



# Affective Neural Response Generation

Nabiha Asghar<sup>1</sup>(✉) , Pascal Poupart<sup>1</sup>, Jesse Hoey<sup>1</sup>, Xin Jiang<sup>2</sup>, and Lili Mou<sup>1</sup>

<sup>1</sup> Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada  
{nasghar, ppoupart, jhoey}@cs.uwaterloo.ca, doublepower.mou@gmail.com

<sup>2</sup> Noah's Ark Lab, Huawei Technologies, Sha Tin, Hong Kong  
xin.jiang@huawei.com

**Abstract.** Existing neural conversational models process natural language primarily on a lexico-syntactic level, thereby ignoring one of the most crucial components of human-to-human dialogue: its affective content. We take a step in this direction by proposing three novel ways to incorporate affective/emotional aspects into long short term memory (LSTM) encoder-decoder neural conversation models: (1) affective word embeddings, which are cognitively engineered, (2) affect-based objective functions that augment the standard cross-entropy loss, and (3) affectively diverse beam search for decoding. Experiments show that these techniques improve the open-domain conversational prowess of encoder-decoder networks by enabling them to produce more natural and emotionally rich responses.

**Keywords:** Dialogue systems · Human computer interaction  
Natural language processing · Affective computing

## 1 Introduction

Human-computer dialogue systems have wide applications ranging from restaurant booking [24] to emotional virtual agents [13]. In a neural network-based dialogue system, discrete words are mapped to real-valued vectors, known as *embeddings*, capturing abstract meanings of words [14]; then an encoder-decoder framework—with long short term memory (LSTM)-based recurrent neural networks (RNNs)—generates a response conditioned on one or several previous utterances. Recent advances in this direction have demonstrated its efficacy for both task-oriented [24] and open-domain dialogue generation [11].

While most of the existing neural conversation models generate syntactically well-formed responses, they are prone to being short, dull, or vague. Latest efforts to address these issues include diverse decoding [22], diversity-promoting objective functions [10], human-in-the-loop reinforcement/active learning [1, 11] and content-introducing approaches [15]. However, one shortcoming of these existing open-domain neural conversation models is the lack of *affect* modeling of natural language. These models, when trained over large dialogue datasets, do not capture the emotional states of the two humans interacting

in the textual conversation, which are typically manifested through the choice of words or phrases. For instance, the attention mechanism in a sequence-to-sequence (Seq2Seq) model can learn syntactic alignment of words within the generated sequences [2]. Also, neural word embedding models like Word2Vec learn word vectors by context, and can preserve low-level word semantics (e.g., “king” – “male”  $\approx$  “queen” – “woman”). However, emotional aspects are not explicitly captured by existing methods.

Our goal is to alleviate this issue in open-domain neural dialogue models by augmenting them with affective intelligence. We do this in three ways.

1. We embed words in a 3D affective space by retrieving word-level affective ratings from a cognitively engineered affective dictionary [23], where affectively similar constructs are close to one other. In this way, the ensuing neural model is aware of words’ emotional features.
2. We augment the standard cross-entropy loss with affective objectives, so that our neural models are taught to generate more emotional utterances.
3. We inject affective diversity into the responses generated by the decoder through *affectively diverse* beam search algorithms, and thus our model actively searches for affective responses during decoding.

We also show that these emotional aspects can be combined to further improve the quality of generated responses in an open-domain dialogue system. Overall, in information-retrieval tasks like question-answering, our proposed models can help retain the users by interacting in a more human way.

