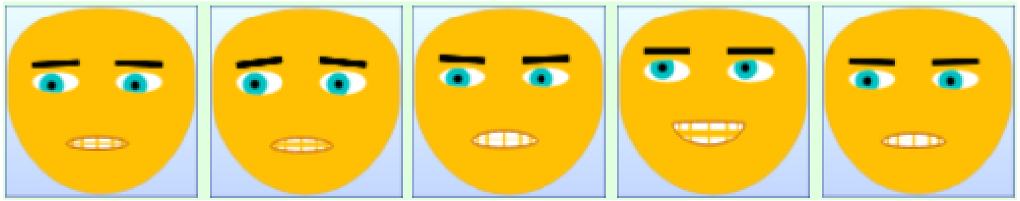


Emotional Dialogue

Qintong Li[1,2]

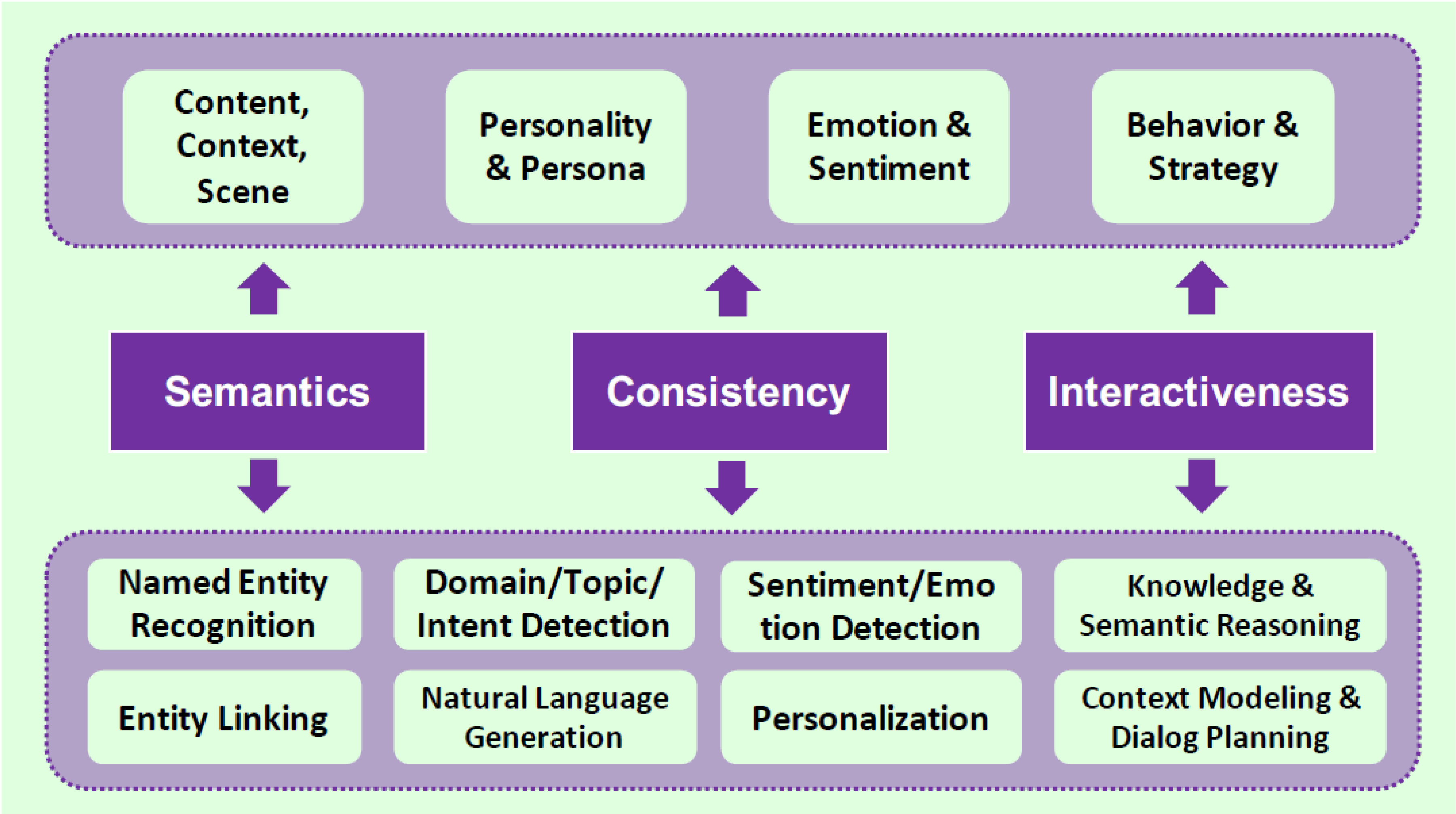
1. Shandong University

2. Tencent AI Lab, NLP Center

Overview

- Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory AAAI18 Tsinghua University
- MOJITALK: Generating Emotional Responses at Scale ACL18 Tsinghua University
- An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss AAAI19 NTU

Why emphasize Emotion?



Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory

Three challenges

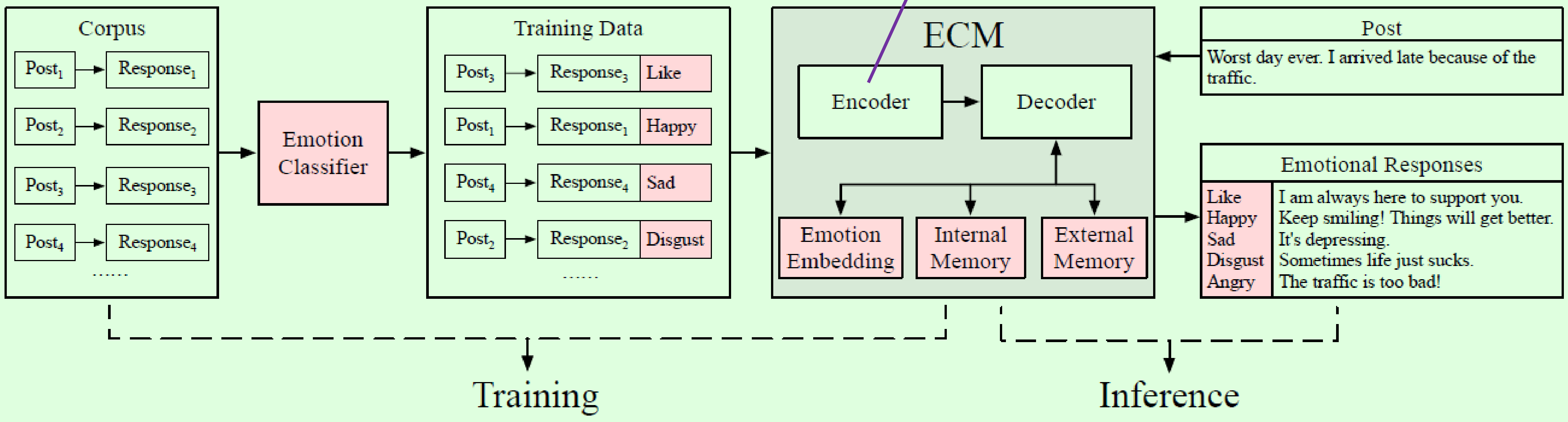
1. Emotion-labeled dataset
2. Balance gramma and emotion
3. How to embed emotion information?

Solutions

1. Pre-train classifier to annotate dataset
2. Emotion category embedding
3. Internal emotion state
4. External emotion memory

User: Worst day ever. I arrived late because of the traffic.
Basic Seq2Seq: You were late. ECM (*Like*): I am always here to support you. ECM (*Happy*): Keep smiling! Things will get better. ECM (*Sad*): It's depressing. ECM (*Disgust*): Sometimes life just sucks. ECM (*Angry*): The traffic is too bad!

$$h_t = \mathrm{GRU}(h_{t-1}, x_t).$$



S2S-decoder: $s_t = \mathrm{GRU}(s_{t-1}, [c_t; e(y_{t-1}); v_e])$.

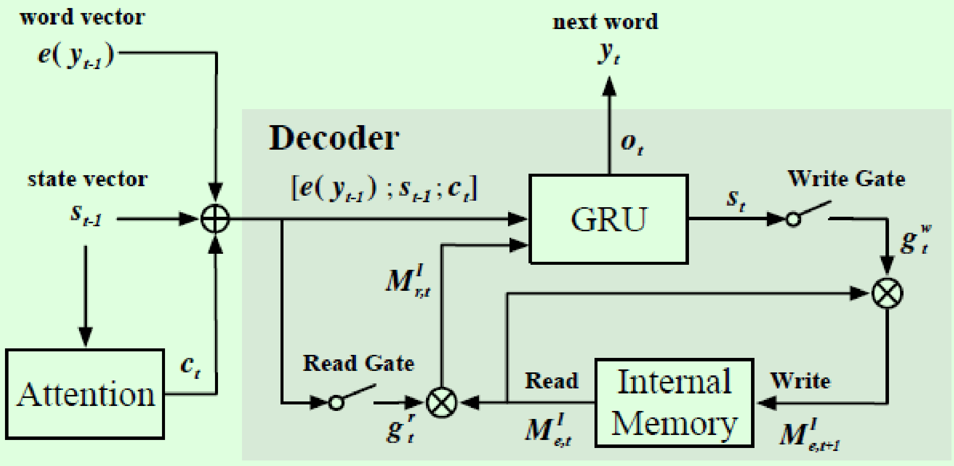


Figure 2: Data flow of the decoder with an internal memory. The internal memory $M_{e,t}^I$ is read with the read gate g_t^r by an amount $M_{r,t}^I$ to update the decoder's state, and the memory is updated to $M_{e,t+1}^I$ with the write gate g_t^w.

$$g_t^r = \mathrm{sigmoid}(\mathbf{W_g^r}[e(y_{t-1}); s_{t-1}; c_t]),$$
$$g_t^w = \mathrm{sigmoid}(\mathbf{W_g^w} s_t).$$
$$M_{r,t}^I = g_t^r \otimes M_{e,t}^I,$$
$$M_{e,t+1}^I = g_t^w \otimes M_{e,t}^I,$$
$$s_t = \mathrm{GRU}(s_{t-1}, [c_t; e(y_{t-1}); M_{r,t}^I]).$$
$$y_t \sim o_t = P(y_t \mid y_1, y_2, \cdots, y_{t-1}, c_t), \quad (3)$$
$$= \mathrm{softmax}(\mathbf{W_o} s_t). \quad (4)$$

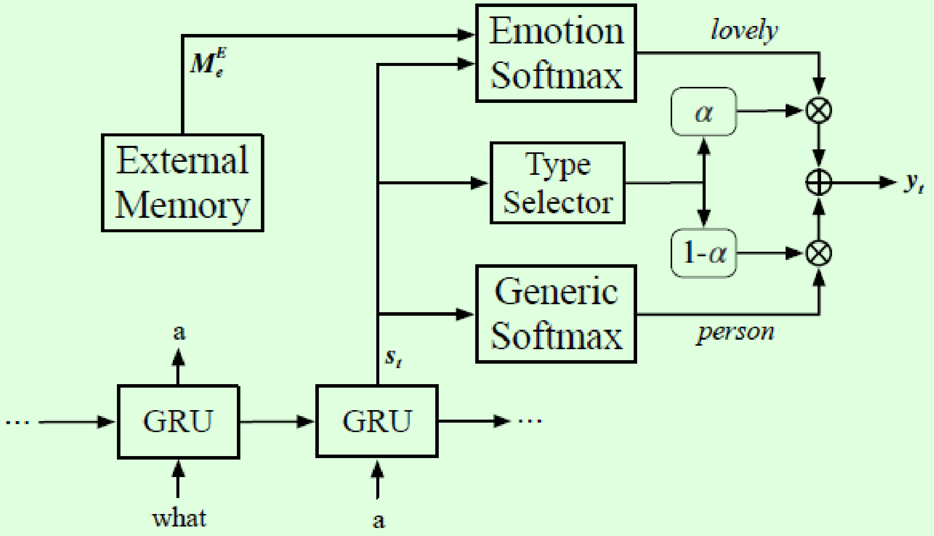


Figure 3: Data flow of the decoder with an external memory. The final decoding probability is weighted between the emotion softmax and the generic softmax, where the weight is computed by the type selector.

