Doctoral Dissertation

Eliciting Emotion Improvements with Chat-based Dialogue Systems

Nurul Fithria Lubis

March 15, 2019

Division of Information Science
Graduate School of Science and Technology
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Division of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Nurul Fithria Lubis

Thesis Committee:

Professor Satoshi Nakamura	(Supervisor)
Professor Yuji Matsumoto	(Co-supervisor)
Professor Wolfgang Minker	(Ulm University)
Associate Professor Sakriani Sakti	(Co-supervisor)
Assistant Professor Koichiro Yoshino	(Co-supervisor)

Acknowledgements

I thank the Japanese Ministry of Education, Culture, Sports, Science and Technology for providing support throughout this study through MEXT scholarship.

I'm profoundly grateful to be a part of the Augmented Human Communication (AHC) Laboratory. I would like to thank Professor Satoshi Nakamura for providing me with the opportunity to conduct this research under his supervision and guidance. I would also like to thank my direct supervisor, Associate Professor Sakriani Sakti, for the continuous support on my research since my very first topic as an internship student in 2013, until the fruition of this doctoral dissertation in 2019. I have learned a lot from our years of working together. I thank Assistant Professor Koichiro Yoshino for the discussion and constructive feedback on my progress reports and written works, and the members of the thesis committee, Professor Yuji Matsumoto and Professor Wolfgang Minker, for the thoughtful review and comments on this thesis. My sincere thanks go to Ms. Manami Matsuda for her bright presence and steadfast support in AHC Lab.

I would like to thank every one of the teachers I've had throughout the many years I've spent in school and university. I'm thankful for the great old pals from the good old days who have supported me during my study in their own ways, and the new friends I've made in Japan and conferences around the globe.

This acknowledgment would not be complete without my family – a mix of strikingly different personalities, yet unanimous when it comes to support. Thank you for the love and encouragement, unbroken even when we are thousands of kilometers apart.

Lastly to Michael, thank you for always being there, for still severely overfitting, and even more happily so.

Eliciting Emotion Improvements with Chat-based Dialogue Systems*

Nurul Fithria Lubis

Abstract

Social interactions can support the treatment of emotion-related problems by aiding a person's emotional process. A number of studies have showed a consistent inclination of humans to talk about and socially share their emotional experiences, especially for an intense and/or negative emotion exposure [70]. Although we have seen encouraging progress in affective human-computer interaction, the potential benefits for users by incorporating emotion in computer interaction are not yet studied in depth. For example, emotion elicitation looks at the change of emotion in dialogue, however its application for emotional improvements is not yet well researched. Furthermore, although there exist technologies that address clinical emotional disturbances, such as depression [23] and distress [24], there is a lack of research on emotion improvement from negative emotional exposures commonly encountered in everyday life.

The goal of this thesis is to diminish these gaps. In particular, I aim for chat-based dialogue systems with an implicit goal of eliciting emotion improvements though dialogue interactions. The concept of eliciting emotional improvements can be examined through two perspectives: *short-term* and *long-term*. Assuming positive emotional state as the goal, short-term elicitation of emotional improvement is reformulated into turn-based positive emotion elicitation. In turn, long-term emotional improvement extends elicitation scope to the entire dialogue. This thesis will be focusing on the short-term improvement elicitation task, exploring dialogue aspects contributing to a successful elicitation and modeling them in a dialogue system. In addition, potential approaches of extension into long-term emotion elicitation are investigated.

*Doctoral Dissertation, Division of Information Science, Graduate School of Science and Technology, Nara Institute of Science and Technology, March 15, 2019.

First, I study emotion processing and negative emotion recovery in human communication through corpus construction and analysis. Second, to endow dialogue systems with an attunement of emotional context in dialogue, I propose novel neural network architectures that allow a dialogue system to track emotion and incorporate this information in the response generation process. Third, novel methods to learn dialogue strategies for short-term positive emotion elicitation in chat-based dialogue systems are proposed. Aspects that contribute to elicitation success are inspected: emotion, dialogue action, and emotional impact. Lastly, I present the result of preliminary study on long-term emotion improvement elicitation. The dialogue structure allowing long-term emotion improvement is identified, and simulations using language modeling techniques are examined. An effort to combine response generation techniques and the counselor simulator are presented as well. The efforts presented in this thesis should serve as the basis for future efforts in supporting emotion improvement through human-computer interactions.

Keywords:

dialogue systems, human computer interactions, neural networks, dialogue response generation, emotion, affective computing

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"I've learned that people will forget what you said, people will forget what you did, but people will never forget how you made them feel."
– Maya Angelou (1928-2014), *Poet*

1.1 Social-Affective Human Communication

Human communication is often remarked as one of the most important key aspects in the advancement of the human race. Tomasello argued that human communication originally evolved from the basic needs of helping and sharing – to request help, exchange information, and social bonding within a group [107]. Even though its function to communicate our wants and needs (e.g., hunger, thirst) are crucial, it has been shown that humans communicate for social reasons the majority of their lifespan [61]. The term social-affective communication refers to interactions between two or more people (social) that involve emotion (affective). Emotion strongly governs the way humans socially communicate with each other and is key to a thriving social connection between people. Moreover, it also works inversely: for a healthy emotional well-being, human communication is crucial.

To understand social-affective human communication better, I believe it is necessary to review emotion and its underlying processes in further detail. The *appraisal theory of emotion* argues that most of our emotional experiences are the result of a cognitive process, unconscious or controlled, of evaluating situations

and events [28, 95]. Among the contributing factors, the social world is argued to be one of the important aspects that influences our appraisal processes [72]. Emotion of others can pose as clue as to how we should appraise a situation, or as an additional stimulus in how we appraise a situation [80].

As put forward by Scherer [96], emotional competence can be broken down into three lower level competences that interact and depend on one another: appraisal, regulation, and communication competences.

1. **Appraisal competence** refers to the person's ability to accurately evaluate a situation. There are two sides of appraisal competence: 1) emotion differentiation, which is the ability to tell various kinds of emotion apart, and 2) internal emotion elicitation, which is the ability to appraise the appropriate emotional response, or the absence thereof, in a given situation.

2. **Regulation competence** refers to a person's ability to appropriately modify their raw emotion in an effective manner. This modification can be influenced by a number of complex factors, such as strategic intention, societal rules, or re-appraisal of the situation.

3. **Communication competence** refers to a person's ability to encode and decode emotion into and from communication clues. This competence dictates the ability of someone to convey their feelings into others so as to be understood, as well as understanding others' emotional states.

In the social sphere, these competences govern two main processes: emotion perception and production.

1. **Emotion perception** refers to the process of recognizing emotion and understanding its implication. Two competences that play important roles in this process are 1) *communication competence* to decode an emotional state based on social clues, and 2) *appraisal competence* to relate it to their environment and situation, and appraise the resulting emotion accordingly.

2. **Emotion production** refers to the adaptive function of emotion that is essential in coping with events related to a person's well being [96]. Two main competences for this action are 1) *regulation competence* to efficiently modify the raw emotion according to re-appraisal, social rules, or strategic

intentions, and 2) *communication competence* to actualize the processed emotion and project it to the environment.

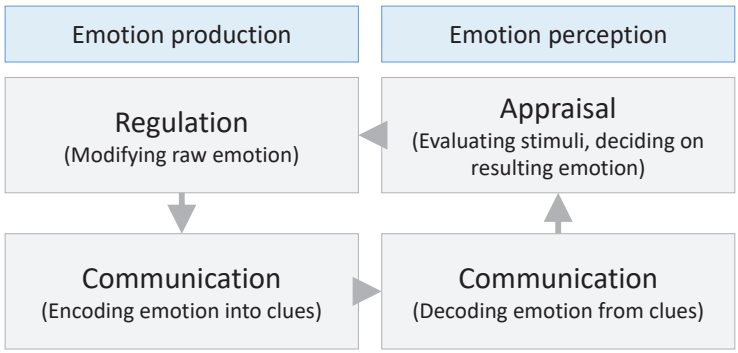


Figure 1.1: Emotional competences in emotion processes

Figure 1.1 illustrates these competences and processes and their underlying loop during social-affective communication. Conscious or not, we are constantly required to utilize these competences in any social interactions.

1.2 Human-Computer Interaction

It is fair to posit that dialogue system is the most natural form of human-computer interaction. If machine were to persuade humans that it possesses intelligence, language would have to be its media of choice to make the case. Both in science and popular culture, this is how humans have dreamed of intelligent machines, from HAL 9000 in the movie "2001: A Space Odyssey" to Samantha in "her."

In this section I will present a brief review of the advancements of dialogue system technologies. I will also discuss how the state-of-the-art systems have addressed the different functions of human communication, and the role and contribution of emotion in human-computer interactions.

1.2.1 Task- and Chat-Oriented Dialogue Systems

Dialogue systems are originally developed to allow interactions with natural language between human and machine. The dialogue systems provide an interface for the users that is more natural and intuitive, i.e. instead of pushing a button

or move a slider, the user simply has to state their intention with words. The main goal is to allow computer users to access information more conveniently and perform tasks more efficiently.

Task-oriented Dialogue Systems

This motivation is the basis of the first family of dialogue systems, the so-called task-oriented dialogue systems. Interactions with such systems are typically constrained in terms of domain (e.g. restaurant, hotel) and task (e.g., search based on cuisine, make a booking). The clearly defined domain and task allow for a specifically designed dialogue flow and actions.

A common approach is to declare a set of variables, or slots (e.g. in restaurant domain: area, price range, type of cuisine), to be filled by the user sequentially according to a dialogue flow. Based on the values assigned to the slots, the system then inquire a database and present the user with the information they need (e.g. list of restaurants in the city center that serves expensive Japanese food). This particular task is called slot-filling, and the system's ability to deduce user's goal throughout the dialogue is coined as dialogue state tracking [119, 44].

Examples of goal-oriented dialogue systems include the DARPA communicator dialogue system for travel planning [56] and the Let's Go Public dialogue system that provides bus schedule information in the Pittsburgh area [87]. This family of dialogue systems mimics the function of human dialogue to communicate wants and needs.

Chat-oriented Dialogue Systems

On the other hand, the so-called chat-oriented dialogue systems focus on conversing with the user without any clearly predefined goal. In essence, the systems try to simulate natural conversation and mimic the social functions of human communication. One can not mention chat-oriented dialogue systems without mentioning ELIZA [115], one of the first chat bots to attempt the Turing Test [108]. Despite its convincing ability to hold a conversation with a user, in essence ELIZA simply relies on rules, pattern matching, and template-based text substitution without any framework to contextualize events and truly understand the content of the conversation. As such, it easily fails outside the scope of its script. Since then, researchers have identified and studied a variety of aspects that sup-

port a coherent, human-like interaction, such as personality [52], memory [4], non-verbal gestures [42]. Many of the aspects contributing to success mentioned above, such as personality and gestures, have very strong links to emotion appraisal and productions. Although they are initially intended simply for user entertainment [3], we have started to explore their potential application for other tasks such as language learning [104].

1.2.2 Affective Dialogue Systems

It is argued that humans also impose the emotional aspect of social communications in their interaction with computers and machines [88]. They treat them politely, laugh with them, and sometimes get angry or frustrated at them. To mimic human interaction and benefit from its emotion-related potential, e.g. to provide emotional support, many works and studies have attempted to equip computers with emotional capabilities to reciprocate with humans in this regard. In this section I will review the landscape of existing works in affective Human-Computer Interaction (HCI), with a particular focus on dialogue systems.

The field of affective computing aims to develop systems capable of recognizing, interpreting, processing, as well as simulating human affects [83]. In other words, research in this field is primarily dedicated to the incorporation of emotion into HCI. As in humans, it is believed that emotional competence in computers will enhance its quality of decision making and providing user assistance. This main goal of emotional reciprocation demands the integration of many capabilities from a range of research topics, such as speech recognition, natural language understanding, emotion recognition, speech synthesis, and computer graphics. Over the years, many advancements have been made toward achieving this goal.

As illustrated in Figure 1.2, two of the most studied affective issues in dialogue systems are:

- **Emotion recognition**, which allows a system to discern the user's emotions and address them in giving a response or performing their tasks [41, 106]. In relation to emotional competences, this is the equivalent of communication competence on emotion perception, where we decode verbal and non-verbal cues into the underlying emotional state of the speaker

Some of the icons used in this thesis is made by Freepik from www.flaticon.com

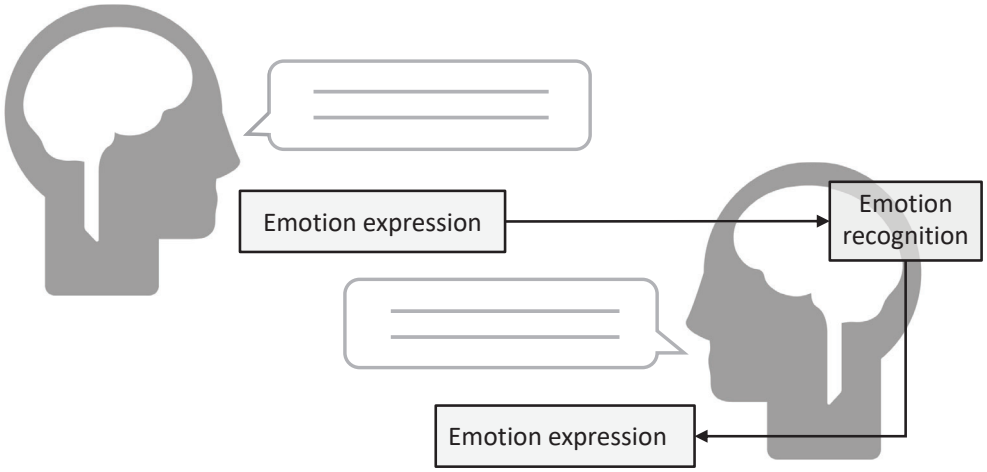


Figure 1.2: Traditional works on emotion competences for affective HCI.

[96] and use the information accordingly. A study showed that when a tutoring system takes information of user's emotional state into account, task success rate can be significantly increased [32].

- **Emotion expression**, which helps convey a message to the user through emotional nuance, such as that in [122]. This is also equivalent to the communication competence, where we encode our feelings and emotional reaction into clues such that the listener can understand what we are feeling. In a listening-oriented system, this has been shown to increase closeness and satisfaction [45].

Researchers have also attempted to endow computer agents with its own emotion model:

- **Emotion modeling** attempts to equip a system with its own emotion appraisal model. The main motivation is to allow decision making and behavioral signal that suggest an underlying emotional process. An embedded emotion model in a system, such as that proposed in [111, 110, 25], allows a system to adjust its behavior [38] and belief [73] to align with that of human's. Emotional triggers and responses that make up emotion appraisal has also been studied in [66, 67]

And lastly, there has also been an increase of interest in the problem of emotion elicitation:

- **Emotion elicitation**, or **emotional triggers**, concerns eliciting a certain emotion from the user using the system's response. A recent study by Hasegawa et al. addresses this issue by predicting and eliciting emotion in online conversation [43]. The model is reported to be able to elicit a number of emotion classes properly by utilizing Twitter data and statistical machine translation techniques.

Figure 1.3 illustrates the landscape of existing works in affective HCI.

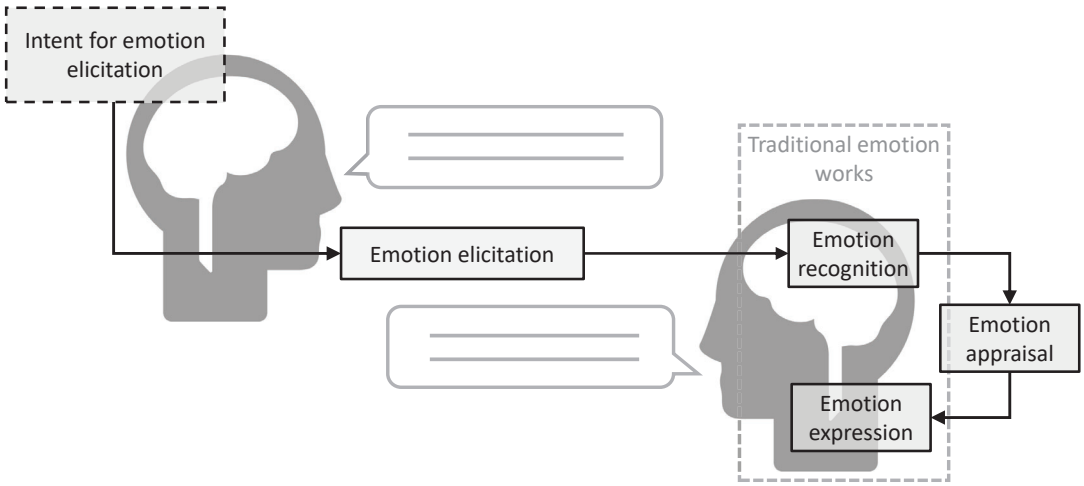


Figure 1.3: Existing works on affective dialogue systems.

At the core of the majority of the efforts in affective computing is the goal to help users meet their emotional needs through HCI, by taking into account their emotional nature [82]. This elevates existing HCI technologies, which are predominantly focused on task completion, to also consider social and emotional aspect of human interactions. Many critics question the possibility of achieving this goal, given our lack of complete understanding of how emotion works in the first place. Picard et al. [82] argued against this skepticism, noting how humans routinely meet some of these needs through non-humans, for example pets, which have presumably less examined understanding of human emotions than we do today.

Indeed, a number of computational models of appraisal, personality, and relations are revealed to be successful in conveying emotional competences to the user. Skowron et al. constructed a dialogue system with positive, negative, and

neutral affective profiles, showing consistent effect to the user compared to the respective profiles in humans [101]. Similarly, an evaluation of a conversational companion reported that the users felt that the companion had a personality; polite, friendly, and patient [5]. Even further, Bickmore et al. [7] reported affirming results in building and maintaining long-term human-computer relationships.

1.3 Limitations and Challenges

Although we have seen positive progress in affective HCI, there is still a large portion of social-affective human communication that is not yet studied. The links between some tasks to its immediate goal is quite clear: emotion recognition to detect user emotional states, expression to convey emotions to user, and emotion modeling for an emotionally natural reactions in agents. However, the potential benefits of incorporating emotion in computer interaction for users are not yet studied in depth. For example, although emotion elicitation looks at the change of emotion in dialogue, its application for emotional improvements is not yet well researched.

A number of studies have showed a consistent inclination of humans to talk about and socially share their emotional experiences, especially for an intense and/or negative emotion exposure [70]. This is argued to be an essential part of the emotional processes [89]. Social-affective interactions can provide social support, giving positive effect to the treatment of emotion-related problems by supporting a person's emotional process. [19, 120].

For average, emotionally healthy users, technologies such as listening oriented systems [76] and a companion conversational agent [16] may be useful and beneficial. However, existing works in this domain have not yet considered negative emotional experiences and the recovery from them. On the other end of the spectrum, there exists efforts in addressing clinical emotional disturbances. Some examples are a system for depression and suicide risk evaluation [23], and an simulated interviewer agent for distress clues assessments [24]. While they address important issues, these works are not applicable for the larger, general audiences as they are focusing on clinical circumstances.

To the best of my knowledge, existing studies have not yet examined negative emotional exposures commonly encountered in everyday life, such as those in situations that may elicit a negative emotional response. For example, reading

the news, or having a debate on social issues. Prompt recovery of such experience will prevent its accumulation into a more serious emotional problem, and an emotionally-competent computer agent could be a valuable assistive technology in addressing this need.

1.4 Thesis Objective and Contribution

I have identified the following gaps in the landscape of affective dialogue systems:

1. The lack of human-computer interaction works that focus on emotional benefits of affective systems for users.

2. The absence of dialogue systems that address negative emotions commonly encountered in everyday life.

The goal of this thesis is to minimize, or possibly, eliminate these gaps. In particular, I aim for **chat-based dialogue systems with an implicit goal of eliciting emotion improvements though dialogue interactions**, which combines the objectives of chat-oriented and goal-oriented dialogue systems. The interaction will take form in a domain-free chat-based manner, focused on achieving emotion improvement through simulating natural conversation and emulating social functions of human communication. However, the system itself has an internal goal, invisible to the user, to improve user emotional state through the dialogue.

Even though studies such as [78, 79] have shown that emotion plays a role in enhancing task-based interactions, I argue that the role of affect and especially emotion in HCI is more apparent in chat-based interactions as they mimic the social functions of human dialogue. User expectation in such a setting is more focused on the quality of conversation itself, and not influenced by the completion of a certain goal. This highlights the conversational ability of the system and allows us to observe user emotions in dialogue with more clarity. Furthermore, findings within this scope will be more easily generalizable to task-oriented and domain-specific dialogue than vice versa, as imposing constraints to a problem is generally easier than removing them.

It is important to note that eliciting emotion improvements does not translate to responding with positive emotion at all times. Figure 1.4 illustrates this difference. In various real-life dialogue scenario, relentless reinforcement of positive

emotion may be perceived as unnatural and can even lead to emotional state decline. The goal is to train the dialogue system to respond in a way that is likely to elicit a more positive emotion, which in some scenarios could mean showing negative emotions, such as relating to one's anger or showing empathy.

(a) Responding to elicit positive emotion.

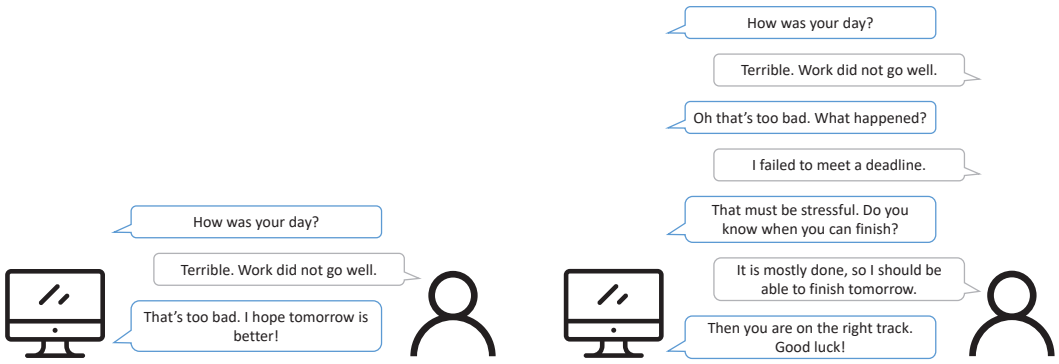
(b) Responding with positive emotion.

Figure 1.4: Emotion improvement elicitation does not mean responding with positive emotion. In some situations, reinforcing positive emotion to elicit similar emotion in the dialogue partner is ineffective, or even emotionally harmful.

The concept of eliciting emotion improvements can be examined in two perspectives: *short-term* and *long-term*. Assuming positive emotional state as the goal, short-term elicitation of emotional improvement is reformulated into turn-based positive emotion elicitation. On the other hand, long-term emotional improvement expands the positive emotion elicitation scope to the entire dialogue. Figure 1.5 illustrates the difference between short-term and long-term emotion improvement elicitation through dialogue.

This thesis will be focusing on the short-term improvement elicitation task, exploring dialogue aspects contributing to a successful elicitation and modeling them in a dialogue system. In addition, potential approaches of extension into long-term emotion elicitation are investigated, as well as the combination of both perspective into a dialogue system. I approach this systematically through the following tasks:

1. **Analysis of social-affective human communication.** *How do 1) positive emotion elicitation, and 2) improvement from negative emotion looks like in human communication?* I construct corpora containing carefully designed conversations to capture the phenomena of interest. In depth analysis of the data is conducted to gain valuable insight into these processes.



(a) Short-term. (b) Long-term.

Figure 1.5: Dialogue examples comparing short-term and long-term emotion improvement elicitation. In short-term elicitation, the system attempts to elicit improvement within a single dialogue turn response. In long-term elicitation, the system considers the entirety of the dialogue to improve user's emotion.

2. **Emotion-sensitive response generation.** *How can we consider emotion in generating a dialogue response?* Attunement of emotional context in dialogue is crucial to successfully elicit emotion. I propose a novel neural network architecture that allows a dialogue system to track emotion and incorporate this information in the response generation process.

3. **Positive emotion elicitation.** *How can we elicit emotional improvement through dialogue response generation?* I propose novel methods for positive emotion elicitation in chat-based dialogue systems. Aspects that contribute to elicitation success are inspected: emotion, dialogue action, and emotional impact.

4. **Eliciting emotion improvements through dialogue.** *Can we identify the structure of and simulate emotion improvements through dialogue?* I analyze one of the constructed corpora to identify the structure of dialogue to support emotional improvements. I then attempt to simulate this process by utilizing language modeling techniques.

From an engineering perspective, the aforementioned tasks brings as their main contributions the design, training method, and implementation of dialogue systems that better reflect real human interactions. First, solution to the proposed problems will allow consideration of emotion and its underlying processes

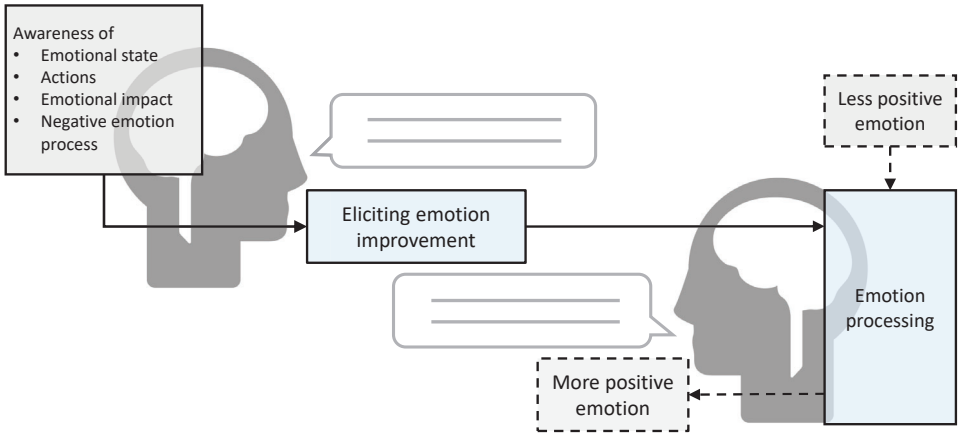


Figure 1.6: I propose to elicit emotion improvements through dialogue system interactions. In short-term elicitation, the emotion processing underlying it is treated as a black box and the system solely focus on eliciting positive emotion. In long-term view, we identify the processes behind emotion improvements.

for HCI that is closer to real human interactions. Furthermore, as with humans, dialogue systems capable of emotion improvement elicitations has the potential to be more natural and more successful in fulfilling its purpose. Lastly, emotion improvement elicitation opens a plethora of possibilities of dialogue system applications such as caring for the elderly, low-cost ubiquitous chat therapy, or providing emotional support in general.

Figure 1.7 shows the roadmap of my research towards emotionally intelligent affective dialogue systems. Several tasks in less complex scenarios (grey area) have been tackled in the past. The research is progressing towards more complex tasks and scenarios in affective HCI.

1.5 Thesis Overview

Figure 1.8 presents the overview of this thesis. The remainder of this thesis is arranged as follows. In Chapter 2, I review the definitions of dialogue and emotion and formalize their scope within this thesis. In Chapter 3, approaches for chat-based dialogue systems are discussed. We review two main families of approaches: response retrieval and response generation, their advantages and disadvantages are presented, especially in relation to the task I try to solve in this thesis.

