# Dialogue Model and Response Generation for Emotion Improvement Elicitation

**Nurul Lubis**[†,‡‡], **Sakriani Sakti**[†,‡], **Koichiro Yoshino**[†,‡,††], **Satoshi Nakamura**[†,‡]
[†]*Division of Information Science, Nara Institute of Science and Technology*, Japan
[‡‡]*Institute for Informatics, Heinrich Heine University, Germany*
[‡]*Advanced Intelligence Project AIP, RIKEN*, Japan
[††]*PRESTO, Japan Science and Technology Agency*, Japan
{nurul.lubis.na4, ssakti, koichiro, s-nakamura}@is.naist.jp

## Abstract

As the technology develops, the potential of agents to improve the emotional-well being of users has been growing as well. Emotional support through human-computer interactions (HCI) has the potential advantage of being ubiquitous, round-the-clock, and accessible. However, existing works which address user emotion are still performed in a restrictive manner and bound to an oversimplified view of the emotional processes. In this paper, we build upon the existing works on positive emotion elicitation and extend them towards emotion processing through an entirety of a dialogue. The contribution of this paper is three-fold: 1) We construct a corpus of spontaneous human conversation carefully designed to highlight emotion improvement elicitation common in day-to-day situations. 2) We analyze the aforementioned corpus to find a shared dialogue structure representing the cognitive process underlying emotional changes and how it takes place in dialogue. 3) We apply this model in a dialogue system framework using a hybrid n-gram and neural network approach.

## 1 Introduction

Emotional support through HCI has the potential advantage of being ubiquitous, round-the-clock, and accessible. There has been evidence that people feel more comfortable self-disclosing and being honest about their emotion to computer agents rather than another human being [1]. Researchers have shown how emotionally intelligent systems could improve emotional well-being of users through various affective tasks, such as caring for the elderly [2], emotional distress assessments [3], or providing emotional support in general.

Despite this rich potential, existing works which address user emotion are still performed in a restricted manner and focused only on specific aspects to manage the complexity of affective interactions. For example, by limiting user into a number of options as dialogue input instead of spontaneous speech [4], focusing only on proactivity of agents [5] or cyber-bullying cases [6]. More recently, there is an increase of interest in positive emotion elicitation through chat-based dialogue systems [7]. Such a system allows domain-free conversations through which it attempts to elicit emotion improvement in the user. This mimics social sharing of emotion, an important facilitator of negative emotion processing [8]. However, state-of-the-art approaches still rely on turn-based response generation [9], i.e. they fall short in facilitating long-term emotion processing through dialogue, such as ones which requires cognitive re-appraisal of emotion [10, 11].

In this paper, we build upon the existing works and extend them towards emotion processing through an entirety of a dialogue. The contribution of this paper is three-fold: 1) We construct a corpus of spontaneous human conversation carefully designed to highlight emotion improvement elicitation

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

common in day-to-day situations. 2) We analyze the aforementioned corpus to find a shared dialogue structure representing the cognitive process underlying emotional changes and how it takes place in dialogue. 3) We simulate this model in a dialogue system interaction using language modeling and response generation techniques. A hybrid n-gram and neural network based approach is employed, which works subsequently on two levels: turn-level and word-level.

## 2 Corpus Construction: Positive Emotion Elicitation by an Expert

Even though various affective conversational scenarios have been considered [12, 13], 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 now, we will refer to this corpus as the counseling corpus.

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 [14, 15]. 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 [16]. It also offers easy replication in constrained environmental settings, such as the recording room. In contrast to previous works such as [17], we look for clips that depict real life situations and issues, i.e., non-fiction and non-films. This is to avoid the unpredictability of subjective emotional response to fictional clips. Furthermore, Non-fictional inducers better reflect real everyday situations. 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.