$$\alpha_t = \mathrm{sigmoid}(\mathbf{v_u}^\top s_t), \quad (11)$$
$$P_g(y_t = w_g) = \mathrm{softmax}(\mathbf{W_g^o} s_t), \quad (12)$$
$$P_e(y_t = w_e) = \mathrm{softmax}(\mathbf{W_e^o} s_t), \quad (13)$$
$$y_t \sim o_t = P(y_t) = \begin{bmatrix} (1-\alpha_t) P_g(y_t = w_g) \\ \alpha_t P_e(y_t = w_e) \end{bmatrix}, \quad (14)$$

Loss function: $$L(\theta) = -\sum_{t=1}^{m} p_t \log(o_t) - \sum_{t=1}^{m} q_t \log(\alpha_t) + \| M_{e,m}^I \|,$$

External Internal

Dataset:

NLPCC emotion classification dataset -> classifier

Classifier -> STC conversation dataset

6 emotion categories:

Angry, Disgust, Happy, Like, Sad, and Other.

Method	Accuracy
Lexicon-based	0.432
RNN	0.564
LSTM	0.594
Bi-LSTM	0.623

Table 2: Classification accuracy on the NLPCC dataset.

Training	Posts	217,905	
	Responses	Angry	234,635
		Disgust	689,295
		Happy	306,364
		Like	1,226,954
		Sad	537,028
		Other	1,365,371
Validation	Posts	1,000	
Test	Posts	1,000	

Table 3: Statistics of the *ESTC Dataset*.

Noise and Classification error

content emotion

Method	Perplexity	Accuracy
Seq2Seq	68.0	0.179
Emb	62.5	0.724
ECM	65.9	**0.773**
w/o Emb	66.1	0.753
w/o IMem	66.7	0.749
w/o EMem	**61.8**	0.731

Table 4: Objective evaluation with perplexity and accuracy.

Method	Overall		Like		Sad		Disgust		Angry		Happy	
	Cont.	Emot.	Cont.	Emot.	Cont.	Emot.	Cont.	Emot.	Cont.	Emot.	Cont.	Emot.
Seq2Seq	1.255	0.152	1.308	0.337	1.270	0.077	**1.285**	0.038	**1.223**	0.052	1.223	0.257
Emb	1.256	0.363	1.348	0.663	1.337	0.228	1.272	0.157	1.035	0.162	1.418	0.607
ECM	**1.299**	**0.424**	**1.460**	**0.697**	**1.352**	**0.313**	1.233	**0.193**	0.98	**0.217**	**1.428**	**0.700**

Table 6: Manual evaluation of the generated responses in terms of *Content* (Cont.) and *Emotion* (Emot.) .

Method (%)	2-1	1-1	0-1	2-0	1-0	0-0
Seq2Seq	9.0	5.1	1.1	37.6	28.0	19.2
Emb	22.8	9.3	4.3	27.1	19.1	17.4
ECM	**27.2**	**10.8**	4.4	24.2	15.5	17.9

Table 5: The percentage of responses in manual evaluation with the score of *Content-Emotion*. For instance, 2-1 means content score is 2 and emotion score is 1.

Pref. (%)	Seq2Seq	Emb	ECM
Seq2Seq	-	38.8	38.6
Emb	60.2	-	43.1
ECM	**61.4**	**56.9**	-

Table 7: Pairwise preference of the three systems.

Post	Model		Response(Original)	Response(Translated)
承诺，就是一个骗子说给一个傻子听的。 A promise is what a cheater says to a fool.	Seq2Seq		承诺，就是一个骗子说给一个傻子听的。	A promise is what a cheater says to a fool.
	ECM	Like	承诺是一个专一的傻子	Promise is a **single-minded** fool
		Happy	我就是一个开心的傻子	I'm just a **happy** fool
		Sad	我就是一个悲伤的傻子	I am a **sad** fool
		Disgust	一个骗子说给一个傻子听	A **cheater** says to a **fool**
		Angry	不管你信不信，反正我信了！	Whatever you believe or not, I trust in it!
知道真相的我眼泪笑下来 Knowing the truth, I laughed myself to tears	Seq2Seq		知道真相的我眼泪掉下来。	After knowing the truth, I burst into tears.
	ECM	Like	太可爱了，太可爱了！	Too **lovely**, it's too **lovely**.
		Happy	笑得我眼泪都出来了。	I can't help laugh myself to tears.
		Sad	知道真相的我眼泪掉下来。	After knowing the truth, I burst into **tears**.
		Disgust	这是真的还是假的？	Is this true?
		Angry	什么意思？没看懂。	What do you mean? I don't understand.

Figure 4: Sample responses generated by Seq2Seq and ECM (original Chinese and English translation, the colored words are the emotion words corresponding to the given emotion category). The corresponding posts did not appear in the training set.

