Unsupervised Counselor Dialogue Clustering for Positive Emotion Elicitation in Neural Dialogue System

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

Abstract

Positive emotion elicitation seeks to improve user's emotional state through dialogue system interaction, where a chat-based scenario is layered with an implicit goal to address user's emotional needs. Standard neural dialogue system approaches still fall short in this situation as they tend to generate only short, generic responses. Learning from expert actions is critical, as these potentially differ from standard dialogue acts. In this paper, we propose using a hierarchical neural network for response generation that is conditioned on 1) expert's action, 2) dialogue context, and 3) user emotion, encoded from user input. We construct a corpus of interactions between a counselor and 30 participants following a negative emotional exposure to learn expert actions and responses in a positive emotion elicitation scenario. Instead of relying on the expensive, labor intensive, and often ambiguous human annotations, we unsupervisedly cluster the expert's responses and use the resulting labels to train the network. Our experiments and evaluation show that the proposed approach yields lower perplexity and generates a larger variety of responses.

1 Intoduction

Emotionally intelligent systems has high potential as assistive technology in 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*, which allows a system to discern the user's emotions and address them in giving a response (Forbes-Riley and Litman, 2012; Han et al., 2015; Tielman et al., 2014), and *emotion simulation*, which helps convey non-verbal aspects to the user for a more believable and human-like interaction, for example to show empathy (Higashinaka et al., 2008) or personality (Egges et al., 2004). Acosta and Ward (2011) have attempted to connect the two competences to build rapport, by recognizing user's emotion and reflecting it in the system response. Although these competences address some of the user's emotional needs (Picard and Klein, 2002), they are not sufficient to provide emotional support in an interaction.

Recently, there has been an increasing interest in eliciting user's emotional response via dialogue system interaction, i.e. *emotion elicitation*. Skowron et al. (2013) have studied the impact of different affective personalities in a text-based dialogue system, while Hasegawa et al. (2013) constructed translation-based response generators with various emotion targets. Despite the positive results, these approaches have not yet paid attention to the emotional benefit for the users. Our work aims to draw on an important overlooked potential of emotion elicitation: its application to improve emotional states, similar to that of emotional support between humans. This can be achieved by actively eliciting a more positive emotional valence throughout the interaction, i.e. *positive emotion elicitation*. This takes form as a chat-oriented dialogue system interaction that is layered with an implicit goal to address user's emotional needs.

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). However, application of this approach towards positive emotion elicitation is still very lacking. Zhou et al. (2017)

have investigated 6 categories to emotionally color the response via the internal state of the decoder. However, this study has not yet considered user's emotion in the response generation process, nor attempted improve emotional experience of user.

Towards positive emotion elicitation, Lubis et al. (2018) have recently proposed a model that encodes emotion information from user input and utilizes it in generating response. However, the resulting system is still limited to short and generic responses with positive affect, echoing the long standing lack-of-diversity problem in neural network based response generation (Li et al., 2016). Furthermore, the reported system has not learn about positive emotion elicitation strategies from an expert as the corpus construction relied on crowd-sourcing workers.

This points to another problem: the lack of data that shows positive emotion elicitation or emotion recovery in everyday situations. Learning from expert responses and actions are essential in such a scenario as these potentially differ from standard chat-based scenarios. With scarcity of large-scale data, additional knowledge from higher level abstraction, such as dialogue action labels, may be highly beneficial. However, such high-level information must rely on human annotations, which are expensive, labor intensive, and often ambiguous.

To answer these challenges, first, we construct a corpus containing recordings of a professional counselor and 30 participants in a positive emotion elicitation scenario. Second, we extract higher level information from the expert's responses via unsupervised clustering and use the resulting labels to train a neural dialogue system. Lastly, we propose a hierarchical neural dialogue system which considers 1) expert's action, 2) dialogue context, and 3) user emotion, in generating a response by encoding them from user input. Our evaluations show that the proposed method yields lower perplexity, elicits a positive emotional impact, and generates longer responses that improves subjective engagement.

2 Corpus Construction: Positive Emotion Elicitation by an Expert

Even though various affective conversational scenarios have been considered (McKeown et al., 2012; Gratch et al., 2014), there is still a lack of resources that show common emotional problems in everyday social settings. Furthermore, a great majority of existing corpora does not involve any professional who is an expert in handling emotional reactions in a conversation.