The data collection is as follows. The dyad 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 [18, 8]. This form of social sharing is argued to be integral in processing the emotional event [18]. 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." Each participant joins two sessions, one with anger inducer and one with sadness. There is at least one week interval between the two sessions. The combined duration of all sessions sums up to 23 hours and 41 minutes of material. On average, a session yields 23.6 minutes relevant data. 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.

We annotate the data with self-reported emotion label. Emotion is defined following the *circumplex model of affect* [19], which states that emotion can be described using two dimensions: *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). We choose to annotate the data with self-reported emotion to focus on the underlying felt emotion, not only parts that are expressed and observable. For each recording, the participants self report their emotional state using the FEELtrace system [20] immediately after the interaction. As they watch the recorded session, they move a cursor along a linear scale on an adjacent window to indicate their emotional

aspect (i.e., valence or arousal, annotated separately) at that point in time. 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 [20].

## 3 Identifying the Structure of Emotion Processing in Dialogue

We aim to identify the general, domain-independent dialogue structure of emotion processing in human communication. Such a structure will provide an essential framework or design in constructing dialogue systems capable of aiding user's emotion processes. The counseling corpus provides a solid basis for this study, as the recorded interactions have been carefully designed to highlight negative emotion processing toward emotion improvement.

The *appraisal theory of emotion* argues that most of our emotional experiences are the result of a cognitive process of evaluating situations and events [21, 22]. Burleson and Goldsmith posit that the appraisal theory can be used to explain how emotion comforting works [23]. Since emotion results from the appraisal of an event, and not the event itself, emotional reaction can be altered through re-appraisal of the event underlying the initial emotion.

In real life, social interactions play a big role in eliciting emotion improvements. A number of studies have reported that emotionally distressed people often feel an improvement as the outcome of socially sharing the event leading to the negative emotion [24, 25, 26]. More recently, an interaction experiment has empirically tested that verbal and non-verbal emotional support from helpers can facilitate the cognitive re-appraisal process of distressing emotions [27]. We aim to replicate this effect in human-computer interaction.

### 3.1 Metholodgy

We manually assess and analyze all sessions in the counseling corpus to find a shared dialogue structure across the sessions. The first step is to collect all the counselor's questions from the corpus as they potentially illustrate the kind of information needed by the counselor to proceed with the interaction and achieve the emotion improvement goal. These questions also allow us to observe the larger picture of the information exchange in the dialogue. When the questions are grouped per session, re-occurrence of question patterns can be observed, showing typical dialogue phases within the corpus. The analysis that follows is then guided by literatures on counseling skills and techniques and counselor's assessment of the conversation, elaborated below. Commonalities between those and the found pattern will ensure validity of the proposed dialogue structure.

#### 3.1.1 Counseling Skills and Techniques

We first study the skills and approaches essential to counselors in conducting an emotionally supportive dialogue. We believe this would give us an insights into the important points which require attention during the dialogue. We found a good amount of resources that explains this set of skills from handbook designed to train to-be counselors, written by experts.

Three of the main skills for counseling are active listening, clarification, and effective questioning:

- **Effective questioning**: Crucial in gathering information, as well as encouraging clarification and re-assessment of the distressing situation. The types of questions asked can be close-ended for quick factual answers; open-ended for gathering detailed information, opinion, and ideas; and probing to encourage continuation and a more in-depth exploration.
- **Active listening**: Essential to signal to the participant that they are listened to. This will encourage further dialogue and willingness to engage from the participant.
- **Clarification**: Disambiguate statements that are unclear. A well executed clarification could pose as an evidence of good understanding from the counselor.

These skills are then utilized by the counselor to expertly execute various counseling techniques, e.g.:

- **Reflecting Feelings**: Restating feeling to show understanding of their emotion.
- **Relating**: Relating to the participant's feelings to further demonstrate understanding.
- **Validating**: Affirm that their reaction in is common and normal given the situation.

