Positive Emotion Elicitation in Chat-based Dialogue Systems

Nurul Lubis, *Non-Member, IEEE,* Sakriani Sakti, *Member, IEEE,* Koichiro Yoshino, *Member, IEEE,* Satoshi Nakamura, *Fellow, IEEE*

***Abstract*—We aim to draw on an important overlooked potential of affective dialogue systems: their application to promote positive emotional states, similar to that of emotional support between humans. This can be achieved by eliciting a more positive emotional valence throughout a dialogue system interaction, i.e., *positive emotion elicitation*. Existing works on emotion elicitation have not yet paid attention to the emotional benefit for the users. Moreover, a positive emotion elicitation corpus does not yet exist despite the growing number of emotion-rich corpora. Towards this goal, firstly, we propose a response retrieval approach for positive emotion elicitation by utilizing examples of emotion appraisal from a dialogue corpus. Secondly, we efficiently construct a corpus using the proposed retrieval method, by replacing responses in a dialogue with those that elicit a more positive emotion. We validate the corpus through crowdsourcing to ensure its quality. Lastly, we propose a novel neural network architecture for an emotion-sensitive neural chat-based dialogue system, optimized on the constructed corpus to elicit positive emotion. Objective and subjective evaluations show that the proposed methods result in dialogue responses that are more natural and elicit a more positive emotional response. Further analyses of the results are discussed in this paper.**

***Index Terms*—natural language processing, emotion elicitation, affective computing, chat-based dialogue system, neural networks**

I. Introduction

It has been argued that humans impose the emotional aspect of social communications in their interaction with computers and machines [1]. As the technology develops, the potential of agents to improve the emotional well-being of users has been growing as well. In particular, emotionally intelligent systems could provide assistance and prevention measures through various affective tasks.

Two of the most studied emotional competences for agents are *emotion recognition* and *emotion simulation*. *Emotion recognition* allows a system to discern the user's emotions and address them in giving a response [2], [3], [4]. On the other hand, *emotion simulation* helps convey non-verbal aspects to the user for a more believable and human-like interaction, for example to show empathy [5] or personality [6]. Acosta and Ward [7] have attempted to connect the two competences to build rapport, by recognizing user's emotion and reflecting it in the system response. Although these competences address some of the user's emotional needs [8], they are not sufficient to provide emotional support in an interaction.

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 [9]. This is argued to be an essential part of the emotional processes [10]. On the contrary, the absence of such support in times of need can result in more serious, longer-term consequences. Unfortunately, there is a lack of studies examining negative emotion commonly encountered in everyday life.

To provide support for healthy users, there exist technologies such as listening oriented systems [11], [12] and companion conversational agents [13]. However, these studies do not consider negative emotional experiences nor recovery from them. On the other hand, there exist efforts in addressing a number of clinical emotional disturbances, such as depression and suicide risk, through affective computing, with speech processing techniques [14], or distress clues assessments [15]. Unfortunately, these works are not applicable for the larger, general audience as they are focused on clinical circumstances.

Our work aims to draw on an important overlooked potential of emotion in dialogue systems: its utilization to improve emotional experience and promote positive emotional states, similar to that of emotional support between humans for prompt recovery from negative emotion in everyday situations. This can be achieved by actively eliciting a more positive emotional valence throughout a chat-based interaction, i.e., *positive emotion elicitation*.

Skowron et al. have studied the impact of different affective personalities in a text-based dialogue system [16], reporting consistent impacts with the corresponding personality in humans. Hasegawa et al. constructed translation-based response generators with various emotion targets to elicit a pre-defined emotional state [17]. A shortcoming of existing works is that they have not yet paid attention to the emotional experience of the users. Instead, they focus on one-dimensional personalities or target emotions on the system side to achieve the elicitation. Furthermore, despite the growing number of emotion-rich corpora [18], [19], [20], a positive emotion elicitation corpus does not yet exist.

Towards positive emotion elicitation in chat-based dialogue systems, we first propose a response retrieval method which exploits emotion appraisal in dialogue for an example-based dialogue system (Section VI). We augment the response selection criteria with emotional parameters to select the response

All authors are with the Division of Information Science, Nara Institute of Science and Technology, Japan, e-mail: {nurul.lubis.na4, ssakti, koichiro, s-nakamura}@is.naist.jp.

Koichiro Yoshino is with PRESTO, Japan Science and Technology Agency, Japan.

Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura are with the Center for Advanced Intelligence Project AIP, RIKEN, Japan.

Manuscript received April 19, 2005; revised August 26, 2015.

that has the most potential to elicit positive emotion. Secondly, we construct a positive emotion elicitation corpus by obtaining new responses that elicit a more positive emotion with the proposed retrieval method (Section VII). We validate the corpus through crowdsourcing to ensure its quality. Lastly, we propose a neural chat-oriented dialogue system that captures user's emotional state and considers it in generating a dialogue response (Section VIII). We achieve positive emotion elicitation without the need of an explicit dialogue strategy by optimizing the emotion-sensitive network with the constructed corpus. The overview is illustrated in Figure 1.