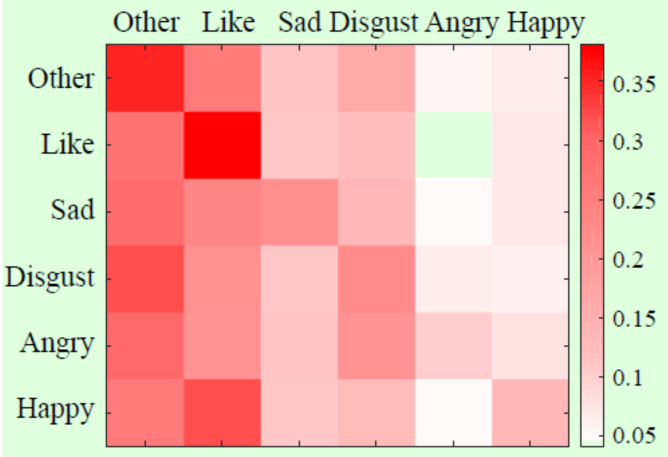


Figure 5: Visualization of emotion interaction.

Analysis of Emotion Interaction and Case Study

- a darker color occurs more frequently than a lighter color
- Like – Happy or Like
- Different types exist
- Other has much more data

Summary

- Strength

1. The first work that addresses the emotion factor in large-scale conversation generation.

- Weakness

1. Category is relatively abstractive
2. Produce responses according to explicit user-input emotions
3. Not consider emotions in input sentences when generating emotional responses (emotion interactions)

MojiTalk: Generating Emotional Responses at Scale

Xianda Zhou
Dept. of Computer Science and Technology
Tsinghua University
Beijing, 100084 China
zhou-xd13@mails.tsinghua.edu.cn

William Yang Wang
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106 USA
william@cs.ucsb.edu

Two challenges

1. the lack of large-scale, manually labeled emotional text datasets
2. coarse-grained classifcation labels make it difficult to capture the nuances of human emotion
3. control the target emotion labels

Solution

1. naturally-occurring emoji-rich Twitter data to construct a dataset using Twitter conversations with emojis in the response.
2. experiment with several extensions to the CVAE model

imitate Twitter users' language style when using those emojis

Figure 1: An example Twitter conversation with emoji in the response (top). We collected a large amount of these conversations, and trained a reinforced conditional variational autoencoder model to automatically generate abstractive emotional responses given any emoji.

Dataset

Not all emojis are used to express emotion and frequency of emojis are unevenly distributed.

😂	184,500	😎	9,505	😌	5,558	👊	2,771
😭	38,479	🙏	9,455	😋	5,114	😴	2,532
😊	30,447	😏	9,298	😐	5,026	😣	2,332
😉	25,018	😒	8,385	💯	4,738	😓	2,293
👍	19,832	😢	8,341	💙	4,623	🙊	1,698
😘	16,934	😔	8,293	😞	4,531	♥️	1,534
😩	17,009	💀	8,144	👏	4,287	😷	1,403
😁	15,563	💖	7,101	😑	4,205	😖	1,258
😍	15,046	😳	6,939	💪	4,066	😠	1,091
😅	14,121	😄	6,769	😫	3,973	🙅	698
💕	13,887	🙌	6,625	😡	3,841	✋	627
👀	13,741	🙈	6,558	😪	3,863	💔	423
❤️	13,147	💜	6,374	✌️	3,236	💟	250
☺️	10,927	😬	6,031	✨	3,072	🔫	243
👌	10,104	😕	5,849	💁	3,088	🎶	154
😜	9,546	😤	5,624	😈	2,969	🎧	130

Table 1: All 64 emoji labels, and number of conversations labeled by each emoji.

Crawl data

- Crawl conversation pairs consisting of an original post and a response on Twitter
- The response to a conversation must include at least one of the 64 emoji labels
- only English tweets without multimedia contents (such as URL, image or video) are allowed

Emoji Labelling

- use the emoji with most occurrences inside the response
- with same occurrences, choose the least frequent one across the whole corpus

596,959/32,600/32,600 conversation pairs for train /validation/test set

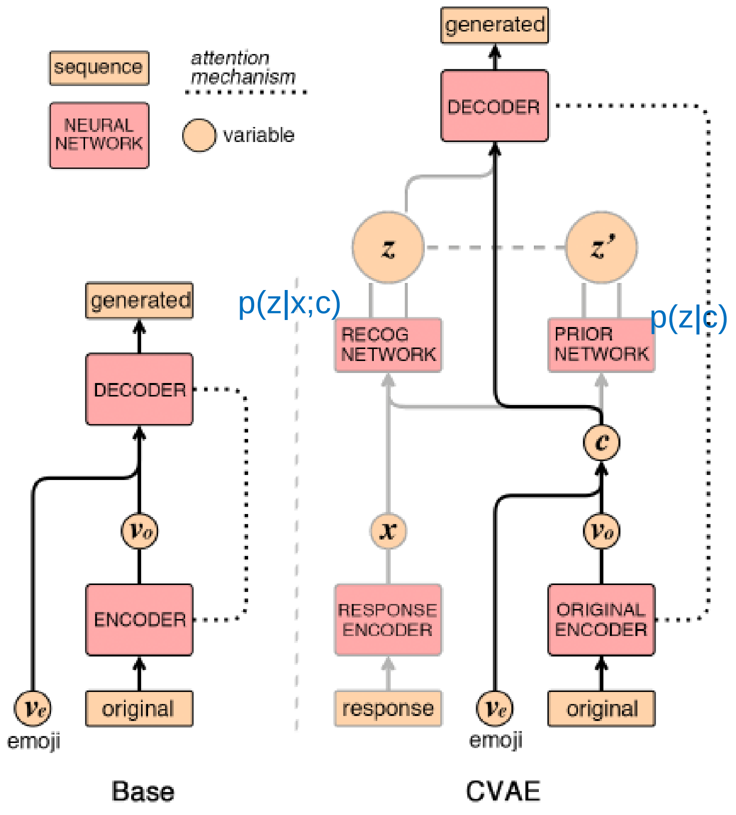


Figure 3: From bottom to top is a forward pass of data during training. **Left**: the base model encodes the original tweets in v_o, and generates responses by decoding from the concatenation of v_o and the embedded emoji, v_e. **Right**: In the CVAE model, all additional components (outlined in gray) can be added incrementally to the base model. A separate encoder encodes the responses in x. Recognition network inputs x and produces the latent variable z by reparameterization trick. During training, The latent variable z is concatenated with v_o and v_e and fed to the decoder.