- **Paraphrasing**: Restate succinctly what the participant has said as a way of confirming.
- **Encouraging/Positive Asset Search**: Focus on the participant's strengths and assets to help them see themselves or the situation in a positive light.
- **Interpretation**: Providing new meaning, reason, or explanation for behaviors, thoughts, or feelings, such that the participant can see their problems in a new way.

#### 3.1.2 Counselor Assessment

At the end of every recording session, we asked the counselor to provide a verbal summary about the session. The summary includes how participant reacts to the emotion inducer, counselor's course of action and the motivation for it, as well as assessment of its effectiveness. Below is an example summary taken from one of the sessions in the counseling corpus.

> *"So she was fairly strongly upset by that video, I think because she relates to both the mothers and babies in the video. But also with that experience as being a mother then she understands the difficulties of the mother* (in the video)*, so she is not perhaps as angry as some people who don't relate to that so well. But also still finds it difficult to understand why she* (the mother in the video) *didn't have any hint or* (bad) *feeling in doing something like that, so she feels secure that she wouldn't do something like that rather than being afraid as well. So we moved on to positive things related to children and she started to talk about her own family, and so it helped to reinforce her idea of herself as a competent mother, doing well for her family and probably with happy children and so on. So I think that helped her to feel okay when she left. Although if she is reminded of the video, she may still have um you know bad feelings again because perhaps she didn't have a chance to really explore all of her feelings around it."*

From the above summary, we can conclude that the counselor understood the participant's emotional reaction, understood the reasoning behind the emotional reaction, understood the how the participant relate to the event, and tried to reinforce the participant's positive asset in relation to the event to elicit emotion improvement. This verbal summary is immensely helpful for observing the expert's strategy throughout the dialogue. It also allows matching between techniques mentioned in literatures, and those that are actually executed in the corpus.

### 3.2 Proposed Dialogue Structure

Analysis of the corpora revealed a common session flow as follows. A sessions starts with greetings and small talk. After the emotion inducer, the counselor assessed participant's feelings and opinion about the event shown in the video inducer. In some sessions, the typical coping strategy of the participant is discussed and followed accordingly. The latter part of the sessions are commonly used to discuss the event in a positive light, brainstorming about ideas for solutions, or discussion about other topics that may elicit an improved emotional state, usually related to participant's personal life.

This observation is refined and matched through comparison with the 1) counseling skills and technique, 2) counselor's assessment of the session, and 3) related work of dialogue model in affective dialogue system [6]. We propose the following dialogue phases and actions as the underlying structure of emotion improvement through dialogue.

#### 3.2.1 Opening phase

The opening phase serves as warm-up prior to addressing emotional topics, and to ensure that the participant is comfortable with proceeding with the dialogue.

**Small talk**: Small talks encompass various small topics, such as how the participant is doing, the weather, biographic information, and recent events within the current week.

#### 3.2.2 Understanding phase

The goal of this phase is to gather information to effectively resolve the distressing event in question. Four main aspects are especially important in determining the solution on the next phase.

**Emotion**: Assessment and understanding of the participant's feelings or reactions toward an emotional event or exposure (e.g. emotion inducer within the counseling corpus).

**Event**: Discussion about the emotional event. Typically, the counselor asks the participant to describe the event and offers comments regarding the event. This allows the assessments of participant's understanding and interpretation of the event.

**Experience**: Discussion about how the participant relates to the event. Whether they have experienced something similar before, or whether it has happened to someone or someplace they know. The participant's experience often very well explains their emotional reaction to the event and how they understood and interpret it.

**Strategy**: Discussion about participant's typical coping mechanism towards the event. For example, whether the participant prefers gathering more information and facing the problem directly, or whether they prefer distancing themselves from the problem. When disclosed, this highly influence the steps taken in the next phase.

#### 3.2.3 Resolution phase

Three main techniques are observed in the data and shown to be effective. These actions are aimed to alleviate participant's emotional discomfort, and directly intended to elicit emotion improvement.