## 2 Related Work

Affectively cognizant virtual agents are attracting interest both in the academia [13] and the industry,<sup>1</sup> due to their ability to provide emotional companionship to humans. Past research has mostly focused on developing hand-crafted speech and text-based features to incorporate emotions in retrieval-based or slot-based spoken dialogue systems [3, 18]. Our work is related to two very recent studies:

- Affect Language Model [6, Affect-LM] is an LSTM-RNN language model which leverages the LIWC [17] text analysis program for affective feature extraction through keyword spotting. It considers binary affective features, namely *positive emotion*, *angry*, *sad*, *anxious*, and *negative emotion*. Our work differs from Affect-LM in that we consider affective dialogue systems instead of merely language models.
- Emotional Chatting Machine [26, ECM] is a Seq2Seq model. It takes as input a prompt and the desired emotion of the response, and produces a response. It has 8 emotion categories, namely *anger*, *disgust*, *fear*, *happiness*, *like*, *sadness*, *surprise*, and *other*. Our approach does not require the input of desired emotion as in ECM, which is unrealistic in applications. Instead, we intrinsically model emotion by affective word embeddings as input, as well as objective functions and inference criterion based on these embeddings.

<sup>1</sup> <https://www.ald.softbankrobotics.com/en/robots/pepper>.

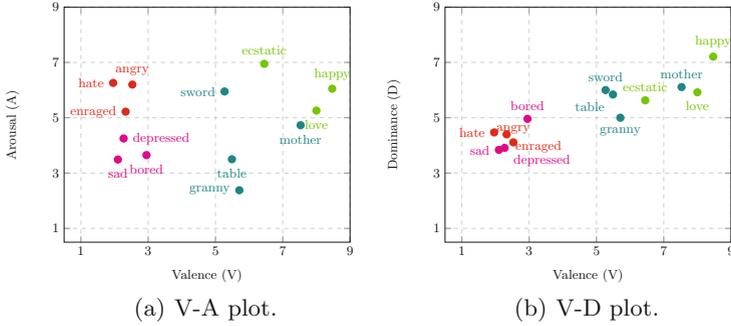


Fig. 1. Relationship between several adjectives, nouns, and verbs on 3-D VAD scale.

### 3 Background

In NLP, **word embeddings** map words (or tokens) to real-valued vectors of fixed dimensionality. Typically, they are learned from the co-occurrence statistics of words in large natural language corpora, and the learned embedding vector space has such a property that words sharing similar syntactic and semantic context are close to each other. However, it is known that co-occurrence statistics are insufficient to capture sentiment/emotional features, because words different in sentiment often share context (e.g., “a *good* book” vs. “a *bad* book”).

A **sequence-to-sequence (Seq2Seq) model** is an encoder-decoder neural framework that maps a variable length input sequence to a variable length output sequence [21]. It consists of an encoder and a decoder, both of which are RNNs (typically with LSTM units). The encoder network sequentially accepts the embedding of each word in the input sequence, and encodes the input sentence as a vector. The decoder network takes the vector as input and sequentially generates an output sequence. Given a message-response pair  $(X, Y)$ , where  $X = x_1, \dots, x_m$  and  $Y = y_1, \dots, y_n$  are sequences of words, Seq2Seq models (parametrized by  $\theta$ ) are typically trained with cross entropy loss (XENT):

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

### 4 The Proposed Affective Approaches

In this section, we propose affective neural response generation, which augments traditional neural conversation models with emotional cognizance. We leverage a cognitively engineered dictionary to propose three strategies for affective response generation, namely affective word embeddings as input, affective training objectives, and affectively diverse beam search. As will be shown later, these affective strategies can be combined to further improve Seq2Seq dialogue systems.

## 4.1 Affective Word Embeddings

As said, traditional word embeddings trained with co-occurrence statistics are insufficient to capture affect aspects. We propose to augment traditional word embeddings with a 3D affective space by using an external cognitively-engineered affective dictionary [23].<sup>2</sup> The dictionary we use consists of 13,915 lemmatized English words, each of which is rated on three traditionally accepted continuous and real-valued dimensions of emotion: Valence (V, the pleasantness of a stimulus), Arousal (A, the intensity of emotion produced by a stimulus), and Dominance (D, the degree of power exerted by a stimulus). Sociologists hypothesize that the VAD space captures almost 70% of the variance in affective meanings of concepts [16]. VAD ratings have been previously used in sentiment analysis and empathetic tutors, among other affective computing applications [8, 19]. To the best of our knowledge, we are the first to introduce VAD to dialogue systems.