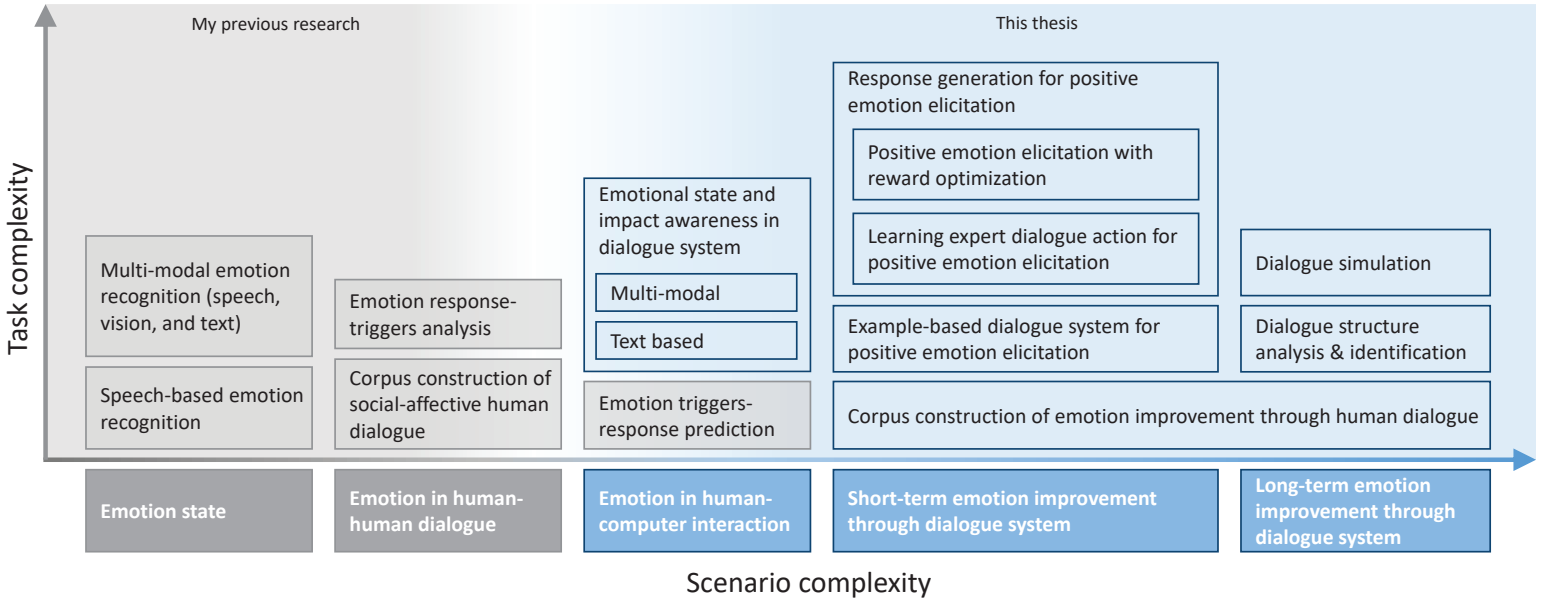


Figure 1.7: Research roadmap. Grey area shows topics from my previous research. Blue area shows tasks to be tackled in this thesis. The research is progressing towards more complex tasks and scenarios in affective HCI. Task complexity is relative to its corresponding scenario.

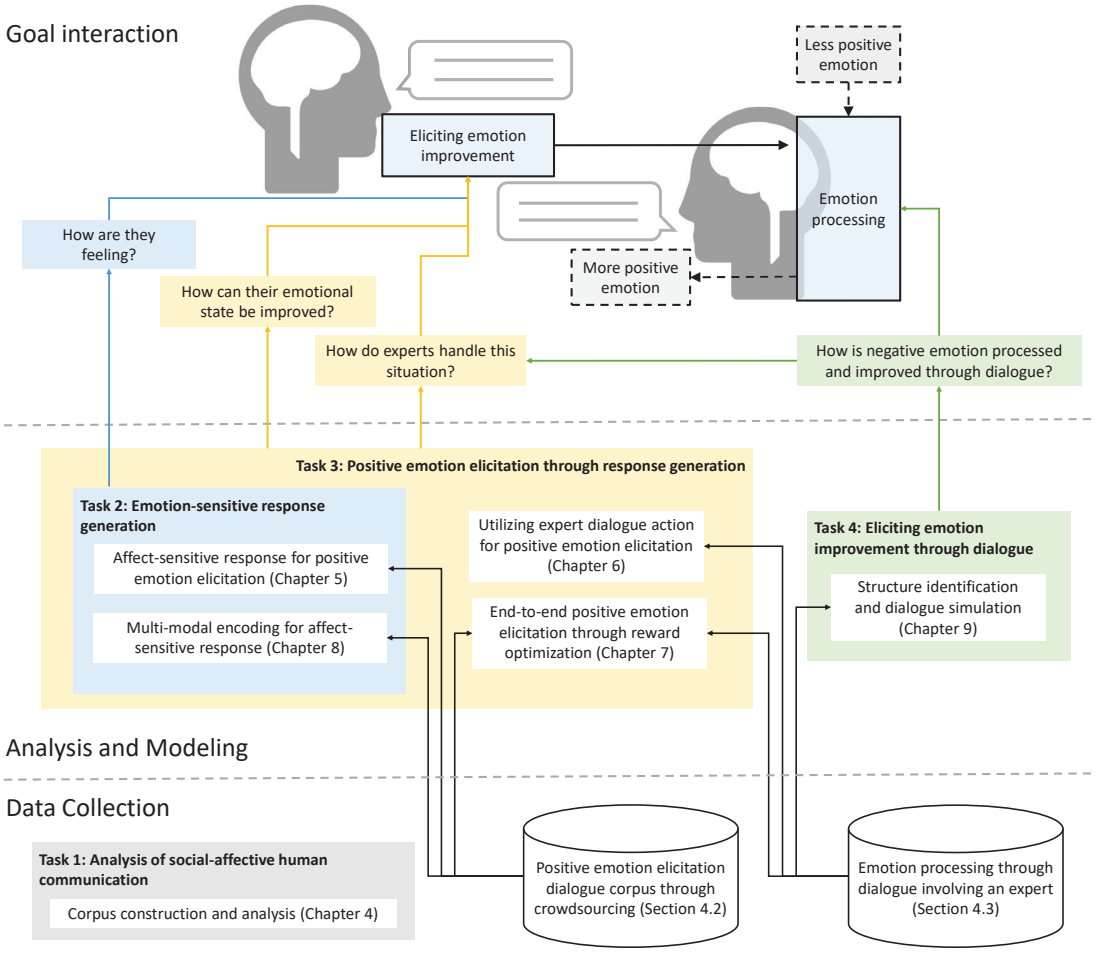


Figure 1.8: Thesis Overview.

In Chapter 4, I identify existing corpora that are potentially suitable for the positive emotion elicitation task. The limitations and missing links are identified, and the efforts to surmount this limitations via corpus construction are elaborated in detail (Task 1). The challenge of emotion awareness in dialogue (Task 2) is tackled in Chapters 5 and 8. In Chapter 5, the system attempts to infer the emotional context of the dialogue based on dialogue history in text form. To further improve system performance, this problem is revisited in Chapter 8 by incorporating acoustic information in inferring the emotional context. Speech has been argued to be the richest channel of communication, containing paralinguistic informations including emotion and affect. To benefit from this source of information, additional acoustic features is utilized in modeling the emotion

context within the dialogue.

The main problem of positive emotion elicitation is tackled in Chapters 5 through 8. Each chapter builds on its preceding, solving its limitations and shortcomings. In Chapter 5, I propose to train an emotion-sensitive dialogue system on responses that elicit emotion improvements, achieving an end-to-end emotion improvement elicitation in chat-based dialogue system. In Chapter 6, I incorporate higher level information from human expert's responses to train the affective dialogue systems to 1) allow the system to distinguish between action taken to elicit emotion improvements, and 2) promote diversity in the generated responses. Unsupervised clustering methods are employed to extract underlying categories of actions and behaviors from the expert's responses, allowing a fully automatic extraction of dialogue information. In Chapter 7, I propose to explicitly utilize emotional impact information to optimize neural dialogue system towards generating responses that elicit positive emotion. Leveraging this information allows us to promote responses that elicit positive emotion, and suppress those that has negative impact. This shifts the elicitation improvement approach to rely on awareness of emotional impact on the system side, and not solely by imitating the training data. As previously mentioned, Chapter 8 presents further efforts to improve the system performance by considering more source of information for emotion encoding, i.e. speech.

Chapter 9 reports preliminary study on long-term emotion improvements through human-human dialogue (Task 4). I investigate the cognitive process underlying emotional changes and how it takes place in a dialogue. The main goal is to identify how an active helper would be able to strategically support and catalyze this process through dialogue interactions. An initial attempt in combining this strategy with the response generation techniques in previous chapters is presented as well. Lastly, Chapter 10 concludes this thesis, with summary and discussion on future direction.

Chapter 2

Emotion and Dialogue

2.1 Computational Models of Emotion

A number of families of emotion theories have been proposed in literatures in psychology and neuroscience. Generally, these theories differ in terms of the aspects of emotion they include and highlight. An understanding of varying views of emotion will be invaluable in deciding for a model that is compatible to the problem that we are trying to solve. In this section I elaborate on two of main classes of emotion models.

2.1.1 Categorical Emotion

The first family of models of emotion is *categorical*. In these models, a finite number of emotion categories are defined. One of the most adapted set of categories was proposed by Ekman, including happiness, anger, sadness, disgust, fear, surprise, and neutral [26]. These emotion are argued to be the most basic emotion that are universal regardless of culture or other social influences. On the other hand, Plutchik proposed the wheels of emotion, containing 8 basic emotion and their derivatives, which are the secondary and tertiary emotions [85]. Plutchik's wheel of emotion takes analogy from the color wheel. It arranges emotion in terms of similarity, with similar emotions placed close to each other, and opposites 180 degrees apart. Primary emotions are the center of the wheel, and mixtures of primary emotions result in secondary emotions, etc. Plutchik's wheel of emotion can be seen in Figure 2.1.

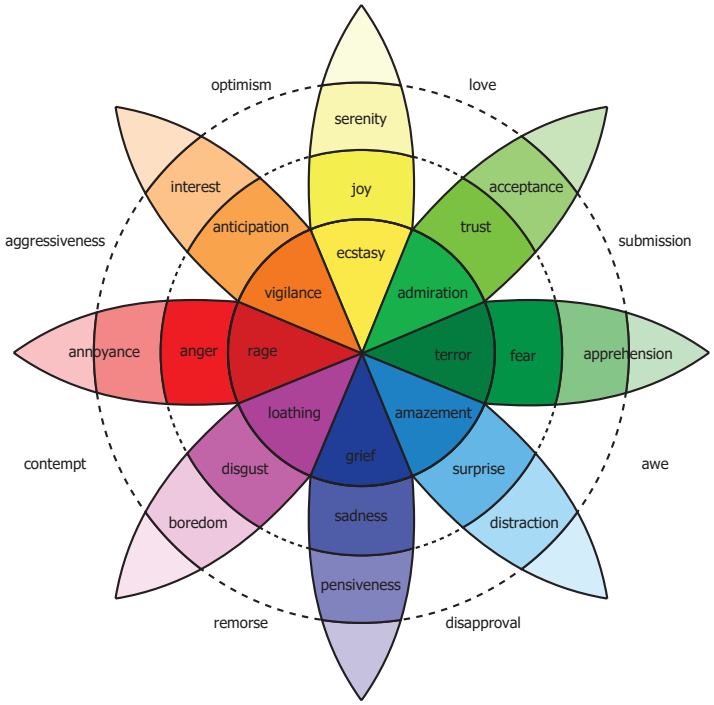


Figure 2.1: Plutchik's wheel of emotion [85].

At the center of the categorical definition of emotion is its evolutionary function. Emotion and its expression are universal and guarantee a coordinated and quick behavioral response [94]. Each emotion category is basic and differs from the others based on several key aspects, such as the way it is signaled, its physiology, its antecedent events [27].

In terms of computational model, categorical emotions have the big advantage of being clearly defined and distinguishable linguistically [94]. However, a unanimous agreement for which set of categories best describe universal experience of emotions is very difficult to achieve. Therefore, comparison of affective computing works can be ambiguous and problematic, even more so when a mapping between sets is required. Moreover, to capture differences and changes in finer-grained fashion, the number of emotion categories can be exceedingly large, which increases ambiguity of subjective perception exponentially. This may lead to data sparsity in less commonly occurring categories as well as poor annotator

agreements in data annotation.

2.1.2 Dimensional Model of Emotion

The second family of emotion is *dimensional*, where emotion is seen as a point in an n-dimensional space, described using affective dimensions as the axes. The concept is pioneered by Wilhelm Wundt in 1905. The longest established affective dimensions are valence and arousal, such as that proposed by Russel in the circumplex model of affect [90]. It has also been argued that additional dimensions, such as dominance and expectancy, are needed to better distinguish certain types of emotions, e.g. fear and anger: power and expectancy [31]. However, the decision on which affective dimensions to use remains strongly tied to the task at hand.

There are a number of advantages in computationally modeling emotion in a dimensional space. First, transitions between emotions become very intuitive as it can be directly mapped in the corresponding space. Second, mixture of emotions can be easily inferred, for example by taking an average position over a period of time. Third, changes of emotional states can be observed in a fine-grained manner through the movements in the dimensional space over time. Moreover, this model is intuitive and easily adaptable and extendable to either discrete or dimensional emotion definitions. The long established dimension are core to many works in affective computing and potentially provides useful information even at an early stage of research, allowing useful comparison to a wide range of related works.

In this work, we define the emotion scope based on the *circumplex model of affect* [90]. 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 2.2 illustrates the valence-arousal dimension in respect to a number of common emotion terms.

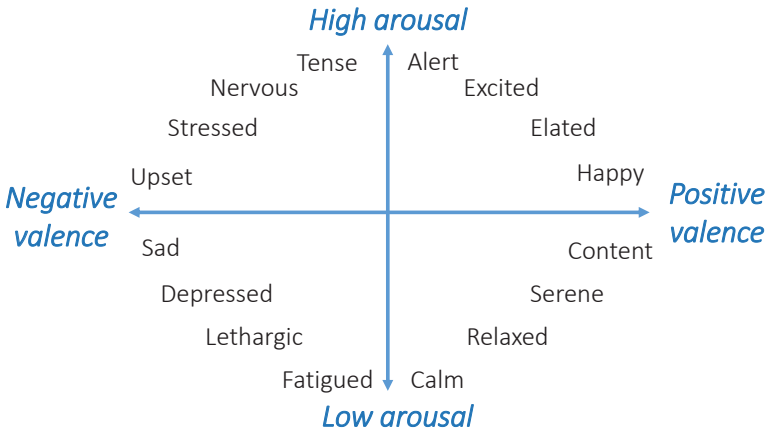


Figure 2.2: Emotion dimensions and common terms.

2.2 Dialogue Definition

2.2.1 Dialogue

Dialogue can be defined as a speech or text based exchange of information between two or more people. It is one of the methods of human communication, which has been discussed in more details in Section 1.1. In particular, we are focusing on dyadic text-based dialogue with chat-based dialogue systems.

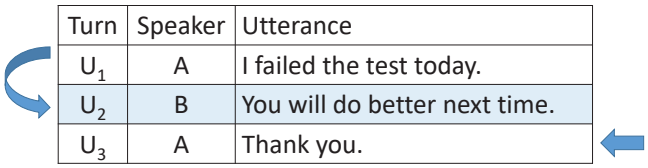
To computationally define dialogue, we utilize the hierarchical sequential data model, adapting that posited by Serban et al. [98] regarding the two-hierarchy view of dialogue. In this view, a dialogue is made up of a sequence of dialogue turns, and a dialogue turn is made up by a sequence of words. More formally, a dialogue D can be viewed 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, $U_m = \{w_{m,1}, ..., w_{m,N_m}\}$.

2.2.2 Dialogue Triples

Throughout the study and experiments, in particular we utilize the dialogue triple format. A *triple* is a sequence of three dialogue turns. That is, $D = \{U_1, U_2, U_3\}$. And following the hierarchical view of dialogue, Each dialogue turn U_m is a sequence of tokens of arbitrary length N_m, i.e $U_m = \{w_{m,1}, ..., w_{m,N_m}\}$.

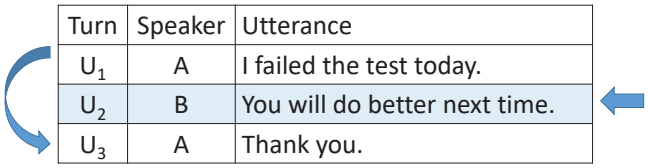
As I am focused on dyadic dialogue, U_1 and U_3 are considered to be uttered by speaker A, and U_2 by speaker B.

The triple format has been previously utilized for considering context in response generation [103], and filtering multi-party conversations into dyadic snippets [54]. In this thesis, the format is particularly useful to observe emotional changes in a dialogue. Figure 2.3 illustrates emotion events that can be observed in a triple. Given the first two dialogue turns, emotional change can be observed in the third turn. Similarly, the second turn can be regarded as the trigger of the emotional change from the first turn to the third.



Turn	Speaker	Utterance
U_1	A	I failed the test today.
U_2	B	You will do better next time.
U_3	A	Thank you.

(a) Emotion response. Emotion state of A in U_3 is a response to the dialogue turns U_1 and U_2.



Turn	Speaker	Utterance
U_1	A	I failed the test today.
U_2	B	You will do better next time.
U_3	A	Thank you.

(b) Emotion trigger. The emotion response of A from U_1 to U_3 is triggered by B's dialogue turn U_2.

Figure 2.3: Observing emotion response and trigger in a triple.

2.3 Emotion Improvements in Dialogue

I align the definition of emotion and dialogue to formalize the emotion-related terms in this thesis. Figure 2.4 illustrates the terms, explained below, with respect to the valence-arousal space.

- **Positive emotion** is any point in the valence-arousal space with positive arousal values.

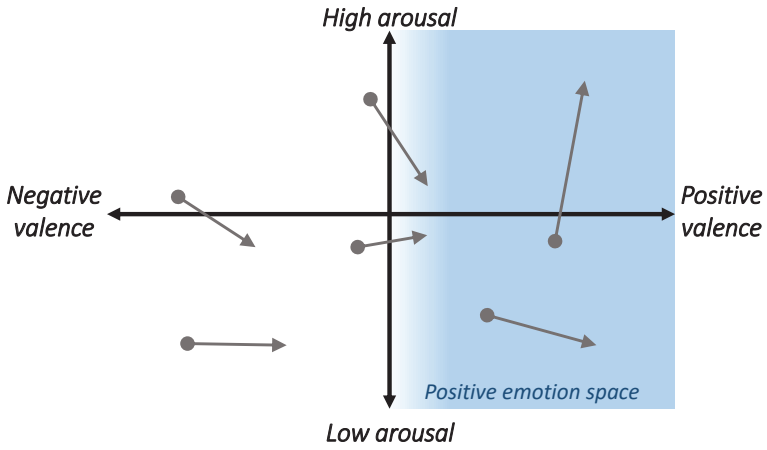


Figure 2.4: Blue area shows space of positive emotion. Arrows are examples of emotion improvements. The arrows are not related to each other. Note that the arrows begin and end in various places on the valence-arousal space. The important point of emotion improvement is movement towards a more positive valence.

- **Emotion improvement** refers to a change of emotion towards a more positive valence value. It is important to note that the end emotion of an improvement need not be positive, as long as the change of emotion leads toward a more positive emotion.
- **Emotion improvement elicitation, or positive emotion elicitation** is an attempt to elicit an emotion improvement or a more positive emotion. which can be realized short-term (at dialogue turn level) or long-term (encompassing an entire dialogue). Note that like above, the focus of the elicitation is the change of emotion, which has to be improved or more positive, and not the end emotion itself.

In chat-based dialogue, there may be multiple responses that are natural and coherent given a single input utterance. In real conversation, each of these responses actually have their own emotional impact to the listener, i.e. the change of emotion it causes in the listener. This phenomena is observable in a dialogue triple, as illustrated in Figure 2.5.

The goal of this thesis is then to build a dialogue system that 1) leans toward responses that have a more positive emotional impact, and 2) is able to elicit emotion improvement through dialogue. As previously mentioned, it is important

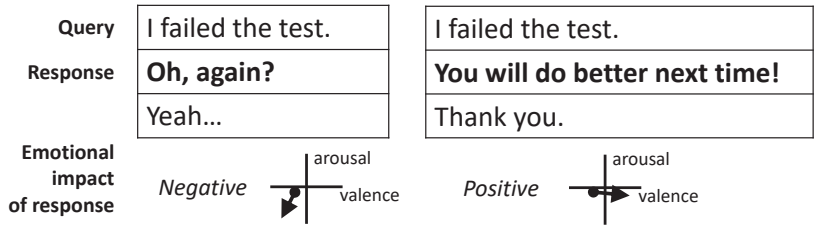


Figure 2.5: Examples of dialogue triples showing the emotion effect of different dialogue responses to an identical utterance. The responses have different emotional impacts, i.e. they elicit different emotional changes in the listener.

to highlight that this does not translate to consistently giving a "happy response." There are situations where such a response is inappropriate and could cause the opposite effect, e.g., in times of grief. Therefore, it is important for the system to reflect proper positive emotion elicitation techniques given semantic and emotional contexts.

Chapter 3

Approaches in Chat-based Dialogue Systems

In this chapter, I will review a number of approaches that have been successfully utilized in constructing chat-based dialogue systems. They are divided into two main approaches: response retrieval and response generation. Each of their related works, advantages, and shortcomings are discussed in detail in the following sections.

3.1 Response Retrieval

The response retrieval method for chat-based dialogue systems relies on a predefined set of responses given to the system. The main question to be solved is the manner in which the system selects, or retrieves, an appropriate response given a user input. Given a large amount of example and rules, a chat-based dialogue system can achieve quite a high performance. However, painstaking labor is required to realize such a system and even then, the system will never be immune to the out-of-example (OOE) problem as it is infeasible to include all possible scenario of human interactions. Two main methods of response retrieval is discussed in this section, including its application and drawbacks.

3.1.1 Rule-based

The most straightforward way of implementing a dialogue system is by defining a set of rules to govern the way it responds to user input. An example is Artificial Intelligence Mark-up Language (AIML), an open-source language that allows simplified dialogue modeling through the definition of patterns in the conversation [113]. AIML contains several elements, one of the most important ones being a *category*, which forms a unit of knowledge. The category then contains a set of patterns and its corresponding template with which the system would respond. Writers can also define variable in the response template that can be used with varying values (e.g. a name variable for the chatbot can be assigned with John, Jane, etc.). A similar approach can also be implemented in ChatScript [117].

Some well known chatbots developed in this manner are Eliza [115], A.L.I.C.E [114], and Suzette [118]. Three of the annual Loebner Prize Competition in AI winners are based on the AIML. Although effective to a certain extent, rule-based chatbots suffer from fundamental limitations. One, rigorous hand-crafting of rules is required. This hinders the scaling of such systems to cover larger domains and variety of conversation flow. Two, the hand-crafted rules are not domain portable. The same conversation in different domains would require an entirely new set of rules. While this approach is useful for deployment in a strictly constrained scenario, not much research advancement potential lies in this direction other than evaluation and analysis studies.

3.1.2 Example-based Dialogue Management

Example-Based Dialogue Management (EBDM) is a data-driven approach of dialogue modeling that uses a semantically indexed corpus of *query-response* pair examples instead of handcrafted rules or probabilistic models [55]. At a given time, the system will return a response of the best example according to a semantic constraint between the query and example query. This circumvents the challenge of domain identification and switching—a task particularly hard in chat-oriented systems where no specific goal or topic is predefined. With the increasing amount of available conversational data, EBDM offers a straightforward and effective approach for deploying a dialogue system in any domain.

In the context of dialogue system, we will use term *query* to refer to user's input, and *response* to refer to system's output

Lasguido et al. have previously examined the utilization of cosine similarity for response retrieval in an example-based dialogue system [54]. In their approach, the similarity is computed between term vectors of the query and the examples. The vector for an utterance T is the size of the database term vocabulary, where each term t is weighted by its TF-IDF score, computed as:

$$\text{TFIDF}(t, T) = F_{t,T} \log \frac{|T|}{DF_t}, \tag{3.1}$$

where $F_{t,T}$ is defined as term frequency of term t in sentence T, and DF_t as total number of sentences that contains the term t, calculated over the example database. Cosine similarity between two vectors a and b is computed as:

$$\cos_{sim}(S_q, S_e) = \frac{S_q \cdot S_e}{\|S_q\| \, \|S_e\|}. \tag{3.2}$$

Given a query, this cosine similarity is computed over all example queries in the database and treated as the example pair scores. The response of the example pair with the highest score is then returned to the user as the system's response. This is one of the main approaches in chat-oriented dialogue systems. This process is illustrated in Figure 3.1.

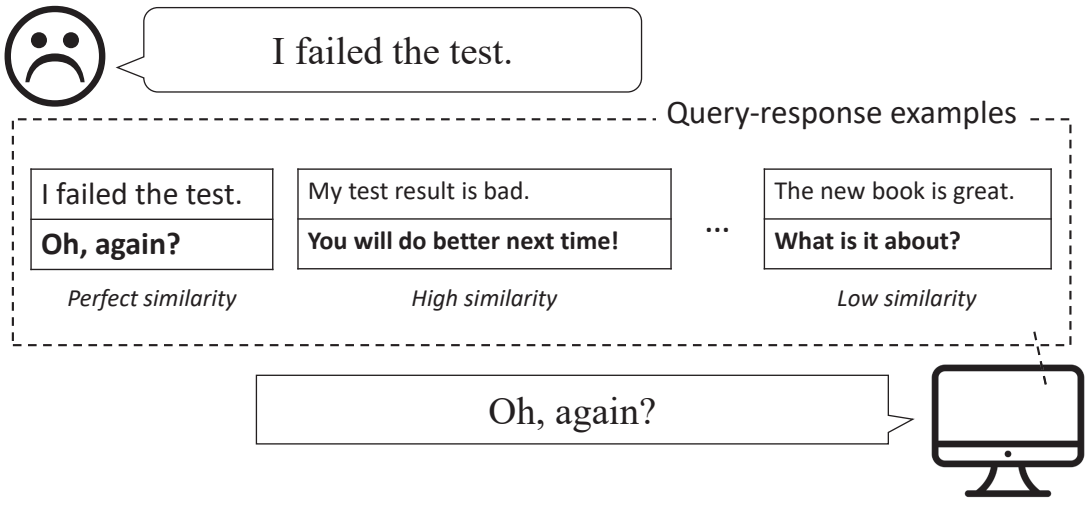


Figure 3.1: Response selection on EBDM.

3.2 Response Generation

In contrast to response retrieval, response generation approach for chat-based dialogue system produces the dialogue response sequentially, from beginning to end. In training a response generator, commonly the objective is to maximize the likelihood of the training data under the model parameters. This comes with a different set of issues, the most prominent being model tendency to generate generic and short responses [57], i.e. ones that have high likelihood under many user utterance input such as "I don't know." However, the highly abstractive modeling of the data circumvents the need of rules and heuristics design. This means that the model are highly scalable and less labor expensive to train. State-of-the-art response generators are based on neural networks. Two main architectures are discussed in this section.

3.2.1 Recurrent Neural Network Encoder-Decoder

A recurrent neural network (RNN) is a neural network variant that can retain information over sequential data. To gain better understanding of RNN, it is helpful to briefly review its more basic form, the artificial neural network (ANN or NN).

An NN is a system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic response to external inputs [74]. The input layer handles the input vector of length K, $\mathbf{x} = \{x_1, x_2, ..., x_K\}$. The i-th hidden layer consist of a collection of N neurons $\mathbf{h_i} = \{h_{i,1}, h_{i,2}, ..., h_{i,N}\}$. Similarly, the output layer consist of neurons that represent each element of the output vector $\mathbf{y} = \{y_1, y_2, ..., y_M\}$.

Figure 3.2 illustrates a neural network with three layers. The input layer consists of three neurons. The one hidden layer consists of six neurons, and the last layer, the output layer, consists of two neurons. Every element in the input layer is connected to every element in the hidden layer, and so on for every two sequential layers. This constitutes a fully-connected NN. Each of this connection is assigned a weight w, and all the weights in an NN, W, make up the weight parameters of the model. The output of an arbitrary fully-connected layer h_i is

$$h_i = a(W_i \cdot h_{i-1} + b_i), \tag{3.3}$$

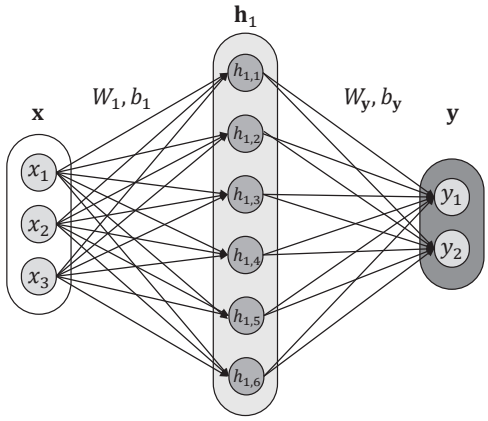


Figure 3.2: A fully connected neural network with one hidden layer.

where a is an activation function. Activation functions are commonly used to transform outputs into a desirable range, e.g. step function for binary output, sigmoid function for outputs in range $[0, 1]$, hyperbolic tangent function for range $[-1, 1]$, or linear function if no transformation is required. A bias factor b is added to allow shifting for a better fit of the data.