**Brainstorming**: The counselor probes the participant to think about how the situation may be improved. The goal is to encourage the participant to come up with problem solving ideas, or actionable solutions regarding the event. Knowing that improvement is possible and the steps that can be taken to achieve it are important stimuli to elicit emotion improvement.

**Distancing**: The counselor tries to put distance between the participant and the event in question. Some of the ways this can be done is by emphasizing participant's current state highlighting some differences so as to disconnect it from the event. Distancing can also be achieved simply by talking about other topics that have a more positive sentiment, or topics that the participant is interested in.

**Positive asset search (PAS)**: The counselor tries to emphasize the positive assets of the participant, and how that asset will help them in overcoming the situation in question, right now and in the future. The information gathered in the understanding phase, on discussion about participant's experience is highly useful in reinforcing their positive assets.

#### 3.2.4 Closing phase

**Goodbye**: The counselor expresses appreciation to the participant for sharing their thoughts. The goodbye may also be accompanied by final positive thought to end the conversation at a more positive note.

### 3.3 Proposed Dialogue Flow

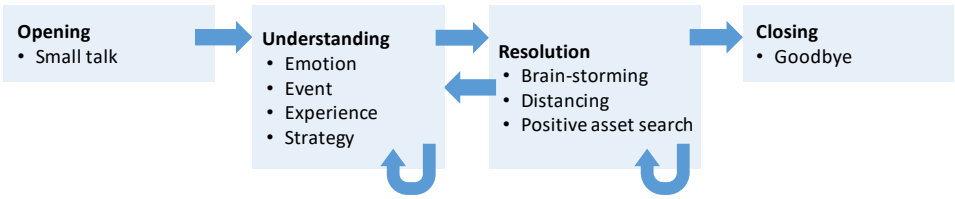


Figure 1: Flow between dialogue phases in the proposed structure.

In the recorded spontaneous interaction, the flow between and within the understanding phase and the resolution phase varies significantly. The information gathering during the understanding phase is not done in any specific order, neither are the actions during the resolution phase. Furthermore, the flow between these two phases is bidirectional. That is, more information may be gathered even though a resolution action is already being performed. Multiple resolution actions may be done within a session. Flow between the dialogue phases are illustrated in Figure 1.

### 3.4 Corpus Analysis

The counselor corpus is manually annotated by an expert who assigns a phase-action label on every dialogue turn. We count the occurrence of each label within the counselor corpus. Some dialogue turns are excluded in this analysis, in particular those that relate to the procedural part of the dialogue (e.g. "You can leave the headset on the chair.", "We will do another questionnaire at the end.").

The statistics revealed that the majority of the conversations are spent on the understanding and resolution phases (88.92% in total). This is expected given that the scenario is carefully designed to focus on negative emotion processing. The opening phase (8.13%) tend to happen over more dialogue turns than the closing (2.96%). It is also observed that the understanding phase have larger portion in the data than the resolution phase (48.42% and 40.51%, respectively).

Action composition within the understanding phase shows that discussion about the event (40.32%) and participant's experience (34.90%) related to it tend to dominate the phase. This shows that while assessment of the felt emotion is essential (occurred 15.69% of the understanding phase), the counselor as an expert put even more effort in understanding the reasoning behind the emotional reaction. In some sessions, coping strategy of the participant is discussed as well (9.09%). The statistics further show that the three resolution actions is equally likely to be employed in the collected dialogue (brainstorming 34.41%, distancing 32.36%, and positive asset search 33.23% of the time).

# 4 Dialogue System Framework

We demonstrate potential application of the proposed model in HCI by utilizing language modeling and response generation approaches. For the experiments, we partition the counseling corpus into 50 recording sessions (5053 triples) for training, 5 (503) for validation, and 5 (508) for testing.