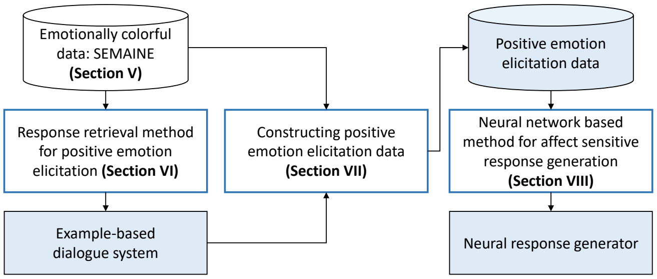

Emotionally colorful data: SEMAINE (Section V)
Positive emotion elicitation data
Response retrieval method for positive emotion elicitation (Section VI)
Constructing positive emotion elicitation data (Section VII)
Neural network based method for affect sensitive response generation (Section VIII)
Example-based dialogue system
Neural response generator

Fig. 1: Overview. White boxes with blue border show proposed methods. Blue shapes show resulting models and data.

II. Related Works

Two main approaches in building chat-based dialogue systems are *response retrieval* and *response generation*. These approaches operate in different ways and therefore have different advantages and disadvantages, which will be discussed in this section. The retrieval method relies on a pre-defined set of example responses. The main question to be solved is the manner in which the system retrieves an appropriate response from its pre-defined set, given a user input. In contrast, the generation approach produces the dialogue response sequentially, from beginning to end, based on the probability distribution it learned from the training data. This imposes an additional challenge in training since the model is also required to learn how to form a correct and grammatical response in addition to it being coherent with the dialogue content.

One particular retrieval method is example-based dialogue modeling (EBDM), which uses a semantically indexed corpus of query-response examples instead of handcrafted rules or probabilistic models [21]. At a given time, the system will return the best response according to semantic similarity between user input and the dialogue examples in its database. This circumvents the challenge of domain identification and switching – a particularly hard task in chat-oriented systems where no specific goal or domain is predefined.

Despite its domain-free property, the EBDM approach has not yet been studied to its full potential for affective dialogue systems. An example is a micro-counseling dialogue system that uses the example-based approach to select a template for slot-filling response generation [3]. However, this approach still relies on a small number of rule-based strategies. Strong limitations of the EBDM approach include the lack of variety of responses and the difficulty in handling inputs that are not included in the example database, i.e., out-of-example cases.

With recent advancements in neural network research, end-to-end approaches have shown promising results for non-goal oriented dialogue systems [22], [23], [24]. Instead of retrieving a response from an example database, neural dialogue systems generate the response utterance, most commonly as a sequence of words, based on the input query. These approaches rid the need for explicit definition of dialogue states and action spaces, allowing for a more dynamic model that mimics human conversation contained in the training corpus. Furthermore, neural dialogue systems may still be able to produce a natural response given a query that does not exist in the training data, solving the out-of-example problem of the example-based approach. These qualities are especially desirable for domain-free, chat-oriented dialogue systems. Unfortunately, as with response retrieval, application of this approach towards incorporating emotion in the dialogue is still very lacking.

Only recently, Zhou et al. published their work addressing the emotional factor in neural network response generation [25]. They examined the effect of internal emotional states on the decoder, investigating 6 categories to emotionally color the response, similar to that of emotion simulation. However, this study has not yet considered user's emotion in the response generation process, nor attempted to utilize emotion to improve user experience. To the best of our knowledge, neural network approaches have not yet been utilized for affect-sensitive response generation and emotion elicitation.

To bridge these gaps, we propose response retrieval and generation methods to incorporate emotion into chat-based dialogue systems. We exploit emotion appraisal in dialogue for an example-based dialogue system. Modifying the traditional response selection criteria with emotional parameters allows us to select the response that has the most potential to elicit positive emotion. For response generation, we propose a neural chat-oriented dialogue system that captures user's emotional state and considers it while generating a dialogue response. We train the latter with responses retrieved by the former to achieve end-to-end positive emotion elicitation.

III. Emotion Definition

In this work, we define emotion using the *circumplex model of affect* [26]. 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 illustrates the valence-arousal dimension in respect to a number of common emotion terms.

This model describes the perceived form of emotion, and is able to represent both primary and secondary emotion. Furthermore, it is intuitive and easily adaptable and extendable to either discrete or other dimensional emotion definitions. The long established dimensions are core to many works in affective computing and potentially provide useful information even at an early stage of research.

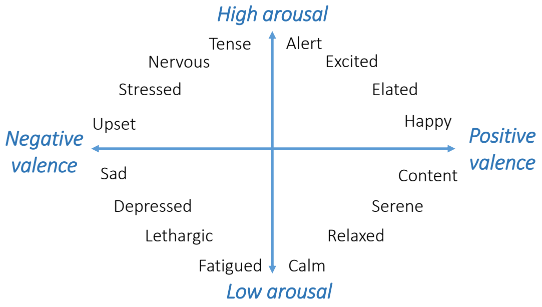


Fig. 2: Emotion dimensions and common terms.