An RNN differs from NN in its ability to retain sequential information, unlike NN which assumes that all inputs are independent of each other. This property is essential especially in problems such as language modeling, since words and languages have inherent sequentiality in them. Figure 3.3 illustrates the connections in an RNN over several time steps.

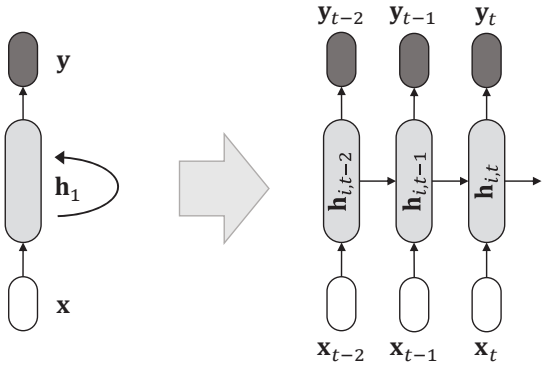


Figure 3.3: Simplified representation of an RNN. Units inside each layer are not shown. Left side shows compacted view, right side unrolled. t denotes time step information.

The hidden layer is equipped with a mechanism to retain and forget information from previous time steps. In an RNN, Equation 3.3 is replaced with a more complex operation to allow this mechanism. Two of the most popular neuron cell for RNN architectures are long short-term memory (LSTM) cell [46] and gated recurrent unit (GRU) cell [17].

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 3.4 presents a schematic view of this process. This architecture was previously proposed as neural conversational model in [112].

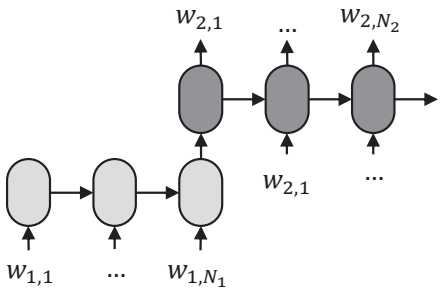


Figure 3.4: RNN in a response generation task. The lower RNN encodes the information from input sentence U_1, the upper RNN decodes the response U_2.

The basic encoder-decoder model has been extended to address various issues in dialogue response generation. Sordoni et al. proposed neural network architectures to consider dialogue context from previous dialogue utterances in generating a response [103]. Li et al. designed scoring functions as training objectives to promote more diverse responses and stretch conversation length [58, 60], as well as investigated persona-based generation in chat-based systems [59]. Furthermore, emotion expression in dialogue systems has also been researched using the neural response generator framework [47, 122].

In practice, there are a number of difficulties and drawbacks in training RNNs on sequential data, one of the biggest being the vanishing gradient problem. When

optimizing an RNN using gradient-based learning methods and backpropagation, each of the parameters in the network receives an update proportional to the partial derivative of the error function wrt. the current parameter. However, in some cases this gradient may vanish and become really small due to the chain rule computation during backpropagation. RNN are prone to this problem especially when modeling long-term dependency in long sequential data. The very small gradient value hinders the optimization of the network, and in the worst case may even completely stop the neural network from further training.

3.2.2 Hierarchical Recurrent Encoder-Decoder

Based on the two-hierarchy view of dialogue, the hierarchical recurrent encoder-decoder (HRED) extends the sequence-to-sequence architecture [98]. 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 3.5 presents an overview of this architecture.

The HRED makes use of the gated recurrent unit (GRU) [17] with hyperbolic tangent activation function. The model is trained to maximize the log-likelihood of the training data using the Adam optimizer [49].

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. This solves the vanishing gradient problem mentioned in the previous section, allowing the model to converge even when handling multi turn dialogues, which will be treated as one very long sequence in traditional RNN architecture. Experimental results reported in [98] shows

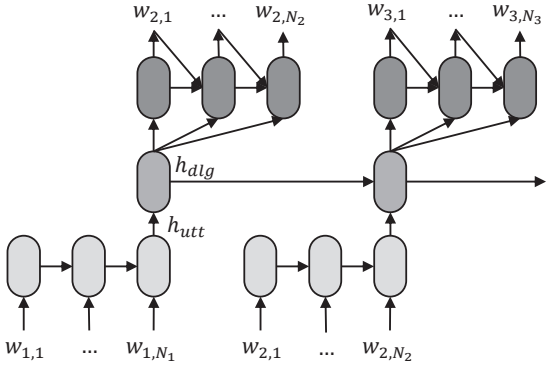


Figure 3.5: HRED architecture. The lower RNN encodes sequences of tokens, the middle RNN encodes sequence of the dialogue turn, and the upper RNN decodes the tokens of the next dialogue turn.

that HRED yields significantly higher performance than RNN on the dialogue response generation task.

Chapter 4

Constructing Emotion-rich Dialogue Corpora

4.1 Existing Corpora

4.1.1 Large Scale Dialogue Corpora

Chat-based dialogue systems have seen an important breakthrough in recent years following the innovations of neural network architectures and improvements of computing hardware capabilities. However success stories in training an end-to-end chat-based dialogue systems, such as [112, 98, 100], all require a huge amount of data to start with – in the hundreds of millions or even billions of words [99]. In other words, large scale dialogue corpora is central to state-of-the-art approaches.

One of the most influential dialog corpora is the Switchoard corpus [35]. The Switchboard corpus contains approximately 2,500 dialogues involving 500 speakers. Two telephone callers are connected and asked to converse about a topic introduced by an automatic operator system. The proceeding conversation is recorded and transcribed. The telephone was a popular mean in collecting large amounts of data involving many people, it was also used in the collection of CALLHOME [13] and CALLFRIEND [14] corpora.

Recently, twitter crawler has been shown to be effective and efficient for constructing large scale corpora, such as that used in [6, 59]. A crawler is utilized to gather s large amount of twitter data, and a dialogue from a set of tweets can be established by sequentially organizing their reply tags. On the other

hand, unlike tweets which often use written slang, emoticons, and abbreviations, movie subtitles can provide millions dialogue turns that are closer to spontaneous human-human conversations.

It is effective to leverage large amounts of available human conversations to expand system's vocabulary of various domain. Banchs et al. collected hundreds of thousands of movie dialogue to construct the Movie-DiC corpus, amounting to around 760K dialogues. By resorting to data that are easier to crawl and collect, Ameixa et al. expanded upon this idea through movie subtitles, constructing a corpus of around 5.5M dialogue units in text form [2]. Large-scale dialogue corpora such as OpenSubtitle [105] and SubTle [2] have been successfully used to train end-to-end chat-based systems [112, 98].

In this work, the SubTle corpus is of particular interest due to two reasons: 1) it is accessible and containing huge amount of dialogue of natural human conversation, and 2) it allows comparison to existing works. The SubTle corpus contains conversational pairs extracted from movie subtitles expanding four genres: horror, science fiction, western, and romance. High-quality movie subtitles are obtained using movie identifiers shared by 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 Query-Answer format.

4.1.2 Affective Corpora

Emotion-rich data is pre-requisite for incorporating emotion in HCI. The majority of existing corpora are constructed for the purpose of recognizing emotion-related phenomena in humans, such as facial expression [71], physiological signals [51], emotion perception [12], and physical motions or gestures [11]. Even though these corpora contain important information for understanding emotion, their scope does not reach the dynamics of emotion in dialogue, at the interpersonal level.

There exist a handful of social affective corpora, each with its own focus and scope. An example is the Vera am Mittag corpus, containing recordings of spontaneous emotional conversation from television talk shows [39]. Talk show recordings are suitable for affective computing research as they provide natural emotion occurrences in a typical social setting [63, 65, 64]. Despite this quality, an explicit conversation goal that could provide emotional benefits for the

participants is still missing from such data.

On the other hand, the Distress Analysis Interview Corpus (DAIC) contains clinical interviews designed for the development of an automated agent for psychological diagnoses [37]. It includes interviews with distressed and non-distressed participants and highlights verbal and non-verbal dialogue actions. However, the corpus does not provide emotion annotation nor employ an expert as the interviewer of the distressed participant.

The SEMAINE database is particularly relevant for this study. In the following subsection, we describe this corpus in detail and highlight the qualities that can be utilized to support the contributions of this thesis.

Induced Emotion in Laboratory Environment: SEMAINE

The SEMAINE Database is an annotated multimodal records of emotionally colored conversation between a person and a limited agent [75]. The corpus is collected from conversations in Wizard-of-Oz setting between two participants, one acts as the user and another acts as a wizard, posing as a Sensitive Artificial Listener (SAL). During the interaction, one restriction of the wizard is that it is unable to answer questions from the user.

A SAL is a limited agent designed to give the impression of attentive listening through verbal and non-verbal cues. In the corpus, there are four characters of SAL: cheerful Poppy, angry Spike, depressed Obadiah, and sensible Prudence. Each SAL responds to the user according to their characteristics, eliciting different reactions from the user, thus yielding an emotionally-colorful conversation.

This set up offers two important advantages. First, it allows the observation of various emotional reactions in dialogue. The emotion contained in the corpus arises spontaneously as induced by the way the SAL behaves. Therefore it can be assumed to be quite true to human emotion appraisal.in contrast to human-human interaction, the data shows how humans treat and react to the shortcomings of an automated agent. The corpus is designed specifically to represent interaction between a human and an automated agents. As such, the nature of the data provided is highly suitable for research in affect for HCI.

Recordings and Annotations A participant posing as a user interacts with all 4 SAL characters, recorded in the span of 4 sessions (one session for each SAL).

There are 24 participants posing as the user, and 4 participants posing as SAL. In total, 95 sessions of interaction is recorded. The relevant sessions amounts to 4 hours of material. The majority of the recordings in the SEMAINE Database are fully transcribed, with time alignment according to the turn taking changes. Disfluencies (e.g. em, uh) are annotated as is, while laughter are assigned a special tag.

On the other hand, the emotion occurrences are annotated using the FEELtrace system [21] to allow recording of perceived emotion in real time. As an annotator is watching a target person in a recording (i.e. visual and audio information), 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 real numbers ranging from -1 to 1, called a *trace*, that shows how a certain emotional aspect fall and rise within an interaction. The numbers in a trace are provided with an interval of 0.02 seconds.

The amount of annotation for the user and the SAL's clips differ in number. A small number of the SAL's clip are annotated by one to three annotators. On the other hand, for the user's clips, the majority have been annotated by 6 raters. The core annotation includes five emotion dimensions: valence, arousal, power, expectation, and intensity. Two reliability analyses, Quantitative Agreement (QA) and correlational analysis, performed by the authors shows that 2/3 of the traces pass the stringent criterion of either analysis, and about 80% reach the level that are normally regarded as acceptable.

4.1.3 Limitations and Proposals

The construction of large scale emotion-rich corpora requires multiple fold of resources compared to constructing dialogue corpora without any emotion information. Extra efforts need to be paid for the design, participant recruitment, as well as annotation of the corpus. I believe the major challenge in constructing one lies in the expensive process of emotion annotation, since multiple annotators are needed for every occurrence. Even when such process takes place, strong agreement between the annotators can not be guaranteed. Consequently, with the absence of large-scale emotion-rich corpora, it becomes very challenging to train an end-to-end affective dialogue system successfully.

In terms of existing affective corpora, I believe three main limitations wrt.

the goal of this thesis are present:

1. Even though various conversational scenarios have been considered, there is still an absence of dialogue corpora showing how to respond in a dialogue so as to promote positive emotions.

2. Existing works either focus on more serious emotional problems, such as distress and depression, or without any focus in minor emotional troubles, leaving a gap in between. There is a lack of resources that show common emotional problems in a everyday social setting.

3. To the best of my knowledge, a great majority of existing corpora does not involve any professional who is an expert in handling emotional reactions in a conversation. Knowledge from such situations is highly potential in constructing assistive technology for emotion-related problems in everyday situations.

To outgrow these limitations, I designed and constructed the following two corpora:

1. A dialogue corpus containing responses that promote positive emotional states. To circumvent the cost of data recording and processing from scratch, I leverage existing emotion-rich corpus and crowdsourcing to construct the corpus.

2. A dialogue corpus which demonstrate emotion processing and improvement through social communication, involving an expert in the conversation. The corpus is designed to 1) contain recordings of dyadic spontaneous social-affective interactions before and after a negative emotion exposure, and 2) involve a professional counselor as an expert in the conversation. 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. This allows the observation of emotion fluctuation in a conversation, and how an external party can guide and facilitate emotion processing through an interaction.

The construction and analysis of these corpora will be discussed in the following sections.

4.2 Constructing Dialogue Corpus with Responses that Elicit Positive Emotion

In this section I will elaborate on the construction and analysis of a dialogue corpora containing responses that promote positive emotions. Such a corpus will allow training an end-to-end dialogue system that elicit positive emotions. I.e., eliciting positive emotion without explicit definition of dialogue strategy, straightforwardly my learning to mimic the data. To circumvent the cost of data recording and processing from scratch, I leverage existing emotion-rich corpus collected in with WoZ scenario, i.e. SEMAINE data, and crowdsourcing to construct the corpus. The crowdsourcing provides a built-in advantage of aligning the resulting corpus to human standards.

4.2.1 Precedure

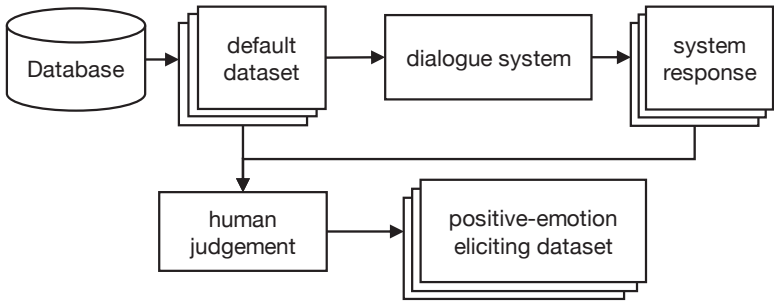


Figure 4.1: Obtaining references that elicit positive emotion.

First, we run the triples through a dialogue system that elicits positive emotion (described below), to obtain new candidate responses that supposedly elicit positive emotion. Subsequently, through crowdsourcing, we ask 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.

Example-based Dialogue Management for Positive Emotion Elicitation

We have recently extended the EBDM framework for the positive emotion elicitation task [68]. To allow the consideration of emotional aspects, we make use of *triples* units in the selection process in place of the *query-response* pairs in the traditional EBDM approach. As discussed in Section 2.2.2, a triple consists of three consecutive dialogue turns that are in response to each other.

To elicit positive emotion, we instead exploit the triple format to observe emotional triggers and responses in a conversation. The triturn format allows the observation of the future response, i.e. the user response to the system response, in the examples. We believe that expected future impact is an aspect that should not be overlooked. This common knowledge is prevalent in humans and strongly guides how we communicate with other people – for example, to refrain from provocative responses and to seek pleasing ones.

Thus, in addition to traditional semantic constraint as described in Section 3.1.2, we formulate two types of emotional constraints: (1) emotion similarity between the query and the example queries, and (2) expected emotional impact of the candidate responses. Figure 4.2 shows how considering expected emotional impact could potentially elicit a more positive effect in the real interaction.

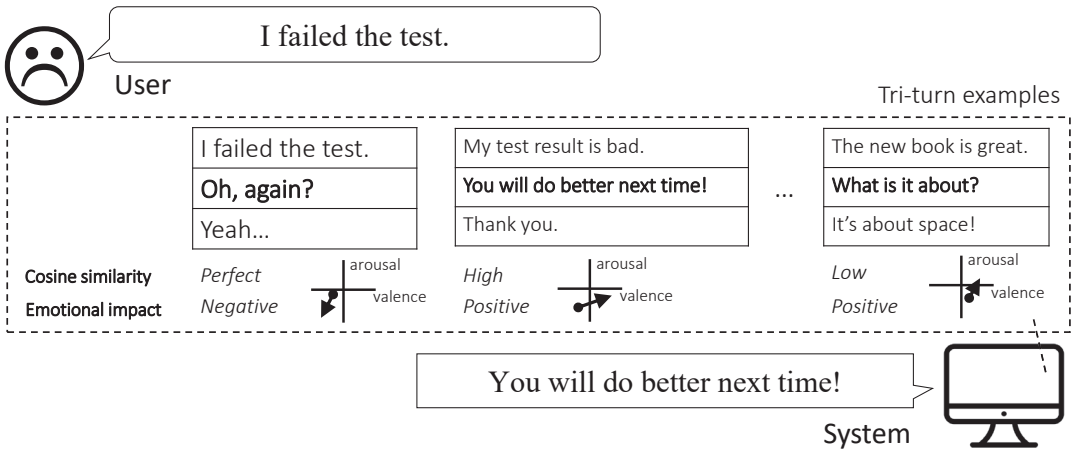


Figure 4.2: Considering expected emotional impact in dialogue response selection

To measure emotion similarity, we compute the Pearson's correlation coefficient of the emotion vector between the query and the example queries. Correlation r_{qe} between two emotion representation vectors for query q and example e of length n is calculated using Equation 4.1.

$$r_{qe} = \frac{\sum_{i=1}^{n}(q_i - \bar{q})(e_i - \bar{e})}{\sqrt{\sum_{i=1}^{n}(q_i - \bar{q})^2}\sqrt{\sum_{i=1}^{n}(e_i - \bar{e})^2}}. \tag{4.1}$$

This similarity measure utilizes real-time valence-arousal values instead of discrete emotion label. In contrast with discrete label, real-time annotation captures emotion fluctuation within an utterance, represented with the values of valence or arousal with a constant time interval, e.g. a value for every second.

As the length of emotion vector depends on the duration of the utterance, prior to emotion similarity calculation, sampling is performed to keep the emotion vector in uniform length of n. For shorter utterances with fewer than n values in the emotion vector, we perform sampling with replacement, i.e. a number can be sampled more than once. The sampling preserves distribution of the values in the original emotion vector. We calculate the emotion similarity score separately for valence and arousal, and then take the average as the final score.

Secondly, we measure the expected emotional impact of the candidate responses. In a triple, emotional impact of a *response* according to the *query* and *future* is computed using Equation 4.2.

$$\text{impact}(response) = \frac{1}{n}\sum_{i=1}^{n} f_i - \frac{1}{n}\sum_{i=1}^{n} q_i, \tag{4.2}$$

where q and f are the emotion vectors of *query* and *future*. In other words, the actual emotion impact observed in an example is the expected emotional impact during the real interaction. For expected emotional impact, we consider only valence as the final score.

Figure 4.3 illustrates the steps of response selection of the baseline and proposed systems. We perform the selection in three steps based on the defined constraints. For each step, a new score is calculated and re-ranking is performed only with the new score, i.e. no fusion with the previous score is performed.

The baseline system will output the *response* of the triple example with the highest semantic similarity score (Equation 3.2). On the other hand, on the proposed system's response selection, we pass m examples with highest semantic

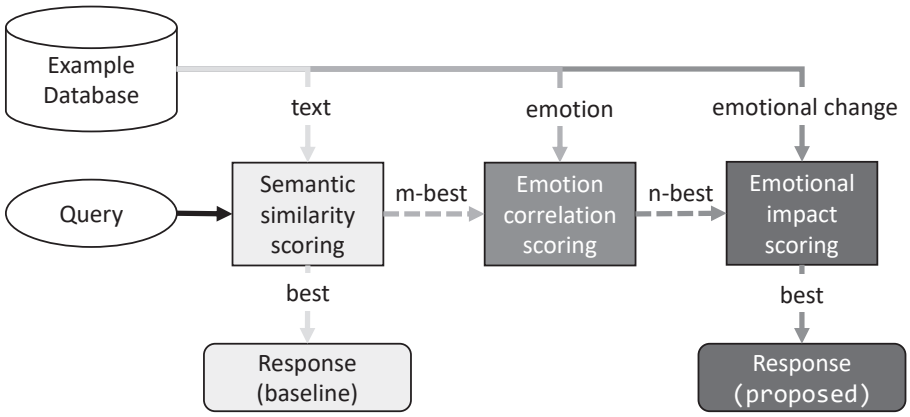


Figure 4.3: Steps of response selection

similarity scores to the next step and calculate their emotion similarity scores (Equation 4.1). From n examples with highest emotion similarity scores, we output the *response* of the triple example with the most positive expected emotional impact (Equation 4.2).

The filtering on each step is done to ensure that the semantic and emotional contexts in the candidate examples match the real interaction such that the response yields as similar an impact as possible with the example. As emotion space is much smaller than that of semantic, many of the examples may achieve a high emotion similarity score. Thus, imposing the emotion constraints in the reduced pool will help achieve a more relevant result. Furthermore, this reduces the computation time since the number of examples to be scored will be greatly minimized. When working with big example databases, this property is beneficial in giving a timely response.

It is important to note that this strategy does not translate to selection of the response with the most positive emotion. On the other hand, it is equivalent to selecting the response that has the most potential in eliciting a positive emotional impact, given a semantic and emotional context. Even though there is no explicit dialogue strategy to be followed, we hope that the data reflects the appropriate situation to show negative emotion to elicit a positive impact in the user, such as relating to one's anger or showing empathy.

We compared the proposed response selection method with the traditional method in terms of coherence, emotional connection, and emotional impact. Subjective evaluation showed that by incorporating emotional state and potential

impact in selecting a response, we can elicit a more positive emotional impact in the user, as well as achieve higher coherence and emotional connection.

4.2.2 Result and Analysis

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. 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.

4.3 Constructing Dialogue Corpus of Emotion Processing and Improvement through Dialogue

Unlike the corpus in Section 4.2 which reflects WoZ dialogues between human and computer, in capturing emotion processing and improvement through dialogue we focus on human-human interaction. Since the emotion improvement elicitation task is not yet studied in dialogue system interactions, it is important to collect data on the behavior of the interlocutors within the scenario, and analyze it beforehand. Information gathered from human-human interaction would guide the dialogue system design, help identify dialogue flow and important aspects pertaining to the improvement elicitation task. Such a bottom-up approach has been used successfully in designing a dialogue system in a number of previous works, e.g. [24], where data-driven design is applied to a dialogue system.

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.

4.3.1 Corpus Design

Recording Scenario

We focus on capturing social support in everyday situations that may trigger a negative emotion, such as reading the news or debating on a social issue. Specifically, we would like to observe how an external party can guide and facilitate emotion processing through an interaction after a negative emotional response. Thus, we arrange for the dyad to consist of an *expert* and a *participant*, each with a distinct role. The *expert* plays the part of the external party who helps facilitate the emotional response of the *participant*.

We design the recording scenario as follows. The session starts with an opening talk as a neutral baseline conversation. Afterwards, we induce negative emotion by showing an emotion inducer to the dyad. The recording then continues with a discussion phase that targets at emotional processing and recovery. In this phase, the expert is given the objective to facilitate the emotional process that follows the emotion induction, and to elicit a positive emotional change through the conversation.

Throughout the process, we ask the participants to assess their emotional states with the use of a questionnaire. In particular, these assessments are collected after briefing, before the emotion inducer, after the emotion inducer, and after the discussion. This allows us to keep track of their emotional states as they occur before and/or after the moments where fluctuations are expected. After the recording, the participants are asked to fill out a post-recording questionnaire to rate their experience of the interaction. Figure 4.4 illustrates the design of a recording session.

To approximate an everyday social situation, we focus on various social issues as the conversation topic, such as politics and environmental issues. More specifically, we target issues where opinions with negative sentiment might arise. We intended for the initial emotion to be generated by an external factor to promote openness in the discussion. Furthermore, a more personal topic is likely to take longer to process and recover from, which is undesirable given the limitation of the interaction circumstances.

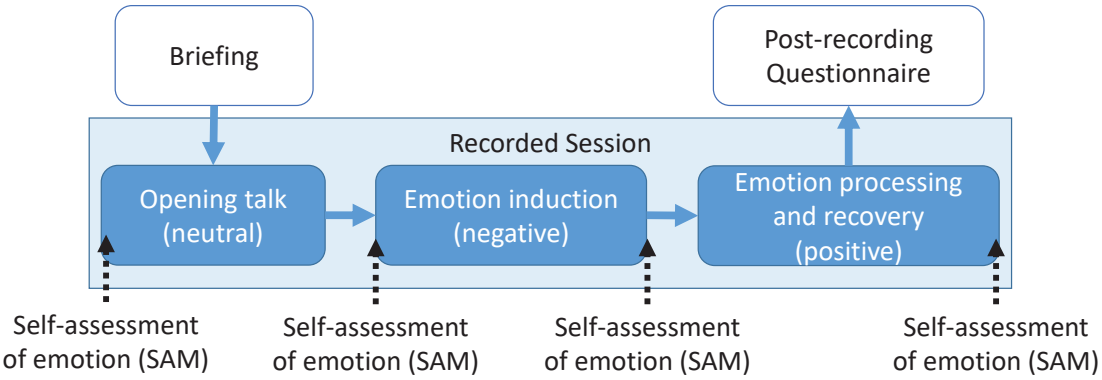


Figure 4.4: The flow of a recording session.

Emotion Inducer

We opt for short video clips (a few minutes in length) as emotion inducers in the sessions. The use of video clips as emotion elicitor is well established and has been studied for several decades [40, 92]. 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 [116]. Furthermore, this technique offers practical replication in constrained environmental settings, such as the recording room. Finally, in terms of ethical concerns, this technique is less personally involved for the inducee compared to others such as autobiographical recollection [93] or the real life method, where inducees are asked to perform various physical tasks [53].

However, unlike the majority of previous studies which uses excerpts of films or movies showing hyper-realistic fictional situations [93], we look for clips that depict real life situations and issues, i.e., non-fiction and non-films. Our concern is the unpredictability from person to person when emotionally responding to clips knowing that it is fictional. Furthermore, the use of a non-fictional inducer reflects real everyday situations better. We intend for the clips to contain enough information and context of a certain subject to serve as conversation topic throughout the recording session.

To fit these requirements, we select short video clips of news reports, interviews, and documentary films as emotion inducers. 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.

Among others, the anger inducers include reports on an unfair working environment, animal cruelty, and domestic violence. The sadness inducers include but are not limited to stories on environmental changes, a person who went through child abuse, and a child bride.

Questionnaires

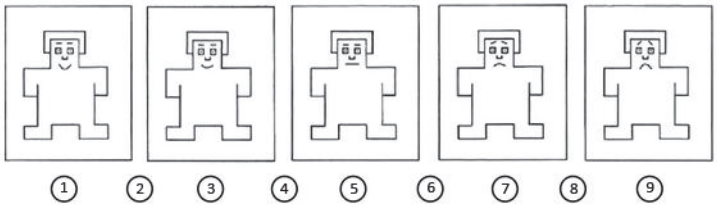
Two questionnaires are designed to measure subjective qualities of the session. The first questionnaire is used to measure participant's emotion in the moment (as opposed to after-the-fact in annotation). The second questionnaire is aimed to measure participant's satisfaction of the session. Both questionnaires are explained in detail below, and samples are attached in Appendix A.

Self-Assessment Manikin (SAM) Questionnaire Throughout the process, we ask the participants to assess their emotional states through the use of a Self-Assessment Manikin (SAM). The SAM is a pictorial rating scheme designed for easy-to-use non-verbal assessment of emotional state and reaction [8]. Following the emotion definition in Section 2, we exclusively use the valence and arousal SAM. The pictorial scale for valence and arousal is depicted in Figure 4.5.

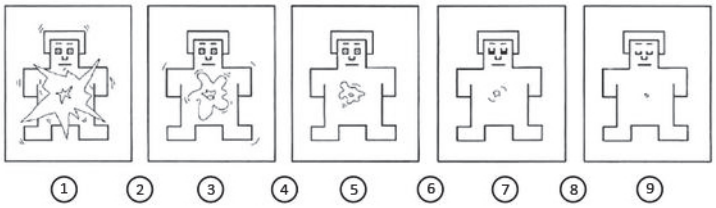
For both dimensions, the scale ranges from 1 (most positive) to 9 (most negative) with matching illustrations of the emotional states. During assessment, the participants are simply asked to choose a number for each dimension that matches their current state of emotion, as written below the illustrations.

