Modelling Context Emotions using Multi-task Learning for Emotion Controlled Dialog Generation*

Deeksha Varshney, Asif Ekbal, Pushpak Bhattacharyya

Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India

*https://aclanthology.org/2021.eacl-main.255/

Outline

- Problem Definition
- Motivation
- Contributions
- Dataset
- Methodology
  - Proposed Model
  - Baselines
- Results
- Analysis
- Conclusion and Future Works
- References

Problem Definition

- **Problem Statement:** Our task is to understand and acknowledge any implied feelings of user by generating emotion controlled responses
  - For *multi-turn conversations*
  - Using *relevant emotion labels.*
- *Mathematically,* let $\mathbf{U} = \mathbf{u}^{(1)},\ldots,\mathbf{u}^{(k)},\ldots,\mathbf{u}^{(K)}$ denote the set of K utterances of our multi-turn conversation.
  - Emotion labels is denoted by $\mathbf{E} = \mathbf{e}^{(1)},\ldots,\mathbf{e}^{(k)},\ldots,\mathbf{e}^{(K)}$.
  - Hence, we try to generate a response 'y' given
    - emotion label $e^{(k+1)}$ and
    - the set of previous 'k' utterances.

Motivation

Agent 1	**Do you like wearing hats? It has so many functions.**	Curious
Agent 2	I don't like them on myself but I know a lot of people that can pull them off.	Neutral
Agent 1	Yes me as well. In the military hats denote a nationality, branch of service, rank or regiment.	Curious
Agent 2	Yes. I love hats! I have a wide variety of hats and wear them for different reasons.	Happy
Agent 1	Yes. . . Even I like it too !! Specially I am on vacation, roaming around I do carry 2–3 hats. And I wear it according to my dressing style.	Happy

- A snippet of two different emotionally inclined conversations with a common query.
- The example shows how two different emotionally inclined responses can lead a conversation in two different directions.

Key contributions

- We propose an effective deep *multitask framework* that performs emotion classification and response generation.
- To handle the imbalanced data distribution, we use Focal Loss (Lin et al., 2017) instead of regular cross entropy loss for emotion classification of utterances.
- To maintain uniformity between the attention weights of different tasks, we utilise consistency loss (Nishino et al., 2019) in addition to the original task-specific losses.

Dataset

- Dataset name: Topical chat dataset[1]
    - a knowledge-grounded open domain conversation dataset.
    - the underlying knowledge spans *8 broad topics .*
    - Fashion, Politics, Books, Sports, General Entertainment, Music, Science & Technology, Movies.

- The knowledge base comprises -
    - Wikipedia articles, Washington Post articles and Reddit fun facts.

- Each utterance has an emotion label associated with it. There are *8 emotion classes.*
    - Angry, Disgusted, Fearful, Sad, Happy, Surprised, Curious to Dive Deeper, Neutral

[1] https://m.media-amazon.com/images/G/01/amazon.jobs/3079_Paper._CB1565131710_.pdf
[1] https://github.com/alexa/alexa-prize-topical-chat-dataset

Dataset Details

Agent 1	Are you afraid of snakes?	Curious
Agent 2	Hi, I am a little! but I was surprised there are none in New Zealand!	Happy
Agent 1	Sounds like a perfect place for me lol, I'm terrified of them	Fearful
Agent 2	Wow! I can understand , I am more terrified of crocodiles but it seems they are closer to birds than to snakes!	Fearful
Agent 1	Some snakes can even fly to catch their prey so that's scary	Curious
Agent 2	Wow, I would like to see that! And did you know its head is designed to swallow prays larger than them	Happy
Agent 1	Yeah I did know that, that's actually a bit disgusting, watching them eat prey	Disgusted
Agent 2	It looks like monkeys are terrified of snakes too!	Happy
Agent 1	They are? monkey are smart, they should stay as far as they can of snakes, dangerous animals	Fearful
Agent 2	Maybe you are terrified of snakes! But do you like dancing?	Happy

- Example conversations from the topical chat dataset showing different context emotion labels.