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

IV. Dialogue Definition

Serban et al. have previously considered the two-hierarchy view of dialogue [23], which we extend and adapt in this study. First, we view a dialogue D as a sequence of dialogue turns of arbitrary length M between two speakers. That is, $D = \{U_1, ..., U_M\}$. Each utterance in the m-th dialogue turn is a sequence of tokens of arbitrary length N_m. That is, $U_m = \{w_{m,1}, ..., w_{m,N_m}\}$.

Throughout the study and experiments, we utilize the dialogue triple format. A *triple* is a sequence of three dialogue turns. That is, $D = \{U_1, U_2, U_3\}$. As we are focused on dyadic dialogue, we consider U_1 and U_3 to be uttered by speaker A, and U_2 by speaker B. The triple format has been previously utilized for considering context in response generation [27], and filtering multi-party conversations into dyadic snippets [28].

In this study, we assume two speaker orders in the triple format. First, user-system-user, to observe emotional triggers and responses in a conversation. The change of emotion observed from U_1 to U_3 can be regarded as the impact of U_2. Second, system-user-system, to observe both past and future dialogue contexts of an emotion occurrence; U_1 and U_3 are the contexts of the emotion occurrence in U_2.

V. Emotionally Colorful Conversational Data

In this study, we utilize the SEMAINE database as an emotion-rich conversational data [18]. The SEMAINE database consists of spontaneous dialogues between a user and a sensitive artificial listener (SAL) in a Wizard-of-Oz fashion. A SAL is a system capable of holding a multimodal conversation with humans, involving speech, head movements, and facial expressions, topped with emotional coloring [29]. This emotional coloring is adjusted according to each of the SAL characters; cheerful Poppy, angry Spike, sad Obadiah, and sensible Prudence. Each character is distinct and unchanging, i.e., Poppy is constantly happy, Spike is constantly angry, etc., independent of the dialogue content.

The corpus consists of a number of sessions, in which a user is interacting with a wizard SAL character. Each user interacts with all 4 characters, with each interaction typically lasting for 5 minutes. The topics of conversation are spontaneous, with a limitation that the SAL can not answer any questions.

The interactions are annotated using the FEELtrace system [30] to allow recording of perceived emotion in real time. As an annotator is watching a target person in a video recording, they would move a cursor along a linear scale on an adjacent window to indicate the perceived emotional aspect (e.g., valence or arousal) of the target. This results in a sequence of real numbers ranging from -1 to 1, called a *trace*, that shows how a certain emotional aspect falls and rises within an interaction. The numbers in a trace are provided with an interval of 0.02 seconds. Statistical analyses of validation experiments have confirmed the reliability and indicated the precision of the FEELtrace system [30].

In total, 95 sessions are provided, amounting to 475 minutes of material. In this study, we consider 66 sessions from the corpus based on transcription and emotion annotation availability; 17 of Poppy's sessions, 16 of Spike, 17 of Obadiah, 16 of Prudence. For every dialogue turn, we keep the speaker information, time alignment, transcription, and emotion traces.

VI. Response Retrieval: Eliciting Positive Emotion Based on Examples of Human Appraisal

EBDM is a data-driven approach that uses a semantically indexed corpus of query-response pair examples, i.e. $\{U_1, U_2\}$, instead of handcrafted rules or probabilistic models [21]. In traditional EBDM, the system will return a response of the best example according to a semantic constraint between user input and example queries, i.e. examples of U_1.

Lasguido et al. have previously examined the utilization of cosine similarity for response retrieval in an example-based dialogue system [28]. 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:

$$\text{TFIDF}(t, T) = F_{t,T} \log \frac{|T|}{DF_t}, \tag{1}$$

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:

$$\cos_{sim}(a, b) = \frac{a \cdot b}{\|a\| \, \|b\|}. \tag{2}$$

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 approach has a number of benefits. First, the TF-IDF weighting allows emphasis of important words. Such quality is desirable in considering emotion in spoken utterances. Second, as this approach does not rely on explicit domain knowledge, it is practically suited for adaptation into an affective dialogue system. Third, the approach is straightforward and highly reproducible. On that account, it serves as the baseline in this section.

A. Proposed Approach

Limitation of existing works on emotion elicitation [16], [17] is the oversight of emotion appraisal, which gives rise to the elicited emotion in the first place. This entails the relationship between an utterance, which acts as stimulus evaluated during appraisal (*emotional trigger*), and the resulting emotion by the end of appraisal (*emotional response*) [31]. By reversing this loop, an appraisal-driven approach attempts to determine the appropriate trigger to a desired emotional response. This 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.

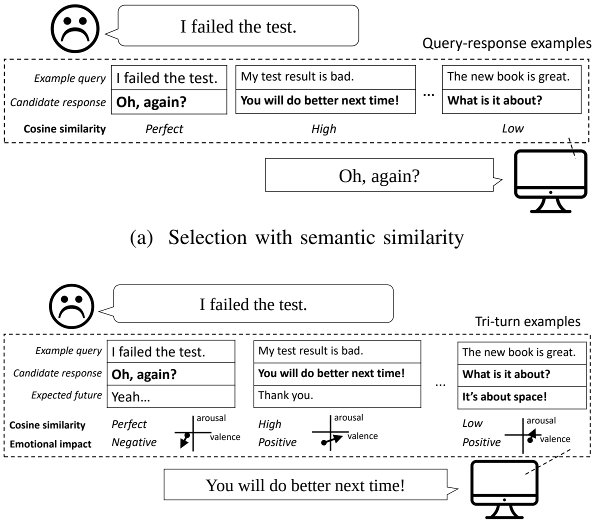
We attempt to elicit a positive emotional change by exploiting examples of appraisal in spontaneous dialogue. We collect triples containing emotional triggers and responses to serve as examples in an EBDM. We augment the traditional response selection criterion with emotional parameters: 1) user's emotional state, and 2) expected[1] future emotional impact of the candidate responses. These parameters represent parts of the information that humans use in social-affective interactions.