Post-Recording Questionnaire We ask the participants to fill out post-recording questionnaires to quantitatively measure 1) the effectiveness of the emotion inducer, and 2) the satisfaction of the session with the counselor. The participants are asked to answer the following questions on a Likert scale ranging from 1 (strongly agree) to 5 (strongly disagree):

- I noticed a negative emotional change in myself after watching the video.
- I noticed a positive emotional change in myself during and after the conversation with the counselor.



(a) Valence SAM.



(b) Arousal SAM.

Figure 4.5: SAM for self-assessment of emotional state and reaction [8].

- The conversation helped me deal with and process my emotion.
- I felt understood by the counselor.
- I enjoyed the conversation with the counselor.
- I found a kind of emotional connection between myself and the counselor.
- I would like to talk again with the counselor in the future.

4.3.2 Data Collection

Participants

We recruit a professional counselor as the *expert* in the recording. The expert obtained a Diploma in Counseling and has been an accredited member of the British Association for Counseling and Psychotherapy. The expert has more than 8 years of professional experience, and has been practicing with the following areas of expertise: general counseling (e.g., depression, anxiety), relational issues, sexuality, childhood trauma, identity, cultural adjustment, and family problems.

As *participants*, we recruit 30 individuals that speak English fluently as first or second language. The group of speakers covers 13 nationalities in total. All of the participants are residing in the same area and are embedded in an international academic environment during the time of our recordings. The group consists of 20 males and 10 females.

Set Up

We record the videos of the dyad with two cameras, each facing a single person for a portrait shot. The two cameras are the SONY Handycam HDR-CX670 and the SONY Handycam HDR-PJ675. We record with 29.97 frames per second and a resolution of 1280x720 pixels. The video recordings are stored with the H.264 video compression standard in the yuv420p color space.

The audio signals are captured with two Crown CM-311A cardioid condenser head-worn vocal microphones, both of them wired to a USB audio interface of the type Roland QUAD-CAPTURE UA-55. We record the speech of the dyad as two mono audio signals, one for each speaker, at a sound rate of 44.1 kHz with 16-bit PCM quality, stored in a single .wav file format. After the recording, the data from the camcorders and the microphones are synchronized manually. The layout of the recording room is illustrated in Figure 4.6.

Recording Procedure

We record 60 sessions in the course of 6 weeks, i.e. 10 sessions are recorded in a week. Each of the 30 participants attended 2 sessions with at least one week period between them. For each participant, one session is assigned to an anger-inducing video clip, and another to a sadness-inducing clip. Each video clip is shown to 3 different participants.

Each session is allocated 30 minutes and the procedure is as follows. The camcorders and the expert's microphone are set up prior to the recording. The expert waits for each participant in the recording room while they are briefed in a separate room. Afterwards, the participant enters the recording room, takes a seat, and the recording assistant places his or her microphone. After the equipment is set up, the assistant retreats behind the room separators.

Every session is opened with a brief talk covering topics such as the general well-being, the current research, study or career progress, and weekend plans.

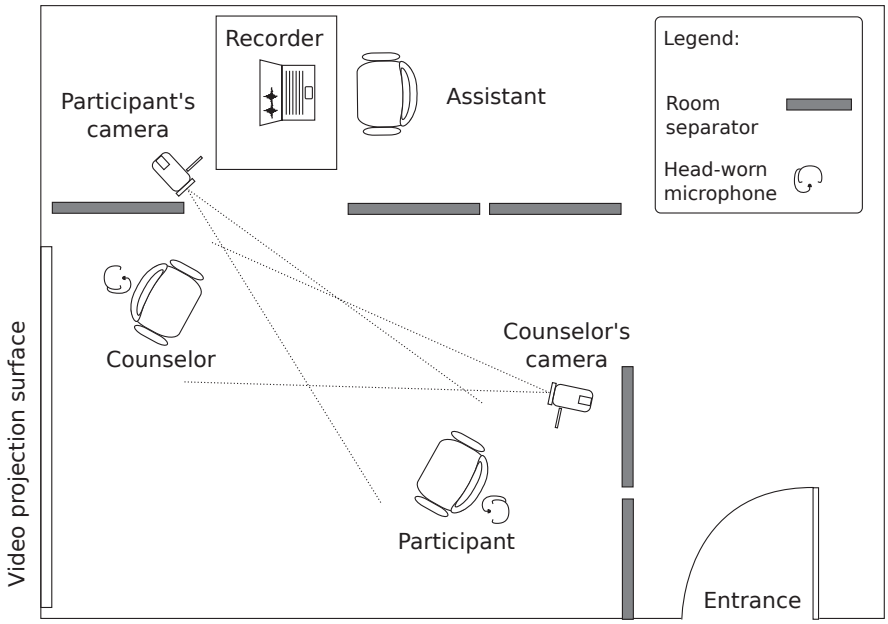


Figure 4.6: The recording room layout.

Afterwards, the expert hands over the SAM questionnaire for self-assessment. Upon completion, the emotion inducer is shown on the video projection surface. The expert and participant watch the video at the same time. The assistant leaves the room after the playback of the video stops. The participant is then asked to fill out another SAM questionnaire.

Afterwards, the conversation between the dyad begins for the remaining of the allocated 30 minutes. The participant leaves the room when either the allocated time has passed, or when the conversation comes to a natural conclusion. The participant continues to the post-recording procedure, which consists of debriefing and filling out the post-recording questionnaire. In the meantime, the expert does verbal assessment regarding the participant and the conversation. The assistant returns to the recording room at this point to secure the data from the current audio and video recordings and to prepare for the next session.

4.3.3 Annotation

Emotion Annotation

The emotion occurrences are annotated using the FEELtrace system [21] to allow the recording of perceived emotion in real time. This annotation tool and scheme has been described previously in Section 4.1.2. Following the emotion definition of Section 2.1.2, we annotate both the valence and arousal dimensions of each recording. A screen capture of the annotation tool in operation on one of the sessions is shown in Figure 4.7.

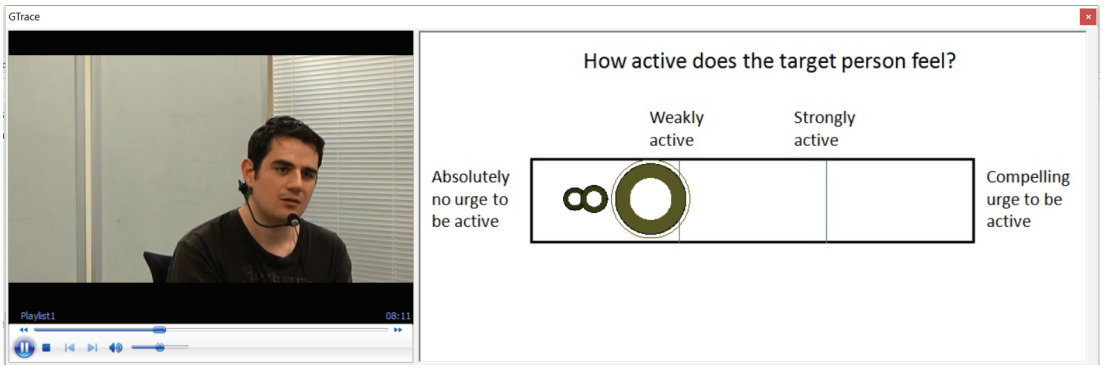


Figure 4.7: Annotating the arousal dimension.

At this stage of development, we focus on annotating the participant's emotional state. We aim to provide two types of annotations: self-reported and perceived. Self-reported emotion is annotated by the subjects themselves (in this case, the participants), while perceived emotion is annotated by another party according to the communication clues that the subject expresses (in this case, the expert).

For each session, the annotation is performed twice: once for valence, and once for arousal. The self-reported emotion is annotated by the participants directly after the recording is finished. Due to the tight recording schedule with the expert, this arrangement is not possible for the perceived emotion annotation. Instead, the expert performs the task off-site post-recording. All 60 sessions have been annotated with self-reported emotion and perceived emotion traces.

Transcription

We transcribe the spoken language of each recorded session. The speech of each speaker is split into separate channels, providing clean non-overlapping speech for the transcription task. We employ a paid Automatic Speech Recognition (ASR) service to obtain an automatic transcription of the data. The automatic transcription is then subject to manual revision and inspection of a professional human transcriber.

During manual revision, we maintain non-speech information that potentially gives emotional state clues. The following parts of speech are given special notations: laughter, back-channel utterances, lip, nose and throat noise. All of the sessions have been automatically transcribed manually corrected.

4.3.4 Result and Analysis

Collected Data

During the course of the recordings, a large quantity of audio and video data has been recorded. After removing the overheads of pre- and post-recording periods, the combined duration of all sessions sums up to 23 hours and 41 minutes of material. On average, one session yielded 23.6 minutes of parallel audio and video data that is relevant for annotation. This time includes the opening talk prior to showing the emotion inducer, the video playback period and the discussion. The shortest and longest sessions are 10 and 33 minutes long, respectively.

SAM and Post-Recording Questionnaire

The details of the SAM and post-recording questionnaires collected throughout the data collection are laid out in Section 4.3.1. The analysis in this subsection is based on all 60 recorded sessions.

Figure 4.8 presents the proportion of ratings for all metrics in the post-recording questionnaire. The rating ranges from 1 (strong agreement) to 5 (strong disagreement). Aside from the first metric, low ratings or agreement on the questionnaire indicates a satisfying session. On average, the participants express an

https://www.popuparchive.com

agreement to varying degrees for all of the evaluated metrics. Looking at the proportions as well as the average of the ratings for each metric, we found that:

- the emotion inducer videos are effective in eliciting a negative emotional response (video_neg),
- the participants reported an agreement towards the positive emotional effect of the conversation (chat_pos),
- the participants feel that the conversation helps them to process their emotion (helps_emo),
- strongest agreement is observed on the enjoyment of the conversation (enjoyed), followed by the feeling of being understood by the expert (understood),
- emotional connection appears to be the most difficult feeling to achieve through the interaction (emo_connect), possibly due to the limited time and lack of continuity of the interaction, and
- in general, the participants express that they would like to interact with the expert again in the future (chat_again).

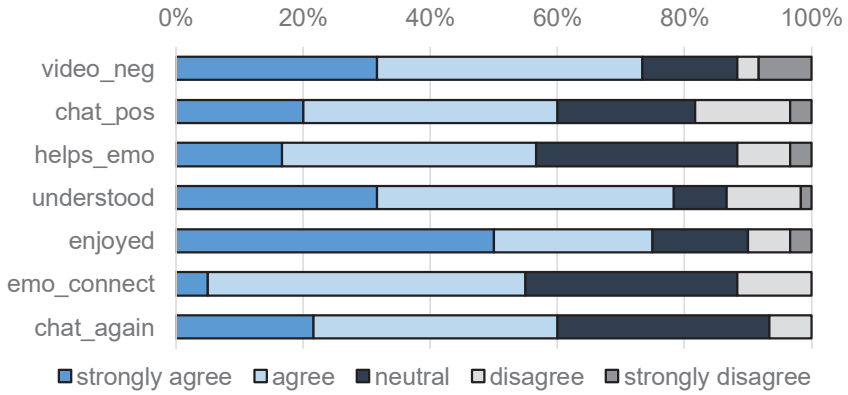
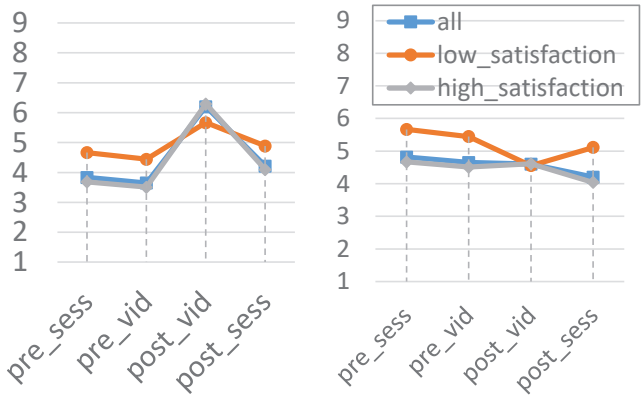


Figure 4.8: The proportion of ratings of the post-recording questionnaire. The statements of the questionnaire are detailed on Section 4.3.1.

Using the rating of all metrics except the first one (video_neg), we divided the participants into two groups: low and high satisfaction. When the average rating

of the metrics is larger than 3, i.e., suggesting disagreement, the participant is put into the low satisfaction group. Otherwise, the participant is put into the high satisfaction group.

For both groups, different trends can be observed in the SAM questionnaire, as shown in Figure 4.9. The results show significant statistical difference between the two groups' pre-recording valence ($p \leq 0.1$). Furthermore, the impact on valence of both the emotion inducer (negative) and the session (positive) is significantly more intense on the high satisfaction group compared to the low satisfaction group ($p < 0.1$). On the other hand, the two groups show opposing emotional changes in terms of arousal. We observe a statistically significant difference between the two groups in terms of arousal change after interacting with the counselor ($p < 0.05$). Figure 4.9(a) also confirms the consistent negative effect of the inducers and the role of the interaction with the counselor in recovering from it.



(a) Valence. (b) Arousal.

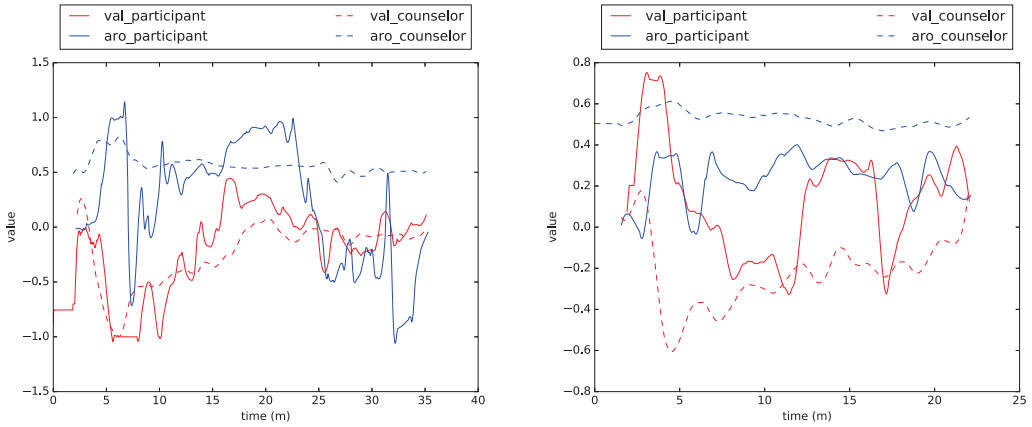
Figure 4.9: Average levels of emotion throughout the recording process. The scale ranges from 1 (strongly positive for valence and strongly activated for arousal) to 9 (strongly negative for valence and strongly deactivated for arousal).

Self-Reported and Perceived Emotion Annotation

The following analysis is based on all of the sessions in the database. We investigate the correlation between the participant's self-report of their emotion and the corresponding emotion as perceived by the expert (Section 4.3.3). We suspect that there are differences between emotion as reported by the person who experiences it and by another person who perceives it from the outside. To quantify

the agreement between the two annotations, we utilize Pearson's correlation coefficient r, computed with Equation 4.1. Pearson's r measures the strength and direction of a linear relationship between two variables. Prior to computation, we apply the Savitzky-Golay filter to smooth the annotation as well as increase the signal-to-noise ratio [91].

We found that the correlations for valence annotation are consistently stronger than that of activation. Strong correlation ($r \geq 0.5$) for valence are observed in 68.33% of the annotated sessions, while only 8.3% of the sessions have strong correlation for arousal. The average correlation is 0.585 for valence and 0.044 for arousal. Annotations from two sessions with respectively strong and weak correlations are depicted in Figure 4.10. We notice that in general the self-reported and perceived annotations are correlated more strongly when the emotion of the participant is more intense, i.e., emotion with more drastic changes, and values that reach the extremes of the scale.



(a) r for valence: 0.78, r for arousal: 0.46.

(b) r for valence: 0.32, r for arousal: 0.08.

Figure 4.10: Emotion of the participants of two sessions as annotated by the participant (**_participant*) and as annotated by the expert (**_counselor*).

4.3.5 Summary

A corpus containing recordings of dyadic social-affective interactions was constructed. The interactions between a professional counselor and 30 participants amounts to 23 hours and 41 minutes of annotated data. The proposed database

differs from existing ones in that it is explicitly designed to allow the observation of emotion changes at interpersonal level, involving an external party that guides the emotional process that follows negative emotion induction. We recruited a professional counselor to fill the role of an expert in facilitating this process.

We are aiming to provide high quality information of relevance to maximize the potential use of the collected data. We have completed the human annotation of self-reported and perceived emotions, as well as the manually refined transcriptions of the conversation. The presented corpus is designed to support affective computing research that focuses on emotion improvements at interpersonal or social level. Later chapters in this thesis present the utilization of this corpus on designing and training a dialogue system for emotion improvements elicitation.

Chapter 5

Affect-Sensitive Dialogue Response Generation for Positive Emotion Elicitation

5.1 Proposal

The HRED architecture discussed in Section 3.2.2 holds a property that is essential in positive emotion elicitation, which is not present in the RNN architecture: retaining dialogue history at turn level. Without the hierarchical structure of HRED, turn-level emotion information may be difficult to capture and utilize in model optimization and response generation. Furthermore, by having encoded dialogue history at turn level, HRED allows the consideration of multiple preceding dialogue turns without the danger of the vanishing gradient problem as the steps taken between each turn is minimized.

I propose to incorporate an *emotion encoder* into the HRED architecture, placed in the same hierarchy as the dialogue encoder. The emotion encoder captures emotion information at dialogue-turn level and maintains the emotion context history throughout the dialogue. Unlike emotion encoding at utterance level, this allows us to consider information from previous dialogue turns when modeling the emotion context. This is important since the same semantic content could signal different emotional states depending on the dialogue context. For example, "It came back," could have either positive or negative emotion, depending whether "it" refers to something negative (e.g. sickness), or positive (e.g. a

pet). Furthermore, this architecture allows parameter sharing between the emotion encoding and utterance decoding, which allows the model to be trained with fewer data.

I propose to incorporate emotion information in generating a dialogue response with an novel architecture described below, called the emotion-sensitive hierarchical recurrent encoder-decoder (Emo-HRED).

5.1.1 Emo-HRED

We utilize RNNs with GRU cells as the building blocks of the Emo-HRED. 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 an utterance representation h_{utt},

$$h_{utt} = h^{utt}_{N_m} = f(h^{utt}_{N_m-1}, w_{m,N_m}), \tag{5.1}$$

where f represents one time step operation of RNN with GRU. h_{utt} is then fed into the dialogue encoder to model the sequence of dialogue turns into dialogue context h_{dlg},

$$h_{dlg} = h^{dlg}_{m} = f(h^{dlg}_{m-1}, h_{utt}). \tag{5.2}$$

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},

$$h_{emo} = f(h^{emo}_{m-1}, h_{dlg}). \tag{5.3}$$

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

$$P_\theta(w_{n+1} = v | w_{\leq n}) = \frac{\exp\big(g(\text{concat}(h_{dlg}, h_{emo}), v)\big)}{\sum_{v'} \exp\big(g(\text{concat}(h_{dlg}, h_{emo}), v')\big)}. \tag{5.4}$$

Figure 5.1 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.

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

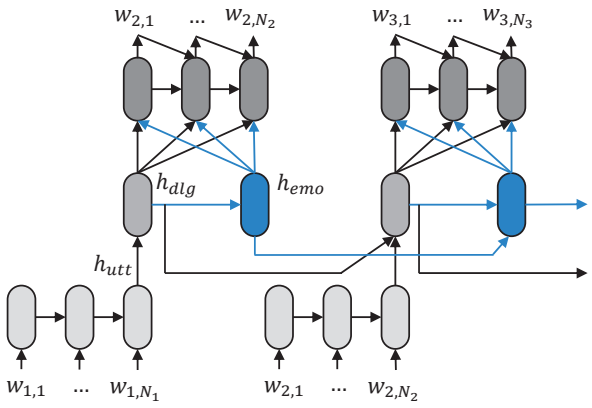


Figure 5.1: Emo-HRED architecture.

training cost to incorporate the prediction error of the emotion encoder $cost_{emo}$ and use this cost with the response generation error $cost_{utt}$ to jointly train the entire network.

We define $cost_{utt}$ as:

$$cost_{utt} = -\frac{1}{N_w}\sum_{n=1}^{N}\log P_\theta(U_1^n, U_2^n, U_3^n), \tag{5.5}$$

i.e. negative log-likelihood of N triples $(U_1^n, U_2^n, U_3^n)_{n=1}^N$ with a total number of tokens N_w under the model parameter θ. On the other hand, $cost_{emo}$ is the prediction error of the current emotion context U_m^{emo} by the emotion encoder. For real-valued emotion label, we consider mean squared error (MSE) as $cost_{emo}$,

$$cost_{emo} = \frac{1}{K}(U_m^{emo} - h_{emo})^2, \tag{5.6}$$

where K is the length of the emotion label. The training cost of the Emo-HRED is a linear interpolation between the response generation error $cost_{utt}$ and the emotion label prediction error $cost_{emo}$ with a decaying weight α,

$$cost = \alpha \cdot cost_{emo} + (1-\alpha)\cdot cost_{utt}. \tag{5.7}$$

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

5.1.2 Pretraining 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 solve this issue, pretraining the Emo-HRED with a large scale conversational corpus is essential to infer content and syntactic knowledge prior to training its emotion-related parameters. Previous works have demonstrated the effectiveness of large scale conversational data in improving the quality of dialogue systems [3, 1, 98].

Furthermore, 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 pretraining 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 pretrained encoders can be used for the affect-sensitive response generation task as is.

5.2 Experiment Set Up

5.2.1 Pretraining

In this study, we make use of SubTle, a large scale conversational corpus for pretraining (Section 4.1.1) of the HRED architecture (Section 3.2.2), where no emotion information is utilized. The pretraining is aimed to learn the syntactic and semantic knowledge for response generation. The use of movie subtitles is particularly suitable as they reflect natural human conversation and are available in large amounts.

The data preprocessing steps are performed as in [98]. The processed SubTle corpus contains 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 pretrained 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 use word embeddings of 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.

5.2.2 Fine-tuning

All the models considered in this study are the result of fine-tuning the pretrained 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.

Model. We propose the Emo-HRED architecture in place of HRED which serves as the baseline. 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. 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.

Selective fine-tuning scheme. We propose selective fine-tuning of the Emo-HRED. In the standard fine-tuning, we fine-tune all the parameters of the model. In the proposed selective fine-tuning scheme, we fix the parameters of the utterance and dialogue encoders, and train only the emotion encoder and utterance decoder (Section 5.1.2). We hypothesize that the selective fine-tuning will produce a more stable model. The SubTle and SEMAINE corpus have a number of major differences that may cause straight-forward fine-tuning to not work optimally: the sizes of the corpora differ significantly (5.5M vs. 2K triples), and the conversations are of different nature (human-human vs. human-wizard, acted speech vs. spontaneous speech).

Positive emotion elicitation data. We propose an implicit positive emotion elicitation strategy via training data. We compare fine-tuning with two datasets: the default SEMAINE dataset, using the dialogue turns as provided by the SEMAINE corpus; and the positive SEMAINE dataset, containing positive emotion

eliciting responses, i.e., U_3 produced through the process previously described (Section 4.2). We hypothesize that the positive corpus will cause the model to elicit more positive emotion. For both datasets, we consider all 66 sessions from the SEMAINE corpus. We partition the data as follows: 58 sessions (1985 triples) for training, 4 (170) for validation, and 4 (194) for test.

5.3 Evaluation

5.3.1 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 [98] and has been previously recommended for evaluating generative dialogue systems [84]. We evaluate the models using the positive SEMAINE test set, as we assume this dataset to be the one that fulfills our main goal of an emotionally positive dialogue. Table 5.1 presents the perplexity of the models fine-tuned with different set ups.

Table 5.1: Model perplexity on positive SEMAINE test set.

No	**Model**	**Fine-tuning**	**Fine-tune data**	**Perplexity**
1	Baseline HRED	standard	SEMAINE	185.66
2			positive SEMAINE	121.44
3		selective	SEMAINE	151.77
4			positive SEMAINE	100.94
5	Proposed Emo-HRED	selective	positive SEMAINE	**42.26**

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 entire network (standard scheme). This supports the hypotheses that we have previously made.

Second, we test the effect of using positive data compared to the default dataset. On models 1-4, we observe consistently lower perplexity on systems

trained on positive data. However, this does not come as a surprise as the models are tested on the positive data.

Lastly, 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 [98]. However, there are differences that are important to highlight, summarized in Table 5.2. 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 5.2: Model comparison.

	[98]	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	42.26

5.3.2 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.

To compare subjective perception between HRED and Emo-HRED, we evaluate 2 models with the best fine-tune set up, i.e. with selective fine-tune and positive SEMAINE data. 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).

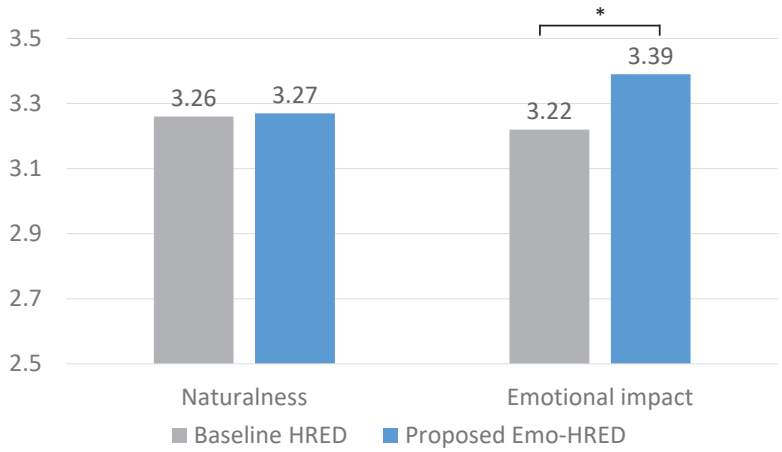


Figure 5.2: Subjective evaluation result. * denotes statistically significant difference (p<0.05).

Figure 5.2 presents the result of subjective evaluation. The proposed method is successful in improving both perceived naturalness and emotional impact. Statistically significant improvement is achieved in the emotional impact metric.

This result shows that the proposed method is able to generate responses that are perceived as more natural and elicit a more positive emotional impact.

5.3.3 Analysis

Table 5.3: 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 gonna go this far (laugh).
U_3 default	yeah.
U_3 positive	that's good yes.
HRED	yeah so you have to be inside really for the best.
Emo-HRED	i'm glad to hear.

Both the subjective and objective evaluations show consistent improvements when the proposed set ups are applied to the dialogue system. In this subsection, we analyze the generated responses to reason for these objective and subjective evaluation results. Table 5.3 shows an example of a test triple along with responses generated by the HRED and Emo-HRED with selective fine-tuning on the positive SEMAINE data.

We found that on average, the Emo-HRED model generated responses that are similar in length compared to that of HRED (8.89 vs. 8.19 words). However, the vocabulary of the Emo-HRED model contains larger proportions of positive-sentiment words. For example, on the HRED and Emo-HRED respectively, the word "good" makes up 4.9% and 13.5% of the evaluated responses, excluding stopwords. Table 5.4 lists top 10 words from these vocabularies in order of frequency.

Table 5.4: 10 most frequent words in the generated responses, excluding stop words. Positive sentiment words are bold-faced.

System	Most frequent content words
HRED	tell, **good**, think, **nice**, well, **sensible**, see, meet, know, really
Emo-HRED	hear, **good**, **glad**, tell, **nice**, **ok**, think, yeah, gonna, em

It may be of interest to note that we also observe the same tendency in the responses we collected from human annotators for the positive SEMAINE 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 positive SEMAINE 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 [58] – 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.

5.4 Summary

In this chapter, I proposed a neural network approach for affect-sensitive response generation and positive emotional impact elicitation. I extend the recently proposed HRED [98] 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 dialogue triples with positive-emotion eliciting dialogue targets (Section 4.2), the affect-sensitive system is influenced 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 significantly eliciting a more positive emotional

impact. Analysis of the evaluations suggests that the proposed method generates responses that contain more positive-sentiment words. This resembles human strategy when promoting positive emotion in a conversation with limited context.