To fill these gaps, we design our corpus to 1) contain recordings of spontaneous dyadic interactions before and after a negative emotion exposure, and 2) involve a professional counselor as an expert. In each interaction, a negative emotion inducer is shown to the dyad, and the goal of the expert is to aid emotion processing and elicit a positive emotional change through the interaction. From this point, we will refer to this corpus as the counseling corpus.

2.1 Negative Emotion Inducer

To induce negative emotion, we opt for short video clips which are a few minutes in length. This method is well established and has been studied for several decades (Gross and Levenson, 1995; Schaefer et al., 2010). One study shows that amongst a number of techniques, the use of video clips is the most effective way to induce both positive and negative emotional states (Westermann et al., 1996). It also offers easy replication in constrained environmental settings, such as the recording room.

However, in contrast to previous works (Schaefer and Philippot, 2005), we look for clips that depict real life situations and issues, i.e., non-fiction and non-films. We select short video clips of news reports, interviews, and documentary films as emotion inducers to avoid the unpredictability of subjective emotional response to fictional clips. Non-fictional inducer also reflects real everyday situations better. We ensure that the clips contain enough information and context to serve as conversation topic throughout the recording session.

We target two emotions with negative valence: anger and sadness. First, we manually selected 34 of videos with varying relevant topics that are provided freely online. Two human experts are then asked to rate them in terms of intensity and the induced emotion (sadness or anger). Finally, we selected 20 videos, 10 of each emotion with varied intensity level where the two human ratings agree.

2.2 Data Collection

We arrange for the dyad to consist of an *expert* and a *participant*, each with a distinct role. The roles are based on the "social sharing of emotion" scenario, which argues that after an emotional event, a person is inclined to initiate an interaction which

centers on the event and their reactions to it (Rime et al., 1991; Luminet IV et al., 2000). This form of social sharing is argued to be integral in processing the emotional event (Rime et al., 1991).

In the interactions, the *expert* plays the part of the external party who helps facilitate this process following the emotional response of the *participant*. We recruit a professional counselor as the *expert* in the recording, an accredited member of the British Association for Counseling and Psychotherapy with more than 8 years of professional experience. As *participants*, we recruit 30 individuals (20 males and 10 females) that speak English fluently as first or second language.

A session starts with an opening talk as a neutral baseline conversation. Afterwards, we induce negative emotion by showing an emotion inducer to the dyad. This is followed by a discussion that targets at emotional processing and recovery, during which the expert is given the objective to facilitate the processing of emotional response caused by the emotion induction, and to elicit a positive emotional change.

In total, we recorded 60 sessions of interactions, 30 with "anger" inducer and 30 with "sadness". The combined duration of all sessions sums up to 23 hours and 41 minutes of material. The audio and video recordings are transcribed, including a number of special notations for non-speech sounds such as laughter, back-channels, and throat noise.

2.3 Emotion Annotation

We follow the *circumplex model of affect* (Russell, 1980) in annotating emotion occurrences in the recordings. 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).

For each recording, the participants self report their emotional state using the FEELtrace system (Cowie et al., 2000) immediately after the interaction. While an annotator is watching a target person in a recording, he or she is moving 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 real numbers ranging from -1 to 1 with a constant time interval, called a *trace*. Statistical analyses of validation experiments have confirmed the reliability and indicated the precision of the FEELtrace system (Cowie et al., 2000).

2.4 Dialogue Triples

Throughout the study and experiments, we utilize the dialogue triple format, i.e. a sequence of three dialogue turns. It has been previously utilized for considering dialogue context (Sordoni et al., 2015), filtering multi-party conversation (Lasguido et al., 2014), and observing emotion appraisal (Lubis et al., 2017). In this study, we exploit it to provide both past and future contexts of an emotion occurrence

We extend and adapt the two-hierarchy view of dialogue (Serban et al., 2016). We view a dialogue D as a sequence of dialogue turns of arbitrary length M between two speakers, i.e. $D = \{U_1, ..., U_M\}$. Each utterance in the m-th dialogue turn is a sequence of tokens of arbitrary length N_m, i.e. $U_m = \{w_{m,1}, ..., w_{m,N_m}\}$. In a triple, $D = \{U_1, U_2, U_3\}$, where U_1 and U_3 are uttered by speaker A, and U_2 by speaker B. In particular, we are interested in triturns with counselor-participant-counselor speaker sequence. It is practical to view U_1, U_2, and U_3 as dialogue *context*, *query*, and *response*, respectively. U_1 and U_3 are the contexts of the emotion occurrence in U_2.

