Modeling the Continuum of Emotions in Neural Dialogue Systems

Nabiha Asghar

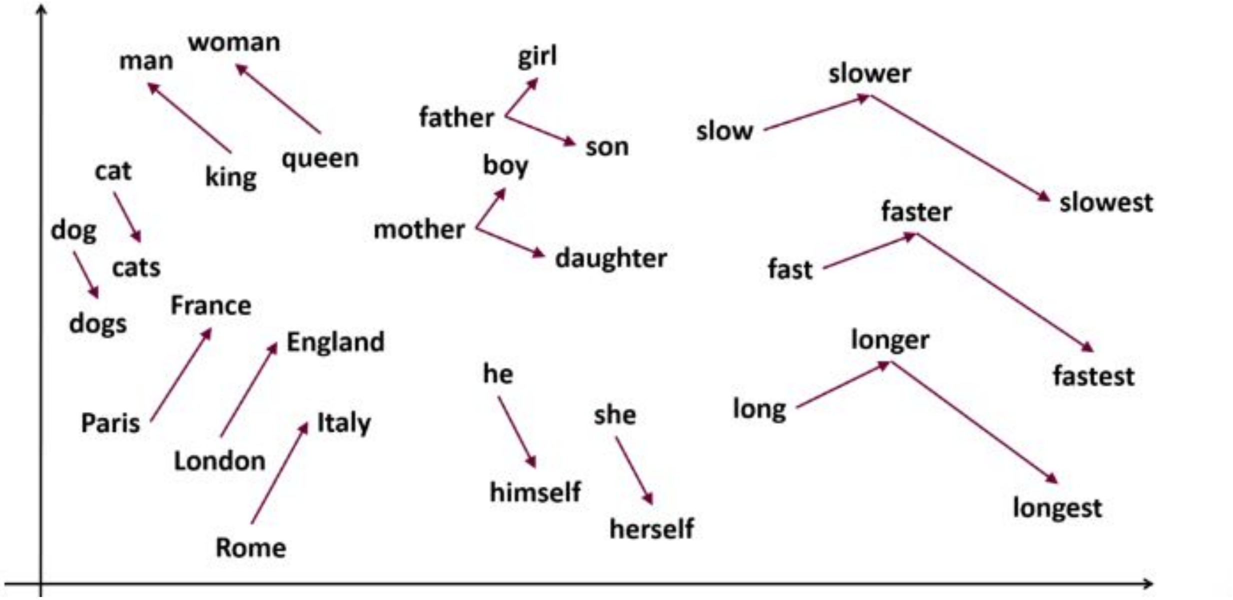
PhD Seminar
19th February 2019

Motivation

- Language → essential for human-human communication
- Humans are rational/logical *and emotional*
- Human actions/decisions are motivated by both emotional *and* non-emotional goals [1,2] [Example: language use]
- Discrete Emotion Theory vs Continuous Emotion Theory
  - Most recent NN research studies discrete emotions

[1] R. Picard. “Affective Computing”. MIT Press, 1997.
[2] Zhu & Thagard. “Emotion and Action”. Philosophical Psychology Vol 15 No 1, 2002.

Word Embeddings (Word2Vec)

