Emotion Alignment in Neural Dialogue Generation

Nabiha Asghar*†◇, Ivan Kobyzev*†, Jesse Hoey†◇, Pascal Poupart†◇, Muhammad Bilal Sheikh

* Equal Contribution

† Cheriton School of Computer Science, University of Waterloo, Canada

◇ Vector Institute for AI, Toronto, Canada

Main Idea

- State-of-the-art neural dialogue systems excel at syntactic and semantic modelling of language, but often have a hard time establishing emotional alignment with the human interactant during a conversation.
- We augment neural dialogue models with Affect Control Theory (ACT) [1], a socio-mathematical model of affect.
- Given the affective *identity* of each of the two interactants, ACT prescribes affective actions for them that are mutually aligned towards minimizing conflict and chaos.
- We integrate the blackbox ACT model with an encoder-decoder neural conversation model for emotionally aligned dialogue generation (Fig 1).

Affect Control Theory (ACT)

- ACT propose that *fundamental sentiments*, $\mathbf{f}$, are representations of interactants' identities and behaviours as vectors in a 3D affective space.
- The basis vectors of affective space are called Evaluation, Potency, and Activity (EPA).
- Social events cause *transient impressions*, $\boldsymbol{\tau}$ (also 3D in EPA space) of identities and behaviours that may deviate from their corresponding $\mathbf{f}$.
- $\boldsymbol{\tau} = \boldsymbol{M}G(\mathbf{f})$, where $\boldsymbol{M}$ is a matrix of statistically estimated prediction coefficients from empirical studies; G is a vector of polynomial features in $\mathbf{f}$.
- $d = \|\mathbf{f} - \boldsymbol{\tau}\|_w^2$, is called *deflection* and is hypothesised to correspond to an aversive state of mind that humans seek to avoid.
- For two given identities of the actors (two EPA vectors) and an initial EPA action by one actor, ACT predicts the optimal response for the second actor through prediction equations which minimize deflection.

Proposed Model

ACT is instantiated with two affective identities, one each for the human participant and the artificial agent.

S2EPA: To map sentences to the EPA space, we modify the output of a pretrained and publicly available BiLSTM network called DeepMoji [2], which produces a probability distribution over a set of 64 *emojis* given an input sentence. We achieve this by manually labeling the 64 emojis with EPA vectors, and taking a weighted average (using the softmax probabilities) of these vectors.

EPA2S: To generate a sentence given an input prompt and a target EPA vector, we explore two models, traditional Seq2Seq with attention [3] and a conditional variational autoencoder (CVAE) [4]. In Seq2Seq, the target EPA and input are passed through the encoder together to produce a fixed-length context vector. This context is passed through the decoder to generate a response. On the other hand, the CVAE model encodes the input into a Gaussian latent space. A sample from this latent space is propagated through a decoder to generate an appropriate response.

For CVAE training (Fig 2), we use a collection of triples of the form $(\boldsymbol{C}, \boldsymbol{\alpha}, \boldsymbol{X})$, where $\boldsymbol{C}$ and $\boldsymbol{X}$ are the prompt and the response respectively, and $\boldsymbol{\alpha}$ is an EPA vector of the response $\boldsymbol{X}$.

$$L_{\texttt{CVAE}}(\boldsymbol{\theta}_C, \boldsymbol{\theta}_U, \boldsymbol{\theta}_D; \boldsymbol{C}, \boldsymbol{X}, \boldsymbol{\alpha}) = \mathrm{KL}\big(q_U(\boldsymbol{z}|\boldsymbol{C}, \boldsymbol{\alpha}, \boldsymbol{X}) \| q_C(\boldsymbol{z}|\boldsymbol{C}, \boldsymbol{\alpha})\big) - \mathbb{E}_{q_U}\big[\log q_D(\boldsymbol{X}|\boldsymbol{z}, \boldsymbol{C}, \boldsymbol{\alpha})\big] \quad (1)$$