## 4.1 N-gram Simulator

A simulator is one way to extend a limited amount of static data into a dynamic model that is capable of generating unlimited amount of dialogue. This approach is commonly used to explore the unbounded dialogue state space when training a dialogue policy [28, 29], since collecting data to cover all possible dialogue states is unfeasible. In this section, we build a model to simulate counselor's actions based on the counseling corpus, allowing us to extend the corpus into a model that can provide interactions in form of a counseling session in an inexhaustible manner.

As with the majority of works in dialogue simulation, we focus on semantic representation of the dialogue at turn-level. The n-gram model has been previously proposed for turn-level user simulation in [28]. N-gram models are suitable for this study since it can be trained easily given any dataset, being purely probabilistic and fully domain-independent. Furthermore, n-gram works effectively at turn-level representations, as this smaller state space yields high coverage even on a small corpus.

First, we define the phase and action labels as the turn-level representation of the counselor's dialogue. This amounts to 10 possible actions on the counselor side; 9 of which have been elaborated in Section 3.2, and an addition of a padding token for the beginning and ending. We use $a$ to denote these tokens. Second, we train an n-gram model using action sequences extracted from the counseling corpus. The counselor simulator outputs the probability of the next action $a_{t+1}$ given the sequence of actions up to that time $(a_1, ..., a_t)$, which is approximated by only considering the last $n-1$ tokens in the sequence $(a_{t-n+1}, ..., a_t)$. All possible next actions are assigned probabilities $P(a_{t+1}) = P(a_{t+1}|a_{t-n+1}, ..., a_t)$. Lastly, the simulator samples the next action based on this probability distribution.

### 4.1.1 Result and Analysis

We evaluate how well the simulator model counselor actions by computing the perplexity of their respective n-gram models. Three values of $n$ were tested: 1 (unigram), 2 (bigram), and 3 (trigram). Respectively, the model perplexities are: 7.86, 1.69, 1.71. There are important differences between this simulation task and language modeling that should be noted before we analyze the perplexities of the simulators. First is the vocabulary size and sequence length. With language, vocabulary size are much larger than typical sequence length. On the contrary, in a dialogue session, the number of possible tokens are very small (in this case 10 action tokens) and the sequence length are much longer (in this case, an average session length is 93 turns).

Furthermore, unlike language, repetition of a token is very natural in a dialogue. For example, constant transition between dialogue phases is unusual in the collected data, and each phase typically lasts multiple dialogue turns. With no context in unigram model, the perplexity of the model is quite high relative to vocabulary size of each model. On the other hand, even only with an additional context of 1, the bigram model can predict the data much better as the tokens are highly repetitive.

### 4.2 Dialogue Response Generation

With recent advancements in neural network research, end-to-end approaches have been reported to show promising results for chat-based dialogue systems [30, 31, 32]. Towards positive emotion elicitation, Lubis et al. have recently proposed a model that encodes emotion information from user input and dialogue history, and utilizing it in generating a response [7] . It has been further improved by considering unsupervisedly found dialogue acts in addition to emotional context [9]. In both works, the dialogue is modelled utilizing a hierarchical recurrent encoder-decoder (HRED) [31] with additional context encoders. Although they show improvements on perceived emotional impact, these approaches have not taken into account the dialogue structure of emotion processing and thus fall short in facilitating long-term positive emotion elicitation.

In this section, we extend the idea of a multi-context HRED (MC-HRED) [9] towards long-term emotion processing by combining it with the counselor n-gram simulator, creating a hybrid MC-HRED. The hybrid MC-HRED conditions the response generation process on the phase-action label, emotional context, and dialogue history. First, 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 *emotion* and *action encoder* takes $h_{dlg}$ and predicts the emotion and action context at dialogue-turn level, i.e. $h_{emo}$ and $h_{utt}$, and maintaining these contexts throughout the dialogue. Lastly, the *utterance decoder* takes $h_{dlg}$, $h_{emo}$ and $h_{utt}$ to predict the probability distribution over the tokens in the next utterance. Figure 2 shows schematic view of the system.