Chapter 6

Utilizing Expert Dialogue Action for Positive Emotion Elicitation

Two main limitations of the previously discussed Emo-HRED (Chapter 5) are: 1) its inability to produce varied and long responses, echoing the long-standing problem of generic responses in end-to-end neural dialogue systems, and 2) its lack of knowledge of expert strategies in positive emotion elicitation, as it has only learned from crowdsourced dialogue responses. In this section, I describe approaches to surpass this limitation by utilizing the patterns of expert behavior in one of the previously constructed corpora.

6.1 Proposal

I propose to train a neural dialogue response generator on the previously constructed corpus containing negative emotion processing with a counselor posing as an expert (Section 4.3, from here on referred to as the *counseling corpus*). I further propose to incorporate higher level information from the expert's responses to train the affective dialogue systems to promote diversity in the generated responses. To circumvent challenges in procuring said information, I propose to employ unsupervised clustering methods to extract underlying categories of actions and behaviors from the expert's responses. The resulting labels are then utilized to train a neural dialogue system. I propose a hierarchical neural dialogue system which considers 1) expert's action, 2) dialogue context, and 3) user emotion, in generating the response.

6.1.1 Unsupervised Expert Dialogue Clustering

In constructing an emotionally intelligent system, learning from expert actions and responses is 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 underlying categories of actions and behaviors, may be highly beneficial. I hypothesize that these labels will reduce data sparsity by categorizing potential 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.").

In this regard, unsupervised clustering offers a number of benefits. First of all, it does not require pre-definition of labels, which would require expert knowledge to do. Second, it is quick to execute, and only require small computational resource. Third, it is easily performed on new data regardless of the size, making the overall approach scalable and reproducible.

Clustering Features

Two embedding methods to obtain the clustering features are considered: Word2Vec [77] and skip-thoughts vectors [50]. The suitability of each of these features for the task are described below.

Word2vec Word2vec is a model that produces word embeddings, trained to reconstruct linguistic contexts of words [77]. It takes a large corpus of text as training data, and maps each unique word in the corpus to a vector in high dimensional space, typically in the hundreds of dimensions. The training objective of the model allows words that have similar meaning or contexts to be mapped with close proximity to each other. In other words, the vector embedding of each

word is able to capture the meaning and context of the word. Various types of relationships can also be recovered by manipulating the vectors, the most famous example being the embeddings of king - man + woman equal to queen.

Word2Vec is highly suitable for the clustering task as it transforms the utterances into computationally convenient vectors while still retaining the meaning of each utterance. Furthermore, the relationship between the meanings are reflected in vector manipulation operations as explained above. By transforming each of the expert's utterances into its Word2Vec embedding vectors, we obtain a representation that is convenient for computational purposes while still preserving the semantic content of the utterances. Clustering based on these embeddings would then mean clustering the utterances based on their semantic content. That means, the found clusters would reveal sentences with similar semantic content, each cluster potentially represents an action that could be taken within the dialogue.

Skip-thoughts vectors Analogous to that of Word2Vec, skip-thoughts model learns to find a generic distributed sentence embedding [50]. It is trained on a large corpus containing sentences extracted from books, with an objective of recovering the surrounding (preceeding and proceeding) sentences based on the encoding result a sentence. Since sentences in a book have a certain continuity to them, the training objective allow sentences that share semantic and syntactic properties to be mapped to similar vector representations. The model training employ a vocabulary expansion method that allows words unseen in the training set to be encoded as well.

The motivation of utilization of this feature is similar with Word2Vec. However, unlike Word2Vec that works at word level, skip-thoughts models work at sentence level. This rids the need of additional operations required to obtain Word2Vec representation at sentence level (typically, by summation or averaging).

Clustering Methods

We propose unsupervised clustering of counselor dialogue to obtain dialogue act labels of expert responses. Two unsupervised clustering techniques are employed: K-means and Dirichlet process Gaussian mixture model (DPGMM). Henceforth,

we refer to the result of the clustering as *cluster label*.

K-means K-means is a clustering algorithm that tries to partition n observations into k clusters. k clusters are formed by finding k observations that serve as the mean, or prototype of each cluster. The rest of the observations are then considered to belong to a cluster with closest mean. The main objective is to minimize the observation variance within a cluster.

More formally, given a set of observations $\{x_1, x_2, ..., x_n\}$, where x_i is a d-dimensional vector, K-means clustering aims to cluster these observations into $k, k \leq n$ clusters $S = \{S_1, S_2, ..., S_k\}$ such that the variance within a cluster S_i is minimized, i.e.,

$$\arg\min_S \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 = \arg\min_S \sum_{i=1}^{k} |S_i| \,\text{variance}(S_i), \tag{6.1}$$

where μ_i is the mean of points in observations in S_i. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster,

$$\arg\min_S \sum_{i=1}^{k} \frac{1}{2\,|S_i|} \sum_{x,y \in S_i} \|x - y\|^2 . \tag{6.2}$$

Dirichlet process Gaussian mixture model (DPGMM) Dirichlet processes (DP) are a family of stochastic processes which are realized as probability distributions, i.e. it is a probability distribution with a range of probability distributions. The DP is most commonly utilized for clustering by inferring a mixture model, such as that in a DPGMM [86].

In contrast to K-means clustering, DPGMM is a non-parametric model, i.e. it attempts to represent the data without prior definition of the model complexity. Given a set of observations (in this case, embedding vectors of expert's utterance), a label is assigned to every data point according to a set of previously initialized mixing weights. A data point is then generated according the GMM components that corresponds to the assigned label. A more comprehensive review of the algorithm can be found in [86].

6.1.2 Hierarchical Neural Dialogue System with Multiple Contexts

We propose providing higher level knowledge about the target response to the model, in form of response cluster labels, to aid its response generation process. To allow this, we require a neural response generator capable of incorporating multiple contexts. We propose a neural dialogue system which generate response based on multiple dialogue contexts, henceforth referred to as the multi-context HRED (MC-HRED). In particular, we are concerned with three dialogue contexts: 1) dialogue history, 2) user emotional state, and 3) expert's action label.

All three contexts are modeled in the same hierarchy, i.e. dialogue turn level. In MC-HRED, the action encoder is trained to utilize dialogue history to predict the dialogue action taken at the next turn. This information is passed to the response generation process to inform the decoder of the type of response to generate. By placing the action encoder at dialogue turn level, consideration of multiple dialogue turns becomes possible, which is essential especially in predicting the action to take after a frequent dialogue action. Furthermore, this architecture design allows parameter sharing between the emotion encoding, action encoding, and utterance decoding. This is desirable as it allows more efficient use of the limited amount of data in training the model.

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},

$$h_{utt} = h^{utt}_{N_m} = f(h^{utt}_{N_m-1}, w_{m,N_m}), \tag{6.3}$$

where f is one time step operation of an RNN. h_{utt} is then fed into the dialogue encoder to model the sequence of dialogue turns into dialogue context h_{dlg},

$$h_{dlg} = h^{dlg}_{m} = f(h^{dlg}_{m-1}, h_{utt}). \tag{6.4}$$

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},

$$h_{enc} = f(h^{enc}_{m-1}, h_{enc}), \tag{6.5}$$

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,

$$P_\theta(w_{n+1} = v | w_{\leq n}) = \frac{\exp(g(\text{concat}(h_{dlg}, h_{emo}, h_{act}), v))}{\sum_{v'} \exp(g(\text{concat}(h_{dlg}, h_{emo}, h_{act}), v'))}. \tag{6.6}$$

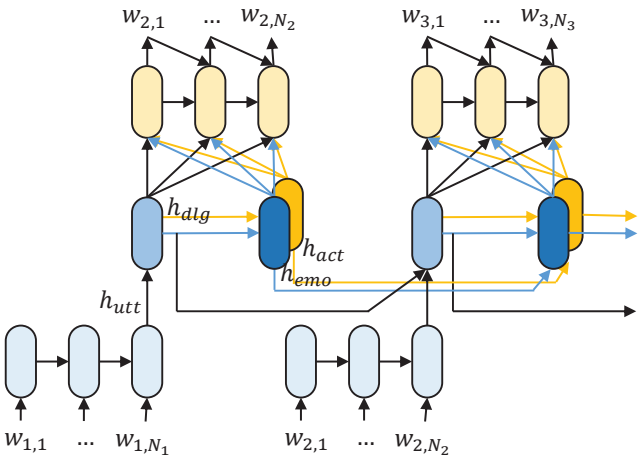


Figure 6.1: MC-HRED architecture. Emotion encoder is shown in dark blue, and action encoder in dark yellow. Blue NNs are relating to input, and yellow NNs to response.

Figure 6.1 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. 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 response generation cost $cost_{utt}$, as well as cross entropy losses of the emotion and action encoders $cost_{emo}$ and $cost_{act}$. First, we define $cost_{utt}$ as:

$$cost_{utt} = -\frac{1}{N_w} \sum_{n=1}^{N} \log P_\theta(U_1^n, U_2^n, U_3^n), \tag{6.7}$$

i.e. negative log-likelihood of N triples $(U_1^n, U_2^n, U_3^n)_{n=1}^N$ with a total number of tokens N_w under the model parameter θ.

The emotion and action encoders have their own respective target vectors. The emotion encoder predicts the emotion label of the current dialogue turn, and the action encoder predicts expert's action label of the target response. For discrete labels of K classes, i.e $c = \{1, ..., K\}$, the cost can be computed as multiclass cross entropy:

$$cost_{enc} = -\sum_{c=1}^{K} U_m^{enc,c} \log(h_{enc}^c), \tag{6.8}$$

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

The training cost of the MC-HRED is a linear interpolation between the response generation error $cost_{utt}$ and the prediction errors of the encoders $cost_{emo}$ and $cost_{act}$ with decaying weights α and β,

$$cost = (1 - \alpha - \beta) \cdot cost_{utt} + \alpha \cdot cost_{emo} + \beta \cdot cost_{act}. \tag{6.9}$$

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

6.2 Experiment Set Up

6.2.1 Unsupervised Clustering

A total of 6384 counselor utterances are collected from the counseling corpus to form the response clusters. A vector representation is computed for each utterance as the basis of the clustering. The features and method used in this study is explained in this section.

Clustering Features

As Word2Vec works on word level, the utterances are transformed into vectors by obtaining the embeddings of each words in the utterance and averaging them. A Word2Vec model pretrained on 100 billion words of Google News is used. The vocabulary size of the model is 3 million. The word and utterance embeddings are of length 300.

Unlike Word2Vec, skip-thoughts vectors work at a sentence level. The sentence embedding can be straightforwardly used in the clustering process. The

publicly available pretrained model is used in the experiment [50], with vocabulary size of 1 million words. The model results in embedding vectors of length 4,800.

Clustering Methods

In our experiment with K-means, we perform hierarchical clustering, starting with an initial K of 8, chosen empirically. We perform K-means clustering the second time on the clusters which are larger than half the full data size. This is to allow better characterization, and in turn, understanding of the resulting cluster.

For the DPGMM, We use the stick-breaking construction to generate the mixing weights. As previously stated, new data point would either join an existing GMM component or start a new one following the mixing weight. In this case, we view data points with the same component label as a cluster. We use diagonal covariance matrices to compensate for the limited amount of data. As it is a non-parametric model, the final component size relies solely on the data.

6.2.2 MC-HRED

Pretraining

As in Chapter 5, in the following experiments, we utilize the HRED trained on the SubTle corpus as our starting model. We follow the data preprocessing method in [98]. 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.

However, the selection of system vocabulary is modified. 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 pretrained by feeding the SubTle dataset sequentially into the network until it converges. 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.

Fine-tuning

All the models considered in this study are the result of fine-tuning the pretrained model with the counseling corpus (Section 4.3). We partitioned the counseling corpus into 50 recording sessions (5053 triples) for training, 5 (503) for validation, and 5 (508) for testing. 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 to test the effectiveness of the proposed method and features.

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.

We obtain the action context as follows. We first extract triples from the counseling corpus, with counselor-participant-counselor speaker order. U_3 of each triples are then transformed into its Word2Vec and skip-thoughts sentence embeddings. Each of the Word2Vec and skip-thoughts embeddings are then fed into the hierarchical K-means and DPGMM clustering. This process results in 2 action labels for each of the counselor's response. Each action label is experimented with, separately, as the action context to the MC-HRED. The cluster

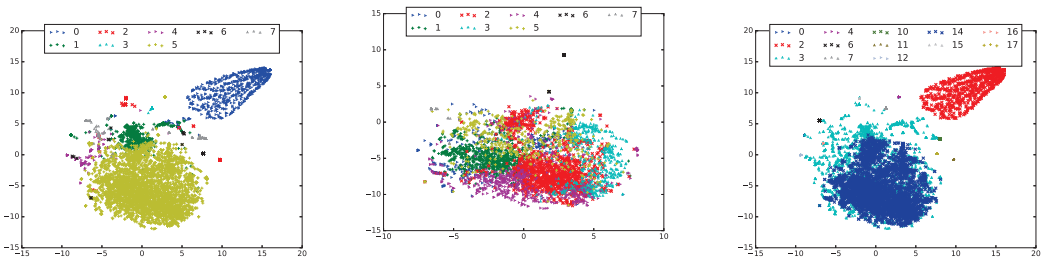
sizes are 15 (K-means) and 13 (DPGMM), and 8 (K-means) and 5 (DPGMM) for the skip-thoughts.

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 [69]. Multi-class cross entropy is used to compute the encoder costs.

6.3 Cluster Analysis

We analyze clusters formed by Word2Vec and skip-thoughts embeddings separately. We visualize the found clusters using T-SNE and elaborate on them in the following sections.

6.3.1 Word2Vec



(a) K-means clustering.

(b) K-means sub-clustering on cluster 5.

(c) DPGMM clustering.

Figure 6.2: T-SNE Representation of the clustering results with Word2Vec embedding. Best viewed in digital format.

K-means clustering shows distinct dialogue acts characteristic in a number of clusters it found. For example, cluster 0 in Figure 6.2(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 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 6.2(a)).

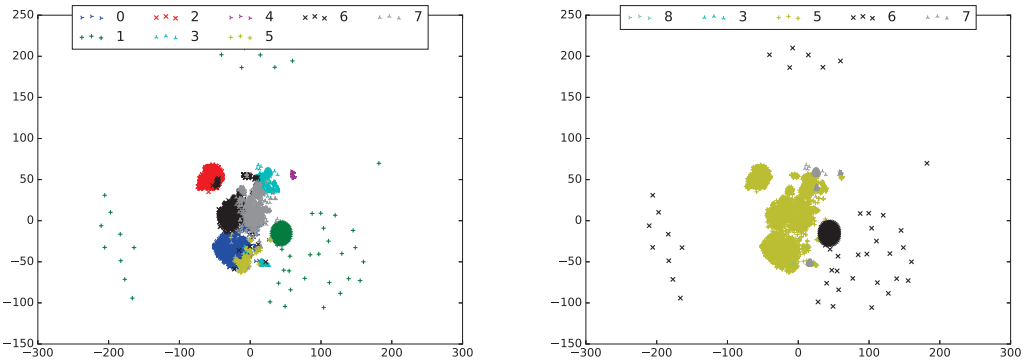
Second clustering on cluster 5 group these utterances into smaller sub-clusters (Figure 6.2(b)). "I" is the most frequent word in sub-cluster 0, and "you" in sub-cluster 1. Some of the actions from the first clustering are re-discovered during the second clustering, such as thanking and appreciation 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.

6.3.2 Skip-thoughts

Figure 6.3 shows the T-SNE visualization of the skip-thoughts vectors and clustering result.

We observe four distinct characteristics in the eight clusters formed by K-means clustering. Backchannel utterances and other non speech sounds are contained in clusters 1, 3, and 4. Questions appear to be separated from other utterances, clustered together in cluster 2. This is also the case for affirmative responses such as "Yeah," "Yes," and "Right." The rest of the clusters, i.e. clusters 0, 6, and 7, contain the rest of the counselor utterances. These clusters are difficult to characterize as the contained words are similar. However, there is a considerable difference between the length of utterances within the clusters. The



(a) K-means clustering.

(b) DPGMM clustering.

Figure 6.3: T-SNE representation of the clustering results with skip-thoughts embedding. Best viewed in digital format.

average lengths for clusters 0, 6, and 7 respectively are 14, 43, and 22 tokens.

The DPGMM clustering with skip-thoughts yields 5 clusters. Like with K-means, backchannels (clusters 3, 6, and 7) and affirmative utterances (cluster 8) can be clustered separately from the rest. However, the rest of the utterances are not separated further apart, clustered together in cluster 5. Cluster 5 is very large, consisting of 5126 utterances, and as such present to distinct characteristic. As previously mentioned, sub-clustering of a DPGMM component is not possible.

Unlike Word2Vec which highlights word meaning in its embedding, the skip-thoughts vector appears to highlight structural differences between the utterances instead. We observe the presence of multiple clusters containing utterances with the same function, with only difference in length, e.g. clusters 3, 6, 7 of DPGMM clustering, which all contain backchannels, differing only in length.

6.4 Evaluation

6.4.1 Objective Evaluation

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. We compute the perplexity for each triple and average it to obtain model perplexity. The

model perplexities are summarized in Table 6.1. 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.	Clustering	Perplexity		
			all	short	long
Emo-HRED	Yes	No	42.60	35.74	61.17
Word2Vec					
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
Skip-thoughts					
Clust-HRED	No	K-means	34.19	30.35	44.35
		DPGMM	30.17	27.78	38.39
MC-HRED	Yes	K-means	36.15	32.54	46.34
		DPGMM	32.24	28.58	39.88

Table 6.1: Comparison of model perplexities.

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, the cluster labels are affected in starkly different ways when combined with emotion labels: Word2Vec K-means gain a significant improvement, while the rest slightly suffers. We found that on long triples, Clust-HRED and MC-HRED yield similar performances when using the most cluster labels. However, when using Word2Vec K-means label, MC-HRED shows significant improvement from Clust-HRED.

We believe the quality of the dialog cluster label is an important aspect in determining the success of combination with emotion information. As discussed in Section 6.3, K-means clustering with Word2Vec features produces clusters with

most informative labels that have distinct characteristics and intents. Multiple contexts in response generation is likely to only be beneficial if there are a certain relationship or correlation between the contexts that can be exploited in modeling the data.

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.

6.4.2 Subjective Evaluation

We evaluate Emo-HRED and the best performing MC-HRED, i.e the model using K-means cluster label with Word2Vec vector. We extract 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 6.4 summarizes the subjective evaluation result.

Figure 6.4: Human subjective evaluation 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 6.2.

Table 6.2: 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.

6.5 Summary

In this chapter I addressed two shortcomings of the proposed Emo-HRED (Section 5), which are 1) its inability to produce varied and long responses, and 2) its lack of knowledge of expert strategies in positive emotion elicitation. To overcome this, I propose to incorporate higher level information from the target responses to train the affective dialogue systems to promote diversity in the generated responses. It is then essential to train the dialogue system on data involving an expert in the dialogue, such as that we have previously constructed which involves a professional counselor (Section 4.3).

The expert's responses are unsupervisedly clustered and the resulting labels used to train a dialogue system. I proposed a novel hierarchical neural architecture for response generation that is conditioned on multiple contexts: 1) expert's action, 2) dialogue context, and 3) user emotion, encoded from user input. The objective evaluation we conducted showed 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.

Chapter 7

End-to-End Positive Emotion Elicitation Through Reward Optimization

In Chapter 5, I have proposed a model that encodes emotion information from user input and utilizes it in generating response for a flexible end-to-end scalable dialogue system. Subsequently in Chapter 6, limitations of the Emo-HRED in terms of training data and response quality are highlighted and solved. While subjective evaluations show affirming results, these proposed approaches still rely on maximizing the likelihood of the target responses to elicit positive emotion. Supported by its emotion encoding capability, the models have to be trained on data showing positive emotion elicitation to achieve its objective. In other words, the models are trained without consideration of emotional impact, i.e. the change of emotional state a response may cause on the listener. Such consideration sets the difference between positive emotion elicitation by pure imitation and by awareness of emotional impact on the system side.

In this chapter, I propose to explicitly utilize emotional impact information to optimize neural dialogue system towards generating responses that elicit positive emotion. Leveraging this information allows us to promote responses that elicit positive emotion, and suppress those that has negative impact. To test the proposed method in a wide range of dialogue situations, we consider two emotion-rich corpora collected in different scenarios: Wizard-of-Oz and spontaneous.

7.1 Proposal

7.1.1 Impact HRED

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

$$h_{utt} = h^{utt}_{N_m} = f(h^{utt}_{N_m-1}, w_{m,N_m}), \tag{7.1}$$

where f represents one time step operation of RNN with GRU. h_{utt} is then fed into the dialogue encoder to model the sequence of dialogue turns into dialogue context h_{dlg},

$$h_{dlg} = h^{dlg}_{m} = f(h^{dlg}_{m-1}, h_{utt}). \tag{7.2}$$

The h_{dlg} is then fed into the emotion encoder, which will then be used to model the emotion context h_{emo},

$$h_{emo} = f(h^{emo}_{m-1}, h_{dlg}). \tag{7.3}$$

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

$$P_\theta(w_{n+1} = v | w_{\leq n}) = \frac{\exp\big(g(\text{concat}(h_{dlg}, h_{emo}), v)\big)}{\sum_{v'} \exp\big(g(\text{concat}(h_{dlg}, h_{emo}), v')\big)}. \tag{7.4}$$

Extending the emotion-sensitive concept of Emo-HRED, we propose to incorporate emotional impact information in the training process to influence the system to generate responses with more positive emotional impact. We linearly interpolate three costs to optimize the network. The first is response generation error $cost_{utt}$,

$$cost_{utt} = -\frac{1}{N_w} \sum_{n=1}^{N} \log P_\theta(U^n_1, U^n_2, U^n_3), \tag{7.5}$$

i.e. negative log-likelihood of N triples $(U^n_1, U^n_2, U^n_3)^N_{n=1}$ with a total number of tokens N_w under the model parameter θ.

The second cost, $cost_{emo}$ is computed from the prediction error of the emotion encoder given its target vector U_m^{emo}, which is the emotion label of the currently processed dialogue turn U_m. For discrete labels of K classes, i.e $c = \{1, ..., K\}$, the cost can be computed as multiclass cross entropy:

$$cost_{emo} = -\sum_{c=1}^{K} U_m^{emo,c} \log(h_{emo}^c). \tag{7.6}$$

Lastly, we consider the emotional impact of the target response, $cost_{impact}$,

$$cost_{impact} = impact_{U_{m+1}} = U_{m+2}^{emo} - U_m^{emo}. \tag{7.7}$$

This information is extracted from data and fed during training as additional label. The final *cost* is defined as

$$cost = \alpha \cdot cost_{emo} + \beta \cdot cost_{utt} \cdot cost_{impact} + (1 - \alpha - \beta) \cdot cost_{utt}. \tag{7.8}$$

As emotional impact is dependent on the utterance, we weight the ground-truth impact collected from the data with the NLL of the target, i.e. $cost_{utt}$. Negative emotional impact will result in high $cost_{impact}$, and inversely for positive impact, thus promoting responses that elicit positive emotion, and suppressing those that have negative impact. We give $cost_{utt}$ and $cost_{impact}$ each a decaying weight α and β. The final cost is then propagated to the network and the parameters are optimized as usual with the optimizer algorithm. As with Emo-HRED, we believe pretraining and selective fine-tuning is essential for achieving a solid Impact-HRED performance.

7.2 Experiment Set Up

7.2.1 Pretraining

In our experiments, we utilize the HRED [98] trained on the SubTle corpus as our starting model. The data preprocessing steps are performed as in [98]. The processed SubTle corpus contained 5,503,741 query-answer pairs in total.

We pretrain two models with different vocabularies. The first model has the 10,000 most frequent tokens in SubTle as the system's vocabulary, and the rest

as unknowns. For the second model, we combine SubTle and the counseling data before selecting the 10,000 most frequent tokens. As the counseling corpus shows only small vocabulary overlap with SubTle, this step is necessary to preserve important tokens in the counseling corpus that are not present in the SubTle corpus.

The models are pretrained 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.

7.2.2 Fine-tuning

We perform fine-tuning with the SEMAINE and counseling corpora, separately. The SEMAINE corpus is used to fine-tune the pretrained model with only SubTle vocabulary, and the counseling corpus to fine-tune the model with combined vocabulary (see Section 7.2.1).

We fine-tune two models for comparison: HRED as the baseline method, and Emo-HRED optimized with emotional impact information as the proposed method. The baseline method is HRED, optimized on NLL of target response without any emotion information. While the proposed model is optimized on Equation 7.8.

As emotion context, we encode the emotion annotation into a one-hot vector representation. 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.

To compute emotional impact (Equation 7.7), we solely consider the average value of valence. This is because our goal is to elicit positive emotion, regardless of the arousal level. However, it is important to note that arousal level still provides useful contextual information for the generation process. For exam-

ple, eliciting positive emotion from anger (negative-high arousal) would require distinct strategy than from sadness (negative-low arousal).

We perform selective fine-tuning as in previous experiments. That is, we fix the utterance and dialogue encoders parameters, and selectively train only the emotion encoder as well as the decoder. To accommodate the emotion context during fine-tuning, we append new randomly initialized parameters to the utterance decoder. These parameters are trained exclusively during the fine-tuning process along with the newly initialized emotion encoder and the pretrained decoder parameters. In this study, we define the emotion encoder as an RNN with GRU cells and sigmoid activation function. We empirically choose both α and β to be 0.3.

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. We use all 60 sessions of the counseling corpus as partition them as follows: 50 recording sessions (5053 triples) for training, 5 (503) for validation, and 5 (508) for testing.

7.3 Evaluation

7.3.1 Objective Evaluation

As with previous experiments, the proposed models are objectively evaluated by computing the model perplexity. We evaluate the fine-tuned models on their respective test sets, taking the average perplexity across the test set as model perplexity. The results are presented in Table 7.1.

Table 7.1: Model perplexity on respective test sets. Best perplixity is bold-faced.

Model	**SEMAINE**	**Counseling**
Baseline	167.44	34.95
Proposed	**29.48**	**30.29**

The dialogue turns on the counseling data tend to be longer due to the nature

of the conversation. Comparison across methods and corpora shows that the proposed method consistently yields lower perplexity compared to the baseline. The baseline method yields contrasting performances across the two corpora, whereas the proposed method shows stable performance. In all evaluations, the perplexity on queries longer than average is consistently higher than on shorter ones.

7.3.2 Subjective Evaluation

Table 7.2: Subjective evaluation scores. Average and standard deviation (in brackets) accross all test triples are shown. * denotes $p < 0.05$ compared with baseline method. Highest scores are bold faced.

Model	SEMAINE		Counseling	
	Naturalness	Emotional impact	Naturalness	Emotional impact
Baseline	3.37 (0.72)	3.10 (0.60)	3.92 (0.44)	3.32 (0.30)
Proposed	**3.83** (0.64) *	**3.46** (0.64) *	**4.05** (0.34)	**3.35** (0.26)

We perform the subjective evaluation through crowdsourcing. We present human judges with a dialogue triple and ask them to rate the response in terms of two criteria: 1) naturalness, which evaluates whether the response is intelligible, logically follows the dialogue context, and resembles real human response, and 2) emotional impact, to measure whether the response elicits a positive emotional impact or promotes an emotionally positive conversation.

We evaluate 100 triples from each the SEMAINE and counseling test sets. Each test query 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. 20 human judgements are collected for a test query. 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. The agreement is given using a Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). Table 7.2 summarizes the subjective evaluation result.

The proposed method is consistently perceived as more natural and having a more positive emotional impact. Significant improvement is observed on the model trained with SEMAINE data, but not with counseling data. We achieve

best naturalness with counseling data, and most positive emotion impact with SEMAINE data.

An important thing to note is the different emotion elicitation strategies between the two corpora. SEMAINE data rely on consistent personalities of the wizards, which is portrayed in a very distinctive manner through the response. On the other hand, the counselor on the counseling data uses counseling techniques to induce emotion processing and regulation. The strategy in the SEMAINE data is more straightforward and thus perceptible in the short context available on subjective evaluation. This difference may explain the higher emotional impact score for the models trained on SEMAINE data. On the other hand, the counselor response is be closer to real human dialogues, which we suspect to be the reason for the higher naturalness score.