The scale of each dimension in the VAD space is from 1 to 9, where a higher value corresponds to higher valence, arousal, or dominance. Thus,  $V \simeq 1, 5$  and 9 corresponds to a word being very negative (*pedophile*), neutral (*tablecloth*) and very positive (*happiness*), respectively. Similarly,  $A \simeq 1, 5$  and 9 corresponds to a word having very low (*dull*), moderate (*watchdog*), and very high (*insanity*) emotional intensity, respectively. Finally,  $D \simeq 1, 5$  and 9 corresponds to a word that is very powerless (*dementia*), neutral (*waterfall*) and very powerful (*paradise*), respectively. The VAD ratings of each word were collected through a survey in [23] over 1800 participants. We directly take them as the 3-dimensional word-level affective embeddings. Some examples of words (including nouns, adjectives, and verbs) and their corresponding VAD values are depicted in Fig. 1.

For words missing in this dictionary, such as stop words and proper nouns, we set the VAD vector to be the neutral vector  $\boldsymbol{\eta} = [5, 1, 5]$ , because these words are neutral in pleasantness (V) and power (D), and evoke no arousal (A). Formally, we define “word to affective vector” (W2AV) as:

$$\text{W2AV}(w) = \begin{cases} \text{VAD}(l(w)), & \text{if } l(w) \in \text{dict} \\ \boldsymbol{\eta} = [5, 1, 5], & \text{otherwise} \end{cases} \quad (2)$$

where  $l(w)$  is the lemmatization of the word  $w$ . In this way, words depicting similar emotions are close together in the affective space, and affectively dissimilar words are far apart from each other. Thus W2AV is suitable for neural processing.

The simplest approach to utilize W2AV, perhaps, is to feed it to a Seq2Seq model as input. Concretely, we concatenate the W2AV embeddings of each word with its traditional word embeddings, the resulting vector being the input to both the encoder and the decoder.

## 4.2 Affective Loss Functions

Equipped with affective vectors, we further design affective training loss functions to explicitly train an affect-aware Seq2Seq conversation model. The philosophy

<sup>2</sup> Available for free at <http://crr.ugent.be/archives/1003>.

of manipulating loss function is similar to [10], but we focus on affective aspects (instead of diversity in general). We have several heuristics as follows.

**Minimizing Affective Dissonance.** We start with the simplest approach: maintaining affective consistency between prompts and responses. This heuristic arises from the observation that typical open-domain textual conversations between two humans consist of messages and responses that, in addition to being affectively loaded, are affectively similar to each other. For instance, a friendly message typically elicits a friendly response and provocation usually results in anger or contempt. Assuming that the general affective tone of a conversation does not fluctuate too suddenly and too frequently, we emulate human-human interactions in our model by minimizing the *dissonance* between the prompts and the responses, i.e. the Euclidean distance between their affective embeddings. This objective allows the model to generate responses that are emotionally aligned with the prompts. Thus, at time step  $i$ , the loss is computed by

$$L_{\text{DMIN}}^i(\theta) = -(1 - \lambda) \log p(y_i | y_1, \dots, y_{i-1}, X) + \lambda \hat{p}(y_i) \left\| \sum_{j=1}^{|X|} \frac{\mathbf{w2AV}(x_j)}{|X|} - \sum_{k=1}^i \frac{\mathbf{w2AV}(y_k)}{i} \right\|_2 \quad (3)$$

where  $\|\cdot\|_2$  denotes  $\ell_2$ -norm. The first term is the standard XENT loss as in Eq. 1. The sum  $\sum_j \frac{\mathbf{w2AV}(x_j)}{|X|}$  is the average affect vector of the source sentence, whereas  $\sum_k \frac{\mathbf{w2AV}(y_k)}{i}$  is the average affect vector of the target sub-sentence generated up to the current time step  $i$ .