Dataset Statistics

- **Total no. of conversations :** 10784
- **Average Number of Turns per Conversation :** 21.9
- **Average Length of Utterance :** 19.7

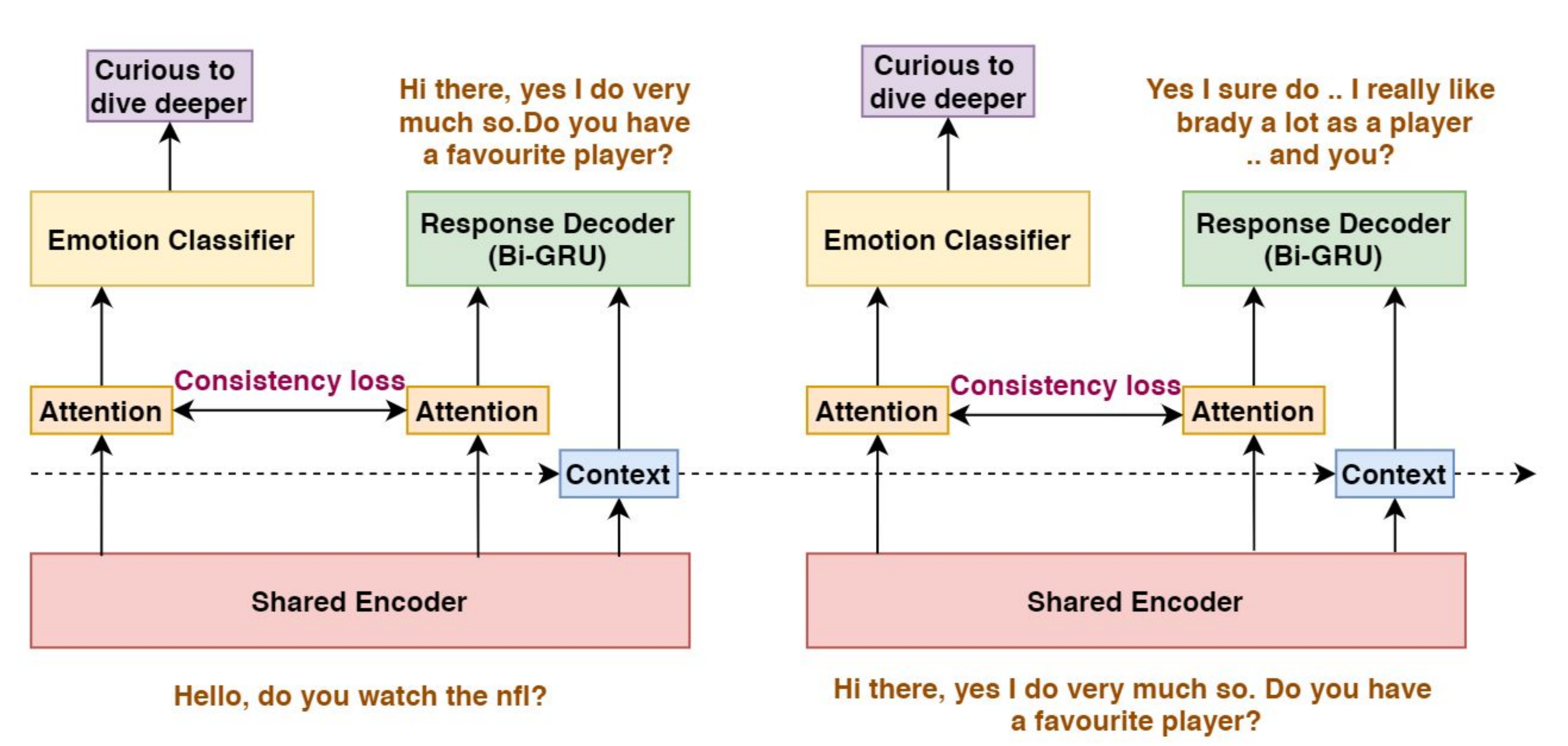
	#Conversation	#Utterances
Train	8628	188378
Valid Frequent	539	11681
Valid Rare	539	11692
Test Frequent	539	11760
Test Rare	539	11770

Note: *Frequent set* contains conversations on entities frequently seen in the training set. *Rare set* contains conversations on entities that were infrequently seen in the training set.

Emotion Distribution

Emotion class	Original count
Happy	36845
Fearful	1174
Surprised	38254
Sad	3070
Angry	1133
Curious to dive deeper	101162
Disgusted	1848
Neutral	51796

Methodology



Methodology: Model Training

- **We minimize the following loss for model training**
  - **L_1 :** A focal loss function for emotion classification,

$$L_1 = -(1 - p_c^{(k)})^{\gamma} log(p_c^{(k)})$$

Where $p_c^{(k)}$ is the emotion class probability and γ is a focusing parameter.