Table 7.3 presents comparison of system responses for two of the test queries. In both examples, the proposed method is able to generate a response with a more positive sentiment than the target while still maintaining dialogue and topic coherence.

Table 7.3: Comparison of system responses. Top example is taken from SEMAINE data, bottom example from counseling data.

U_1	ok i see so what do you do at your work.
U_2	ah i read papers i carry out experiments i go for long lunches.
U_3 (Target)	and do you have an organized approach to your experiment.
Baseline	i try to insult you too.
Proposed	you're lucky.
U_1	oh thanks so much for sharing your ideas.
U_2	you're welcome.
U_3 (Target)	ah maybe we should finish here.
Baseline	so i guess we've run out of time now.
Proposed	how did you feel about it?

7.4 Summary

This chapter presented a method in optimizing neural dialogue systems with emotional impact information to generate responses that elicit positive emotion. Exploiting emotional impact information allows us to promote responses that elicit positive emotion, and suppress those with negative impact. This is similar to endowing the system with awareness of emotional impact, and giving in an objective to simulate responses with positive impact. Experiments with Wizard of Oz and spontaneous data show that the proposed method yields lower perplexity, as well as judged to be more natural and likely to elicit a more positive emotion.

Chapter 8

Multi-modal Emotion Encoding for Affect-Sensitive Response Generation

In this chapter we experiment with extending the Emo-HRED architecture to utilize multi-modal information for emotion encoding. In particular, we examine the use of acoustic features in addition to the text features. The following sections describe the extended architecture and experiments, continued with an in depth analysis of the evaluation results.

8.1 Proposal

Speech has been argued to be the richest channel of communication, containing paralinguistic informations including emotion and affect. Various emotion recognition studies have shown the usefulness of speech features in predicting emotional states [64, 22]. To benefit from this source of information, we consider utilizing additional features in modeling the emotion context within Emo-HRED. In particular, by using a fully connected neural network to encode the dialogue turn's acoustic feature aud_m into h_{aud} and feeding it into the emotion encoder,

$$h_{aud} = aud_m \cdot W_{aud} + b_{aud}. \tag{8.1}$$

In this case, Equation (5.3),

$$h_{emo} = h^m_{emo} = f\big(h^{emo}_{m-1}, h_{dlg}\big), \tag{5.3}$$

is modified as follows to incorporate acoustic feature information:

$$h_{emo} = h^m_{emo} = f\big(h^{emo}_{m-1}, \mathrm{concat}(h_{dlg}, h_{aud})\big), \tag{8.2}$$

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

$$P_\theta(w_{n+1} = v | w_{\leq n}) = \frac{\exp\big(g(\mathrm{concat}(h_{dlg}, h_{emo}), v)\big)}{\sum_{v'} \exp\big(g(\mathrm{concat}(h_{dlg}, h_{emo}), v')\big)}. \tag{8.3}$$

Figure 8.1 shows a schematic view of this architecture. Audio encoder is shown at the lowest hierarchy.

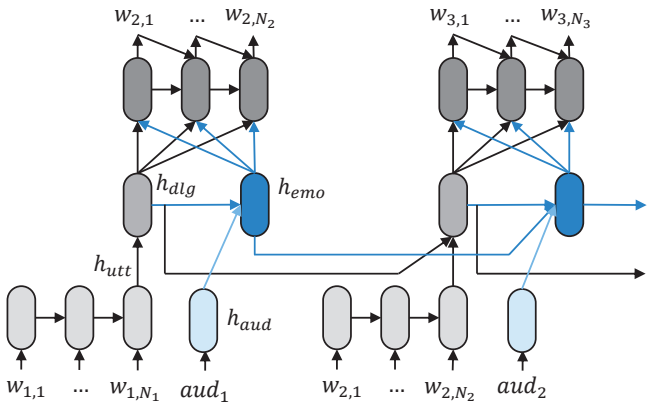


Figure 8.1: Emo-HRED architecture with audio encoder for emotional context encoding.

As with Emo-HRED in Chapter 5, 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 $cost_{emo}$ and use this cost with the response generation error $cost_{utt}$ to jointly train the entire network.

We define $cost_{utt}$ as:

$$cost_{utt} = -\frac{1}{N_w} \sum_{n=1}^{N} \log P_\theta(U^n_1, U^n_2, U^n_3), \tag{8.4}$$

i.e. negative log-likelihood of N triples $(U_1^n, U_2^n, U_3^n)_{n=1}^N$ with a total number of tokens N_w under the model parameter θ. On the other hand, $cost_{emo}$ is the prediction error of the current emotion context U_m^{emo} by the emotion encoder. For real-valued emotion label, we consider MSE as $cost_{emo}$,

$$cost_{emo} = \frac{1}{K}(U_m^{emo} - h_{emo})^2, \tag{8.5}$$

where K is the length of the emotion label. The training cost of the Emo-HRED is a linear interpolation between the response generation error $cost_{utt}$ and the emotion label prediction error $cost_{emo}$ with a decaying weight α,

$$cost = \alpha \cdot cost_{emo} + (1 - \alpha) \cdot cost_{utt}. \tag{8.6}$$

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

8.2 Experiment Set Up

8.2.1 Pre-training

Pre-training in this experiment is identical to that of Chapter 5. We utilize the HRED trained on the SubTle corpus (Section 4.1.1) as our starting model. The data pre-processing steps are performed as in [98]. The processed SubTle corpus contains 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.

In addition to the model parameters, during pre-training we also learn the word embeddings of the tokens. We use word embeddings of 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.

8.2.2 Fine-tuning

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. We experi-

ment with a larger amount of set ups to gain deeper insight from the evaluations and analyses.

Model. We propose the extended Emo-HRED architecture in place of HRED which serves as the baseline. 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. 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. We use mean-squared error to compute the training cost of emotion encoder in Emo-HRED.

Selective fine-tuning scheme. We propose selective fine-tuning of the Emo-HRED. In the standard fine-tuning, we fine-tune all the parameters of the model. In the proposed selective fine-tuning scheme, we fix the parameters of the utterance and dialogue encoders, and train only the emotion encoder and utterance decoder. We hypothesize that the selective fine-tuning will produce a more stable model. The SubTle and SEMAINE corpus have a number of major differences that may cause straight-forward fine-tuning to not work optimally: the sizes of the corpora differ significantly (5.5M vs. 2K triples), and the conversations are of different nature (human-human vs. human-wizard, acted speech vs. spontaneous speech).

Positive emotion elicitation data. We propose an implicit positive emotion elicitation strategy via training data. We compare fine-tuning with two datasets: the default SEMAINE dataset, using the dialogue turns as provided by the SEMAINE corpus; and the positive SEMAINE dataset, containing positive emotion eliciting responses, i.e., U_3 produced through the process previously described. We hypothesize that the positive corpus will cause the model to elicit more positive emotion. For both datasets, we consider all 66 sessions from the SEMAINE corpus. We partition the data as follows: 58 sessions (1985 triples) for training, 4 (170) for validation, and 4 (194) for test.

Audio encoder. We propose utilizing acoustic features in combination with dialogue state for emotion encoding. We suspect that some affective information that is essential in determining emotional context is lost when the observation is limited to text only. When acoustic features are included, Equation 8.2 is used in

place of Equation Equation 5.3. We extract the INTERSPEECH 2009 emotion challenge baseline features [97] using the OpenSMILE toolkit [29]. The audio encoder is of size 200, randomly initialized and exclusively trained during fine-tuning. Audio information is solely used as additional information for emotion encoding. Since speech recognition is out of the scope of this paper, we use the transcription as the text input for Emo-HRED.

Emotion encoder prediction. We investigate the effect of different emotion inputs for the utterance decoder during fine-tuning: feeding the target emotion vector into utterance decoder, or using the prediction of emotion encoder instead. Note that in either scenario, the emotion prediction is used for evaluation. We suspect that using the prediction of emotion encoder leads to better optimization; when the emotion target is used during training, the error of emotion encoder is not propagated to the output, creating a disconnect between generated response and emotion prediction.

8.3 Evaluation

We perform two evaluations to confirm the effectiveness of the proposed method. First, we perform objective evaluation of the systems by computing the model perplexity. Second, we perform subjective evaluation to measure the naturalness and emotional impact of the generated responses. The result of both evaluations are summarized in Table 8.1. To obtain deeper insight into the evaluation results, analyses are provided at the end of this section.

8.3.1 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 rely on probabilistic approaches [98] and has been previously recommended for evaluating generative dialogue systems [84]. We evaluate the models using the test set of the positive SEMAINE data, as we assume this dataset to be the one that fulfills our main goal of an emotionally positive dialogue. The "Perplexity" column of Table 8.1 presents the perplexity of the models with different fine-tuning set ups. In the interest of space and readability, we iteratively choose the best option of the proposed set ups. That is, we keep a set up fixed when it has shown consistent

Table 8.1: Emo-HRED evaluation results. Each of the proposed methods is incrementally compared. Objective evaluation is measured in "Perplexity." Subjective evaluation is measured in "Naturalness" and "Emotional impact." Best number for each metric is bold-faced. On subjective evaluation, * denotes significant difference ($p<0.05$) with best model (No. 10). Highlighted systems (No. 3, 4, 8, and 10) are further analyzed in the following subsection.

No.	**Model**	**Selective fine-tune**	**Positive data**	**Audio encoder**	**Use prediction**	**Perplexity**	**Naturalness**		**Emotional impact**	
							avg	std. dev	avg	std. dev
1	Baseline HRED	No	No	No	No	185.66	n/a			
2			Yes			121.44				
3		Yes	No			151.77	2.71 *	0.31	2.56 *	0.29
4			Yes			100.94	3.26 *	0.22	3.22 *	0.25
5	Proposed Emo-HRED	Yes	No	No	No	41.30	3.38 *	0.39	3.22 *	0.35
6			Yes			42.26	3.27 *	0.38	3.39 *	0.30
7			No	Yes		42.38	3.49 *	0.34	3.36 *	0.28
8			Yes			37.42	3.72	0.25	3.70	0.21
9		Yes	Yes	No	Yes	35.92	3.43 *	0.31	3.51 *	0.29
10				Yes		**20.35**	**3.75**	0.22	**3.78**	0.17

improvement on a number of systems.

We test the effect of the parameter update on both positive and default datasets by keeping the model fixed to HRED. We observe significant improvements when fine-tuning only the decoder compared to fine-tuning the entire network (rows No. 1-4). 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. We train the rest of the set ups with the selective fine-tuning scheme.

We test the impact of the emotion encoder by comparing HRED and Emo-HRED (rows No. 3-6). 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 of 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.

We found that the models trained with positive SEMAINE data tends to yield lower perplexity than those trained with the default dataset (rows No. 1-8), with only an exception between rows 5 and 6. The model perplexity further shows that the incorporation of audio information for emotion encoding allows significant improvement when positive data is used (rows No. 6 and 8), but not when default data is used (rows No. 5 and 7). This suggests that the audio information could allow emotion encoder to form a representation of the emotional context that further helps model the data better. We see consistent improvement by using the prediction of the emotion encoder for utterance decoding during fine-tuning (rows No. 6 to 9, and 8 to 10), reaching the best perplexity of 20.35.

We further test the models utilizing audio encoder for emotion encoding (rows No. 7, 8, and 10), by using the speech input for the utterance encoding as well. Instead of taking the transcription as input, we utilize an automatic speech recognizer (ASR), and use the automatic transcription result of user turn during utterance encoding. We user Google Speech-to-Text service as the ASR system.

Table 8.2: Test perplexity of models utilizing audio. Test on human transcription means that speech is only used for emotion encoding. With ASR, speech is also used prior to utterance encoding.

No.	**Model**	**Selective fine-tune**	**Pos. data**	**Audio encoder**	**Use prediction**	**Perplexity**	
						Human Trans.	**ASR**
7	Proposed Emo-HRED	Yes	No	Yes	No	42.38	45.52
8			Yes		No	35.92	41.46
10			Yes		Yes	20.35	21.97

The Google ASR system yields 45.71% word error rate (WER) on the user test utterances. Two of the major challenges are 1) recognizing very short utterances such as backchannel which appears frequently in dialogue, and 2) processing accented speech, where some the participants are L2 speakers currently living in Belfast area. Table 8.3.1 presents model perplexity on these two test conditions. We observe only small degradation in overall dialogue system performance. This signals that the proposed systems are quite robust to speech recognition errors.

8.3.2 Subjective Evaluation

We perform human subjective evaluation via crowdsourcing. We exclude systems not fine-tuned with the selective scheme due to the poor quality of the generated responses. We present human judges with a dialogue triple and ask them to rate the response on 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 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).

Columns "Naturalness" and "Emotional impact" in Table 8.1 show the results of the subjective evaluation. With identical fine-tune set ups, compared to HRED, Emo-HRED is consistently perceived as more natural and eliciting a more positive emotional impact (rows No. 3-6). We observe consistent improvements when comparing models trained on the default and positive SEMAINE data (rows No. 3 and 4, 5 and 6, 7 and 8). The subjective ratings further show that utilization of audio encoder gives perceptible improvements (rows No. 5-8). When emotion prediction is used during fine-tuning, significant improvements are observed for models without emotion encoder (rows No. 6 and 9), and slight improvement is observed for models with encoder (rows No. 8 and 10). We also notice that the standard deviation of the ratings to diminish as the model improves. We perform t-test to measure whether the improvement of the best model in subjective evaluation is significant. For both naturalness and emotional impact, model No. 10 shows significant improvement compared to all models, except model No. 8.

One of the point I put forth in Section 1.4 is that eliciting emotion improvements does not translate to responding with positive emotion at all times. To prove that the proposed model is superior to a system that aims to constantly output "happy responses," an additional subjective evaluation was conducted. As the system with "happy responses", I consider an EBDM system with only Poppy and Prudence, i.e. characters with positive valence, responses in the example database. The selection criteria of the system is cosine similarity between user query and example query. We perform identical subjective evaluation of the system and compare the result with baseline HRED, best proposed Emo-HRED. The results are presented in Figure 8.2. It is shown that the proposed system is perceived as significantly more natural and elicit a more positive impact, compared to a system that is designed to always output positive responses.

8.3.3 Analysis

Generated Responses

We analyze the generated responses to reason for the objective and subjective evaluation results. In this analysis, we consider models No. 3 and 4 as baseline models, and 8 and 10 as best proposed models. These models are highlighted in Table 8.1.

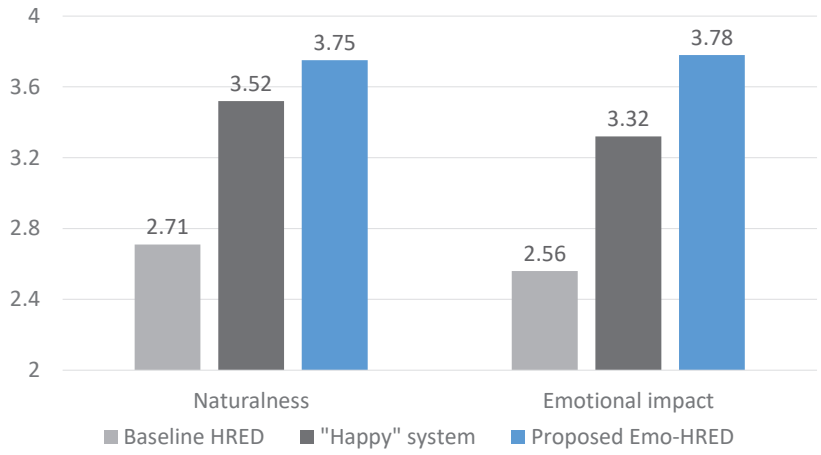


Figure 8.2: Subjective evaluation results for baseline HRED (Model No. 3), "happy" system, and best proposed Emo-HRED (Model No. 10). All score differences are statistically significant ($p<0.05$).

Table 8.3 shows an example of a test triple along with responses generated by the models. This example demonstrates the disconnect between perplexity and subjective perception of a dialogue response. While the response from model No. 10 is more similar to the target response, both are arguably natural and have potential in eliciting positive emotional impact. This could explain the different trends between subjective and objective evaluation results (Table 8.1).

Table 8.3: Comparison of system responses for a triple in the 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 gonna go this far [laugh].
U_3 SEMAINE	yeah.
U_3 positive SEMAINE	that's good yes.
Model No. 3	cause you don't really want to go out with people.
Model No. 4	yeah so you have to be inside really for the best.
Model No. 8	it's good to hear that.
Model No. 10	that's good i hope that.

We found that on average, the Emo-HRED models generated responses that are shorter compared to that of HRED (5.54 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 systems No. 3, 4, 8, and 10 respectively, the word "good" makes up 2.4%, 4.9%, 25.6%, and 26.3% of the evaluated responses, excluding stop-words. Table 8.4 lists the top 10 words from these vocabularies in order of frequency, as well as the positive SEMAINE vocabulary for reference.

Table 8.4: 10 most frequent words in the responses, excluding stop words. Positive sentiment words are bold-faced.

No. 3	tell, well, **like**, **good**, think, make, else, go, get, know
No. 4	tell, **good**, think, **nice**, well, **sensible**, see, meet, know, really
No. 8	**good**, able, yeah, tell, well, hear, oh, ok, **nice**, aha
No. 10	**good**, **hope**, **happy**, makes, yeah, **nice**, meet, well, ok, aha
positive SEMAINE	**good**, think, **laugh**, oh, well, **like**, things, else, **excellent**, tell

These analyses align with that presented in Chapter 5, and thus pose as supporting evidence of the effectiveness of the proposed approach.

Emotion Encoding Performance

To further analyze the role of emotion in the response generation, we perform analysis of the emotion encoding performance. We compare emotion of user turn as predicted by the emotion encoder with the ground truth emotion context obtained from the corpus annotation (see Section 8.2). Real-valued emotion label is used in these experiments, therefore to measure the emotion recognition performance, we calculate the mean squared error (MSE) of the emotion context prediction on the test set. The result is presented on Table 8.5.

We compute Pearson's correlation coefficient r to find how the MSE of the emotion recognition correlates to the objective and subjective evaluations of the models. It is shown that emotion prediction error: 1) has weak positive correlation with perplexity ($r = 0.28$), 2) has weak positive correlation with perceived naturalness ($r = 0.15$), and 3) has weak negative correlation with perceived emotion impact ($r = -0.25$). It is important to note that for MSE and perplexity, lower score is better, and for perceived naturalness and emotional impact, higher score is better. The weak correlations to all three evaluation metrics suggest that

Table 8.5: Average emotion recognition MSE on test set. All of the models are Emo-HRED with selective-fine-tune. Model No. refers to that in Table 8.1.

Model No.	Positive data	Audio encoder	Use prediction	MSE
5	No	No	No	0.375
6	Yes			0.114
7	No	Yes		0.187
8	Yes			0.320
9	Yes	No	Yes	0.118
10		Yes		0.142

the emotion context prediction performance may not impact the performance of the systems directly.

We further test this hypothesis by enforcing a confidence level, in form of error threshold level in considering emotion information. An emotion error threshold of 0.1 means that when emotion prediction error is above 0.1, the model will not consider emotion information in generating a response. In the extremes, a threshold of 0 means that emotion information is never considered, and a threshold of 1 means emotion information is always considered.

Figure 8.3 shows system performance with a number of emotion error threshold, as well as the percentage of data that violates the threshold. The model with threshold of 1 is equivalent to Model No.6 (Emo-HRED), and threshold of 0 is Model No. 4 (HRED). These two models are trained with identical set-up except for the neural network used, i.e. with or without emotion.

The numbers reveal that consideration of emotion always benefit the model's performance. This signals that the improvement of Emo-HRED is not determined by the emotion recognition performance during test time. As posited in Chapter 5, the relationship between semantic and emotional content is not arbitrary, and thus utilizing them in combination could benefit the learning process of the model in general, and not only when the emotion recognition is perfect. Although, it may be important to note that squared-error of 0.15 is relatively low given the emotion value range of $[-1, 1]$.

This finding also relates to the analysis on Section 4.3.4 regarding expert's perception of participant's emotion. Especially for arousal, in reality perception by the expert may not exactly reflect the truth feeling of the participant. However, emotion perceptiveness is still an invaluable tool in addressing emotional topics

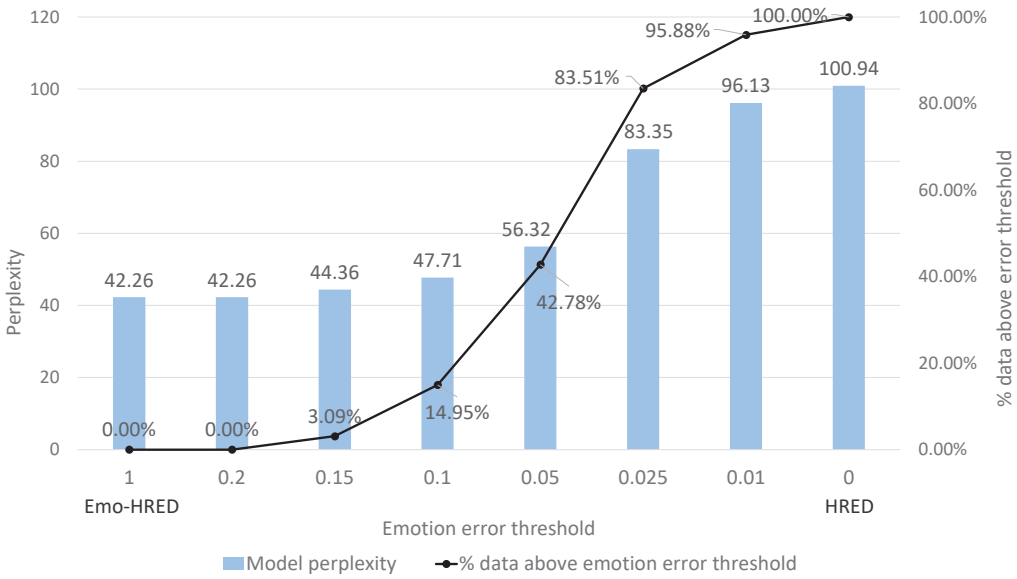


Figure 8.3: Perplexity on models with different emotion error threshold. Line chart shows the percentage of data that violates the threshold. The numbers reveal that consideration of emotion always benefit the model's performance.

and conducting an affective dialogue successfully. It is the same case for emotional context in a dialogue system as shown by the experiments. Although more works is required to achieve perfect emotion context prediction performance, awareness of emotional context is shown to be beneficial for the performance of the system in general.

8.4 Summary

In this chapter we experiment with extending the Emo-HRED architecture to utilize multi-modal information for emotion encoding. In particular, we examine the use of acoustic features in combination with the text features. Speech has been argued to be the richest channel of communication, containing paralinguistic informations including emotion and affect. Various emotion recognition studies have shown the usefulness of speech features in predicting emotional states [64, 22]. To benefit from this source of information, we consider utilizing additional features in modeling the emotion context within Emo-HRED.

Extended experiments, additional evaluation results and analyses are presented in this chapter. Audio information are shown to be quite consistently useful to aid the emotion encoding process, lowering model perplexity as well as improving perceived subjective quality. We also see additional gain in performance by utilizing predicted emotion label during model training. We found that the insights gained from the evaluation and analyses align with that presented in Chapter 5.

Chapter 9

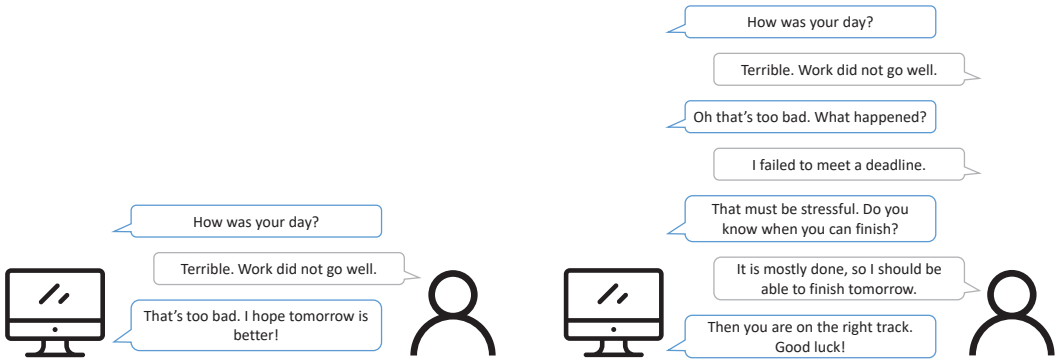
Long-term Emotion Improvement Elicitation through Human-Computer Interaction: Preliminary Study

At the heart of emotion processing is the goal of emotional improvement. In Chapters 5 through 8, we attempt to directly induce this improvement through positive emotion elicitation at dialogue turn level. In this chapter, we conduct preliminary study and experiments on emotion improvement in a dialogue spanning multiple turns. This difference is illustrated in Figure 9.1.

The motivation for long-term emotion improvement elicitation is as follows. A number of studies have reported that emotionally distressed people often feel an improvement as the direct outcome of socially sharing the event leading to the negative emotion [81, 15, 120]. However, it has also been argued that improvement of emotion might not always be directly and immediately attainable as some emotions require cognition, i.e. a kind of cognitive process is necessary for some emotional changes to occur [33, 18].

In this chapter, I investigate the above cognitive process underlying emotional changes and how it takes place in a dialogue. The main goal is to identify how an active helper would be able to strategically support and catalyze this process through dialogue interactions. An initial attempt in combining this strategy with the aforementioned response generation techniques is presented towards the end

of the chapter.



(a) Short-term. (b) Long-term.

Figure 9.1: Dialogue examples comparing short-term and long-term emotion improvement elicitation. This chapter focuses on long-term emotion improvement elicitation, illustrated in (b). In short-term elicitation ((a), Chapters 5 to 8), we attempt to elicit improvement within a single dialogue turn response. On the other hand, long-term elicitation attempts to achieve this through a dialogue spanning multiple turns. In long-term improvement elicitation, deeper understanding of emotion processing through dialogue is necessary.

9.1 The Role of Dialogue in Emotion Improvement

The *appraisal theory of emotion* argues that most of our emotional experiences are the result of a cognitive process, unconscious or controlled, of evaluating situations and events [28, 95]. At the center of this appraisal process is how we evaluate an event in relation to our well-being [30]. This largely determines our emotional reaction towards an event, i.e. negative reactions to events which threaten our well-being, and positive reactions to ones which nourish it.

An important aspect of appraisal is that it is personally characterized: the emotions resulting from an identical event vary from person to person [102]. This means, the relationship between an event and an emotional reaction is not absolute nor deterministic. Even for a single person, it is not uncommon that event-emotion relationships change over time.

Burleson and Goldsmith posit that the appraisal theory can be used to explain how emotion comforting works [10]. 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 facilitating this process. A conversational partner (helper) may help a person clarify their reasoning, thoughts, and feelings relating to the upsetting event. Such a process often results in the change of view, feelings, and coping efforts towards the event in question.

Burleson and Smith proposed a dialogue flow and topics for a conversation aimed to provide emotional support [10]:

- Understanding of the current emotion. Reasoning for its occurrence should be carefully laid out. Assessment of its appropriateness given the event is then discussed.
- Current coping strategy should be discussed as well, and its effectiveness assessed.
- If current coping strategy appears to be suboptimal, alternative strategies should be explored.

For the purpose of the analysis in this chapter, I define two distinct roles in this scenario: **the participant**, i.e. the distraught person who would directly benefit from the comforting process, and **the counselor**, i.e. the person who interacts with the participant to actively catalyze their re-appraisal process. The role of a counselor in such a scenario is inherently limited. Ultimately, the distraught person is the one responsible for their own re-appraisal and emotional improvement. However, dialogue itself is essential in facilitating said process. The process as a whole is gradual, it is discursively constructed by both parties through the dialogue [36]. 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 thoughts, events, and emotion [48].

9.2 Identifying the Structure of Emotion Processing in Dialogue

I aim to identify the 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 processing. 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 (Section 4.3). We focus on the counselor's actions and try to identify the steps taken throughout the dialogue.

9.2.1 Metholodgy

I manually assess and analyze all sessions in the counseling corpus to find a shared dialogue structure across the sessions within the corpus. Direct identification of dialogue phases and actions by simply reading the transcriptions proved to be perplexing and give ambiguous results. To alleviate this problem, I resort to a deductive course of analysis, explained below.