We define the end of a dialogue turn as either 1) natural end of the sentence, or 2) turn taking by the other speaker, whichever comes first. Back channels in the middle of a speaker's utterance are not considered as turn taking since they instead signal active listening. This also prevents overly fragmented dialogue turns. The backchannels are instead appended into the next dialogue turn once one of the criteria above is met. We extract a total of 6,064 dialogue triples from the collected data. All U_2 are aligned with self-report emotion annotation by the participants.

3 Recurrent Encoder-Decoder for Dialogue Systems

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. Afterwards, a *decoder* recurrently pre-

dicts the output sequence conditioned by h_t and its output from the previous time step. This architecture was previously proposed as neural conversational model in (Vinyals and Le, 2015).

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. An *utterance encoder* recurrently processes each token in the utterance, encoding it into a vector respresentation h_{utt}. This information is then passed on to the *dialogue encoder*, which encodes the sequence of dialogue turns into h_{dlg}. The *utterance decoder*, or the response generator, takes h_{dlg}, and then predicts the probability distribution over the tokens in the next utterance.

Recently, the HRED architecture has been extended to Emo-HRED for the positive emotion elicitation task, exploiting the hierarchical view of dialogue to observe the conversational context of an emotion occurrence (Lubis et al., 2018). Emo-HRED incorporates an *emotion encoder* which predicts user emotional state and passes this information to the response generation process. The emotion encoder is placed in the same hierarchy as the dialogue encoder, capturing emotion information at dialogue-turn level h_{emo} and maintaining the emotion context history throughout the dialogue. Improved naturalness and a more positive emotional impact were reported in the evaluations of Emo-HRED, however the resulting system is still limited to short and generic responses with positive affect. This echoes the long standing lack-of-diversity problem in neural network based response generation (Li et al., 2016), which is also shared by other models previously discussed.

4 Proposed Method

4.1 Unsupervised Clustering of Counselor Dialogue

In constructing an emotionally intelligent system, learning from expert actions and responses are essential. Although statistical learning from raw data has been shown to be sufficient in some cases, it might not be so for positive emotion elicitation task. Due to the absence of large scale data, additional knowledge from higher level abstraction, such as dialogue action labels, may be highly beneficial. We hypothesize that these labels will reduce data sparsity by categorizing counselor responses and emphasizing this information in the training and generation process.

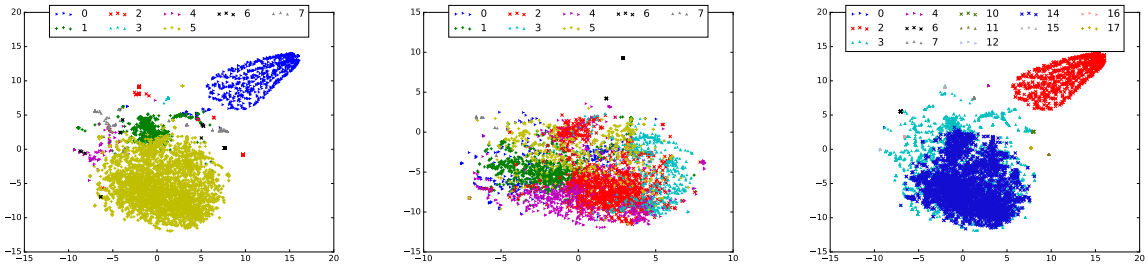
However, procuring such labels is not a trivial task. Human annotation is not a practical solution as it is expensive, time-consuming, and labor intensive. Especially with subjective aspects such as dialogue act labels, they are often less reliable due to low annotator agreement. On the other hand, training an automatic classifier from data with standard dialogue act labels will not cover actions with specific emotion-related intent that are present in the collected data. For example, empathy towards negative affect ("That's sad.") and positive affect ("I'm happy to hear that.").

