Eliciting Positive Emotional Impact in Dialogue Response Selection

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, Satoshi Nakamura

Abstract Introduction of emotion into human-computer interaction (HCI) have allowed various system's abilities that can benefit the user. Among many is emotion elicitation, which is highly potential in providing emotional support. To date, works on emotion elicitation have only focused on the intention of elicitation itself, e.g. through emotion targets or personalities. In this paper, we aim to extend the existing studies by utilizing examples of human appraisal in spoken dialogue to elicit a positive emotional impact in an interaction. We augment the widely used example-based approach with emotional constraints: 1) emotion similarity between user query and examples, and 2) potential emotional impact of the candidate responses. Text-based human subjective evaluation with crowdsourcing shows that the proposed dialogue system elicits an overall more positive emotional impact, and yields higher coherence as well as emotional connection.

1 Introduction

To a large extent, emotion determines our quality of life [5]. However, it is often the case that emotion is overlooked or even treated as an obstacle. The lack of awareness of the proper care of our emotional health has led to a number of serious problems, including incapabilities in forming meaningful relationships, sky-rocketing stress level, and a large number of untreated cases of emotion-related disturbances. In dealing with each of these problems, outside help from another person is invaluable.

The emotion expression and appraisal loop between interacting people creates a rich, dynamic, and meaningful interaction. When conducted skillfully, as performed by experts, a social-affective interaction can provide social support, reported to give positive effect with emotion-related problems [2]. Unfortunately, an expert is a limited and costly resource that is not always accessible to those in need. In this regard, an emotionally-competent computer agent could be a valuable assistive technology in addressing the problem.

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

A number of works have attempted to equip automated systems with emotion competences. Two of the most studied issues in this regard are emotion recognition, decoding emotion from communication clues; and emotion simulation, encoding emotion into communication clues. These allow the exchange of emotion between user and the system. Furthermore, there also exist works on replication of human's emotional factors in the system, such as appraisal [4] and personality [1, 6], allowing the system to treat an input as a stimuli in giving an emotional response.

On top of this, there has been an increasing interest in eliciting user's emotional response. Skowron et al. have studied the impact of different affective personalities in a text-based dialogue system [14]. They reported consistent impacts with the corresponding personality in humans. On the other hand, Hasegawa et al. constructed translation-based response generators with various emotion target, e.g the response generated from the model that targets "sadness" is expected to elicit sadness [7]. The model is reported to be able to properly elicit the target emotion.

Emotion elicitation can constitute a universal form of emotional support through HCI. However, existing works on emotion elicitation have not yet observed the appraisal competence of humans that gives rise to the elicited emotion. This entails the relationship between an utterance, which acts as stimuli evaluated during appraisal (*emotional trigger*), and the resulting emotion by the end of appraisal (*emotional response*) [12]. By examining this, it would be possible to reverse the process and 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.

In this paper, we attempt to elicit a positive emotional change in HCI by exploiting examples of appraisal in human dialogue. Figure 1 illustrates this idea in relation to existing works. We collected dialogue sequences containing emotional triggers and responses to serve as examples in a dialogue system. Subsequently, 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.

The proposed system improves upon the existing studies by harnessing information of human appraisal in eliciting user's emotion. We eliminate the need of multiple models and the definition of emotion targets by aiming for a general positive affective interaction. The use of data-driven approach rids the need of complex modeling and manual labor. Text-based human subjective evaluation with crowdsourcing shows that in comparison to the traditional response selection method, the proposed one elicits an overall more positive emotional impact, and yields higher coherence as well as emotional connection.

[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.

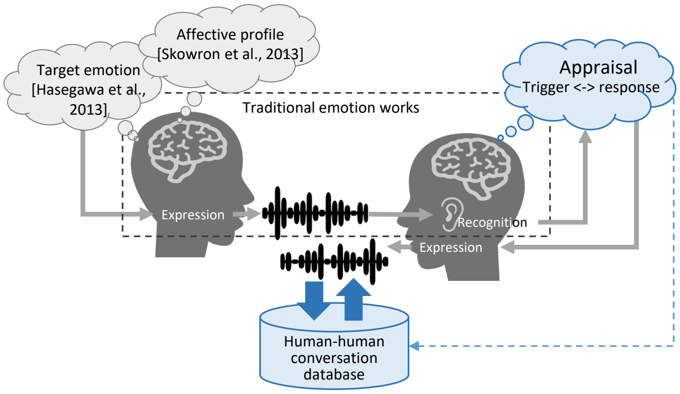


Fig. 1 Overview of proposed approach to elicit a positive emotional change in HCI using examples of appraisal in human dialogue. Relation to existing works is shown.