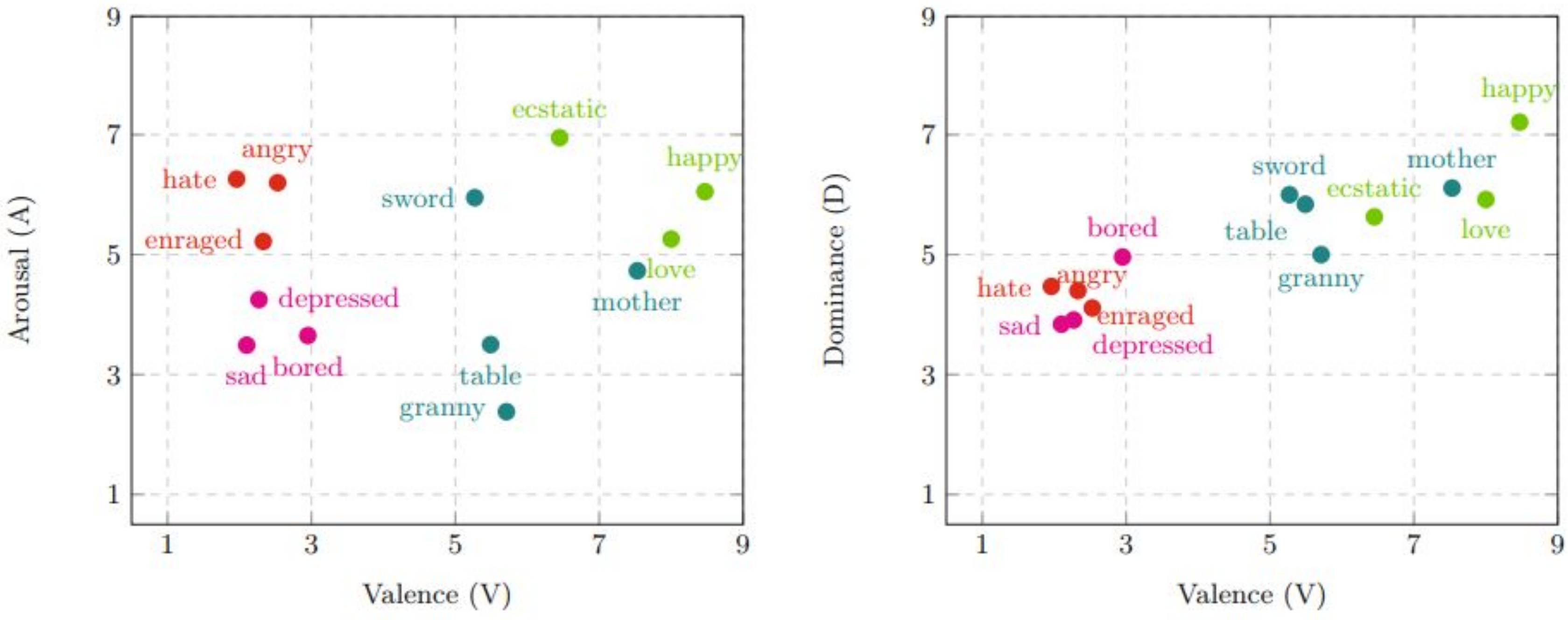


Co-occurrence statistics insufficient to capture emotional features

Words different in sentiment often share context (e.g., “a good book” vs. “a bad book”).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. “Distributed representations of words and phrases and their compositionality”. NIPS 2013

Affective Word Embeddings



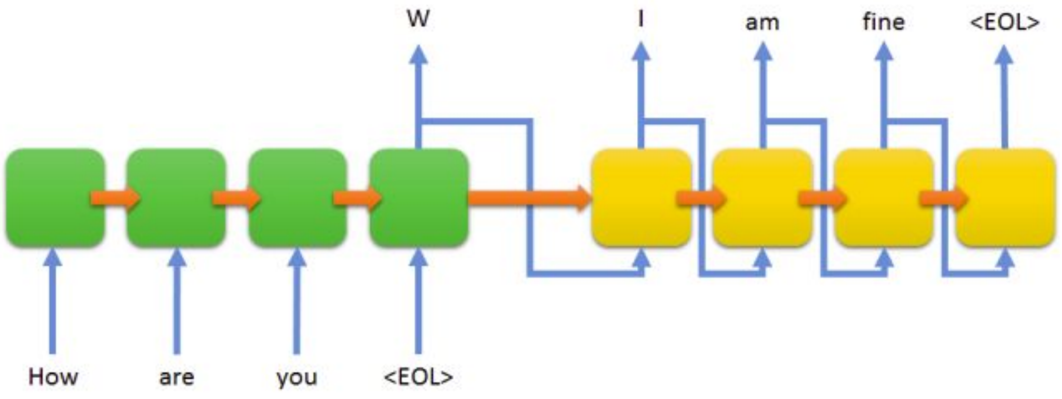
Valence -- Arousal -- Dominance ⇒ (VAD space) ⇒ Scale [0, 9]