We propose unsupervised clustering of counselor dialogue to obtain dialogue act labels of expert responses. We collected a total of 6384 counselor utterances from the counseling corpus. We transform the utterances into vectors by obtaining the embeddings of the words in the utterance and averaging them. We use a word2vec model pretrained on 100 billion words of Google News (Mikolov et al., 2013). The word and utterance embeddings are of length 300. We then apply two clustering methods to the vectorized utterances: K-Means and Dirichlet process Gaussian mixture model (DPGMM).

With K-means, we perform hierarchical clustering, starting with an initial K of 8. We perform K-means clustering the second time on the clusters which are larger than half the full data size. In contrast, DPGMM is a non-parametric model, i.e. it attempts to represent the data without prior definition of the model complexity. We use the stick-breaking construction for the DPGMM. A new data point would either join an existing cluster or start a new cluster following some probabilities. We use diagonal covariance matrices to compensate for the limited amount of data. Henceforth, we refer to the result of the clustering as *cluster label*.

Cluster Analysis

We visualize the found clusters using T-SNE in Figure 1. K-Means clustering shows distinct dialogue acts characteristic in a number of clusters it found. For example, cluster 0 in Figure 1(a) consists of various utterances signaling active listening, such as follow up questions and short back channels. On the other hand, cluster 2 and 6 contains utterance showing confirmation or agreement, such as utterances containing the words "yeah," "right," and "yes." We also obtain



(a) K-Means clustering. (b) K-Means sub-clustering on cluster 5. (c) DPGMM clustering.

Figure 1: T-SNE Representation of the clustering results.

smaller clusters for appreciation or thanking and non-speech sounds, such as laughter and breathing. The rest of the utterances which are relatively longer are grouped together in a very large cluster with 4220 members (cluster 5 in Figure 1(a)).

Second clustering on cluster 5 group these utterances into smaller sub-clusters (Figure 1(b)). "I" is the most frequent word in sub-cluster 0, and "you" in sub-cluster 1. Some of the actions from the first clustering are re-discovered during the second clustering, such as thanking and appreciation in sub-cluster 7, and confirmation in sub-cluster 6. The largest sub-cluster is sub-cluster 2 with 1324 members which contain longer utterances, a combination of opinion, questions, and other sentences. In total, we obtained 15 clusters from K-means clustering.

On the other hand, the DPGMM clustering results in 13 clusters. DPGMM clustering yield a similar result, giving one huge cluster for longer sentences and smaller clusters populated with for back channel, non-speech sounds, thank you, and agreement. However, there are several differences between the results from DPGMM and K-means that are worth mentioning. First, we notice that the characteristic of each cluster is less salient compared to that of K-Means; e.g. numerous back channels can be found in several other clusters. Second, the class size distribution is more uneven: there are 6 clusters with less than 100 members, in contrast to only 1 with K-Means. Third, unlike K-Means, re-clustering of the biggest cluster is not possible as it is already represented by one component in the model.

4.2 Hierarchical Neural Dialogue System with Multiple Contexts

We propose providing higher level knowledge about the response to the model, in form of response cluster labels (Section 4.1), to aid its response generation. We propose a neural dialogue system which generate response based on multiple dialogue contexts: 1) dialogue history, 2) user emotional state, and 3) expert's action label. Henceforth we call this model the multi-context HRED (MC-HRED)

The information flow of the MC-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 MC-HRED, the h_{dlg} is then fed into the emotion and action encoders, which will then be used to encode the emotion context h_{emo} as well as the expert action label h_{act}.

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

where $enc = \{emo, act\}$.

The generation process of the response, U_{m+1}, is conditioned by the concatenation of the three contexts: dialogue history, emotion context, and the expert action label.

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

Figure 2 shows a schematic view of this architecture. For each the emotion and action encoders, we consider an RNN with gated recurrent unit (GRU) cells and sigmoid activation function.