2 Example-based Dialogue Modeling (EBDM)

EBDM is a data-driven approach of dialogue modeling that uses a semantically indexed corpus of *query-response*[2] pair examples instead of handcrafted rules or probabilistic models [9]. At a given time, the system will return a response of the best example according to a semantic constraint between the query and example queries. This circumvents the challenge of domain identification and switching—a task particularly hard in chat-oriented systems where no specific goal or domain is predefined beforehand. With increasing amount of available conversational data, EBDM offers a straightforward and effective approach for deploying a dialogue system in any domain.

Lasguido et al. have previously examined the utilization of cosine similarity for response retrieval in an example-based dialogue system [8]. In their approach, the similarity is computed between TF-IDF weighted term vectors of the query and the examples. The TF-IDF weight of term t is computed as:

$$\mathrm{TF}-\mathrm{IDF}(t,T) = F_{t,T} \log \frac{|T|}{DF_t}, \tag{1}$$

where $F_{t,T}$ is defined as term frequency of term t in a sentence T, and DF_t as total number of sentences that contains the term t, calculated over the example database. Thus, the vector for each sentence in the database is the size of the database term vocabulary, each weighted according to Equation 1.

Cosine similarity between two sentence vectors S_q and S_e is computed as:

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

[2] 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

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 study.

3 Emotion Definition

In this work, we define the emotion scope based on the *circumplex model of affect* [11]. 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 dimension are core to many works in affective computing and potentially provides useful information even at an early stage of research.

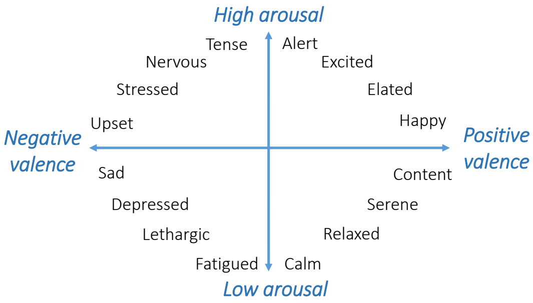


Fig. 2 Emotion dimensions and common terms.

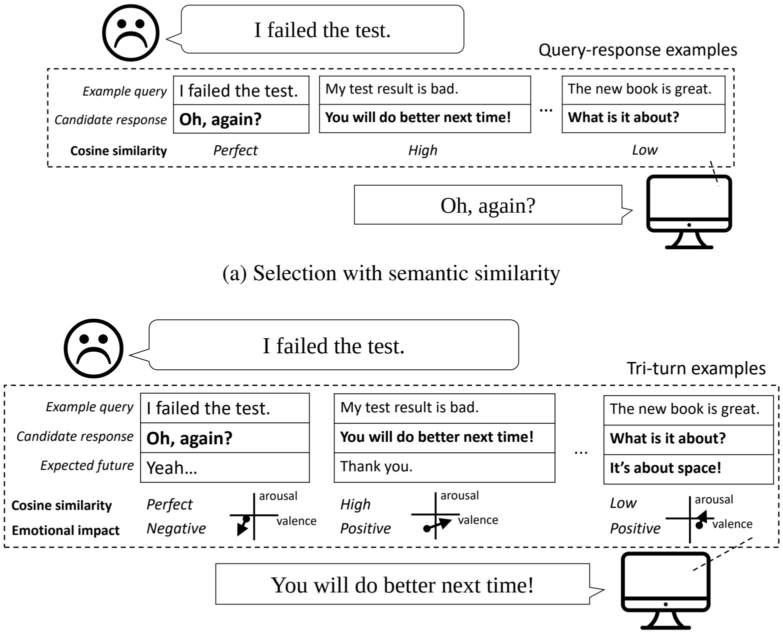
Henceforth, based on this scope of emotion, the term *positive emotion* refers to the emotions with positive valence. Respectively, a *positive emotional change* refers to the change of position in the valence-arousal space where the value of valence after the movement is greater than that of before.

4 Proposed Dialogue System

To allow the consideration of the emotional parameters aforementioned, we make use of *tri-turn* units in the selection process in place of the *query-response* pairs in the traditional EBDM approach. A tri-turn consists of three consecutive dialogue turns that are in response to each other, it is previously utilized in collecting *query-response* examples from a text-based conversational data to ensure that an example is dyadic [8]. In this work, we instead exploit the tri-turn format to observe emotional triggers and responses in a conversation.

Within this work, the first, second, and third turns in a tri-turn are referred to as *query*, *response*, and *future*, respectively. The change of emotion observed from *query* to *future* can be regarded as the impact of *response*.