Goal

- Leverage word-level affect to generate emotional responses in dialogue

- Design:
    - Affective word embeddings
    - Affective loss functions
    - Affectively diverse “decoding” of response during inference
    - Use “Affect Control Theory” to generate emotional responses

Recurrent NN Model (Seq2Seq)



Encoder RNN　　　　Decoder RNN

Given a message response pair (X, Y) where $X = x_1, \cdots, x_m$ and $Y = y_1, \cdots, y_n$

$$L_{\text{XENT}}(\theta) = -\log p(Y|X) = -\sum_{i=1}^{n} \log p(y_i | y_1, \cdots, y_{i-1}, X)$$

Sutskever, Vinyals, and Le. "Sequence to sequence learning with neural networks." NIPS 2014.

Affective Word Embeddings

- Use cognitively engineered VAD dictionary [5]
- Define

$$\mathtt{W2AV}(w) = \begin{cases} \mathrm{VAD}(l(w)), & \text{if } l(w) \in dict \\ \boldsymbol{\eta} = [5, 1, 5], & \text{otherwise} \end{cases}$$

- Concatenate W2AV with Word2Vec
- Input the result to encoder and the decoder

[5] Warriner, Kuperman, and Brysbaert. “Norms of valence, arousal, and dominance for 13,915 English lemmas”. Behavior Research Methods, 45(4), 2013.

Affective Loss Functions

Minimizing (Maximizing) Affective Dissonance

$$L^{i}_{\mathtt{DMIN}}(\theta) = -(1-\lambda)\log p(y_i|y_1,\cdots,y_{i-1},X) + \lambda\,\hat{p}(y_i)\left\| \sum_{j=1}^{|X|} \frac{\mathtt{W2AV}(x_j)}{|X|} - \sum_{k=1}^{i} \frac{\mathtt{W2AV}(y_k)}{i} \right\|_2$$

Average affect vector of the source sentence

Average affect vector of the target sub-sentence generated up to the current time step i

Affective Loss Functions

Maximizing Affective Content

$$L^i_{\mathtt{AC}}(\theta) = -(1-\lambda)\log p(y_i|y_1,\cdots,y_{i-1},X) - \lambda\,\hat{p}(y_i)\left\|\mathtt{W2AV}(y_i) - \boldsymbol{\eta}\right\|_2$$

These loss functions are not directly differentiable (because W2AV(x) is not continuous function, it's a dictionary. It is not a learnable function). So we relax it with predicted probability.

Beam Search

- BS maintains top B most likely (sub)sequences
- At time t, augments the top-B subsequences from time t − 1 with all possible actions available
- Retain the top-B most likely branches up to time t (prune the rest)

$V \rightarrow$ vocabulary of tokens

$X \rightarrow$ input sequence

$\mathbf{y}_{i,[t-1]} \rightarrow$ i'th beam stored at time t-1

$Y_{[t-1]} = \{\mathbf{y}_{1,[t-1]}, \cdots, \mathbf{y}_{B,[t-1]}\} \rightarrow$ set of beams stored at time t-1

$Y_{[t-1]} \times V \rightarrow$ set of all possible extensions of the beams from time t-1

$$Y_{[t]} = y^{1^*}_{1..t}, \cdots, y^{B^*}_{1..t} = \underset{\substack{\mathbf{y}_{1,[t]}, \cdots, \mathbf{y}_{B,[t]} \\ \in Y_{[t-1]} \times V}}{\arg\max} \sum_{b=1}^{B} \sum_{i=1}^{t} \log p(y_{b,i} | \mathbf{y}_{b,[i-1]}, X)$$

Diverse Beam Search

- Incorporate diversity among candidate outputs
- Divide top-B beams into G groups
- Measure the dissimilarity between group g and previous groups 1, ⋯ , g − 1 if token y_t is selected to extend any beam in group g

$$Y^g_{[t]} = \underset{\substack{\mathbf{y}^g_{1,[t]},\cdots,\mathbf{y}^g_{B',[t]} \\ \in Y^g_{[t-1]} \times V}}{\arg\max} \sum_{b=1}^{B'} \sum_{i=1}^{t} \log p(y^g_{b,i} | \mathbf{y}^g_{b,[i-1]}, X) + \lambda_g \Delta(Y^1_{[t]}, \cdots, Y^{g-1}_{[t]})[y^g_{b,t}]$$

Vijayakumar, Ashwin K., et al. "Diverse beam search: Decoding diverse solutions from neural sequence models." AAAI 2018.