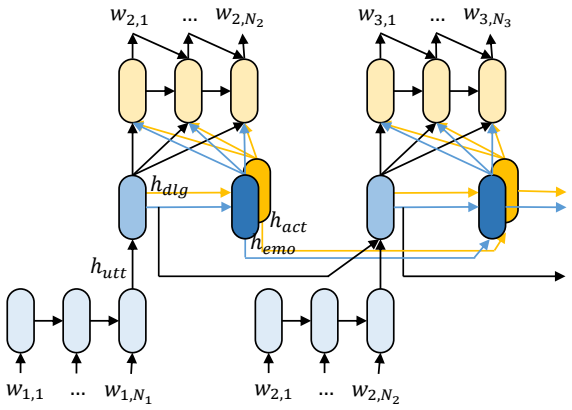


Figure 2: MC-HRED architecture. Emotion encoder is shown in dark blue, and action encoder in dark yellow. Blue NNs are relating to input, and yellow NNs to response.

Both encoders are trained together with the rest of the network. Each encoder has its own target vector, which is the emotion label of the currently processed dialogue turn U_m^{emo} and expert action label of the target response U_m^{act}. We modify the definition of the training cost to incorporate the cross entropy losses of the emotion and action encoders.

$$cost_{enc} = ((1 - U_m^{enc}) \cdot \log(1 - f(h_{enc}))) - (U_m^{enc} \cdot \log f(h_{enc})), \quad (5)$$

where $enc = \{emo, act\}$.

The training cost of the MC-HRED is a linear interpolation between the response generation error $cost_{utt}$ (i.e. negative log-likelihood of the generated response) and the prediction errors of the encoders $cost_{emo}$ and $cost_{act}$ with weights α and β which decays after every epoch.

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

The final cost is then propagated to the network and the parameters are optimized as usual with the optimizer algorithm.

5 Experimental Set Up

Figure 3 illustrates the experimental set up of this work. Each of the steps will be explained in this section. The scope of this study is limited to text data.

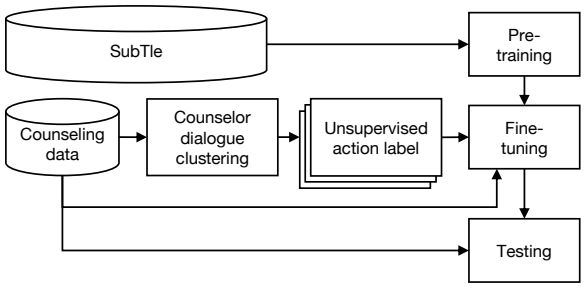


Figure 3: The flow of the experiment.

5.1 Pre-trained model

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 (Ameixa et al., 2014), a large scale conversational corpus collected from movie subtitles, 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.

In our experiments, we utilize the HRED trained on the SubTle corpus as our starting model. We follow the data pre-processing method 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. We select the 10,000 most frequent token from the combination of SubTle and the counseling data as system vocabulary. The purpose is twofold: to help widen the intersection of words between the two corpora, and to preserve special token from the counselor corpus such as laughter and other non-speech sounds.

The model is pre-trained by feeding the SubTle 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.

5.2 Fine-tuning

All the models considered in this study are the result of fine-tuning the pre-trained model with the

counseling corpus (Section 2). The triples from the corpus are fed sequentially into the network. To investigate the effectiveness of the proposed methods, we train multiple models with combinations of set ups.

We consider two different models: Emo-HRED as baseline model and MC-HRED as the proposed model. Emo-HRED considers only dialogue history and emotional context during the response generation, while MC-HRED considers expert action context in addition to the dialogue history and emotional context. For completeness, we also train a model that only utilized dialogue history and action context, which we will call Clust-HRED for convenience.

As emotional context, we encode the self-report emotion annotation into a one-hot vector as follows. We first obtain the average valence and arousal values of an utterance. We then discretize these values respectively into three classes: positive, neutral, and negative. The intervals for the classes are $[-1, -0.07]$ for negative, $(-0.07, 0.07)$ for neutral, and $[0.07, 1]$ for positive. We then encode this class information into a one-hot vector of length 9, one element for each of the possible combinations of valence and arousal classes, i.e. positive-positive, positive-neutral, neutral-negative, etc. Preliminary experiments showed that on the counselor corpus, this representation leads to a better performance compared to fixed-length sampling of the emotion trace.

As action context, we simply encode the cluster label of U_3, obtain as in Section 4.1, into a one-hot vector. We experimented with two cluster label sets, one produced by hierarchical K-Means clustering (15 clusters), and one by DPGMM clustering (13).