CVAE is trained by maximizing a variational lower bound on the conditional likelihood of x given c

$$p(x|c) = \int p(x|z,c)p(z|c)dz$$

The lower bound to logp(x|c):

$$-\mathcal{L}(\theta_D, \theta_P, \theta_R; x, c) = \mathrm{KL}(q_R(z|x,c)||p_P(z|c)) - \mathbb{E}_{q_R(z|x,c)}(\log p_D(x|z,c))$$

Reparameterization trick to sample latent variables

a. During training, **z** by the recognition network is passed to the decoder and trained to **approximate z'** by the prior network
b. During testing, the target response is absent, and z' by the prior network is passed to the decoder

Control the emotion of our generation more explicitly ---- RL+CVAE

1. Train an emoji classifier to produce reward for the policy training
2. Get the generated response x' by passing x and c through the CVAE
3. x' to classifier and get the probability of the emoji label as reward R

$$\mathcal{J}(\theta) = \mathbb{E}_{p(x|c)}(R_\theta(x,c)) \qquad \nabla\mathcal{J}(\theta) = (R-r)\nabla\sum_t^{|x|}\log p(x_t|c, x_{1:t-1})$$

Modified policy gradient

1. Adjust rewards according to the position of the emoji label
2. Train Reinforced CVAE by a hybrid objective of REINFORCE and variational lower bound objective

$$\nabla\mathcal{J}'(\theta) = \alpha(R-r)\nabla\sum_t^{|x|}\log p(x_t|c, x_{1:t-1})$$

$$\min_\theta \mathcal{L}'' = \mathcal{L}' - \lambda\mathcal{J}'$$

General

Model	Perplexity	Emoji Accuracy Top1	Emoji Accuracy Top5
	Development		
Base	127.0	34.2%	57.6%
CVAE	37.1	40.7%	75.3%
Reinforced CVAE	38.1	42.2%	76.9%
	Test		
Base	130.6	33.9%	58.1%
CVAE	36.9	41.4%	75.1%
Reinforced CVAE	38.3	42.1%	77.3%

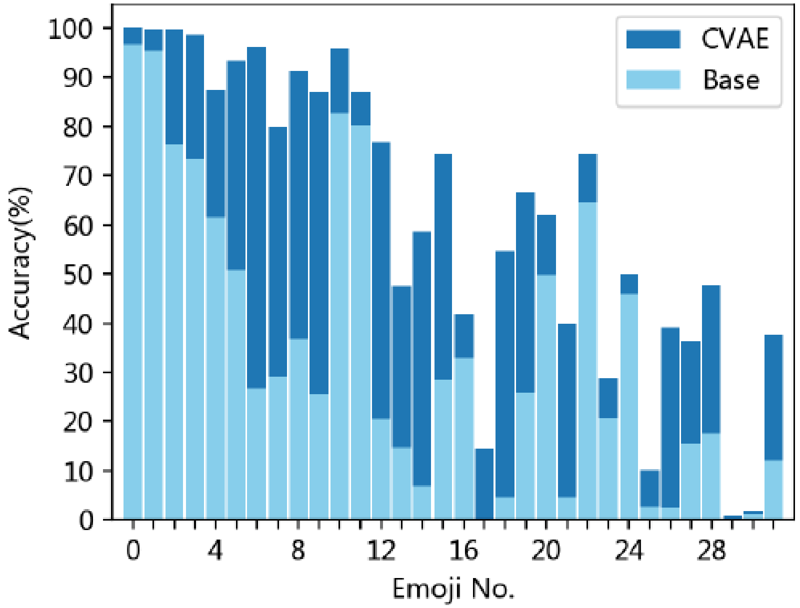
Table 2: Generation perplexity and emoji accuracy of the three models.

Generation Diversity

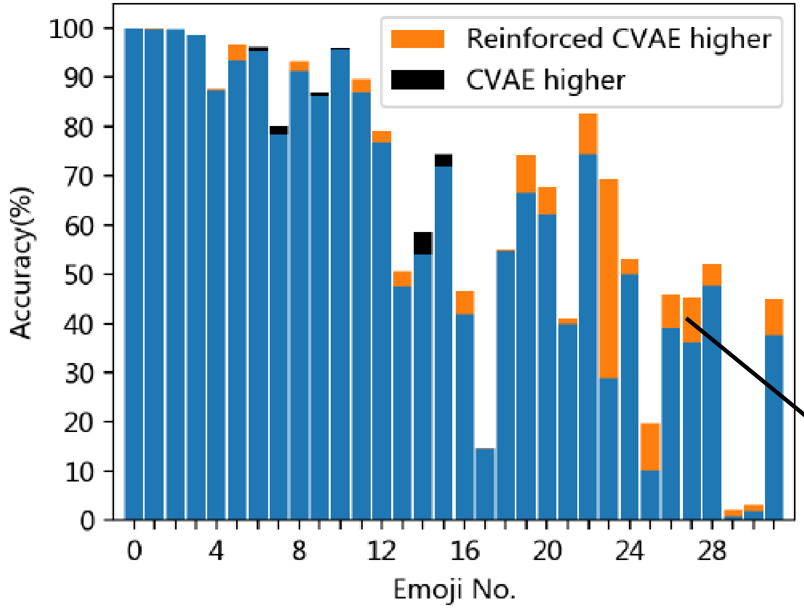
Model	Unigram	Bi-	Tri-
Base	0.0061	0.0199	0.0362
CVAE	0.0191	0.131	0.365
Reinforced CVAE	0.0160	0.118	0.337
Target responses	0.0353	0.370	0.757

Table 3: Type-token ratios of the generation by the three models. Scores of tokenized human-generated target responses are given for reference.

