M. 2013. Predicting and eliciting addressee's emotion in online dialogue. In *Proceedings of Association for Computational Linguistics (1)*, 964–972.

Higashinaka, R.; Dohsaka, K.; and Isozaki, H. 2008. Effects of self-disclosure and empathy in human-computer dialogue. In *Proceedings of Spoken Language Technology Workshop*, 109–112. IEEE.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lasguido, N.; Sakti, S.; Neubig, G.; Toda, T.; and Nakamura, S. 2014. Utilizing human-to-human conversation examples for a multi domain chat-oriented dialog system. *Transactions on Information and Systems* 97(6):1497–1505.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, 110–119.

Lubis, N.; Sakti, S.; Yoshino, K.; and Nakamura, S. 2017. Eliciting positive emotional impact in dialogue response selection. In *Proceedings of International Workshop on Spoken Dialogue Systems Technology*.

McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M.; and Schroder, M. 2012. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Transactions on Affective Computing* 3(1):5–17.

Nio, L.; Sakti, S.; Neubig, G.; Yoshino, K.; and Nakamura, S. 2016. Neural network approaches to dialog response retrieval and generation. *IEICE Transactions on Information and Systems*.

Picard, R. W., and Klein, J. 2002. Computers that recognise and respond to user emotion: theoretical and practical implications. *Interacting with computers* 14(2):141–169.

Pietquin, O., and Hastie, H. 2013. A survey on metrics for the evaluation of user simulations. *The knowledge engineering review* 28(1):59–73.

Russell, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39(6):1161.

Schröder, M.; Bevacqua, E.; Cowie, R.; Eyben, F.; Gunes, H.; Heylen, D.; Maat, M. T.; McKeown, G.; Pammi, S.; Pantic, M.; et al. 2012. Building autonomous sensitive artificial listeners. *Transactions on Affective Computing* 3(2):165–183.

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

Skowron, M.; Theunis, M.; Rank, S.; and Kappas, A. 2013. Affect and social processes in online communication–experiments with an affective dialog system. *Transactions on Affective Computing* 4(3):267–279.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Tielman, M.; Neerincx, M.; Meyer, J.-J.; and Looije, R. 2014. Adaptive emotional expression in robot-child interaction. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 407–414. ACM.

Vinyals, O., and Le, Q. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Waldinger, R. J.; Cohen, S.; Schulz, M. S.; and Crowell, J. A. 2015. Security of attachment to spouses in late life: Concurrent and prospective links with cognitive and emotional well-being. *Clinical Psychological Science* 3(4):516–529.

Zech, E., and Rimé, B. 2005. Is talking about an emotional experience helpful? Effects on emotional recovery and perceived benefits. *Clinical Psychology & Psychotherapy* 12(4):270–287.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.