  - **L_2:** We use the negative log-likelihood loss for dialog generation

$$L_2 = -\sum_{t=1}^{m} log P(\hat{y}_{t+1}/y_{<t})$$

Methodology: Model Training

- **L_{cl}: Consistency Loss**

$$L_{cl} = \sum_{i=1}^{I} |\max_j e_{p,ij}^{(k)} - \max_j e_{q,ij}^{(k)}|+$$

  - Where $e_{p,ij}^{(k)}$ is the attention weight for every k-th utterance for the p-th task. To compare the two attention weights, a ramp function denoted by $|x|+$ is used.
  - **Importance:** It is used to maintain consistency between different attention weights. If each decoder focuses on the same word in the input text, the model can generate a more consistent output.

Experiments (Baselines)

- **HRED:**
  - Our first baseline is based on the hierarchical encoder-decoder model
  - In this model, the encoder RNN encodes the subsequent words of the utterances, and the context RNN encodes the conversation history.
- **HRED-A:**
  - We extend the HRED model with word-level attention on the encoder side to focus on the relevant words of the input sequence.
- **HRED-SA:**
  - In this model, we use the transformer encoder to encode the utterances of the multi-turn conversations.

Experiments (Baselines) (cont.)

- **EmoHRED-A-FL-CL:**
  - We extend the HRED-A model to EmoHRED-A-FL-CL, a deep multi-task learning framework that jointly performs the task of both response generation and emotion analysis.
  - We add *focal loss* and *consistency loss* to the existing dialog generation loss.
- **Ablation Models**
  - **EmoHRED-SA-FL** - To prove the effectiveness of our consistency loss we remove the consistency loss from our proposed model.
  - **EmoHRED-SA** - We also show the strength of the focal loss we remove FL from EmoHRED-SA-FL model.

Automatic Evaluation Metrics

- **BLEU Score:** BLEU measures the n-gram overlap between a generated response and a gold response.
- **F1-score:** We also compute unigram F1-score between the predicted sentences and the ground truth sentences[1].
- **Perplexity(PPL) :** It is a measurement of how well a probability distribution or probability model predicts a sample. A low perplexity indicates the probability distribution is good at predicting the sample.
- **N-gram diversity (Div.)** : It is used as a measure of informativeness and diversity of sentences.
    - Div=1/M[# unique n-grams /# words in predicted response]

Where M is the total number of samples in the test set.

[1] https://github.com/facebookresearch/ParlAI/blob/master/parlai/core/metrics.py

Manual Evaluation Metrics

- **Fluency (0-2):** It measures the grammatical correctness of the predicted response.
- **Adequacy (0-2):** It measures whether the generated response is contextually relevant.
- **Emotional Content (0-1):** It checks whether the generated response reflects the target emotion.

Note: **0-2 scale:**
'0' indicates an incomplete or incorrect response,
'1' indicates acceptable responses and
'2' indicates a perfect response.

0-1 scale:
'0' indicates the incorrect emotion
'1' indicates the correct emotion.

Results: Automatic and Manual Evaluation

Models	PPL (Freq/Rare)	BLEU % (Freq/Rare)	F1% (Freq/Rare)	Div(n=1) (Freq/Rare)	Div(n=2) (Freq/Rare)	Fluency (Freq/Rare)	Adequacy (Freq/Rare)	Emotional Content (Freq/Rare)
HRED	45.61 / 70.30	2.4 / 1.9	0.14 / 0.10	0.88 / 0.87	0.89 / 0.88	1.65 / 1.60	0.85 / 0.70	0.50 / 0.45
HRED-A	41.42 / 71.31	2.3 / 1.8	0.15 / 0.11	**0.91 / 0.90**	**0.90 / 0.90**	1.70 / 1.65	0.90 / 0.84	0.52 / 0.54
HRED-SA	36.63 / 54.87	2.1 / 1.8	0.21 / 0.15	0.83 / 0.82	0.84 / 0.84	1.70 / 1.65	0.98 / 0.88	0.60 / 0.55
EmoHRED-A-FL-CL	36.08 / 51.06	2.1 / 1.7	0.23 / 0.12	0.87 / 0.87	0.87 / 0.88	1.85 / 1.80	1.45 / 1.35	0.74 / 0.64
EmoHRED-SA-FL-CL	**35.45 / 50.45**	**2.6 / 2.1**	0.23 / **0.19**	0.88 / 0.87	0.89 / 0.88	**1.95 / 1.90**	**1.50 / 1.45**	**0.80**/ 0.60
EmoHRED-SA-FL	36.34 / 54.82	2.3 / 1.9	**0.25** / 0.13	0.86 / 0.82	0.86 / 0.84	1.80 / 1.80	1.01 / 0.95	0.64 / **0.65**
EmoHRED-SA	36.04 / 52.98	2.3 / 1.8	0.24 / 0.13	0.88 / 0.83	0.83 / 0.84	1.83 / 1.81	0.93 / 0.81	0.53 / 0.51