To take emotional aspects into account, in addition to the semantic constraint in traditional EBDM, 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. To allow observation of these constraints, we use a database of triples in place of query-response pairs used in traditional EBDM. Within the context of triples in an example database, U_1 is equivalent to the example query, U_2 to a candidate response, and U_3 to the future context. Figures 3a and 3b illustrate the general idea of the baseline and proposed approaches.

First, we measure *emotion similarity* by computing the Pearson's correlation coefficient between emotion vectors of the query and the example queries in the database. Pearson's r_{ab} between two vectors a and b of length n is calculated as

$$r_{ab} = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}}. \tag{3}$$

In this case, a and b symbolize the emotion representations of query and example query. We utilize the trace provided in the SEMAINE database as emotion representation (Section V). 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.



(a) Selection with semantic similarity

(b) Selection with semantic similarity and emotion parameters

Fig. 3: Response selection in baseline and proposed systems.

Secondly, we measure the *expected emotional impact* of the candidate responses. In a triple, emotional impact of a response U_2 is computed according to the query U_1 and future U_3 using Equation (4).

$$\text{impact}(U_2) = \frac{1}{n}\sum_{i=1}^{n} emo_{U_3\,i} - \frac{1}{n}\sum_{i=1}^{n} emo_{U_1\,i}, \tag{4}$$

where emo_{U_1} and emo_{U_3} are the emotion vectors of U_1 and U_3. In other words, the actual emotion impact observed in an example is the expected emotional impact during the real interaction. As positivity of emotion is described by valence, for the expected emotional impact, we consider only valence as the final score.

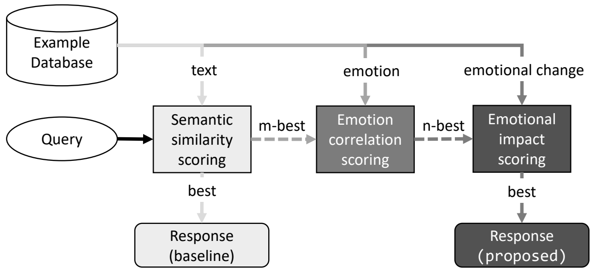


Fig. 4: Steps of response selection.

Figure 4 illustrates the steps of response selection by 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, calculated with Equation (2). In contrast, in the proposed system's

[1] Within the scope of the proposed method, we use the word *expected* for its literal meaning, as opposed to its usage as a term in probability theory.

response selection, we pass m examples with highest semantic similarity scores to the next step and calculate their emotion similarity scores with Equation (3). From n examples with highest emotion similarity scores, we output the *response* of the triple example with the most positive expected emotional impact, calculated using Equation (4).

The scores are processed incrementally considering the importance of each aspect. In selecting a response, we consider semantic similarity to be most important as it ensures coherence of the response. If an example is not semantically relevant to the input, we should not consider it as a response regardless of its emotional scores. Similarly, emotion similarity is important to guarantee consistent emotional impact with that of the data; an utterance can result in different emotional responses given different initial emotional states. Imposing the emotional impact constraints in the reduced pool of semantically and emotionally similar examples will help achieve a more relevant result. Additionally, this reduces the computation time since the number of examples to be scored will be greatly minimized after each step. When working with big example databases, this property is beneficial for a timely response.

Consequently, there are three important points to note regarding the proposed approach. First, U_3, the future context of each triple, is not considered as a definite prediction of user response when interacting with the system. Instead, each triple acts as an example of human's appraisal in a conversation with certain semantic and emotional contexts. In real interaction, given similar semantic and emotional contexts with an example triple's U_1, when the system outputs U_2, we expect the user to experience an emotional change similar to that of U_3. The semantic and emotion similarities between user input and U_1 are important in ensuring an emotional response that is as close as possible to the example.

Second, this strategy does not translate to selection of the *response* that expresses the most positive emotion. Our goal is not to consistently output a "happy response." There are situations where such a response is inappropriate and could cause the opposite effect, e.g., in times of grief. Instead, we would like to reflect proper positive emotion elicitation given semantic and emotional contexts. Even though there is no explicit dialogue strategy to be followed, we expect the data to reflect 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. Lastly, we note that response similarity with the target is not the objective of our retrieval method. Our goal is not to replicate the target response through the example database, but instead to retrieve a response that has most potential to elicit a more positive emotion.