Word-level Affective Diversity

$$\Delta_W(Y^1_{[t]}, \cdots, Y^{g-1}_{[t]})[y^g_{b,t}] = -\sum_{j=1}^{g-1}\sum_{c=1}^{B'} \text{sim}\left(\texttt{W2AV}(y^g_{b,t}), \texttt{W2AV}(y^j_{c,t})\right)$$

$y^g_{b,t}$ → token under consideration at the current time step t for beam b in group g

$y^j_{c,t}$ → token chosen for beam c in a previous group j at time t

This metric ensures that the word affect at time t is diversified across groups

Sentence-level Affective Diversity

$$\Delta_S(Y_{[t]}^1, \cdots, Y_{[t]}^{g-1})[y_{b,t}^g] = -\sum_{j=1}^{g-1} \sum_{c=1}^{B'} \mathrm{sim}\left(\Psi(\mathbf{y}_{b,[t]}^g), \Psi(\mathbf{y}_{c,[t]}^j)\right)$$

$$\text{where} \qquad \Psi(\mathbf{y}_{i,[t]}^k) = \sum_{w \in \mathbf{y}_{i,[t]}^k} \mathtt{W2AV}(w)$$

- Computes the cumulative dissimilarity (given by the function Ψ) between the current beam and all the previously generated beams in other groups
- Bag of affective words

Experiments

- User Study (5 human judges rated responses)
  - Syntactic Coherence
  - Naturalness
  - Emotional Appropriateness
  - Syntactic Diversity
  - Affective Diversity

Experiments

Table 1. The effect of affective word embeddings as input.

Model	Syntactic coherence	Natural	Emotional approp.
Word emb. (baseline)	1.48	0.69	0.41
Word+Affective emb.	**1.71** ↑	**1.05** ↑	**1.01** ↑

Table 2. The effect of affective loss functions.

Model	Syntactic coherence	Naturalness	Emotional approp
$L_{\mathtt{XENT}}$ (baseline)	1.48	0.69	0.41
$L_{\mathtt{DMIN}}$	**1.75** ↑	0.83 ↑	0.56 ↓
$L_{\mathtt{DMAX}}$	1.74 ↑	0.85 ↑	0.58 ↑
$L_{\mathtt{AC}}$	1.71 ↑	**0.95** ↑	**0.71** ↑

Table 3. Effect of affectively diverse decoding. H-DBS refers to Hamming-based DBS used in [22]. WL-ADBS and SL-ADBS are the proposed word-level and sentence-level affectively diverse beam search, respectively.

Model	Syntactic diversity	Affective diversity	# Emotionally approp. responses
BS (baseline)	1.23	0.87	0.89
H-DBS	1.47 ↑	0.79 ↓	0.78 ↓
WL-ADBS	**1.51** ↑	1.25 ↑	1.30 ↑
SL-ADBS	1.45 ↑	**1.31** ↑	**1.33** ↑

Table 4. Combining different affective strategies.

Model	Syntactic coherence	Naturalness	Emotional approp.
Traditional Seq2Seq (baseline)	1.48	0.69	0.41
Seq2Seq+Affective embeddings	1.71 ↑	1.05 ↑	1.01 ↑
Seq2Seq+Affective emb. & Loss	**1.76** ↓	1.03 ↓	1.07 ↑
Seq2Seq+Affective emb. & Loss & Decoding	1.69 ↓	**1.09** ↑	**1.10** ↓

Better Heuristic?