The first step is to collect all the counselor's questions from the corpus. The reasoning is that these questions should illustrate the kind of information needed by the counselor to proceed with the interaction and achieve the emotion improvement goal. Furthermore, the questions 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, potentially showing typical dialogue phases within the corpus.

The analysis that follows is then guided by relevant findings and existing studies. Commonalities between those and the found pattern will ensure validity of the proposed dialogue structure. The following subsections elaborate three main guiding principles of the analysis.

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** is crucial in gathering information from the participant, and in the process encouraging them to clarify and re-assess 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 questions to encourage the participant to continue speaking and do a more in-depth exploration.
- **Active listening** is very important to signal to the participant that they are listened to and understood. This will encourage further dialogue and ensure cooperation from the participant.
- **Clarification** of any ambiguous statements. 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, such as:

- **Reflecting Feelings** Restating the feeling of the participant in order to show understanding of their emotion.
- **Relating** Statements related to the participant's feelings to show that the counselor understand their situation.
- **Validating** Let the participant knows what their reaction in such situation is normal. Explain what the feelings are common in such a 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.

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. Explanations in parentheses have been added for clarity.

> *"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 she might still 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 to the inducer,
- understood the reasoning behind the emotional reaction,
- understood the how the participant relate to the event in the inducer, 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.

Related Works

Van der Zwaan et al. has adapted the topics proposed by Burleson and Smith for application in intelligent agents to support cyber-bullying victims [109]. To better structure the dialogue, the topics are mapped into 5-phase dialogue model of chat- and telephone- based counseling. The dialogue model proposed can be found in Table 9.1. Given some dialogue overlap between supporting cyber-bullying and eliciting emotion improvement, this dialogue model provides a meaningful comparison in the identification process.

Table 9.1: Conversation model proposed by van der Zwaan et al. [109].

No	Conversation phase	Topics
1	Welcome	Hello
2	Gather information	Event (general)
		Emotional state
		Personal goal
		Event (details)
		Coping (current) (if need to cope)
3	Determine conversation objective	Conversation objective
4	Work out objective	Coping (future) (if need to cope)
		Advice (depending on conversation objective)
5	Round off	Bye

9.2.2 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 discusses and followed accordingly. The later part of the sessions are commonly used to discuss the event in a positive light if possible, 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 three aspects discussed above:

1) counseling skills and technique, 2) counselor's assessment of the session, and 3) related work of dialogue model in support for cyber-bullying victim scenario.

Based on the above thorough analysis and careful refinement, I propose the following dialogue phases and actions as the underlying structure of emotion improvement through dialogue.

Table 9.2: Proposed dialogue model for long-term negative emotion processing.

No	Conversation Phase	Actions
1	Opening	Small talk
2	Understanding	Emotion
		Event
		Experience
		Strategy
3	Resolution	Brainstorming
		Distancing
		Positive asset search
4	Closing	Goodbye

1. Opening

The opening phase serves as warm-up prior to addressing emotional topics, as well as 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.

2. Understanding

The goal of this phase is for the counselor 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 of the participant's feelings or reactions toward an emotional event or exposure (in the counseling corpus case, the emotion inducer video).

Event. Discussion about the emotional event. Typically, the counselor asks the participant to describe the event, offers comments regarding the event. This allows the counselor to assess participant's understanding of the event, as well as their interpretation of it.

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, the strategy highly influence the step taken in the next phase.

3. Resolution

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, as well as steps that can be taken to achieve it, is highly likely to elicit emotion improvement.

Distancing. The counselor tries to put some 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. 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 phase 2: Experience is highly useful in reinforcing participant's positive assets.

4. Closing

The dialogue is round up with the 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.

Dialogue Flow

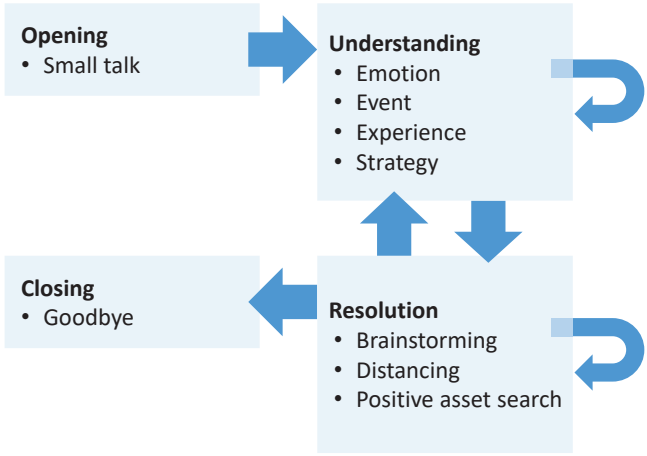


Figure 9.2: Flow between dialogue phases in the proposed dialogue structure.

In the recorded spontaneous interaction, the flow between and within the understanding phase and the resolution phase varies significantly. The information

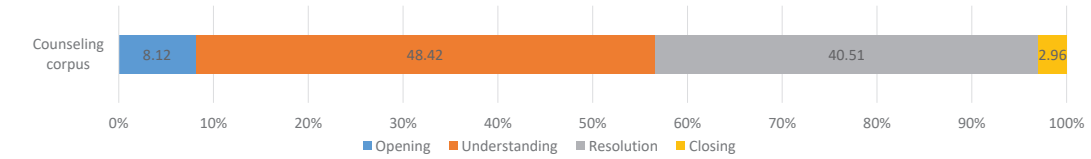
gathering during the understanding phase is not done in any specific order, neither are the actions during the resolution phase. Furthermore, the flow between these two phases are is bidirectional. That is, more information may be gathered even though a resolution action is already being performed. Multiple resolution actions may be done within a session. Flow between the dialogue phases are illustrated in Figure 9.2.

9.3 Corpus Analysis

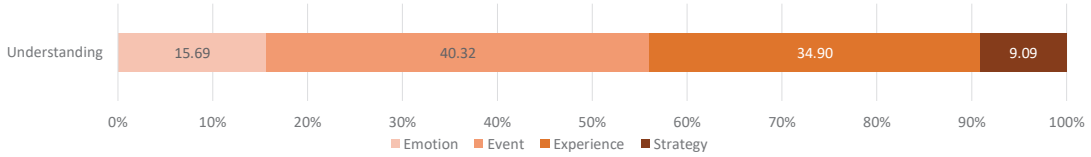
The counselor corpus is manually annotated by assigning phase and action labels on every dialogue turn. We count the occurrence of the phases and actions within the counselor corpus. Some dialogue turns are excluded in this analysis, in particular those that relate to the procedural part of the dialogue (e.g. "You can leave the headset on the chair.", or "We will do another questionnaire at the end.").

The statistics are visualized in Figure 9.3. Figure 9.3(a) shows that the majority of the conversations are spent on the understanding and resolution phases. This is to be expected given that the scenario is carefully designed to focus on negative emotion processing. The opening phase tend happen over more dialogue turns than the closing. Likewise, the understanding phase have larger portion in the data than resolution.

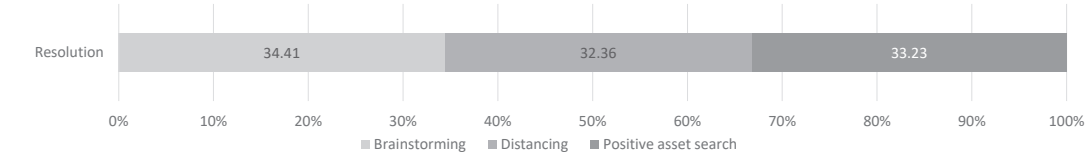
Action composition within the understanding phase (Figure 9.3(b)) shows that discussion about the event and participant's experience related to it tend to dominate the phase. This shows that while assessment of the felt emotion is essential, the counselor as an expert put even more effort in understanding the reasoning behind the emotional reaction. In some sessions, coping strategy of the participant is discussed as well. Figure 9.3(c) shows that the three resolution actions is equally likely to be employed in the dialogue collected in the counseling corpus.



(a) Entire counseling corpus.



(b) Phase 2. Understanding.



(c) Phase 3. Resolution.

Figure 9.3: Composition of phases and actions in the counseling corpus.

9.4 Simulating Long-Term Emotion Improvement Elicitation

We simulate long-term emotion improvement elicitation by utilizing language modeling technique on the counseling corpus. The method, experimental result, and analysis are presented in this chapter.

9.4.1 N-gram Simulator

To model the probability of the sequence, I utilize the n-gram model commonly used for language modeling. A sentence is made of a sequence of words. Analogously, we can view a session as a sequence of actions or states of the counselor and participant. This allows us to model the session with the same concepts used in language modeling. The n-gram model has been previously proposed for user simulation in [34], operating at semantic level as well. N-gram models are suitable for our preliminary study in long-term improvement elicitation since it can be trained easily given any dataset. It is purely probabilistic and fully

domain-independent.

An n-gram is a count for the occurrence of a specific sequence of tokens with a maximum length length of n. An n-gram approximates the probability of the next token in a sequence of length M with the probability of the next token given a context of only $(n-1)$ tokens, i.e.,

$$P(token_1, ..., token_m) = \prod_{i=1}^{m} P(token_i | w_1, ..., token_{i-1}) \approx \prod_{i=1}^{m} P(token_i | token_{n-1}, ..., token_{i-1}). \quad (9.1)$$

The conditional token probability are computed with n-gram counts,

$$P(token_i | token_1, ..., token_{i-1}) = \frac{\text{count}(token_{1-n+1}, ..., token_i)}{\text{count}(token_{1-n+1}, ..., token_{i-1})}. \quad (9.2)$$

In a language model which deals with sentences, each word in a sentence is treated as a token. In modeling a dialogue, we consider turn-level information, or semantic representation, the token.

The simulators for the counselor and participant are trained separately. First, we define the tokens for each of these models. This is done by first defining semantic-level representation of the counselor's and participant's dialogue turns. For the counselor's turn, we consider the phase and action labels as the representation. This results in 9 possible actions on the counselor side, as has been elaborated in Section 9.2.2.

On the other hand, for the participant's turn, we consider their valence and arousal emotion labels. The values are ranging from -1 to 1, with a time step of 0.1 to discretize the values. This yields a total of 441 possible participant state. Since this state space will be large relative to the amount of data, to avoid data sparsity valence and arousal are modeled separately. This results in 21 tokens for each of the valence and arousal n-gram models.

Figure 9.4 illustrates the training and simulation flow. To construct the counselor simulator, we train an n-gram model using action sequences extracted from the counseling corpus. The counselor simulator outputs the probability of the next action (denoted as a_{t+1}) given the sequence of actions up to that time

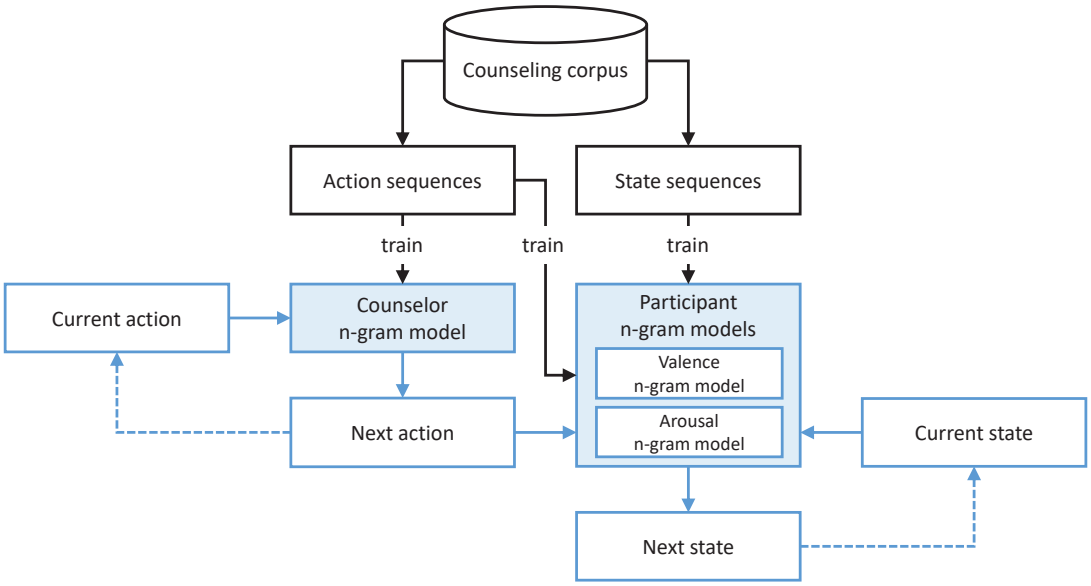


Figure 9.4: Overview of simulation based on counseling data.

$(a_1, ..., a_t)$, which is approximated by only considering the last $n-1$ tokens in the sequence $(a_{n-1}, ..., a_t)$. All possible next actions are assigned probabilities,

$$P(a_{t+1}) = P(a_{t+1}|a_{n-1}, ..., a_t), \tag{9.3}$$

from which the next action will be sampled.

The participant simulator is constructed with the same concept. However, the state transition is conditioned on the counselor's action in addition to the states of previous time steps. This is equivalent to modeling the participant's reaction to counselor's actions in the corpus. The probabilities of the next possible states ($s_{t+1} = val_{t+1}, aro_{t+1}$, treated separately) is computed as

$$P(s_{t+1}) = P(s_{t+1}|(s_{n-1}, a_{n-1}), ..., (s_t, a_t)), \tag{9.4}$$

from which the next possible state is sampled.

9.4.2 Result and Analysis

We evaluate how well the simulator model counselor actions and user state by computing the perplexity of their respective n-gram models. Three values of n

were tested: 1 (unigram), 2 (bigram), and 3 (trigram). The results are presented in Table 9.3.

Table 9.3: Perplexity of n-gram modeling of counselor action, user's valence, and user's arousal.

N	Perplexity		
	Action	Valence	Arousal
1 (unigram)	7.85	15.50	12.83
2 (bigram)	1.69	3.86	4.44
3 (trigram)	1.71	3.50	3.92

There are important differences between this simulation task and language modeling that are worth noting 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 actions and 22 emotion states) 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. Emotion state of the user is likely to remain stable over a few dialogue turns, and constant erratic changes of emotion states is unnatural. Similarly, constant transition between dialogue phases is also unusual in the collected data. These key differences explain the perplexity result.

With no context in unigram model, the perplexity of the model almost 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 since the tokens are highly repetitive.

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

Current model has not yet taken into account user's state in determining flow between the dialogue phases. In other words, it still solely relies on the steps taken by counselor in the data. At the moment, the addition of user state in conditioning the counselor simulator lead to severe sparsity in the data. The n-gram model in itself is still very simple and naive. Especially when data is limited, as in this case, generation of longer sequences can result in unnatural sequences because there is no behavioral constraint to the generation process. Although this can be solved by employing bigger n, adding longer context into the modeling (e.g. 4-gram or 5-gram) will require exponentially more data. Two possible direction in solving these limitations are: 1) expanding of the corpus, or 2) designing a method that is robust in scarce data scenario.

The simulation presented in this chapter relies on mimicking the behavior of the counselor. In other words, in improving user's emotional state through dialogue, the model solely attempts to follow the examples of expert's behavior. In the future, it would be useful to investigate whether the model is able to learn how to elicit emotion improvement and act accordingly based on its internal decision making or dialogue policy. Towards this direction, optimization methods and reinforcement learning approaches may be some of the most promising means to explore.

9.5 Combining Short- and Long-term Positive Emotion Elicitation

To achieve a full-fledged dialogue system, we combine the MC-HRED and counselor's n-gram simulator approaches into one system. The action context encoder of MC-HRED is suitable for modeling the manually defined phases and

action labels introduced in this chapter. The resulting system will pose as a response generator which maps the long-term positive emotion eliciting simulator (i.e. counselor's n-gram models) into sequence of words as a response to a dialogue context.

The experiment flow is as follows. First, we pre-train an HRED model as in Chapter 6 to obtain the starting model. Second, we selectively fine-tune MC-HRED using the manually defined and annotated counselor's phases and actions as action context (in place of the automatic cluster labels). This is to let the model learn the relationship between a counselor's phase-action label and its corresponding responses. During testing, given a dialogue context, we then utilize the phase-actions generated by the n-gram counselor simulator in the MC-HRED response generation process. This results in a dialogue response at sentence level. Figure 9.5 shows schematic view of the system.

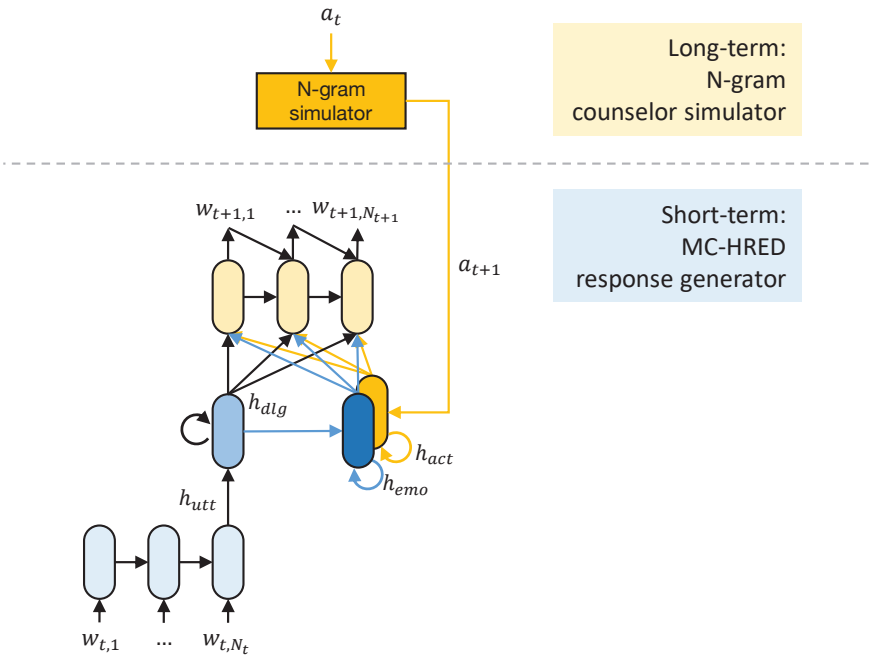


Figure 9.5: Hybrid MC-HRED, combining MC-HRED and n-gram simulator. MC HRED and simulator are trained separately and then combined in the end system. In this case, $n = 2$. When $n = 3$, context a_{t-1}, a_t is used, and when $n = 1$ is used, no action context is passed.

Table 9.4 presents the perplexity of the full-fledged hybrid MC-HRED systems

Table 9.4: Model perplexity on counseling test set.

	Separate systems		**Full-fledged system**	
	Model	Perplexity	Model	Perplexity
High-level (sequence of actions)	Unigram	7.85	Unchanged, used as is	
	Bigram	1.69		
	Trigram	1.71		
Low-level (sequence of words)	MC-HRED with Word2Vec K-Means labels	29.57	Hybrid MC-HRED with unigram simulator	49.74
			Hybrid MC-HRED with bigram simulator	49.62
			Hybrid MC-HRED with trigram simulator	49.78
			Hybrid MC-HRED topline with truth label	49.41

and its components as separate systems. We observe higher perplexity on the full-fledged system compared to the separate MC-HRED trained and tested with the automatic cluster label. This difference can be explained by the origin of the action labels themselves.

As elaborated in Chapter 6, the automatic cluster label is the result of clustering embedding vectors of the counselor's responses. That means, there is direct relationship between the K-Means cluster label and the target response, which could be highly useful for the system to model the data (i.e. yielding lower perplexity). On the other hand, the manually defined and annotated label is created with focus on modeling the dialogue structure. That is, to dialogue flow within one interaction aimed at negative emotion processing. Consequentially, it is highly useful in modeling the counselor's behavior, as has been shown in Section 9.4. However, this means the relationship between the label and the utterances belonging to that label may not be as clearly captured, and in turn bringing less benefit in modeling the training data (i.e. yielding higher perplexity).

In the generated responses, more distinct differentiation of dialogue actions can be observed in the hybrid MC-HRED system. Dialogue samples comparing the systems responses are included in Table 9.5. These examples show that the hybrid system is able to properly decide on the appropriate actions, such as asking

questions to gain more understanding, when responding to a user query. It is also important to note that although these responses at times differ from the target response (and thus yielding higher perplexity), it could be more beneficial for emotion processing through dialogue in the long-term.

Table 9.5: Comparison of system responses for two triples in test set.

U_1	mm. mm. you mean like they said in the video to avoid those kind of circus or something you mean.
U_2	uh-huh. uh-huh. yes but i don't usually go to.
U_3 (Target)	yes yes.
MC-HRED	yes.
Hybrid MC-HRED	do you think it looks like you're doing well?
U_1	mm. at first when he mentioned he was they were expected to work an extra eight hours for free i think i actually grinned.
U_2	it was like. i was um ah somewhat like amazed that wow this actually happens.
U_3 (Target)	mm.
MC-HRED	mm mm.
Hybrid MC-HRED	mm. it's interesting (laughter).

The hybrid MC-HRED system demonstrates one way the short- and long-term positive emotion elicitation approaches can be combined into a full-fledged dialogue system. Future efforts should be aimed at improving the both the response generator and the counselor simulator, as well as devising a more sophisticated combination scheme. Furthermore, it is important to note that the proposed dialogue model is still based on analysis of human communication and not yet tested in real HCI situations. It has been argued that user may have different expectations and reactions when communicating with other humans and machines. In this regard, I believe user study with Wizard-of-Oz set up needs to be conducted in the future to see whether the proposed dialogue model is suitable in HCI as is, or whether modifications are necessary to take into account the possibly existing differences between human communication and HCI in emotion improvement elicitation.

9.6 Summary

In this chapter, I investigate the process underlying emotional changes and how it takes place in a dialogue. The main goal is to identify how an active helper would be able to strategically support and catalyze this process through dialogue interactions. This structure is identified and explained in detail in Section 9.2.2. The structure is identified within the counseling corpus, which is the basis of the analysis in Section 9.3. We train simulators to simulate a dialogue for eliciting emotional improvement, discussed in Section 9.4. Lastly, an effort to combine response generation techniques and the counselor simulator are presented in Section 9.5.

In the future, it is important to conduct user study to test whether the dialogue model is suitable for emotion improvement elicitation in HCI, as well as to observe its emotional impact on the user. As the limited amount of data is the main constraint in modeling the dialogue with finer-grain details, exploration to and development of methods robust to data scarcity will greatly impact future research efforts.

Chapter 10

Conclusion and Future Works

10.1 Conclusion

This thesis presented works on supporting negative emotion processing with dialogue system interaction. This work has been motivated by the lack of 1) human-computer interaction works that focus on emotional benefits of affective systems for users, and 2) dialogue systems that address negative emotions commonly encountered in everyday life. This work aimed to mimic the important role of dialogue in emotion processing, transferring it into dialogue systems to enable such a process through HCI.

The concept of negative emotion processing, or emotion improvement elicitation, can be observed in two perpectives: *short-term* and *long-term*. This thesis focused on the short-term improvement elicitation task, as well as investigating the long-term elicitation task in a preliminary study and trial experiments in addition. Prior to the experiments, I constructed corpora containing emotion improvement elicitation in human dialogues. The corpora has been carefully designed to capture emotion improvement as perceived by humans, as well as expert strategy in long-term dialogue. The collected data served as the basis of the experiments, model training, and analyses throughout the remainder of the thesis.

Assuming positive emotional state as the goal, short-term negative emotion processing is reformulated into turn-based positive emotion elicitation. Complexity of the problem is incrementally increased by incorporating various dialogue aspects that relate to the elicitation goal, such as dialogue action and emotional

impact. In Chapters 5 through 8, novel neural network architectures that consider emotion for response generation in chat-based scenarios have been proposed. Experiments and evaluations on each chapter provide an empirical result which shows that the proposed models yield improvement of naturalness, emotional impact, and engagement through the generated response. While this thesis have conducted experiments mainly on two corpora, i.e. SEMAINE and the counseling corpora, the proposed approaches are entirely domain independent and generalizable to other conversational corpora. It can be trained and tested on any domain-independent data, provided emotion annotation of the dialogue is available.

Analysis of the generated response reveal what the models have learned. When trained with responses that humans consider to elicit positive emotion (Chapters 5 and 8), the models tend towards shorter responses with words that have a positive affective content. Although this might remind us of the generic responses problem in neural response generators, 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. Results from Chapters 6 and 7 show that we can expand on this strategy by highlighting dialogue action and emotional impact information, as well as using data involving expert for training. This allows the model to produce utterances signaling other dialogue actions (e.g. asking questions such as "how did you feel about it?"), as well as producing words that are less positive yet still yielding positive emotional impact (e.g. a remark on an event such as "it's a big thing.").

While it has been shown that simply having a dialogue about a negative emotional experience can directly provide emotional benefit, some emotional changes can only be achieved through cognitive re-appraisal of the event that caused it [33, 18]. To explore this process, long-term elicitation of emotion improvement is inspected by extending the positive emotion elicitation scope to the entire dialogue. This task is the topic of Chapter 9, the goal of which is to identify a dialogue structure to allow a system to play the role of the helper in the long-term negative emotion processing of a distraught person.

Guided by existing works, both in affective computing and clinical psychology, analysis of the previously constructed counseling corpus resulted in a dialogue model of long-term negative emotion processing (Section 9.2). Combining the defined structure and the counseling corpus, simulators are trained to simulate a

dialogue for eliciting emotional improvement, discussed in Section 9.4. Lastly, an effort to short- and long-term improvement elicitation appraoches were presented in Section 9.5.

In summary, this thesis has made the following contributions:

- The construction of conversational corpora demonstrating positive emotion elicitation and negative emotion processing through dialogue. The corpora serves a basis for analysis of emotion processes in social-affective human dialogue, also presented in this thesis.

- Novel architectures and approaches for affective chat-based dialogue systems with an implicit goal of user emotion improvement. The proposed approaches endow the systems with awareness of important dialogue aspects such as emotion, dialogue action, and emotion impact. The systems are capable of generating responses that are subjectively perceived to elicit more positive emotion, as well as more natural and engaging.

- Preliminary study on long-term emotion improvement elicitation with dialogue systems. I identified dialogue structure of negative emotion processing in human dialogue, in compliance with expert actions in such a scenario. A language modeling technique is utilized to model and simulate expert's actions.

There are still a number of limitations in this work that pose interesting research questions to be tackled in the future. Current work has not yet:

- Considered user state in the long-term emotion improvement elicitation. As the simulator are modeled solely on counselor's actions, the current system is not yet able to adapt with user state as it unfolds in a conversation. As previously mentioned, this can be tackled by either expanding the training data, or exploring learning options that performs effectively on sparse data.

- Conducted a study to see whether the proposed dialogue model is suitable in HCI as is, or whether modifications are necessary to take into account the possibly existing differences between human communication and HCI in emotion improvement elicitation.

- Conducted in-depth user study to investigate the emotional effect of the system. At the moment, evaluations are done in a turn-basis. This allow us to judge the effectiveness of short-term emotion improvement elicitation, however to better understand the long-term elicitation success and impacts, user study through dialogue interaction is necessary.

10.2 Future Works

Looking at the complexity and richness of human social-affective communication, there lies many interesting and purposeful topics still waiting to be researched in the future. Figure 10.1 illustrates the potential research direction of this topic. Particularly, in relation to the limitations of this thesis as described above, the following topics seem a logical direction to pursue.

Wizard-of-Oz study to confirm the data-driven dialogue design This thesis has tackled the first step in designing a dialogue structure for long-term emotion improvement elicitation which has not yet been studied in HCI scenario. However, it has been argued that humans may behave differently when communicating with machines as opposed to with other humans, due to different expectations. Therefore in the future, it is essential to conduct user study to test whether the proposed dialogue model holds true for emotion improvement elicitation in HCI. Previous study such as [24] has shown through a Wizard-of-Oz study that careful and thorough analysis of human interaction can be adequate in informing the design of a dialogue system aimed at a particular task. To prove the same of the proposed dialogue flow, it needs to be tested whether a system can elicit emotion improvement in user when following the defined dialogue flow and generated expert actions. A Wizard-of-Oz set up is suitable for this proof-of-concept purpose as it allows us to assume perfect system behavior while circumventing technical challenges in building the actual system.