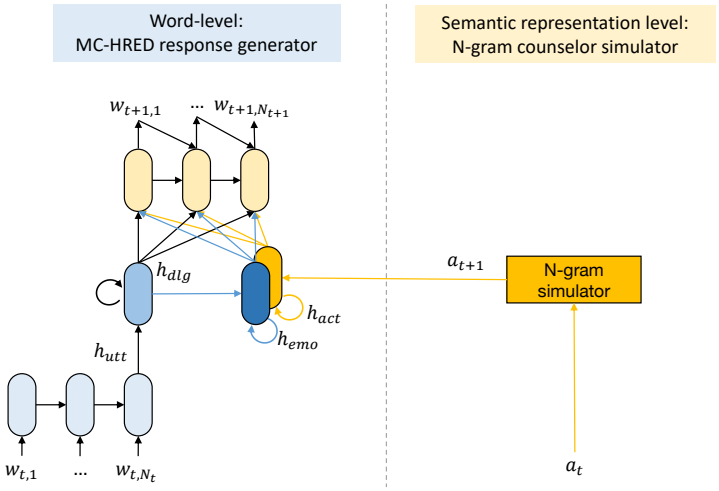


Figure 2: Schematic of a hybrid MC-HRED, combining MC-HRED and n-gram simulator with $n = 2$. When $n = 3$, context $a_{t-1}, a_t$ is used, and when $n = 1$ is used, no action context is passed.

As action context in hybrid MC-HRED, we use the proposed counselor's phases and actions, manually annotated and then encoded in a one-hot vector. This differs to that of [9], which uses automatic cluster labels as action context. As emotion context, we process the self-report emotion annotation as follows. We first obtain the average valence and arousal values of an utterance. We then discretize each of these values into three classes: positive, neutral, and negative with intervals $[-1, -0.07]$ for negative, $(-0.07, 0.07)$ for neutral, and $[0.07, 1]$ for positive. All the combinations of valence and arousal classes are then encoded into a one-hot vector of length 9, i.e. positive-negative, negative-neutral, 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 [9].

To train the response generator, we follow the training procedure of MC-HRED using dialogue triples as described in [9]. First, we *pre-train* an HRED model to obtain the starting model using a large scale conversational data [33]. Second, we *selectively fine-tune* the model using dialogue triples extracted from the counselor corpus, i.e. optimizing only the emotion encoder, action encoder, and the decoder. The loss function is a linear interpolation of 1) negative log likelihood of target response, 2) emotion prediction error by the emotion encoder, and 3) action prediction error by the action encoder. Training with the emotion prediction objective allows the emotion encoder to predict the emotion from dialogue context and pass it to the decoder seamlessly. The action context allows the model to

learn the relationship between the phase-action label and its corresponding responses. During testing, given a dialogue context, we utilize the phase-actions generated by the n-gram simulator, and thus the dialogue can take a different counseling route than that provided in data.

#### 4.2.1 Result and Analysis

We compare the proposed hybrid MC-HRED with Emo-HRED [9], a recent response generator model with awareness of emotional context in dialogue. Comparison with Emo-HRED as the baseline allows us to focus on the effect of the phase-action labels on response generation. Both models are trained in a similar fashion, differing only on the context used for response generation.

We first measure the model perplexity, reported in Table 1 along with the difference in the contexts used. The proposed models yield slightly higher perplexities compared to the baselines, and no significant difference between the simulators used.

Table 1: Results from objective and subjective evaluations.

| History | Emotion | Action | Model | Perplexity | Naturalness | Emotional impact |
|---|---|---|---|---|---|---|
| o | o | x | Emo-HRED | 42.60 | 3.56 | 3.26 |
| o | o | unigram | Hybrid MC-HRED | 49.74 | - | - |
| o | o | bigram | | 49.62 | - | - |
| o | o | trigram | | 49.78 | 3.77 | 3.51 |

Next, we conduct a subjective evaluation through crowdsourcing. Two models are evaluated: Emo-HRED as baseline and the proposed hybrid MC-HRED with the longest simulator context, i.e. trigram. With random order, we present human judges with 100 dialogue snippets from the test set, each followed by the system generated response. For each response, we ask the workers to state their agreement regarding naturalness and emotional impact of the response using a Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). Each snippet consists of 4 turns to give raters more dialogue context in evaluating the response. Two last columns of in Table 1 summarizes the subjective evaluation result, showing that the proposed model substantially improves perceived naturalness and emotional impact.