- use Affect Control Theory to model emotional relationship between prompt and response
- Socio-mathematical theory of interaction between two human identities
- Example: friend-friend vs friend-enemy

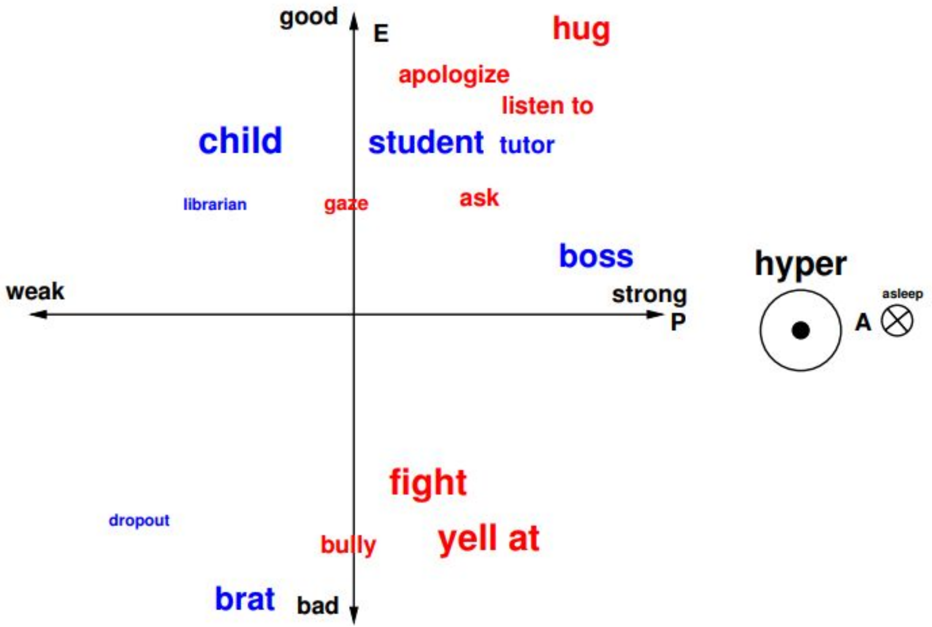
Affect Control Theory (ACT)

- Emotions are points in 3D continuous space $[-4.3, 4.3]^3$
- EPA space: **E**valuation (good/bad), **P**otency (strong/weak), **A**ctivity (excited/calm)
- Fundamental sentiments (**F**) vs Transient impressions (**T**)
- **f**(mother) = [2.9, 1.5, 0.6]
- “A mother hugs a child”, **t**(mother) = [3.5, 1.9, 0.85]
- “A mother hits a child”, **t**(mother) = [-1.0, 3.5, 2.2]

D. R. Heise. “Expressive Order: Confirming Sentiments in Social Actions”. Springer, 2007.

Affect Control Theory (ACT)

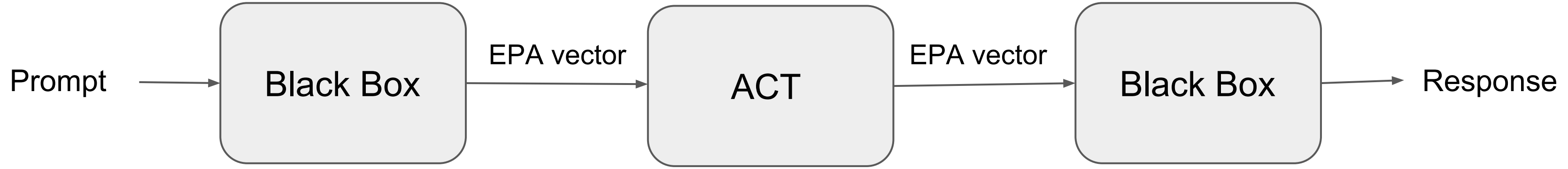
- Emotions are points in 3D continuous space $[-4.3, 4.3]^3$
- EPA space: **E**valuation (good/bad), **P**otency (strong/weak), **A**ctivity (excited/calm)
- Fundamental Sentiments **F**
- Transient Impressions **T**
- Deflection: $D = \sum_i w_i(\mathrm{f}_i - \tau_i)^2$