- **Observation:**
  - High unigram and bi-gram diversities, demonstrating that the models learn to decode fluent and informative responses with great diversity.
  - We observe significant improvement in BLEU and F1-scores when compared with the baseline models which support our multi-task learning architecture.
  - Our proposed model out-performs the other baseline models in terms of human metrics.

Generated examples

Utterance 1	**Are you afraid of snakes?**	**Neutral**
EmoHRED-SA-FL-CL	i do! it's very interesting. i used to play video games when i was a kid i don't have any time.	Happy
HRED	i do. do you ?	Curious
Utterance 2	**i enjoy the getting together with friends part of it but i could take or leave the actual sports events except when i am watching my children play. how about you what's your favorite sport?**	**Happy**
EmoHRED-SA-FL-CL	i do too.	Neutral
HRED	enjoy watching football games as well.	Happy
Utterance 3	**i would say football**	**Neutral**
EmoHRED-SA-FL-CL	nice! speaking of football, did you know women were not allowed to play at the same time ?	Curious
HRED	i don't really like football, but i don't think i've ever heard of it, what about you?	Curious
Utterance 4	**that is the favorite sport in my household as well! tom brady is quite the dedicated player both on and off the field, don't you think?**	**Curious**
EmoHRED-SA-FL-CL	i don't know much about him.	Neutral
HRED	i do not have a favorite team but i don't really follow it.	Neutral
Utterance 5	**yes, he is arguably the greatest steal in nfl history**	**Neutral**
EmoHRED-SA-FL-CL	i can't believe he has never eaten a strawberry ha ha. have you ever heard of him ?	Surprised
HRED	i think it's cool that the nfl has no written rule against female players. i don't know how that's possible.	Surprised

Error Analysis (⅓)

- **Common phrases: Some common phrases are repeated in the generated response.**
  For instance, '*i don't think i've ever heard about it though*', '*i don't know much about it so i don't know much about it either.*' and '*i 'm not sure either. i've never been there*'.

  Observation:
  - Due to *data scarcity and less diversity in the data*, the models may only have learned to predict the most frequent utterances.
  - Since the dialogues are *inherently ambiguous*, predicting them accurately would require more data.

Error Analysis (⅔)

- **Repetition:**
  The proposed model (EmoHRED-SA-FL-CL), in a few cases, go on repeating the information present in the predicted response.
  Predicted Response: that's terrible. *i'll have to check that out. i'll have to check it out!*.

  Observation: This lowers the count of unique uni-gram words in the generated response i.e the F1-score.

Error Analysis (3/3)

- **Emotional inconsistencies:**
  - In some cases, the proposed model (EmoHRED-SA-FL-CL) is unable to produce responses of particular emotion labels
    - due to less occurrence of instances from those classes (angry, sad, fearful and disgusted).
  - The less frequent emotion classes like *anger, sad, fearful and disgusted* get confused with the recurring classes like *curious to dive deeper and surprised.*
    - Also, instances from 'Happy' and 'Surprised' emotion classes get mixed up with each other.
  - For example, the predicted response for Utterance 5 should have the emotion 'Happy' but it gets confused with the emotion 'Surprised' and generates an irrelevant response.

Summary, Conclusion and Future Work

- In this paper,
  - we have proposed a new *deep learning framework for modeling emotion-grounded conversations* using emotion labels as the guiding attributes.
- Extensive experiments show that
  - the predicted responses expressed high levels of emotional accuracy and content adequacy.
- In general, we show
  - how a related task of emotion recognition along with appropriate loss functions can ensure *emotional relevance of the generated response and improves user engagement*.
- In the future,
  - we intend to use *pre-trained language* models for the task of dialog generation using emotion labels.
  - we also aim to extend our model to *handle knowledge-grounded conversations*.

Thank You

Any Questions?