We present the examples of generation results by the model in the appendices. The examples show that the proposed hybrid MC-HRED is able to elicit different phases of emotion processing in dialogue, such as moving from discussion about the event (understanding phase) in the dialogue context, to brainstorming (resolution phase) in the generated response. Such responses at times differ from the target response (and thus yielding higher perplexity), however it could be more beneficial for emotion processing through dialogue in the long-term. Furthermore, we observe that responses generated by the hybrid MC-HRED reflects the phase-action label it was conditioned on.

## 5 Conclusion

In this paper, we investigate the process underlying emotional changes and how it takes place in a dialogue. We build upon the existing works on positive emotion elicitation and extend them towards emotion processing through an entirety of a dialogue. The hybrid MC-HRED system demonstrates one way the turn-level and word-level positive emotion elicitation approaches can be combined into a full-fledged dialogue system. Future efforts should be aimed at improving both the response generator and the counselor simulator, such as through learning a dialogue policy or considering richer dialogue context, as well as devising a more sophisticated combination scheme. Although the evaluation reported in this paper already uses longer dialogue context than previous works, we acknowledge that real user interaction needs to be carried to properly evaluate the system capability of facilitating long-term emotion processing. Lastly, we believe user study with Wizard-of-Oz set up is necessary to confirm whether the proposed dialogue model is suitable in HCI as is, or whether modifications are required to take into account the possibly existing differences between human communication and HCI for emotion improvement elicitation.

## Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237, as well as funding provided by the Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award endowed by the Federal Ministry of Education and Research.

## References


[1] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014.

[2] Juliana Miehle, Ilker Bagci, Wolfgang Minker, and Stefan Ultes. A social companion and conversational partner for the elderly. In *Advanced Social Interaction with Agents*, pages 103–109. Springer, 2019.

[3] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[4] Timothy Bickmore and Daniel Schulman. Practical approaches to comforting users with relational agents. In *CHI'07 extended abstracts on Human factors in computing systems*, pages 2291–2296. ACM, 2007.

[5] Lazlo Ring, Barbara Barry, Kathleen Totzke, and Timothy Bickmore. Addressing loneliness and isolation in older adults: Proactive affective agents provide better support. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 61–66. IEEE, 2013.

[6] Janneke M van der Zwaan, Virginia Dignum, and Catholijn M Jonker. A conversation model enabling intelligent agents to give emotional support. In *Modern Advances in Intelligent Systems and Tools*, pages 47–52. Springer, 2012.

[7] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 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, 2018.

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

[9] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. Unsupervised counselor dialogue clustering for positive emotion elicitation in neural dialogue system. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 161–170. Association for Computational Linguistics, Melbourne, Australia, July 2018.

[10] Nico H Frijda. Moods, emotion episodes, and emotions. 1993.

[11] Gerald L Clore. Why emotions require cognition. *The nature of emotion: Fundamental questions*, pages 181–191, 1994.

[12] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 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, 2012.

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

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

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

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

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

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

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

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

[21] Phoebe C Ellsworth and Klaus R Scherer. Appraisal processes in emotion. *Handbook of affective sciences*, 572:V595, 2003.

[22] Klaus R Scherer, Angela Schorr, and Tom Johnstone. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.

[23] Brant R Burleson and Daena J Goldsmith. How the comforting process works: Alleviating emotional distress through conversationally induced reappraisals. In *Handbook of communication and emotion*, pages 245–280. Elsevier, 1996.