B. Experimental Set Up

We gather the examples from the SEMAINE corpus by extracting triples with user-system-user speaker sequence. An analysis shows that Poppy and Prudence tend to draw the user into the positive-valence region of emotion, resembling a *positive emotional impact*. The same is not true for sessions with Spike and Obadiah. Thus, we exclusively use sessions of Poppy and Prudence to construct the example database. We partition the recording sessions in the corpus into training and test sets. The training set and test set comprise 29 (15 Poppy, 14 Prudence) and 4 (2 Poppy, 2 Prudence) sessions, respectively. We construct the example database exclusively from the training set, containing 1105 triples.

We average as many annotations as provided in a session to obtain the final valence and arousal label. We sample the emotion trace of every dialogue turn into 100-length vectors to keep the length uniform. We utilize the transcription and emotion annotation provided within the corpus as information of the triples to isolate the errors of automatic speech and emotion recognition. For the n-best filtering, we empirically chose 10 for the semantic similarity constraint and 3 for the emotion.

C. Evaluation

To evaluate the proposed method, we perform subjective evaluation to qualitatively measure perceived differences between the two response selection methods. From the test set, we extract 198 test queries. For each test query, we generate responses using the baseline and proposed systems explained in Section VI-A. Queries with identical responses from the two systems are excluded from the evaluation. We further filter the queries based on utterance length, to give enough context to the evaluators; and emotion labels, to give variance in the evaluation. In the end, 50 queries are selected.

We perform subjective evaluation of the systems with crowdsourcing. The queries and responses are presented in form of text. We ask the evaluators to compare the systems' responses in respect to the test queries. For each test query, the responses from the systems are presented with random ordering, and the evaluators are asked three questions, adapted from [16]. 1) Which response is more coherent? Coherence refers to the logical continuity of the dialogue, 2) Which response has more potential in building emotional connection between the speakers? Emotional connection refers to the potential of continued interaction and relationship development, and 3) Which response gives a more positive emotional impact? Emotional impact refers to the potential emotional change the response may cause.

50 judgments are collected per query. Each judgment is weighted with the level of trust of the worker[2]. The final judgment of each query for each question is based on the total weight of the overall judgements – the system with the greater weight wins.

TABLE I: Percentage of wins of each system on all metrics. Numbers in brackets show average agreement on queries where the system wins for that metric.

System	Coherence	Emotional impact	Emotional connection
Baseline	0.34 (0.14)	0.30 (0.19)	0.34 (0.17)
Proposed	0.66 (0.38)	0.70 (0.32)	0.66 (0.36)

Table I presents the evaluation results. It is shown that in comparison to the baseline system, the proposed system is

[2] A worker's trust score is equal to their percentage of correct answers on a quiz, consisting of triples with gold standard answers. In this study, we only employ workers with high level of trust.

TABLE II: Candidate responses re-ranking based on three consecutive selection constraints: 1) semantic similarity with example queries, 2) emotion similarity with example queries, and 3) expected emotional impact of the candidate response. *: baseline response, **: proposed response.

Query: Em going to London tomorrow. *(valence: 0.39, arousal: -0.11)*

Candidate responses (U_2)	ranking steps		
	semantic	**emotion**	**impact**
* And where in Australia?	1		
[laugh]	2		
Organized people need to have holiday.	3	1	
It would be very unwise for us to discuss possible external examiners.	4		
[laugh]	5		
It's good that sounds eh like a good thing to do, although you wouldn't want to em overspend.	6		
That sounds interesting you've quite a lot going on so you need to manage your time.	7	2	
Yes.	8		
Mhm.	9		
** That sounds nice.	10	3	1

perceived as more coherent, having more potential in building emotional connection, and giving a more positive emotional impact. We investigate this result further by computing the agreement of the final judgment using Fleiss' Kappa formula. We separate the queries based on the winning system and compute the overall agreement of the 50 judgments accordingly. It is revealed that the queries where the proposed system wins have far stronger agreement than those where the baseline system wins. Agreement between 0.10 and 0.20 can be interpreted as slight agreement, and that between 0.21 and 0.40 as fair agreement [32].

D. Analysis

We analyze the consequence of re-ranking and the effect of emotion similarity in the response selection using queries extracted from the test set. Table II presents the 10-best semantic similarity ranking, re-ranked and filtered into 3-best emotion similarity ranking, and the candidate response that passed the filtering with the best emotional impact. The table shows that the proposed method can select one of the candidate responses that despite being not the best in semantic similarity score, has a higher score in terms of emotion similarity and expected impact compared to responses with semantically more similar examples.

Furthermore, the proposed selection method is able to generate different responses to identical textual input with different emotional contexts. Table III demonstrates this quality. This shows that the system is able to adapt to the user's emotion in giving a response to elicit positive emotion.

VII. Constructing Positive-Emotion Eliciting Data

Despite the existence of numerous emotion-rich data [18], [19], [20], there is not yet a conversational data demonstrating positive emotion elicitation. Such data is valuable for data-driven approaches of dialogue systems. However, the collection of emotional corpora is vastly expensive and labor intensive. Collection of spontaneous emotional data is a sensitive matter, potentially raising moral and ethical issues [33]. On the other hand, portrayal of emotion often presents a mismatch with real-life emotion occurrences.