Generalization into goal-oriented and domain-specific systems Consideration of emotion in dialogue is beneficial not only for chat-oriented systems, but also for goal-oriented systems. In goal oriented dialogues, emotion awareness can be useful in increasing task success [32] as well as recovery from any dialogue breakdown or task failures. For example, a user that is upset when a restaurant

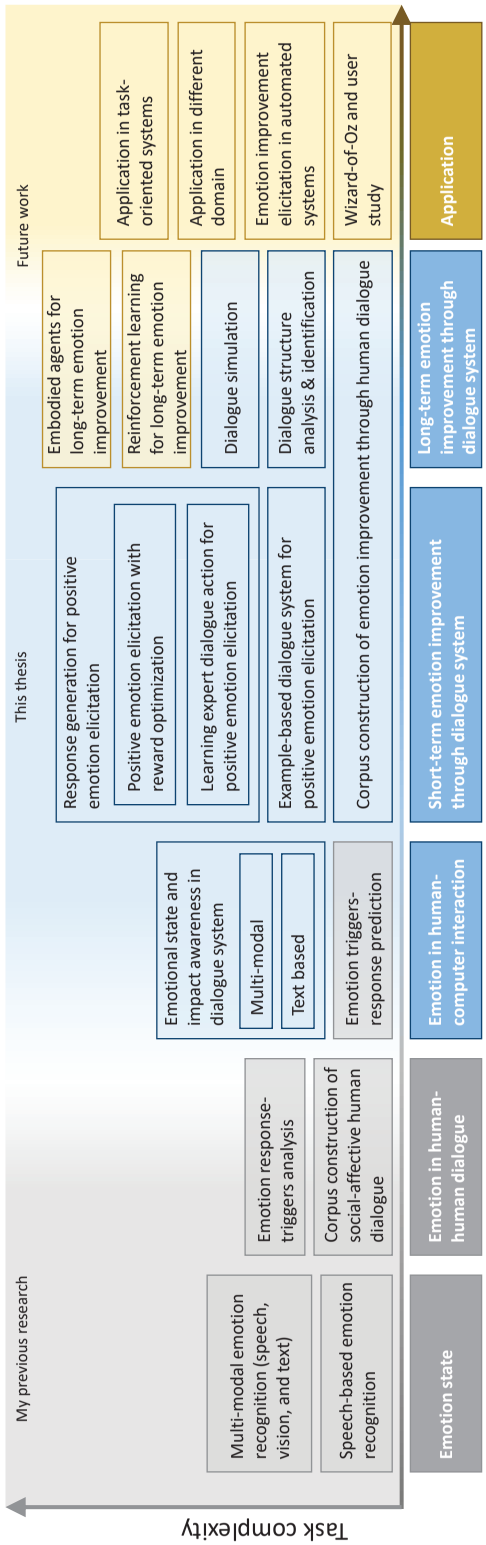


Figure 10.1: Research roadmap towards future works.

they would like to reserve is fully booked, can benefit from emotion improvement elicitation to counter the negative effect of unfulfillment of their goal. A number of recent works have proposed end-to-end methods for goal-oriented dialogue systems [121, 9, 62, 20]. These systems employ encoding-decoding mechanism that is similar to end-to-end chat-oriented systems with additions such as entity indexing and lexicalization. End-to-end approaches can be generalized to domain-specific systems as each dialogue domain actually differs a lot not in the dialogues contained within, but rather the vocabulary or domain descriptors [121]. This allows the system to impose the domain constraint at higher level of abstraction, in the same way that emotion is incorporated in Emo-HRED and MC-HRED. The shared end-to-end paradigm would be a valuable bridge in generalizing the proposed emotion improvement elicitation methods to goal-oriented systems. Although goal-specific dialogue data is available in large amounts, significant effort required in procuring the emotion annotation for them would be a major bottleneck in moving towards this direction.

Deep reinforcement learning for long-term emotion improvement elicitation Learning a strategy within the proposed long-term emotion processing dialogue model seem a particularly interesting challenge to tackle in the future. A dialogue policy will allow the system to take actions adaptively, with real-time considerations of a currently occurring dialogue state. With a well defined dialogue states and actions, deep reinforcement learning may be the a promising avenue for learning of dialogue policy in the future. The main challenge in pursuing this would be the limited amount of data, exploration to and development of methods robust to data scarcity will greatly impact future research efforts.

Integration of dialogue strategy into an embodied conversational agent Human-like appearance may have a significant effect in user perception of a dialogue system. I believe especially when addressing emotion, it may be important the user is able to relate to the system they are talking to. An embodied conversational agent would also be able to elicit emotion improvement through more means, such as non-verbal gestures and expression.

Long-term user study to observe the effect of emotion improvement elicitation with dialogue systems It is also important to conduct user study

to assess the impact of such a system, especially over extended period of time. How does computer user feel about allowing a dialogue system to influence their emotional state? Does this feeling change over time with continual use of the system? In what kind of situation would the system be most useful, and when does it potentially lead to unhealthy user behaviors? These are the questions that are important to be answered before emotion improvement elicitation with dialogue system can be applied outside the research environment.

Appendix

A Questionnaires from Counseling Corpus Data Collection

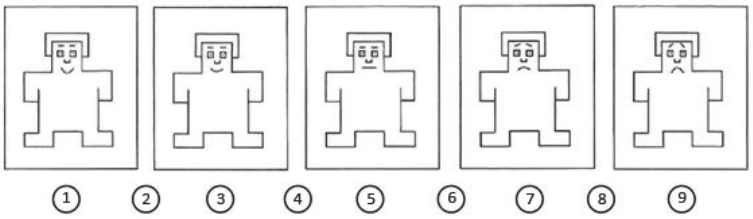
A.1 Pre-recording Questionnaire

Pre-Recording Questionnaire

* Required

1. **Name** *

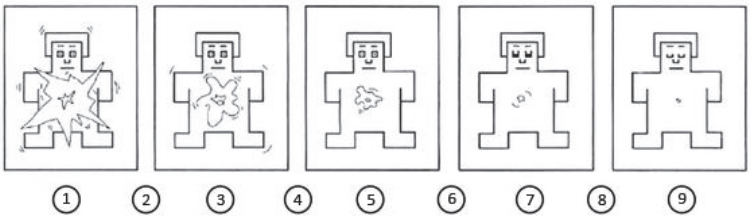
.....



2. **According to the picture above, how positive do you feel?** *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	
Strongly positive	◯	◯	◯	◯	◯	◯	◯	◯	◯	Strongly negative



3. **According to the picture above, how activated do you feel?** *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	
Strongly active	◯	◯	◯	◯	◯	◯	◯	◯	◯	Strongly calm

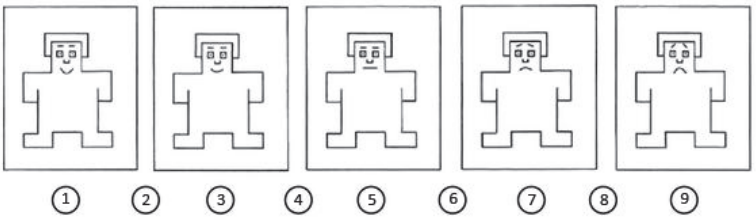
A.2 Video Questionnaire

Video Questionnaire

* Required

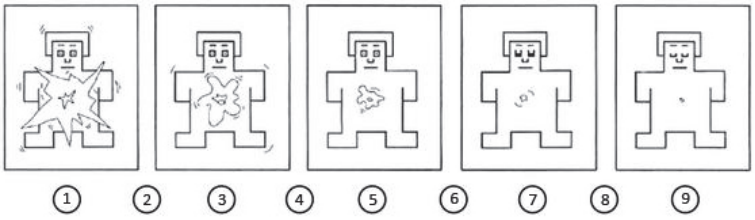
1. **Name** *

.....



2. **According to the picture above, how positive do you feel?** *
Mark only one oval.

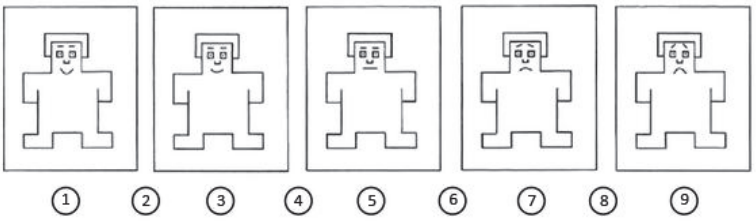
	1	2	3	4	5	6	7	8	9	
Strongly positive	◯	◯	◯	◯	◯	◯	◯	◯	◯	Strongly negative



3. **According to the picture above, how activated do you feel?** *
Mark only one oval.

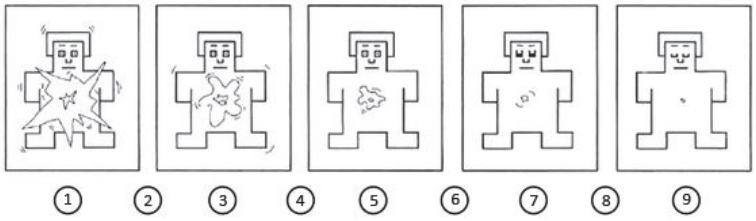
	1	2	3	4	5	6	7	8	9	
Strongly active	◯	◯	◯	◯	◯	◯	◯	◯	◯	Strongly calm

Please watch the video on the screen, and then continue to the next section



4. **According to the picture above, how positive do you feel?** *
Mark only one oval.

	1	2	3	4	5	6	7	8	9	
Strongly positive	◯	◯	◯	◯	◯	◯	◯	◯	◯	Strongly negative



5. **According to the picture above, how activated do you feel?** *
Mark only one oval.

	1	2	3	4	5	6	7	8	9	
Strongly active	◯	◯	◯	◯	◯	◯	◯	◯	◯	Strongly calm

Powered by
Google Forms

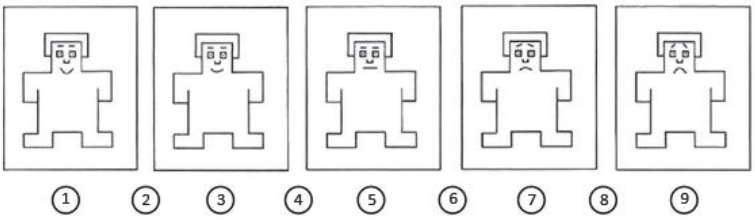
A.3 Post-recording Questionnaire

Post-Recording Questionnaire

* Required

1. **Name** *

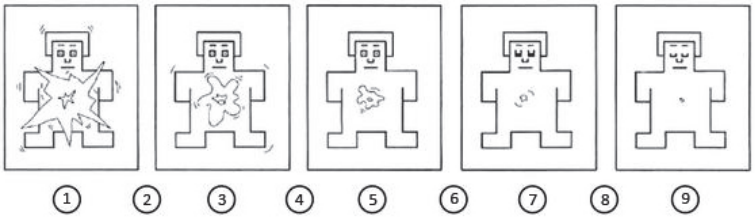
.....



2. **According to the picture above, how positive do you feel?** *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	
Strongly positive	◯	◯	◯	◯	◯	◯	◯	◯	◯	Strongly negative



3. **According to the picture above, how activated do you feel?** *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	
Strongly active	◯	◯	◯	◯	◯	◯	◯	◯	◯	Strongly calm

4. **I noticed a negative emotional change in myself after watching the video** *

Mark only one oval.

	1	2	3	4	5	
Strongly agree	◯	◯	◯	◯	◯	Strongly disagree

5. **I noticed a positive emotional change in myself during and after the conversation with the counselor** *
Mark only one oval.

	1	2	3	4	5	
Strongly agree	◯	◯	◯	◯	◯	Strongly disagree

6. **The conversation helped me deal with and process my emotion** *
Mark only one oval.

	1	2	3	4	5	
Strongly agree	◯	◯	◯	◯	◯	Strongly disagree

7. **I felt understood by the counselor** *
Mark only one oval.

	1	2	3	4	5	
Strongly agree	◯	◯	◯	◯	◯	Strongly disagree

8. **I enjoyed the conversation with the counselor** *
Mark only one oval.

	1	2	3	4	5	
Strongly agree	◯	◯	◯	◯	◯	Strongly disagree

9. **I found a kind of emotional connection between myself and the counselor** *
Mark only one oval.

	1	2	3	4	5	
Strongly agree	◯	◯	◯	◯	◯	Strongly disagree

10. **I would like to talk again with the counselor in the future** *
Mark only one oval.

	1	2	3	4	5	
Strongly agree	◯	◯	◯	◯	◯	Strongly disagree

11. **More specifically, my impression about the conversation is...**

.....

.....

.....

.....

B Subjective Evaluation Instruction

Rate Naturalness And Emotion Impact Of A Dialogue Response (V2)

Instructions

Steps

In this task, you will be asked to judge the quality of a response in a dialogue.

1. **Read the dialogue** snippet between speaker A and B. Each snippet consists of three dialogue turns.
2. **Evaluate A's response** in the snippet (last dialogue turn, highlighted in yellow) in terms of two criteria
   - Naturalness:
     - Does this response **make sense** given the first two dialogue turns?
     - Does this response **logically follow** the the first two dialogue turns?
     - Is this response **intelligible**? Does it have **meaning**?
     - Does it resemble **human response** in real life conversation?
   - Emotional impact
     - Will this response **cause speaker B to have a more positive emotion** after hearing it?
     - Does this response **promote an emotionally positive conversation**?
3. Give numerical **rating** of the two criteria on a scale from 1 (strongly disagree) to 5 (strongly agree). For each criteria, adjust your rating accordingly based on the above questions.
   - Naturalness:
     - If you answer yes to all of the above questions, rate 5
     - If you answer no to all of the above questions, rate 1
     - For rating between 1 and 5, adjust your rating according to the questions. More no answers lean towards 1 (disagree), more yes answers lean towards 5 (agree).
   - Emotional impact
     - If you answer yes to all of the above questions, rate 5
     - If the response does not cause any emotional impact, rate 3 (neutral)
     - If the response causes negative emotion instead, rate 1
     - if the response is NOT intelligible, put yourself in B's position and imagine receiving that response in the dialogue, then answer the questions above.

Rules & Tips

- The term <person> refers to anonymized names of the speaker in the dialogue. For a more natural reading, you can replace this with any name that you want.
- The queries and responses are transcribed from spontaneous speech. So, please expect disfluencies, filler words, unfinished sentences, etc.
- The utterance "aha" has identical sound with "uh-huh."

Speaker	*Dialogue*
A	hello i ' m poppy have i met you before .
B	hm no don ' t think so .
A	**ok .**

A gave a NATURAL response (required)

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

A's response elicits POSITIVE EMOTION in B (required)

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

References

[1] David Ameixa, Luisa Coheur, Pedro Fialho, and Paulo Quaresma. 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, 2014.

[2] 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.

[3] Rafael E Banchs and Haizhou Li. 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, 2012.

[4] Jeesoo Bang, Hyungjong Noh, Yonghee Kim, and Gary Geunbae Lee. Example-based chat-oriented dialogue system with personalized long-term memory. In *2015 International Conference on Big Data and Smart Computing (BigComp)*, pages 238–243. IEEE, 2015.

[5] David Benyon, Björn Gambäck, Preben Hansen, Oli Mival, and Nick Webb. How was your day? evaluating a conversational companion. *Transactions on Affective Computing*, 4(3):299–311, 2013.

[6] Fumihiro Bessho, Tatsuya Harada, and Yasuo Kuniyoshi. Dialog system using real-time crowdsourcing and twitter large-scale corpus. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 227–231. Association for Computational Linguistics, 2012.

[7] Timothy W Bickmore and Rosalind W Picard. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327, 2005.

[8] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[9] Pawel Budzianowski, Inigo Casanueva, Bo-Hsiang Tseng, and Milica Gasic. Towards end-to-end multi-domain dialogue modelling. 2018.

[10] 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.

[11] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[12] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2017.

[13] Alexandra Canavan, David Graff, and George Zipperlen. Callhome American English speech. *Linguistic Data Consortium*, 1997.

[14] Alexandra Canavan and George Zipperlen. Callfriend American English-non-southern dialect. *Linguistic Data Consortium, Philadelphia*, 10:1, 1996.

[15] 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.

[16] Marc Cavazza, Raul Santos De La Camara, and Markku Turunen. How was your day?: a companion eca. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume*

1, pages 1629–1630. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

[17] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

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

[19] Amy N Cohen, Constance Hammen, Risha M Henry, and Shannon E Daley. Effects of stress and social support on recurrence in bipolar disorder. *Journal of affective disorders*, 82(1):143–147, 2004.

[20] Stefan Constantin, Jan Niehues, and Alex Waibel. An end-to-end goal-oriented dialog system with a generative natural language response generation. *arXiv preprint arXiv:1803.02279*, 2018.

[21] 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.

[22] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[23] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[24] 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.

[25] Joao Dias, Samuel Mascarenhas, and Ana Paiva. Fatima modular: Towards an agent architecture with a generic appraisal framework. In *Emotion Modeling*, pages 44–56. Springer, 2014.

[26] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[27] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, pages 45–60, 1999.

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

[29] Florian Eyben, Martin Wöllmer, and Björn Schuller. OPENsmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[30] Susan Folkman and Richard S Lazarus. *Ways of coping questionnaire*. Consulting Psychologists Press, 1988.

[31] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.

[32] Kate Forbes-Riley and Diane Litman. 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, 2012.

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

[34] 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.

[35] John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE, 1992.

[36] Daena J Goldsmith. *Communicating social support.* Cambridge University Press, 2004.

[37] 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.

[38] Jonathan Gratch and Stacy Marsella. Tears and fears: Modeling emotions and emotional behaviors in synthetic agents. In *Proceedings of the fifth international conference on Autonomous agents*, pages 278–285. ACM, 2001.

[39] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. The Vera am Mittag German audio-visual emotional speech database. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 865–868. IEEE, 2008.

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

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

[42] Björn Hartmann, Maurizio Mancini, and Catherine Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In *International Gesture Workshop*, pages 188–199. Springer, 2005.

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

[44] Matthew Henderson, Blaise Thomson, and Steve Young. Deep neural network approach for the dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 467–471, 2013.

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

[46] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[47] Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 49–54, 2018.

[48] 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.

[49] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[50] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[51] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[52] Sumedha Kshirsagar. A multilayer personality model. In *Proceedings of the 2nd international symposium on Smart graphics*, pages 107–115. ACM, 2002.

[53] Carney Landis. Studies of emotional reactions. II. general behavior and facial expression. *Journal of Comparative Psychology*, 4(5):447, 1924.

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

[55] Cheongjae Lee, Sangkeun Jung, Seokhwan Kim, and Gary Geunbae Lee. Example-based dialog modeling for practical multi-domain dialog system. *Speech Communication*, 51(5):466–484, 2009.

[56] Esther Levin, Shrikanth Narayanan, Roberto Pieraccini, Konstantin Biatov, Enrico Bocchieri, Giuseppe Di Fabbrizio, Wieland Eckert, Sungbok Lee, A Pokrovsky, Mazin Rahim, et al. The at&t-darpa communicator mixed-initiative spoken dialog system. In *Sixth International Conference on Spoken Language Processing*, 2000.

[57] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

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

[59] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[60] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[61] Janice Light. "communication is the essence of human life": reflections on communicative competence. *Augmentative and Alternative Communication*, 13(2):61–70, 1997.

[62] Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. *arXiv preprint arXiv:1804.06512*, 2018.

[63] Nurul Lubis, Dessi Lestari, Sakriani Sakti, Ayu Purwarianti, and Satoshi Nakamura. Construction and analysis of Indonesian emotional speech corpus. In *Proc Oriental COCOSDA*, 2014.

[64] Nurul Lubis, Dessi Lestari, Sakriani Sakti, Ayu Purwarianti, and Satoshi Nakamura. Construction of spontaneous emotion corpus from indonesian tv talk shows and its application on multimodal emotion recognition. *IEICE TRANSACTIONS on Information and Systems*, 101(8):2092–2100, 2018.

[65] Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. Construction and analysis of social-affective interaction corpus in English and Indonesian. In *O-COCOSDA/CASLRE, 2015 International Conference*, pages 202–206. IEEE, 2015.

[66] Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, Ayu Purwarianti, and Satoshi Nakamura. Emotion and its triggers in human spoken dialogue: Recognition and analysis. In *Proceedings of International Workshop on Spoken Dialogue Systems*, 2014.

[67] Nurul Lubis, Sakriani Sakti, Graham Neubig, Koichiro Yoshino, Tomoki Toda, and Satoshi Nakamura. A study of social-affective communication: Automatic prediction of emotion triggers and responses in television talk shows. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA, December 2015.

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

[69] 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.

[70] 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.

[71] Michael J Lyons, Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba, and Julien Budynek. The Japanese female facial expression (JAFFE) database, 1998.

[72] Antony SR Manstead and Agneta H Fischer. Social appraisal: The social world as object of and influence on appraisal processes. *Appraisal processes in emotion: Theory, methods, research*, pages 221–232, 2001.

[73] Stacy Marsella and Jonathan Gratch. A step toward irrationality: using emotion to change belief. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pages 334–341. ACM, 2002.

[74] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[75] 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.

[76] Toyomi Meguro, Yasuhiro Minami, Ryuichiro Higashinaka, and Kohji Dohsaka. Learning to control listening-oriented dialogue using partially observable markov decision processes. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(4):15, 2013.

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

[78] Magalie Ochs, Catherine Pelachaud, and David Sadek. Emotion elicitation in an empathic virtual dialog agent. In *Prooceedings of the Second European Cognitive Science Conference (EuroCogSci)*, 2007.

[79] Magalie Ochs, Catherine Pelachaud, and David Sadek. An empathic virtual dialog agent to improve human-machine interaction. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pages 89–96. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[80] Brian Parkinson, Agneta H Fischer, and Antony SR Manstead. *Emotion in social relations: Cultural, group, and interpersonal processes*. Psychology Press, 2004.

[81] 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.

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

[83] Rosalind W Picard and Roalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

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

[85] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[86] Carl Edward Rasmussen. The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000.

[87] Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. Let's go public! taking a spoken dialog system to the real world. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[88] Byron Reeves and Clifford Nass. *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press, 1996.

[89] 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.

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

[91] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[92] 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.

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

[94] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

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

[96] KR Scherer. Component models of emotion can inform the quest for emotional competence. *The science of emotional intelligence: Knowns and unknowns*, pages 101–126, 2007.

[97] Björn Schuller, Stefan Steidl, and Anton Batliner. The INTERSPEECH 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315, 2009.

[98] 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.

[99] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*, 2015.

[100] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301, 2017.

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

[102] Craig A Smith, Richard S Lazarus, et al. Emotion and adaptation. *Handbook of personality: Theory and research*, pages 609–637, 1990.

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

[104] Iain AD Stewart and Portia File. Let's chat: A conversational dialogue system for second language practice. *Computer Assisted Language Learning*, 20(2):97–116, 2007.

[105] Jörg Tiedemann. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248, 2009.

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

[107] Michael Tomasello. *Origins of human communication*. MIT press, 2010.

[108] Alan M Turing. Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer, 2009.

[109] 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.

[110] H Van Dyke Parunak, Robert Bisson, Sven Brueckner, Robert Matthews, and John Sauter. A model of emotions for situated agents. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 993–995. ACM, 2006.

[111] JD Velsquez. Modeling emotions and other motivations in synthetic agents. *Aaai/iaai*, pages 10–15, 1997.

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

[113] Richard Wallace. The elements of aiml style. *Alice AI Foundation*, 2003.

[114] Richard S Wallace. The anatomy of alice. In *Parsing the Turing Test*, pages 181–210. Springer, 2009.

[115] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[116] 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.

[117] Bruce Wilcox, What Files are Where, and Simple Patterns. Chatscript basic user's manual. *URL: https://github. com/bwilcox-1234/ChatScript/blob/master/WIKI/ChatScript-Basic-User-Manual. md (Aufruf am 05.01. 2018)*, 2017.

[118] Bruce Wilcox, Sue Wilcox, BA Psych, and Dipl Fine Arts. Suzette, the most human computer. *Agent's Processing, Cognition: https://www. chatbots. org/images/uploads/researchJpapers/9491. p df*, 2010.

[119] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, 2013.

[120] 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.

[121] Tiancheng Zhao and Maxine Eskenazi. Zero-shot dialog generation with cross-domain latent actions. *arXiv preprint arXiv:1805.04803*, 2018.

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

List of Publications

Journals

- "Positive Emotion Elicitation in Chat-based Dialogue Systems." Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura., IEEE Transactions on Audio, Speech and Language Processing. *(accepted, awaiting publishing)*

- "Construction of Spontaneous Emotion Corpus from Indonesian TV Talk Shows and Its Application on Multimodal Emotion Recognition." Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura., Transactions on Information and Systems, Institute of Electronics, Information and Communication Engineers (IEICE): Vol.E101-D No.8, pp.2092-2100. 2018.

- "Emotional Triggers and Responses in Spontaneous Affective Interaction: Recognition, Prediction, and Analysis." Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura., Transactions of the Japanese Society for Artificial Intelligence 33, no. 1 (2018): DSH-D_1-10, 2018.

International Conferences (Peer-reviewed)

- "Optimizing Neural Response Generator with Emotional Impact Information." Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura., Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018.

- "Unsupervised Counselor Dialogue Clustering for Positive Emotion Elicitation in Neural Dialogue System." Nurul Lubis, Sakriani Sakti, Koichiro

Yoshino, Satoshi Nakamura., Proceedings of the 19th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). Association for Computational Linguistics (ACL), 2018.

- "Eliciting Positive Emotion through Affect-Sensitive Dialogue Response Generation: A Neural Network Approach." Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura., Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18). Association for the Advancement of Artificial Intelligence (AAAI), 2018.

- "Processing Negative Emotions through Social Communication: Multimodal Database Construction and Analysis." Nurul Lubis, Michael Heck, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura., Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). Association for the Advancement of Affective Computing (AAAC), 2017.

- "Eliciting Positive Emotional Impact in Dialogue Response Selection." Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura., Proceedings of the 2017 International Workshop on Spoken Dialogue System Technologies (IWSDS). 2017.

- "Construction of Japanese Audio-Visual Emotion Database and Its Application in Emotion Recognition." Nurul Lubis and Randy Gomez and Sakriani Sakti and Keisuke Nakamura and Koichiro Yoshino and Satoshi Nakamura and Kazuhiro Nakadai, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), 2016.

- "Emotion and its triggers in human spoken dialogue: Recognition and analysis." Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, Ayu Purwarianti, and Satoshi Nakamura. Situated Dialog in Speech-Based Human-Computer Interaction. Springer International Publishing, 2016.

- "A Study of Social-Affective Communication: Automatic Prediction of Emotion Triggers and Responses in Television Talk Shows." Nurul Lubis, Sakriani Sakti, Graham Neubig, Koichiro Yoshino, Tomoki Toda, Satoshi Nakamura. Proceedings of the 2015 IEEE ASRU. IEEE, 2015.

- "Construction and Analysis of Social-Affective Interaction Corpus in English and Indonesian." Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. Proceedings of the 18th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA). 2015.
- "Emotion recognition on Indonesian television talk shows." Nurul Lubis, Dessi Lestari, Ayu Purwarianti, Sakriani Sakti, and Satoshi Nakamura. Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2014.
- "Construction and analysis of Indonesian emotional speech corpus." Nurul Lubis, Dessi Lestari, Ayu Purwarianti, Sakriani Sakti, and Satoshi Nakamura. Proceedings of the 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA). IEEE, 2014.

Domestic Conferences (Non Peer-reviewed)

- "Multimodal Database of Negative Emotion Recovery in Dyadic Interactions: Construction and Analysis." Nurul Lubis, Michael Heck, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura. ASJ, Sep, 2018.
- "Positive Emotion Elicitation in an Example-Based Dialogue System." Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura. SIG-SLP, Feb, 2018.
- "Dialogue Modeling for Eliciting Positive Emotion." Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura. ASJ, Sep, 2017.
- "Constructing a Japanese Multimodal Corpus From Emotional Monologues and Dialogues." Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura. IEICE-SP, Dec, 2016.
- "Predicting Emotional Responses from Spontaneous Social-Affective Interaction Data." Nurul Lubis, Sakriani Sakti, Graham Neubig, Koichiro Yoshino, Satoshi Nakamura. ASJ, Mar, 2016.

- "Recognition and Analysis of Emotion in Indonesian Conversational Speech." Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, Dessi Lestari, Ayu Purwarianti, and Satoshi Nakamura. SIG-SLP, Dec, 2014.