Controllability of Emotions



CVAE v. Base



CVAE v. RL+CVAE

emoji-specified policy training

Human Evaluation

Setting	Model v. Base	Win	Lose	Tie
reply	CVAE	42.4%	43.0%	14.6%
reply	Reinforced CVAE	40.6%	39.6%	19.8%
emoji	CVAE	48.4%	26.2%	25.4%
emoji	Reinforced CVAE	50.0%	19.6%	30.4%

Table 4: Results of human evaluation. Tests are conducted pairwise between CVAE models and the base model.

decide which one better reply the original tweet

pick one better fits given emoji

Content	sorry guys , was gunna stream tonight but i 'm still feeling like crap and my voice disappeared . i will make it up to you		
Target Emotion	😘	🙏	😢
Base	i 'm sorry you 're going to be missed it	i 'm sorry for your loss	i 'm sorry you 're going to be able to get it
CVAE	hope you are okay hun !	hi jason , i 'll be praying for you	im sorry u better suck u off
Reinforced CVAE	hope you 're feeling it	hope you had a speedy recovery man ! hope you feel better soon , please get well soon	dude i 'm so sorry for that i wanna hear it and i 'm sorry i can 't go to canada with you but i wanna be away from canada

Content	add me in there my bro 🙏		
Target Emotion	😜	💪	💔
Base	i 'm not sure you 'll be there	i 'm here for you	i 'm not ready for you
CVAE	you know , you need to tell me in your hometown !	you will be fine bro , i 'll be in the gym for you	i can 't wait 😔
Reinforced CVAE	you might have to get me hip hop off .	good luck bro ! this is about to be healthy	i 'm still undecided and i 'm still waiting

Summary

- Strength

1. The first work that uses emoji-rich Twitter data for emotional response generation. (fine-grained emoji label)

- Weakness

1. Produce responses according to explicit user-input emotions
2. Not consider emotions in input sentences when generating emotional responses (emotion interactions)
3. Multi-turn
4. Exp is enough?

An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss

Peixiang Zhong,[1,2] Di Wang,[1] Chunyan Miao[1,2,3]
[1]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly
[2]Alibaba-NTU Singapore Joint Research Institute
[3]School of Computer Science and Engineering
Nanyang Technological University, Singapore
peixiang001@e.ntu.edu.sg, {wangdi, ascymiao}@ntu.edu.sg

Two challenges

1. Capture the emotion of a sentence. negators and intensifiers often change its polarity and strength
2. Embed emotions naturally in responses with correct grammar and semantics