TABLE III: Baseline and proposed responses for identical text with different emotional contexts. The proposed system can adapt to user emotion, while the baseline method outputs the same response.

Query : Thank you. *(valence: 0.13, arousal: -0.18)*
Baseline : Thank you very much that **Proposed :** And I hope that everything goes exactly according to plan.

Query : Thank you. *(valence: 0.43, arousal: 0.05)*
Baseline : Thank you very much that **Proposed :** It is always a pleasure talking to you you're just like me.

Our goal is to construct a corpus that demonstrates positive emotion elicitation. To efficiently collect responses that elicit positive emotion, we enhance the emotion-rich SEMAINE corpus through the procedure illustrated in Figure 5.

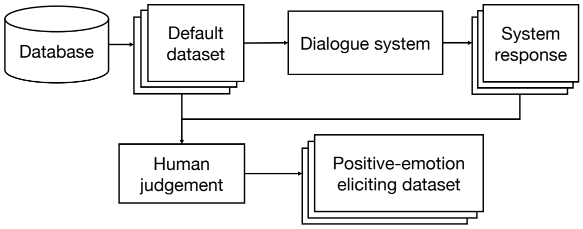


Fig. 5: Obtaining references that elicit positive emotion.

First, we extract system-user-system triples and run them through the EBDM with the proposed response retrieval method (Section VI-A) to obtain new candidate responses that potentially elicit positive emotion. As noted previously, the response retrieval method makes a careful distinction between the expressed emotion of a response and the emotion it may elicit in the listener. Through this response selection method we aim to reflect proper positive emotion in the corpus instead of solely selecting "happy responses."

Subsequently, we validate the data through crowdsourcing. We ask human judges to decide which *response* in a triple, i.e., U_3, elicits a more positive emotional impact in the triple, the default or the system generated one. If neither is 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.

We utilize 66 sessions based on the availability transcription and emotion annotation. We fed a total of 2,349 triples to the entire process. 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 judgment process with an average trust score of 0.93. The average consensus of the voting is 0.78.

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 is 6.09 words.

VIII. Response Generation: Emotion-sensitive Neural Dialogue System

The most commonly used neural architecture for dialogue response generation is the recurrent neural network (RNN). An RNN is a neural network variant that can retain information over sequential data. In response generation, first, an *encoder* summarizes an input sequence into a vector representation. An input sequence at time t is modeled using the information gathered by the RNN up to time $t-1$, contained in the hidden state h_t. Afterwards, a *decoder* recurrently predicts the output sequence conditioned on h_t and its output from the previous time step. This architecture was previously proposed as neural conversational model in [22].

Based on the two-hierarchy view of dialogue (Section IV), the hierarchical recurrent encoder-decoder (HRED) extends the sequence-to-sequence architecture [23]. 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. This information is then passed on to the *dialogue encoder*, which encodes the sequence of dialogue turns h_{dlg}. The *utterance decoder*, or the response generator, takes h_{dlg}, and then predicts the probability distribution over the tokens in the next utterance. The advantages of this architecture are: 1) summarization of dialogue history by the dialogue encoder, containing common knowledge between the two speakers, and 2) reduced computational steps between utterances, allowing a more stable optimization during the training phase.

A. Proposed approach

The HRED architecture holds a property which is essential in positive emotion elicitation: retaining dialogue history at turn level. This allows the observation of both past and future contexts of an emotion occurrence, i.e., U_1 and U_3 are the contexts of the emotion occurrence in U_2. We 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. We propose to incorporate this information in generating a dialogue response with an emotion-sensitive hierarchical recurrent encoder-decoder (Emo-HRED).

1) Architecture: We utilize RNNs with gated recurrent unit (GRU) cells as parts 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}.

$$h_{utt} = h^{utt}_{N_m} = f(h^{utt}_{N_m-1}, w_{m,N_m}), \tag{5}$$

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}.

$$h_{dlg} = h^{dlg}_{m} = f(h^{dlg}_{m-1}, h_{utt}). \tag{6}$$

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}.

$$h_{emo} = f(h^{emo}_{m-1}, h_{dlg}). \tag{7}$$

We consider utilizing additional information in modeling the emotion context. 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. In such a case, Equation (7) is modified as follows.

$$h_{emo} = f\big(h^{emo}_{m-1}, \text{concat}(h_{dlg}, h_{aud})\big). \tag{8}$$

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

$$P_\theta(w_{n+1} = v | w_{\leq n}) = \frac{\exp\big(g(\text{concat}(h_{dlg}, h_{emo}), v)\big)}{\sum_{v'} \exp\big(g(\text{concat}(h_{dlg}, h_{emo}), v')\big)}. \tag{9}$$

Figure 6 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.

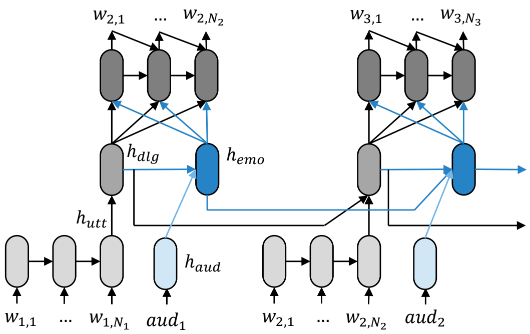


Fig. 6: Emo-HRED architecture.