In addition to semantic constraint as described in Section 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 3 illustrates the general idea of the baseline (a) and proposed (b) approaches.



(a) Selection with semantic similarity

(b) Selection with semantic similarity and emotion parameters

Fig. 3 Response selection in baseline and proposed systems.

First, we measure *emotion similarity* by computing the Pearson's correlation coefficient of the emotion vector between the query and the example queries, i.e. the first turn of the tri-turns. Correlation r_{qe} between two emotion representation vectors for query q and example e of length n is calculated using Equation 3,

$$r_{qe} = \frac{\sum_{i=1}^{n}(q_i - \bar{q})(e_i - \bar{e})}{\sqrt{\sum_{i=1}^{n}(q_i - \bar{q})^2}\sqrt{\sum_{i=1}^{n}(e_i - \bar{e})^2}}. \tag{3}$$

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 tri-turn, emotional impact of a *response* according to the *query* and *future* is computed using Equation 4.

$$\text{impact}(response) = \frac{1}{n}\sum_{i=1}^{n} f_i - \frac{1}{n}\sum_{i=1}^{n} q_i, \tag{4}$$

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.

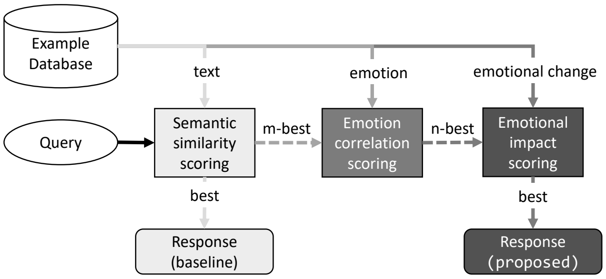


Fig. 4 Steps of response selection.

Figure 4 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 tri-turn example with the highest semantic similarity score (Equation 2). On the other hand, on the proposed system's response selection, we pass m examples with highest semantic similarity scores to the next step and calculate their emotion similarity scores (Equation 3).

From *n* examples with highest emotion similarity scores, we output the *response* of the tri-turn example with the most positive expected emotional impact (Equation 4).

The semantic and emotion similarities are important in ensuring an emotional response that is as close as possible with the example. The two similarity scores are processed incrementally considering the huge difference of their space sizes. The two-dimensional emotion space is much smaller than that of the semantic, making it more likely for emotions to have a high similarity score. Thus, imposing the emotion constraints in the reduced pool of semantically similar examples 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.

There are two important points to note regarding the proposed approach. First, the *future* of each tri-turn is not considered as a definite prediction of user response when interacting with the system. Instead, each tri-turn 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 tri-turn's *query*, when the system outputs the *response*, the user may experience an emotional change consistent with that of the *future*.

Second, this strategy does not translate to selection of the *response* with the most positive emotion. Instead, it is equivalent to selecting the *response* that has the most potential in eliciting a positive emotional response, regardless of the emotion it actually contains. 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.

5 Experimental Set Up

5.1 Emotionally Colorful Conversational Data

To achieve a more natural conversation, we utilize an emotionally colored corpus of human spoken interaction to build the dialogue system. In this section, we describe in detail the SEMAINE database and highlight the qualities that make it suitable for our study.

The SEMAINE database consists of dialogues between a user and a Sensitive Artificial Listener (SAL) in a Wizard-of-Oz fashion [10]. A SAL is a system capable of holding a multimodal conversation with humans, involving speech, head movements, and facial expressions, topped with emotional coloring [13]. This emotional coloring is adjusted according to each of the SAL characters; cheerful Poppy, angry Spike, sad Obadiah, and sensible Prudence.

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 emotion occurrences are annotated using the FEELtrace system [3] 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 fall and rise within an interaction. The numbers in a trace are provided with an interval of 0.02 seconds.

In this study, we consider 66 sessions from the corpus based on transcription and emotion annotation availability; 17 Poppy's sessions, 16 Spike, 17 Obadiah, 16 Prudence. For every dialogue turn, we keep the speaker information, time alignment, transcription, and emotion traces.

5.2 Set Up

As will be elaborated in Section 6.1, in the SEMAINE Corpus, Poppy and Prudence tend to draw the user into the positive-valence region of emotion as opposed to Spike and Obadiah. This resembles a *positive emotional impact*, where the final emotional state is more positive than the initial. 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 tri-turns.

As described in the emotion definition, in this study, we exclusively observe valence and arousal from the emotion annotation. We average as many annotations as provided in a session to obtain the final emotion label. We sample the emotion trace of every dialogue turn into 100-length vectors to keep the length uniform, as discussed in Section 4.