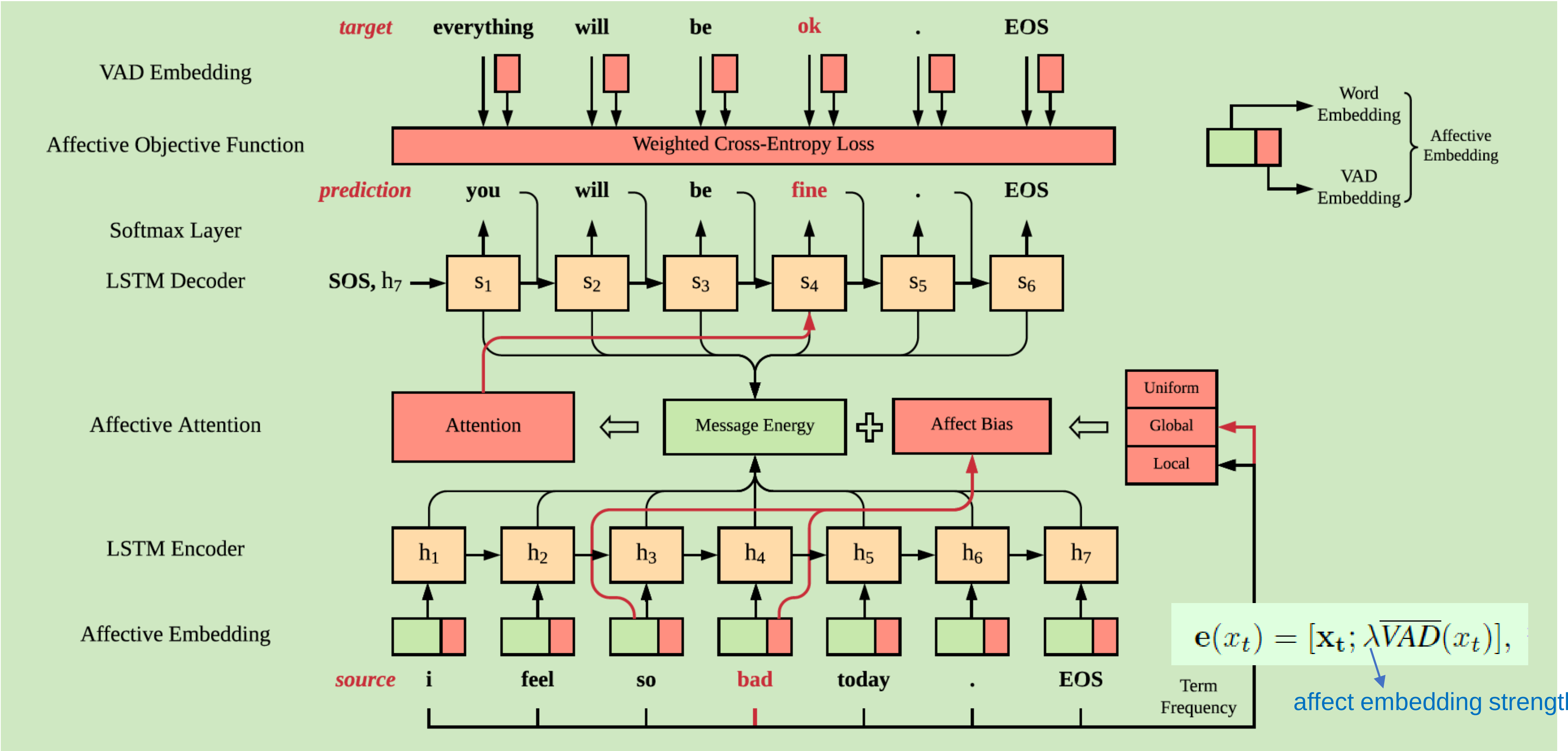
Solution

1. A novel biased attention mechanism that explicitly considers negators and intensifiers in attention computation
2. train Seq2Seq model with a weighted cross-entropy loss that encourages the generation of affect-rich words without degrading language fluency

Dimensions	Values	Interpretations
Valence	3 - 7	pleasant - unpleasant
Arousal	3 - 7	low intensity - high intensity
Dominance	3 - 7	submissive - dominant

Table 1: Interpretations of clipped VAD embeddings.

Nice: V 6.95; A 3.53; D 6.47



$$\mathbf{e}(x_t) = [\mathbf{x_t}; \lambda \overline{VAD}(x_t)],$$

affect embedding strength

Figure 2: Overall architecture of our proposed AR-S2S. This diagram illustrates decoding "fine" and affect bias for "bad".

affect bias

$$e_{t't} = \mathbf{h}_t{}^T \mathbf{s}_{t'} + \eta_t,$$

affective attention coefficient

term importance

$$\eta_t = \gamma \|\mu(x_t)(1+\beta) \otimes \overline{VAD}(x_t)\|_2^2,$$

$$\beta = \tanh(\mathbf{W}^b \mathbf{x_{t-1}}),$$

Negators and Intensifiers. Such as **"not bad"**

term frequency

$$\mu(x_t) = \begin{cases} 1 & \text{ui} \\ a/(a+p(x_t)) & \text{gi} \\ \frac{\log(1/(p(x_t)+\epsilon))}{\sum_{t=1}^{t=T}\log(1/(p(x_t)+\epsilon))} & \text{li} \end{cases},$$

To encourage the generation of affect-rich words

$$\Psi_{t'} = -|V| \frac{1+\delta\|\overline{VAD}(y_{t'})\|_2}{\sum_{\hat{y}_{t'} \in V}(1+\delta\|\overline{VAD}(\hat{y}_{t'})\|_2)} \log(p(y_{t'})), \quad (11)$$

affective loss coefficient

Datasets:

Train: OpenSubtitles
Valid: Cornell Movie Dialog Corpus
Test: DailyDialog (the first two utterances)

Message	Model	Response
Mommy, can i stay up until eleven of the clock?	S2S (MCT)	Of course you can stay up late.
	AR-S2S (MCT)	Of course you can, **sweetheart**.
You are home late today, david. How was school?	S2S (MCT)	It was fine.
	AR-S2S (MCT)	**Great fun** today.
Do you like singing?	S2S (PT)	Yes, i do.
	S2S-Asghar (PT)	I do not know.
	AR-S2S (PT)	I **love** music.
I'm pretty sure that jim will turn out to be a good lawyer.	S2S (PT)	He will turn out to be a good lawyer.
	S2S-Asghar (PT)	I'm sure he will.
	AR-S2S (PT)	The **best** lawyer in the world.

Table 5: Sample responses for models in both MCT and PT. Text in bold are affect-rich words.

Experiment 1: Model Component Test (MCT)

Experiment	Model	#Params	PPL†	PPL‡
MCT (5M pairs) Model Component Test	S2S	99M	42.5	124.3
	S2S-UI	99M	40.4	116.4
	S2S-GI	99M	40.7	120.3
	S2S-LI	99M	40.4	117.0
	S2S-AO	99M	40.2	115.7
	AR-S2S	99M	**39.8**	**113.7**
PT (3M pairs) Preference Test	S2S	66M	41.2	130.6
	S2S-Asghar	66M	46.4	137.2
	AR-S2S	66M	**40.3**	**121.0**

Table 2: Model test perplexity. Symbol † indicates in-domain perplexity obtained on 10K test pairs from the OpenSubtitles dataset. Symbol ‡ indicates out-domain perplexity obtained on 10K test pairs from the DailyDialog dataset.

Model (%)	+2	+1	0	Score	Kappa
S2S	22.4	47.0	30.6	0.918	0.544
S2S-UI	**30.0**	48.6	21.4	**1.086 (+18.3%)**	0.458
S2S-GI	28.6	46.6	24.8	1.038 (+13.1%)	0.413
S2S-LI	29.4	47.2	23.4	1.060 (+15.5%)	0.525
S2S-AO	25.0	46.0	29.0	0.960 (+4.3%)	0.482
AR-S2S	29.6	44.8	25.6	1.040 (+13.3%)	0.487

Table 3: Human evaluations on content quality (MCT).

Model (%)	+2	+1	0	Score	Kappa
S2S	19.0	33.2	47.8	0.712	0.613
S2S-UI	23.6	36.0	40.4	0.832 (+16.9%)	0.483
S2S-GI	26.0	34.2	39.8	0.862 (+21.1%)	0.652
S2S-LI	24.6	36.4	39.0	0.856 (+20.2%)	0.706
S2S-AO	22.6	37.6	39.8	0.828 (+16.3%)	0.602
AR-S2S	**26.8**	37.2	36.0	**0.908 (+27.5%)**	0.625

Table 4: Human evaluations on emotion quality (MCT).