Experiments

Model	Syntactic Coherence	Naturalness	Emotional Approp.
Traditional Seq2Seq (baseline)	1.48	0.69	0.41
ACT with `S2EPA` & `EPA2S-Seq2Seq` (*friend-friend*)	**1.59** ↓	0.73 ↓	0.39 ↓
ACT with `S2EPA` & `EPA2S-CVAE` (*friend-friend*)	1.57 ↓	0.68 ↓	0.47 ↓
ACT with `S2EPA` & `EPA2S-Seq2Seq` (*enemy-enemy*)	1.54 ↓	**0.82** ↑	0.49 ↓
ACT with `S2EPA` & `EPA2S-CVAE` (*enemy-enemy*)	1.55 ↓	0.73 ↓	**0.59** ↑

Comparing the different ACT conversation models. Up arrows indicate significant improvement over baseline ($p = 0.05$).

Line	C	Target α (ACT) & Closest ACT Labels	Defl.	`EPA2S-Seq2Seq`	`EPA2S-CVAE`
1	i hate you	[2.52, 2.52, -0.41] (care for, caress)	17.09	that's not the point	you must be tired now
2	i think i'm in love	[3.13, 1.70, 1.39] (thank, kiss)	1.84	i'm glad you like it	i wouldn't do you if i were you
3	i have no fear of failure	[3.72, 1.90, 1.3] (thank, propose marriage to)	4.36	well that's me	i will ride with you love
4	how about a drink?	[3.37, 1.68, 0.92] (reward, thank)	4.06	sure that's nice	i have money
5	i've been thinking about you	[3.12, 1.96, 1.31] (thank, kiss)	1.87	okay	i like you
6	there is nothing for me here anymore	[3.55, 1.99, 0.45] (embrace, propose marriage to)	9.05	i don't think so	it is better this way

The full ACT conversational model with ACT identities *friend-friend*.

Line	C	Target α (ACT) and Closest ACT Labels	Defl.	`EPA2S-Seq2Seq`	`EPA2S-CVAE`
1	i hate you	[-0.27, 0.35, 0.77] (bellow at)	2.21	i am not your friend	man can you scream
2	you are despicable	[-0.18, 0.55, 0.58] (disagree with)	4.32	i don't care for you	you can calm down
3	what the hell are you doing	[-0.29, 0.35, 0.74] (bellow at)	3.56	i can ask you it	i need to leave
4	i quit.	[-0.17, 0.39, 0.75] (giggle at)	5.29	well that's me	it is too late
5	please don't talk with food in your mouth	[-0.09, 0.48, 0.64] (disagree with)	6.30	not now	go away dog
6	i insist on being told exactly what you have in mind	[-0.17, 0.33, 1.12] (be sarcastic toward)	4.01	yeah you know me	i am singing for you

The full ACT conversational model with ACT identities *enemy-enemy*.

Figures

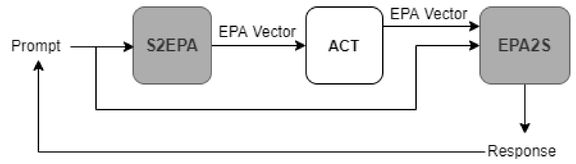

Figure 1: Full dialogue system pipeline

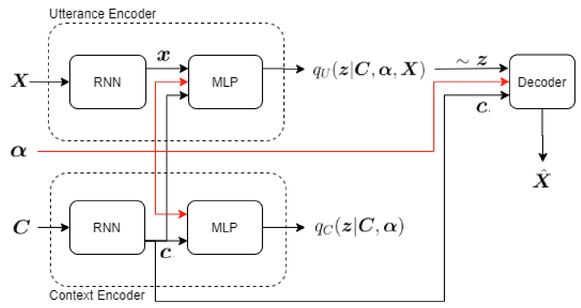

Figure 2: CVAE training architecture

References

[1] David R. Heise. *Expressive Order: Confirming Sentiments in Social Actions*. Springer, 2007.

[2] Felbo et al. *Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm*. EMNLP, pp. 1615-1625, 2017.

[3] Sutskever et al. *Sequence to sequence learning with neural networks*. NeurIPS, pp. 3104-3112, 2014.

[4] Sohn et al. *Learning structured output representations using deep conditional generative models*. NeurIPS, pp. 3483-3491. 2015.