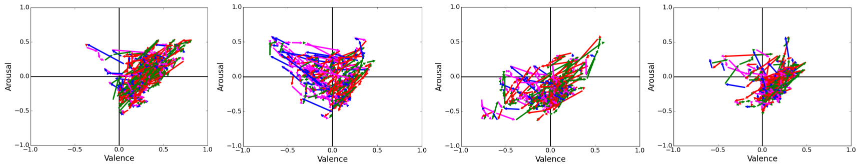
In this phase of the research, we utilize the transcription and emotion annotation provided from the corpus as information of the tri-turns to isolate the errors of automatic speech and emotion recognition. For the n-best filtering, we chose 10 for the semantic similarity constraint and 3 for the emotion considering the size of the corpus.

6 Analysis and Evaluation

6.1 Emotional Impact Analysis

We suspect that the distinct characteristic of each SAL affects the user's emotional state in different ways. To observe the emotional impact of the dialogue turns in the data, we extract tri-turn units from the selected 66 sessions of the corpus. As SEMAINE contains only dyadic interactions, a turn can always be assumed as a response to the previous one.

We investigate whether the characteristics of the SAL affect the tendencies of emotion occurrences in a conversation by analyzing the extracted tri-turns. From all the tri-turns extracted from the subset, we compute the emotional impacts and plot them onto the valence-arousal axes, separated by the SAL to show emotion trends of each one. Figure 5 presents this information.



(a) Poppy (b) Spike (c) Obadiah (d) Prudence

Fig. 5 Emotional changes in SEMAINE sessions separated by SAL character. X- and y-axes show changes in valence and arousal, respectively. Arrows represent emotional impact, with initial emotion as starting point and and final as ending. The direction of the arrows shows the emotional change that occurs. Up and down directions show the increase and decrease of arousal. Right and left directions show the increase and decrease of valence.

The figure shows different emotional paths taken during the conversation with distinct trends. In Poppy's and Prudence's sessions, most of the emotional occurrences and transitions happen in positive-valence and positive-arousal region with occasional movement to the negative-valence region. On the other hand, in Spike's sessions, movement to the negative-valence positive-arousal region is significantly more often compared to the others. The same phenomenon occurs with negative-valence negative-arousal region in Obadiah's. This tendency is consistent to the characteristic portrayed by each SAL, as our initial intuition suggests.

6.2 Human Evaluation

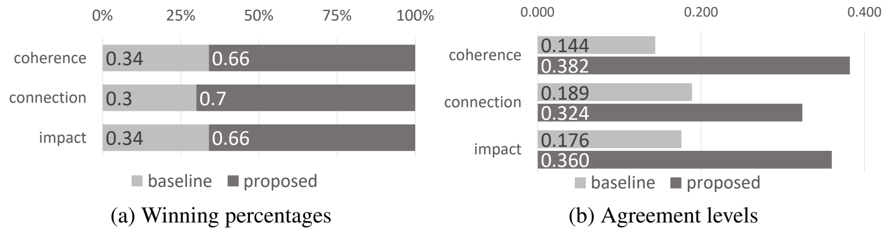
We perform an offline subjective evaluation to qualitatively measure perceived differences between the two response selection methods. From the test set, we extract 198 test queries. For each test query, we generate responses using the baseline and proposed systems. Queries with identical responses from 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 query 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 [14]:

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.

3. Which response gives a more positive emotional impact? Emotional impact refers to the potential emotional change the response may cause.

50 judgements are collected per query. Each judgment is weighted with the level of trust of the worker[3]. The final judgement of each query for each question is based on the total weight of the overall judgements—the system with the greater weight wins.



(a) Winning percentages

(b) Agreement levels

Fig. 6 Human evaluation result.

Figure 6(a) presents the winning percentage of each system for each criterion. It is shown that in comparison to the baseline system, the proposed system is 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 judgement using the Fleiss' Kappa formula. This result is presented in Figure 6(b). We separate the queries based on the winning system and compute the overall agreement of the 50 judgements respectively. It is revealed that the queries where the proposed system wins have far stronger agreement than that where the baseline system wins. Higher agreement level suggests a stronger win, where bigger majority of the evaluators are voting for the winner.

6.3 Discussion

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 1 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. Note that, as described in Section 4: 1) the semantic and emotion similarity scores are computed between the query and example queries (i.e. first turn of the tri-turns), 2) the impact scores are computed from the example queries and example future (i.e. first and third turns of the tri-turns) and 3) the candidate responses are the second turn of the tri-turns. The table shows that the proposed method can select one of the candidate responses that even though is not the best

[3] The level of trust is provided by the crowdsourcing platform we employ in this evaluation. In this evaluation, we employ workers with high-ranking level of trust.

in semantic similarity score, has a higher score in terms of emotion similarity and expected impact.