Analysis of Affective Attention

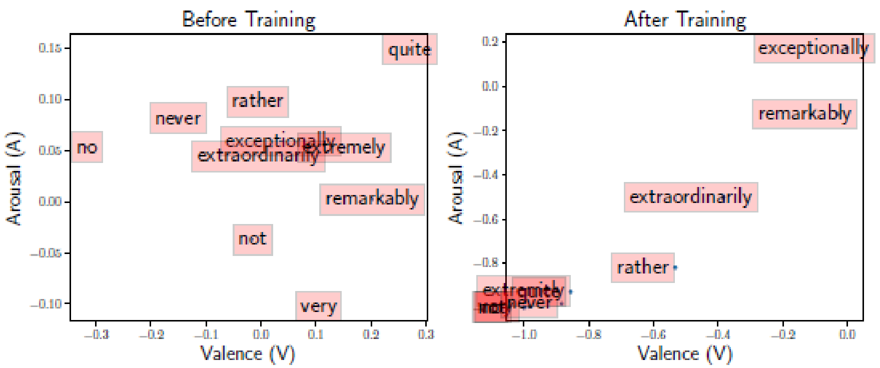


Figure 4: Learned parameter β (see equation (9)) in Valence (V) and Arousal (A) dimensions for several common negators and intensifiers. Left sub-figure: before AR-S2S is trained. Right sub-figure: after AR-S2S is trained.

Different "term importance" have different impacts on the attention strengths

"good" is common in global

Figure 5: Learned attention on a sample input sentence from the testing dataset. From top to bottom, the models are S2S, S2S-UI, S2S-GI and S2S-LI, respectively. Darker colors indicate larger strength.

Analysis of Affective Objective Function

	Threshold for l_2 Norm of VAD		
Model	3	2	1
S2S	25	104	190
S2S-AO ($\delta = 0.5$)	36	138	219
S2S-AO ($\delta = 1$)	50	154	234
S2S-AO ($\delta = 2$)	69	177	256

Table 6: Number of distinct affect-rich words (MCT).

	Threshold for l_2 Norm of VAD		
Model	3	2	1
S2S	21	83	157
S2S-Asghar	31	120	217
AR-S2S	52	173	319

Table 7: Number of distinct affect-rich words (PT).

Experiment 2: Preference Test (PT)

Model (%)	**Content**	**Emotion**	**Kappa**
S2S	64	26	0.522/0.749
S2S-Asghar	66 (+3.1%)	32 (+23.1%)	0.554/0.612
AR-S2S	**80 (+25.0%)**	**49 (+88.5%)**	0.619/0.704

Table 8: Human preference test (PT).

Experiment 3: Sensitivity Analysis

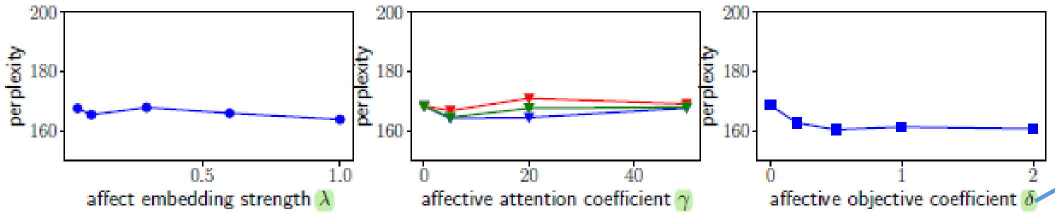


Figure 6: Sensitivity analysis for affect embedding strength λ, affective attention coefficient γ, and affective objective coefficient δ on model perplexity. The blue, red and green curves (*best viewed in color*) in the middle sub-figure represent μ_{ui}, μ_{gi} and μ_{li} (see equation (10)), respectively.

- fairly robust
- affect-rich words are less common than generic words in our training corpus and placing extra weights on them improves the overall prediction performance

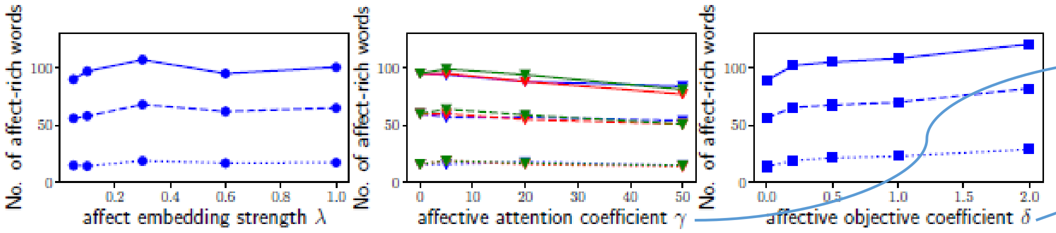


Figure 7: Sensitivity analysis for affect embedding strength λ, affective attention coefficient γ, and affective objective coefficient δ on the number of distinct affect-rich words in randomly selected 1K test responses. The solid, dashed and dotted curves correspond to l_2 norm threshold of 1, 2 and 3, respectively. The blue, red and green curves (*best viewed in color*) in the middle sub-figure represent μ_{ui}, μ_{gi} and μ_{li} (see equation (10)), respectively.

- Decrease because of limited word space
- the number of distinct words consistently increases

Summary

- Strength

1. produces affect-rich responses without performance degradation in language fluency
2. Sufficient experiences in emotion and content

- Weakness

1. Not consider dynamic emotion flow of context in multi-turn settings.
2. The overall emotion state
3. The more emotional words, the better ??

Public Emotional Dialogue Dataset

- DailyDialog: Multi-turn with emotion category label
- EmotionLines: Multi-turn with emotion category label(TV and Fb)
- EmpatheticDialog: Multi-turn based on emotion label and situation

Thanks