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. All models are fine-tuned selectively. That is, we fix the utterance and dialogue encoders parameters, and selectively train only the proposed encoders as well as the decoder. This has been shown to result in a more stable model when fine-tuning with a small amount of data (Lubis et al., 2018).

We partitioned the counseling corpus into 50 recording sessions (5053 triples) for training, 5 (503) for validation, and 5 (508) for testing.

6 Evaluation and Analysis

6.1 Perplexity

We calculate model perplexity, which measures the probability of exactly regenerating the reference response in a triple. Since the target responses are assumed to be expert's response, its reproduction by the model is desirable. Perplexity has also been previously recommended for evaluating generative dialogue systems (Pietquin and Hastie, 2013).

We compute the perplexity for each triple and average it to obtain model perplexity. The model perplexities are summarized in Table 1. We compute the average test triple length (59.6 tokens), and group the test triples into two: those with below average length as "short" (294 triples), and those above as "long" (186). Average perplexities are shown for the entire test set (all), the short group, and the long group, separately.

Model	Emo.	Action	Perplexity		
			all	short	long
Emo-HRED	Yes	No	42.60	35.74	61.17
Clust-HRED	No	K-Means	39.57	32.30	57.37
		DPGMM	30.57	24.79	42.25
MC-HRED	Yes	K-Means	**29.57**	**23.23**	**38.73**
		DPGMM	32.04	25.00	42.34

Table 1: Model Perplexity of different architectures.

We obtain model with the lowest perplexity when emotion and K-Means labels are both utilized in the training and response generation process. For all models, the perplexity of long triples is consistently higher than that of short ones. More significant improvement is observed on long test triples.

Looking at the perplexity on all test triples, interestingly, the two cluster labels are affected in starkly different ways when combined with emotion labels: K-Means gain significant improvement, while DPGMM slightly suffers. We found that on long triples, Clust-HRED and MC-HRED yield similar performances when using the DPGMM cluster label. In contrast, when using K-means label, MC-HRED shows further improvement from Clust-HRED.

We separate the test triples based on the average model perplexity to analyze their properties.

Aside from triple length, no other significant difference was observed. This signals that the ability to capture context is one of the defining characteristic of a strong model for this task.

6.2 Human Subjective Evaluation

We present human judges with a dialogue triple and ask them to rate the response in terms of three criteria: 1) naturalness, which evaluates whether the response is intelligible, logically follows the dialogue context, and resembles real human response, 2) emotional impact, to measure whether the response elicits a positive emotional impact or promotes an emotionally positive conversation, and 3) engagement, to evaluate whether the proposed response shows involvement in the dialogue and promotes longer conversation by inviting more response.

We evaluate Emo-HRED and the best performing MC-HRED utilizing K-Means clustering labels. 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 three statements: 1) A gives a natural response, 2) A's response elicits a positive emotional impact in B, and 3) A's response in engaging. The agreement is given using a Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). Figure 4 summarizes the subjective evaluation result.

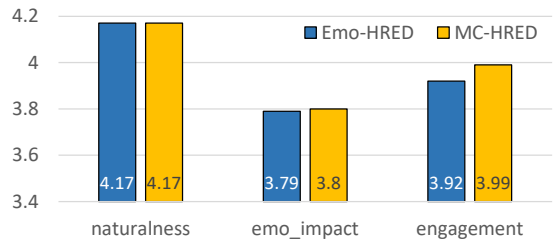


Figure 4: Human subjective rvaluation result.

We observe slight improvement on MC-HRED in the emotional impact and a more notable one in the engagement metric. On average, the responses generated by MC-HRED are 2.53 words longer compared to that of Emo-HRED. From the ratings, we also found that engagement is moderately correlated with response length, with an average Pearson r of 0.41. This signals that MC-HRED is able to produce longer sentences which results in higher engagement, while still maintaining naturalness and emotional impact. Dialogue samples comparing the systems responses are included in Table 2.

Table 2: Comparison of system responses for two triples in test set.