[24] James W Pennebaker, Emmanuelle Zech, Bernard Rimé, et al. Disclosing and sharing emotion: Psychological, social, and health consequences. *Handbook of bereavement research: Consequences, coping, and care*, pages 517–543, 2001.

[25] Scott E Caplan, Beth J Haslett, and Brant R Burleson. Telling it like it is: The adaptive function of narratives in coping with loss in later life. *Health Communication*, 17(3):233–251, 2005.

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

[27] Susanne M Jones and John G Wirtz. How does the comforting process work? an empirical test of an appraisal-based model of comforting. *Human Communication Research*, 32(3):217–243, 2006.

[28] Kallirroi Georgila, James Henderson, and Oliver Lemon. User simulation for spoken dialogue systems: Learning and evaluation. In *Ninth International Conference on Spoken Language Processing*, 2006.

[29] Florian Kreyssig, Inigo Casanueva, Pawel Budzianowski, and Milica Gasic. Neural user simulation for corpus-based policy optimisation for spoken dialogue systems. *arXiv preprint arXiv:1805.06966*, 2018.

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

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

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

[33] David Ameixa, Luísa Coheur, and Rua Alves Redol. From subtitles to human interactions: introducing the subtle corpus. Technical report, Tech. rep., INESC-ID (November 2014), 2013.

# Appendices

## A Dialog Generation Result

We present the examples of generation results from the models. Below is an example of a generated dialogue by the trigram counselor simulator:

- Opening for 2 dialogue turns,
- understanding: Emotion for 10 dialogue turns,
- understanding: Experience for 2 dialogue turns,
- understanding: Event for 15 dialogue turns,
- resolution: Positive asset search for 16 dialogue turns,
- understanding: Event for 13 dialogue turns,
- understanding: Experience for 7 dialogue turns,
- resolution: Distancing for 8 dialogue turns,
- resolution: Positive asset search for 5 dialogue turns, and
- closing for 4 dialogue turns.

## B Dialog Response Examples

Examples comparing system responses in a long dialogue context are presented in Table 2. The examples show that the proposed hybrid MC-HRED is able to elicit different phases of emotion processing in dialogue, such as moving from discussion about the event (understanding phase) in the dialogue context, to brainstorming (resolution phase) in the generated response. Such responses at times differ from the target response (and thus yielding higher perplexity), however it could be more beneficial for emotion processing through dialogue in the long-term. Meanwhile, responses by Emo-HRED simply follows the dialogue topic set by the context. Table 3 shows more example responses generated by the hybrid MC-HRED that are well characterized by the phase-action label it was conditioned on.

Table 2: Comparison of system responses by Emo-HRED and the proposed hybrid MC-HRED in a long dialogue context from the test set.

| | |
|---|---|
| Counselor | (laugh) yes . but you might be right I mean maybe we can't tell whether he really had to work that long. |
| Participant | yes. |
| Counselor | or whether it's his impression that he has to work that long. |
| Participant | mm that's true. |
| Counselor | but somehow I guess I'm sure the society probably has what the company has created that impression. |
| Participant | yes yes. |
| Emo-HRED | it's quite interesting to understand that. |
| Proposed hybrid MC-HRED | yes. Do you think there's something that you can do about it? |

Table 3: Example responses generated by hybrid MC-HRED with various phase-action labels.

| **Phase** | **Action** | **Generated Response** |
|---|---|---|
| Intro | Small talk | oh thank you. so are you having a good week? |
| Under-standing | Emotion | it's kind of sad. |
| | Event | yes it's quite a heavy video. |
| | Experience | so you feel like it's so important to you. |
| Resolu-tion | Brain-storming | do you think that's something that you can do about it? |
| | Distancing | so you just came here two weeks ago? |
| | PAS | it's definitely a way to understand. |
| Closing | Goodbye | anyway thank you for telling me about your opinions. |