Affect Control Principle: actors work to minimize deflection, i.e. experience transient impressions that are consistent with their fundamental sentiments

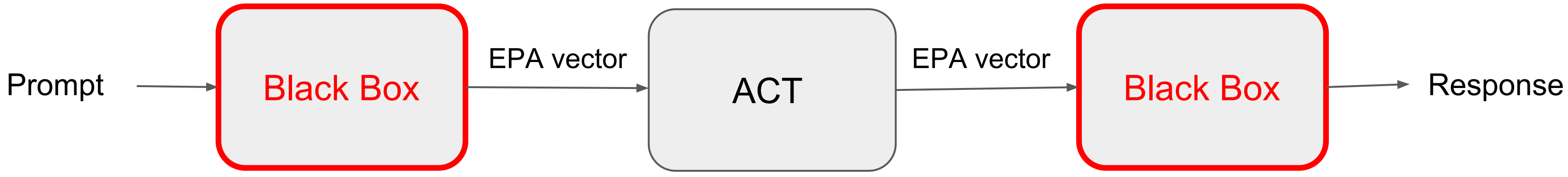
Model

- ACT takes EPA vectors as input, produces EPA vectors (actions)
- Need a way to convert sentences into EPA, and vice versa
- Pipeline:



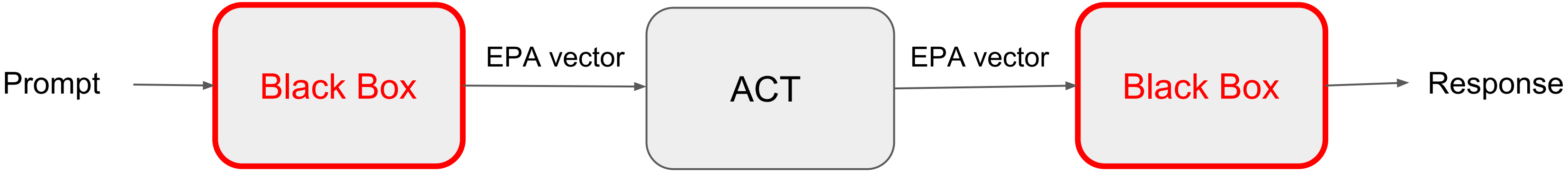
Model

- ACT takes EPA vectors as input, produces EPA vectors (actions)
- Need a way to convert sentences into EPA, and vice versa
- Pipeline:



Research Questions

1. How to convert prompt sentence into EPA vector?
2. How to convert EPA action to output sentence?

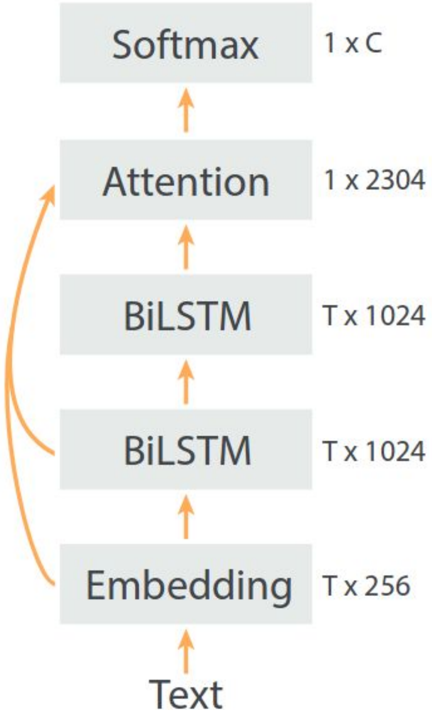


Sentence to EPA

- We can train our own neural network, but labelling is too tedious

Sentence to EPA

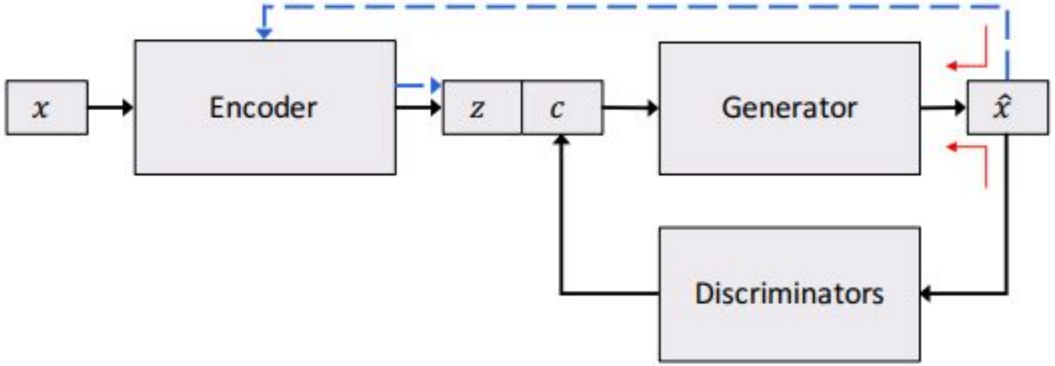
- We can train our own neural network, but labelling is too tedious
- Use **DeepMoji** - a pre-trained RNN model (billions of tweets). Predicts emojis (64 classes) given input sentence.
    - Asked 2 human judges to label the 64 emojis with EPA vectors. Average the labels.
    - This gives us an EPA for each emoji.
    - Given some input sentence, query DeepMoji and take weighted average of output



Felbo, Bjarke, et al. "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm." *EMNLP 2017.*

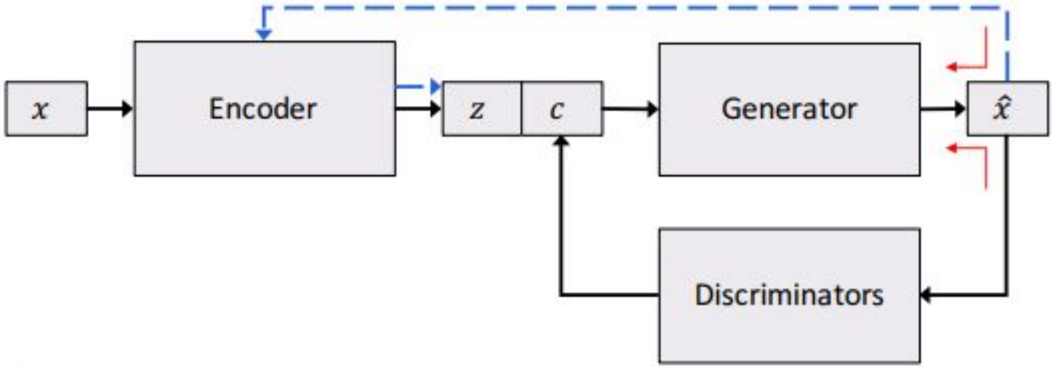
EPA to Sentence

- Variational AutoEncoder (VAE) and discriminator to learn disentangled representation of emotions in text. Then query Generator with EPA



Hu, Zhiting, et al. "Toward controlled generation of text." *ICML* 2017.

Model



Generator Objective:

$$\min_{\boldsymbol{\theta}_G} \mathcal{L}_G = \mathcal{L}_{\text{VAE}} + \lambda_c \mathcal{L}_{\text{Attr},c} + \lambda_z \mathcal{L}_{\text{Attr},z}$$

$$\mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}_G, \boldsymbol{\theta}_E; \boldsymbol{x}) = \text{KL}(q_E(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})) - \mathbb{E}_{q_E(\boldsymbol{z}|\boldsymbol{x}) q_D(\boldsymbol{c}|\boldsymbol{x})} [\log p_G(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})]$$

$$\mathcal{L}_{\text{Attr},c}(\boldsymbol{\theta}_G) = -\mathbb{E}_{p(\boldsymbol{z})p(\boldsymbol{c})} \left[\log q_D(\boldsymbol{c}|\widetilde{G}_\tau(\boldsymbol{z}, \boldsymbol{c})) \right]$$

$$\mathcal{L}_{\text{Attr},z}(\boldsymbol{\theta}_G) = -\mathbb{E}_{p(\boldsymbol{z})p(\boldsymbol{c})} \left[\log q_E(\boldsymbol{z}|\widetilde{G}_\tau(\boldsymbol{z}, \boldsymbol{c})) \right]$$

Experiments - ACT

Prompt (Friend)	Response (Friend)	Response (Enemy)
I missed you buddy	Your pic is so cool!	Who doesn’t do that
Let’s hang out together	Love it been so long	I am horrified by you
Take care I love you	I’m going to miss this	I cannot.