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

Candidate responses	ranking steps		
	semantic	**emotion**	**impact**
* And where in Australia?	1		
[laugh]	2		
Organised 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

Table 1 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.

Furthermore, the proposed selection method is able to generate different responses to identical textual input with different emotional contexts. Table 2 demonstrates this quality. This shows system's ability to adapt to user's emotion in giving a response. These qualities can contribute towards a more pleasant and emotionally positive HCI.

Query : Thank you. *(valence: 0.13, arousal: -0.18)*	**Query :** Thank you. *(valence: 0.43, arousal: 0.05)*
Baseline : Thank you very much that **Proposed :** And I hope that everything goes exactly according to plan.	**Baseline :** Thank you very much that **Proposed :** It is always a pleasure talking to you you're just like me.

Table 2 Baseline and proposed responses for identical text with different emotional contexts. The proposed system can adapt to user emotion, while baseline method outputs the same response.

7 Conclusions

We presented a novel attempt in eliciting postive emotional impact in dialogue response selection by utilizing examples of human appraisal in spoken dialogue. We use tri-turn units in place of the traditional query-response pairs to observe emotional triggers and responses in the example database. We augment the response selection criteria to take into account emotion similarity between query and the example query, as well as the expected future impact of the candidate response.

Human subjective evaluation showed that the proposed system can elicit a more positive emotional impact in the user, as well as achieve higher coherence and emotional connection. The data-driven approach we employ in this study is straightforward and could be efficiently replicated and extended. With the increasing access to data and the advancements in emotion recognition, a large unlabeled corpus of conversational data could be used to extensively expand the example database.

In the future, we look forward to try a more sophisticated function for estimating the emotional impact. We hope to test the proposed idea further in a setting closer to real conversation, e.g. by using spontaneous social interactions, considering interaction history, and using real-time emotion recognition. We also hope to apply this idea to a more complex dialogue models, such as Partially Observable Markov Decision Process (POMDP) and to learn an explicit dialogue with machine learning approaches.

Acknowledgement

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

References

1. Ball, G., Breese, J.: Emotion and personality in a conversational agent. Embodied conversational agents pp. 189–219 (2000)
2. Cohen, A.N., Hammen, C., Henry, R.M., Daley, S.E.: Effects of stress and social support on recurrence in bipolar disorder. Journal of affective disorders **82**(1), 143–147 (2004)
3. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: 'FEELTRACE': An instrument for recording perceived emotion in real time. In: ISCA tutorial and research workshop (ITRW) on speech and emotion (2000)
4. Dias, J., Mascarenhas, S., Paiva, A.: Fatima modular: Towards an agent architecture with a generic appraisal framework. In: Emotion Modeling, pp. 44–56. Springer (2014)
5. Diener, E., Larsen, R.J.: The experience of emotional well-being. (1993)
6. Egges, A., Kshirsagar, S., Magnenat-Thalmann, N.: Generic personality and emotion simulation for conversational agents. Computer animation and virtual worlds **15**(1), 1–13 (2004)
7. Hasegawa, T., Kaji, N., Yoshinaga, N., Toyoda, M.: Predicting and eliciting addressee's emotion in online dialogue. In: Proceedings of Association for Computational Linguistics (1), pp. 964–972 (2013)
8. Lasguido, N., Sakti, S., Neubig, G., Toda, T., Nakamura, S.: Utilizing human-to-human conversation examples for a multi domain chat-oriented dialog system. Transactions on Information and Systems **97**(6), 1497–1505 (2014)
9. Lee, C., Jung, S., Kim, S., Lee, G.G.: Example-based dialog modeling for practical multi-domain dialog system. Speech Communication **51**(5), 466–484 (2009)
10. McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M.: 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)
11. Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology **39**(6), 1161 (1980)
12. Scherer, K.R., Schorr, A., Johnstone, T.: Appraisal processes in emotion: Theory, methods, research. Oxford University Press (2001)
13. Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., Maat, M.T., McKeown, G., Pammi, S., Pantic, M., et al.: Building autonomous sensitive artificial listeners. Transactions on Affective Computing **3**(2), 165–183 (2012)
14. Skowron, M., Theunis, M., Rank, S., Kappas, A.: Affect and social processes in online communication–experiments with an affective dialog system. Transactions on Affective Computing **4**(3), 267–279 (2013)