Affect-sensitive Dialogue Response Generation for Positive Emotion Elicitation

Nurul Lubis[1] Sakriani Sakti[1,2] Koichiro Yoshino[1,2,3] Satoshi Nakamura[1,2]
[1]Nara Institute of Science and Technology
[2]RIKEN AIP
[2]JST PRESTO
{nurul.lubis.na4, ssakti, koichiro, s-nakamura}@is.naist.jp

1 Introduction

An emotionally-competent computer agent could offer valuable assistance in various affective tasks. For example caring for the elderly, low-cost ubiquitous chat therapy, and providing emotional support in general. In this paper, we build an end-to-end chat-oriented dialogue system that can dynamically mimic affective human interactions by utilizing a hierarchical neural network architecture. End-to-end approaches have been reported to show promising results for non-goal oriented dialogue systems. However, application of this approach towards incorporating emotion in the dialogue is still very lacking.

Our contributions in this paper are 1) a neural-network based chat-oriented dialogue system that captures user's emotional state and considers it in generating a dialogue response, trained to elicit positive emotion through the interaction, 2) a dialogue corpus that reflects a positive emotion elicitation strategy for model training to influence its affective tendency. This allows positive emotion elicitation without any elaborate dialogue strategy.

2 Proposed Model

We build our model extending the recently proposed hierarchical recurrent encoder-decoder architecture [4]. We propose to incorporate an *emotion encoder* into the architecture. The emotion encoder is placed in the same hierarchy as the dialogue encoder, capturing emotion information at dialogue-turn level and maintaining the emotion context history throughout the dialogue. Figure 1 shows a schematic view of this architecture.

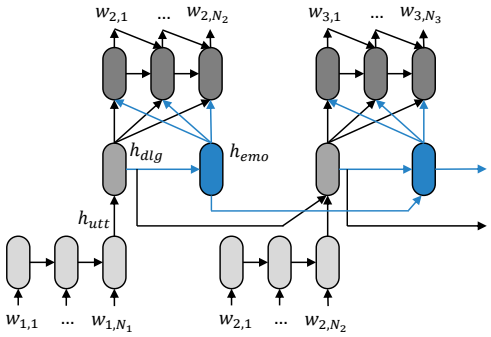

Figure 1: Proposed architecture

After reading the input sequence, the dialogue turn is encoded into utterance representation h_{utt}. h_{utt} is then fed into the recurrent dialogue encoder to capture the information up to the current dialogue turn, into dialogue context h_{dlg}. In Emo-HRED, the h_{dlg} is then fed into the emotion encoder, which models the emotion context h_{emo}. The response generation process is conditioned by the concatenation of the dialogue and emotion contexts.

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 $cost_{emo}$ by the emotion encoder. The final cost is then propagated to the network and the parameters are optimized as usual.

We propose to pre-train the Emo-HRED with a large scale conversational corpus to infer content and syntactic knowledge prior to training its emotion-

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

Table 1: Model perplexity on test set. In subjective evaluation, $*$ denotes significant difference ($p < 0.05$).

Model	Parameter update	Perplexity	Naturalness	Emotional Impact
Baseline HRED	`standard`	121.44	n/a	
	`selective`	100.94	3.26	3.22
Proposed Emo-HRED	`selective`	42.26	3.27	3.39 *

related parameters due to the limited amount of emotion rich data. We further propose selective fine-tuning, limiting the parameter updates to the emotion encoder and utterance decoder only. As emotion is not yet involved during encoding, we hypothesize that the pre-trained encoders can be used for the affect-sensitive response generation task as is.

3 Constructing Positive Data

We aim to use emotion information to elicit positive emotion through dialogue. We realize this implicitly by relying on the training data. We enhance the existing emotion-rich SEMAINE corpus [3] to contain positive-emotion eliciting responses. First, for all triples extracted from the data, i.e. three consecutive dialogue turns with agent-user-agent speaker order, we obtain new responses that elicit positive emotion using an EBDM with modified selection criteria [2]. We ask human judges to decide which response elicits a more positive emotional impact, the default or the system generated one. For each triple, we obtain at least 3 human judgements, or more when ties occur. The final response is obtained by majority voting.

4 Experiment and Evaluation

We utilize the HRED trained on the SubTle corpus [1] as our starting model. The data pre-processing steps are performed as in [4]. This dataset is fed sequentially into the network until it converges. In addition to the model parameters, we also learn the word embeddings of the tokens. The model is trained to optimize the log-likelihood of the training triples using the Adam optimizer.

Two aspects are considered in fine-tuning the above model. First, we consider two different models: HRED as baseline model and Emo-HRED as the proposed model.Second, we consider two different parameter update schemes: `standard` and `selective`. In the `standard` scheme, we fine-tune all the parameters of the model. In the `selective` scheme, we fix the utterance and dialogue encoders parameters. The fine-tuning is done with the constructed positive data We hypothesize that the positive corpus will cause the model to elicit more positive emotion.

Table 1 shows the evaluation result of the different set ups. In objective evaluation, we measure the model perplexity on a held out test set. We observe significant improvements when `selective` update scheme is employed in place of the `standard` scheme. With identical starting model and fine-tune set up, the Emo-HRED architecture converges to significantly better models compared to the HRED. In subsequent subjective evaluation through crowdsourcing, we found that the responses generated by the proposed model are perceived as 1) more natural, and 2) elicit a significantly more positive emotional response ($p < 0.05$) compared to the baseline.

5 Conclusion

We propose a dialogue response generator architecture for eliciting positive emotion through dialogue interactions. Both the subjective and objective evaluation show consistent incremental improvements when each of the proposed set ups is applied to the dialogue system. They also show that a system performs best when all of the proposed set ups are applied at the same time. In the future we hope to further improve the proposed system, both in terms of response quality and user emotional experience.

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

Acknowledgements

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

References

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

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

[3] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Transactions on Affective Computing*, Vol. 3, No. 1, pp. 5–17, 2012.

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

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.