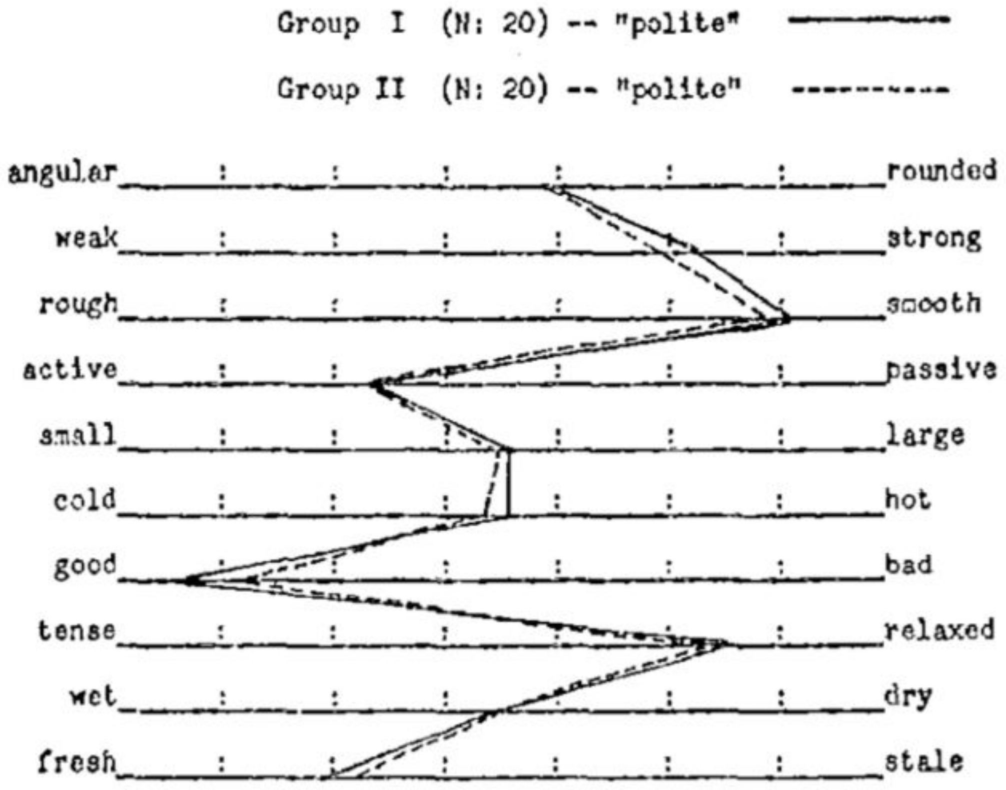
- Constructing sentence given an EPA can be hard (imagine trying to come up with a sentence suitable for (-1, 2, -3).
- Not easy to disentangle emotions from text

Conclusion

Discussed:

- Continuous vs Discrete Emotion Theory
- Modeling continuous emotions in Seq2Seq framework
- Affective word embeddings, loss functions, diverse decoding
- (Attempting to) Disentangle emotions in latent space
- Use socio-mathematical emotion theory to generate emotional responses

Osgood’s Semantic Differential



Observations:

- People within one culture answer more or less similarly.
- On average, 50% of variation in semantic differential ratings can be explained by three principal components:

E: good, nice................bad, awful

P: strong, powerful................weak, powerless

A: active, excited................passive, calm

C. Osgood. “The nature and measurement of meaning. Psychological bulletin”. 49(3):197, 1952