The emotion encoder has its own target vector, which is the emotion label of the currently processed dialogue turn U^{emo}_m. The emotion encoder uses hyperbolic tangent activation function. We modify the definition of the training cost to incorporate the prediction error of the emotion encoder and use this cost to jointly train the entire network. We define

$cost_{emo}$ as the mean squared error between the target emotion context and the predicted emotion context.

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

$$cost = \alpha \cdot cost_{emo} + (1 - \alpha) \cdot cost_{utt}. \qquad (10)$$

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 that extent, 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.

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.

B. Experimental Set Up

1) Pretraining: Previous works have demonstrated the effectiveness of large scale conversational data in improving the quality of dialogue systems [34], [35], [23]. In this study, we make use of SubTle, a large scale conversational corpus, to learn the syntactic and semantic knowledge for response generation. The use of movie subtitles is particularly suitable as they reflect natural human conversation and are available in large amounts.

The SubTle corpus [35] 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.

In our experiments, we utilize the HRED trained on the SubTle corpus as our starting model. The data preprocessing steps are performed as in [23]. 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.

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 VIII-A2). 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 VII). 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) is used in place of Equation (7). We extract the INTERSPEECH 2009 emotion challenge baseline features [36] using the OpenSMILE toolkit [37]. 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.

TABLE IV: 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.	1. Model	2. Selective fine-tune	3. Positive data	4. Audio encoder	5. 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

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.

C. Evaluation and Analysis

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 is summarized in Table IV. To obtain deeper insight into the evaluation results, analyses are provided at the end of this section.

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 [23] and has been previously recommended for evaluating generative dialogue systems [38]. 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 IV 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 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 pretrained 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.

The fine-tune set up of row No. 1 is equivalent to the SubTle bootstrap approach proposed in [23]. However, there are differences that are important to highlight, summarized in Table V, which made it not possible to straightforwardly compare our results with those reported in [23]. Nonetheless, this comparison 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.

2) Human 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

TABLE V: Model comparison.

	Serban, et al. [23]	This research	
Pretraining	SubTle bootstrap		
Fine-tune and test data	MovieTriples	Positive SEMAINE	
# triples	245,492	2,349	
Architecture	HRED Bidirectional	Emo-HRED	
Emotion	No	Yes	
Audio	No	No	Yes
Perplexity	26.81	35.92	20.35

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 IV 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 (p<0.05) compared to all models, except model No. 8.

Finally, we perform the same subjective evaluation to the model proposed in Section VI and compare the results with the best proposed Emo-HRED (model No.10). The EBDM-based model achieves a score of 3.32 ± 0.30 and 3.20 ± 0.25 for naturalness and emotional impact, respectively. T-test reveal statistically significant difference (p<0.05) between the two models in terms of subjective perception, i.e. the responses generated by the Emo-HRED approach is perceived as more natural and elicit a more positive emotional impact.

3) Analysis: 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 IV.

Table VI 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 IV).

TABLE VI: 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 VII lists the top 10 words from these vocabularies in order of frequency, as well as the positive SEMAINE vocabulary for reference.

TABLE VII: 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

It is interesting 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 `pos_sem` dataset and

Emo-HRED model could explain the lower perplexity when one of them is employed, and lowest when both are.

To clarify, this is not to say that short, generic responses are always desirable. This is a standing problem for neural network based response generation [39] – 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.

IX. Conclusion

In this paper we presented a research effort towards positive emotion elicitation in chat-based dialogue systems. First, we utilized examples of human appraisal in an example based dialogue system. We augment the traditional response selection criteria to take into account the emotional context of the dialogue and the emotional impact of the candidate responses. Secondly, we employed the proposed retrieval method to efficiently alter an emotion rich conversational data into a positive emotion elicitation data by replacing the dialogue responses with ones that elicit a more positive emotion. The data is validated through crowdsourcing to ensure its quality. Lastly, we proposed a hierarchical neural dialogue system with an emotion encoder to capture the emotional context of the dialogue. This information is then used in the response generation process to produce an affect-sensitive response. We optimize the network on the constructed corpus to train a neural response generator that elicits positive emotion. Objective and subjective evaluations show that the proposed methods result in dialogue responses that are more natural and elicit a more positive emotional response.

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

Acknowledgment

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

References

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

[2] K. Forbes-Riley and D. 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*. Association for Computational Linguistics, 2012, pp. 217–226.

[3] S. Han, Y. Kim, and G. G. Lee, "Micro-counseling dialog system based on semantic content," in *Natural Language Dialog Systems and Intelligent Assistants*. Springer, 2015, pp. 63–72.

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

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

[6] A. Egges, S. Kshirsagar, and N. Magnenat-Thalmann, "Generic personality and emotion simulation for conversational agents," *Computer animation and virtual worlds*, vol. 15, no. 1, pp. 1–13, 2004.

[7] J. C. Acosta and N. G. Ward, "Achieving rapport with turn-by-turn, user-responsive emotional coloring," *Speech Communication*, vol. 53, no. 9-10, pp. 1137–1148, 2011.