U_1	oh how do you feel about that one.
U_2	yes i heard the story.
U_3 (Target)	you heard it before.
Emo-HRED	right.
MC-HRED	it's a big thing.
U_1	are you a student here?
U_2	uh yes, actually I just got, er that's my lab over there in social computing yes (laughter).
U_3 (Target)	oh really. so you've been watching us going by.
Emo-HRED	oh okay.
MC-HRED	(laughter) it's nice to meet you.

7 Conclusion

We construct a corpus containing recordings of a counselor and 30 participants following a negative emotional exposure to learn expert responses in a positive emotion elicitation scenario. We unsupervisedly cluster the expert's responses and use the resulting labels to train a dialogue system. We proposed a novel hierarchical neural architecture for response generation that is conditioned on 1) expert's action, 2) dialogue context, and 3) user emotion, encoded from user input.

The objective evaluation we conducted show that the proposed model yields lower perplexity on a held-out test set. Subsequent human subjective evaluation shows that MC-HRED is able to produce longer sentences which improve engagement while still maintaining response naturalness and emotional impact. In the future, we would like to consider emotional impact explicitly for the emotion elicitation in lieu of a data-driven approach of positive emotion elicitation. We would also like to consider other modalities such as speech, for a richer emotion encoding. We acknowledge that evaluation through real user interaction needs to be carried in the future to test the system in a more realistic scenario.

Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

References

Jaime C Acosta and Nigel G Ward. 2011. Achieving rapport with turn-by-turn, user-responsive emotional coloring. *Speech Communication*, 53(9-10):1137–1148.

David Ameixa, Luisa Coheur, Pedro Fialho, and Paulo Quaresma. 2014. Luke, i am your father: dealing with out-of-domain requests by using movies subtitles. In *International Conference on Intelligent Virtual Agents*, pages 13–21. Springer.

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

Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder. 2000. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

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

Kate Forbes-Riley and Diane Litman. 2012. Adapting to multiple affective states in spoken dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226. Association for Computational Linguistics.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Citeseer.

James J Gross and Robert W Levenson. 1995. Emotion elicitation using films. *Cognition & emotion*, 9(1):87–108.

Sangdo Han, Yonghee Kim, and Gary Geunbae Lee. 2015. Micro-counseling dialog system based on semantic content. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 63–72. Springer.

Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. 2013. Predicting and eliciting addressee's emotion in online dialogue. In *Proceedings of Association for Computational Linguistics (1)*, pages 964–972.

Ryuichiro Higashinaka, Kohji Dohsaka, and Hideki Isozaki. 2008. Effects of self-disclosure and empathy in human-computer dialogue. In *Proceedings of Spoken Language Technology Workshop*, pages 109–112. IEEE.

Nio Lasguido, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2014. Utilizing human-to-human conversation examples for a multi domain chat-oriented dialog system. *Transactions on Information and Systems*, 97(6):1497–1505.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2017. Eliciting positive emotional impact in dialogue response selection. In *Proceedings of International Workshop on Spoken Dialogue Systems Technology*.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *Proceedings of The Thirty-Second AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.

Olivier Luminet IV, Patrick Bouts, Frédérique Delie, Antony SR Manstead, and Bernard Rimé. 2000. Social sharing of emotion following exposure to a negatively valenced situation. *Cognition & Emotion*, 14(5):661–688.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Lasguido Nio, Sakriani Sakti, Graham Neubig, Koichiro Yoshino, and Satoshi Nakamura. 2016. Neural network approaches to dialog response retrieval and generation. *IEICE Transactions on Information and Systems.*

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

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

Bernard Rime, Batja Mesquita, Stefano Boca, and Pierre Philippot. 1991. Beyond the emotional event: Six studies on the social sharing of emotion. *Cognition & Emotion*, 5(5-6):435–465.

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

Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, 24(7):1153–1172.

Alexandre Schaefer and Pierre Philippot. 2005. Selective effects of emotion on the phenomenal characteristics of autobiographical memories. *Memory*, 13(2):148–160.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

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

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Myrthe Tielman, Mark Neerincx, John-Jules Meyer, and Rosemarijn Looije. 2014. Adaptive emotional expression in robot-child interaction. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 407–414. ACM.

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

Rainer Westermann, Gunter Stahl, and F Hesse. 1996. Relative effectiveness and validity of mood induction procedures: analysis. *European Journal of social psychology*, 26:557–580.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.