[8] R. W. Picard and J. Klein, "Computers that recognise and respond to user emotion: theoretical and practical implications," *Interacting with computers*, vol. 14, no. 2, pp. 141–169, 2002.

[9] O. Luminet IV, P. Bouts, F. Delie, A. S. Manstead, and B. Rimé, "Social sharing of emotion following exposure to a negatively valenced situation," *Cognition & Emotion*, vol. 14, no. 5, pp. 661–688, 2000.

[10] B. Rime, B. Mesquita, S. Boca, and P. Philippot, "Beyond the emotional event: Six studies on the social sharing of emotion," *Cognition & Emotion*, vol. 5, no. 5-6, pp. 435–465, 1991.

[11] T. Meguro, Y. Minami, R. Higashinaka, and K. Dohsaka, "Learning to control listening-oriented dialogue using partially observable markov decision processes," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 10, no. 4, p. 15, 2013.

[12] T. Yamaguchi, K. Inoue, K. Yoshino, K. Takanashi, N. G. Ward, and T. Kawahara, "Analysis and prediction of morphological patterns of backchannels for attentive listening agents," in *Proc. 7th International Workshop on Spoken Dialogue Systems*, 2016, pp. 1–12.

[13] M. Cavazza, R. S. De La Camara, and M. 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*. International Foundation for Autonomous Agents and Multiagent Systems, 2010, pp. 1629–1630.

[14] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[15] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. 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*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1061–1068.

[16] M. Skowron, M. Theunis, S. Rank, and A. Kappas, "Affect and social processes in online communication–experiments with an affective dialog system," *Transactions on Affective Computing*, vol. 4, no. 3, pp. 267–279, 2013.

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

[18] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[19] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 865–868.

[20] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The humaine database: addressing the collection and annotation of naturalistic and induced emotional data," in *Affective computing and intelligent interaction*. Springer, 2007, pp. 488–500.

[21] C. Lee, S. Jung, S. Kim, and G. G. Lee, "Example-based dialog modeling for practical multi-domain dialog system," *Speech Communication*, vol. 51, no. 5, pp. 466–484, 2009.

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

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

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

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

[26] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

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

[28] N. Lasguido, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Utilizing human-to-human conversation examples for a multi domain chat-oriented dialog system," *Transactions on Information and Systems*, vol. 97, no. 6, pp. 1497–1505, 2014.

[29] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. T. Maat, G. McKeown, S. Pammi, M. Pantic *et al.*, "Building autonomous sensitive artificial listeners," *Transactions on Affective Computing*, vol. 3, no. 2, pp. 165–183, 2012.

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

[31] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal processes in emotion: Theory, methods, research.* Oxford University Press, 2001.

[32] J. R. Landis and G. G. Koch, "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers," *Biometrics*, pp. 363–374, 1977.

[33] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

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

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

[36] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge." in *INTERSPEECH*, vol. 2009, 2009, pp. 312–315.

[37] F. Eyben, M. Wöllmer, and B. Schuller, "OPENsmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[38] O. Pietquin and H. Hastie, "A survey on metrics for the evaluation of user simulations," *The knowledge engineering review*, vol. 28, no. 1, pp. 59–73, 2013.

[39] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of NAACL-HLT*, 2016, pp. 110–119.

Nurul Lubis received the B.E degree (with distinction) in 2014 from Bandung Institute of Technology, Indonesia and the M.Eng degree in 2017 from Nara Institute of Science and Technology (NAIST), Japan. She is currently a doctoral candidate at Augmented Human Communication Laboratory, NAIST, Japan. She is a recipient of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) scholarship. She was a research intern at Honda Research Institute Japan, Co. Ltd. Her research interest include affective computing, emotion in spoken language, and affective dialogue systems.

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005-2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003-2007), A-STAR and U-STAR (2006-2011). In 2009-2011, she served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia. In 2011-2017, she was an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She served also as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015-2016, under JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation. Currently, she is a research associate professor at NAIST, as well as a research scientist at RIKEN, the Center of for Advanced Intelligent Project AIP, Japan. She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. She is also the officer of ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL) and a board Member of Spoken Language Technologies for Under-Resourced Languages (SLTU). Her research interests include statistical pattern recognition, graphical modeling framework, deep learning, multilingual speech recognition & synthesis, spoken language translation, affective dialog system, and cognitive communication.

Koichiro Yoshino received his B.A. degree in 2009 from Keio University, M.S. degree in informatics in 2011, and Ph.D. degree in informatics in 2014 from Kyoto University, respectively. From 2014 to 2015, he was a research fellow (PD) of Japan Society for Promotion of Science. From 2015 to 2016, he was a research assistant professor of the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST). Currently, he is an assistant professor of NAIST. He is also a researcher of PRESTO, JST, concurrently. He is working on areas of spoken and natural language processing, especially on spoken dialogue systems. Dr. Koichiro Yoshino received the JSAI SIG-research award in 2013. He is a member of IEEE, ISCA, IPSJ, and ANLP.

Satoshi Nakamura is Professor of Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan, Project Leader of Tourism Information Analytics Team of RIKEN, Center for Advanced Intelligence Project AIP, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampolli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.