In other words, we penalize the distance between the average affective embeddings of the source and the target sentences. Notice that this affect distance is not learnable and that selecting a single predicted word makes the model indifferentiable. Therefore, we relax hard prediction of a word by its predicted probability  $\hat{p}(y_i)$ .  $\lambda$  is a hyperparameter balancing the two factors.

**Maximizing Affective Dissonance.** Admittedly, minimizing the affective dissonance does not always make sense while we model a conversation. An over-friendly message from a stranger may elicit anger or disgust from the recipient. Furthermore, responses that are *not* too affectively aligned with the prompts may be perceived as more interesting, by virtue of being less predictable. Thus, we design an objective function  $L_{\text{DMAX}}$  that *maximizes* the dissonance by flipping the sign in the second term in Eq. 3. (Details are not repeated here.)

**Maximizing Affective Content.** Our third heuristic encourages Seq2Seq to generate affective content, but does not specify the polarity of sentiment. This explores the hypothesis that most of the casual human responses are not dull or emotionally neutral. Concretely, we maximize the affective content of the model’s responses, so that it avoids generating generic responses like “yes,” “no,” “I don’t know,” and “I’m not sure.” That is, at the time step  $i$ , the loss function is

$$L_{\text{AC}}^i(\theta) = -(1 - \lambda) \log p(y_i | y_1, \dots, y_{i-1}, X) - \lambda \hat{p}(y_i) \left\| \mathbf{w2AV}(y_i) - \boldsymbol{\eta} \right\|_2 \quad (4)$$

The second term is a regularizer that discourages non-affective words. We penalize the distance between  $y_i$ 's affective embedding and the affectively neutral vector  $\boldsymbol{\eta} = [5, 1, 5]$ , so the model pro-actively chooses emotionally rich words.

### 4.3 Affectively Diverse Decoding

In this subsection, we propose affectively diverse decoding that incorporates affect into the decoding process of neural response generation.

Traditionally, beam search (BS) has been used for decoding in Seq2Seq models because it provides a tractable approximation of searching an exponentially large solution space. However, in the context of open-domain dialogue generation, BS is known to produce nearly identical samples like “*This is great!*” and “*This is so great!*”. Diverse beam search (DBS) [22] is a recently proposed variant of BS that explicitly considers diversity during decoding; it has been shown to outperform BS and other diverse decoding techniques in many NLP tasks.

Below, we describe BS, DBS, and our proposed affective variants of DBS.

**Beam Search (BS) and Diverse Beam Search (DBS).** BS maintains top- $B$  most likely (sub)sequences, where  $B$  is known as the *beam size*. At each time step  $t$ , the top- $B$  subsequences at time step  $t - 1$  are augmented with all possible actions available; then the top- $B$  most likely branches are retained at time  $t$ , and the rest are pruned. Let  $V$  be the set of vocabulary tokens,  $X$  be the input sequence,  $\mathbf{y}_{i,[t-1]}$  be the  $i$ th beam stored at time  $t - 1$ , and  $Y_{[t-1]} = \{\mathbf{y}_{1,[t-1]}, \dots, \mathbf{y}_{B,[t-1]}\}$  be the set of beams stored by BS at time  $t - 1$ . Then at time  $t$ , the BS objective is

$$Y_{[t]} = y_{1..t}^1, \dots, y_{1..t}^B = \arg \max_{\substack{\mathbf{y}_{1,[t]}, \dots, \mathbf{y}_{B,[t]} \\ \in Y_{[t-1]} \times V}} \sum_{b=1}^B \sum_{i=1}^t \log p(y_{b,i} | \mathbf{y}_{b,[i-1]}, X) \quad (5)$$

subject to  $\mathbf{y}_{i,[t]} \neq \mathbf{y}_{j,[t]}$ , where  $Y_{[t-1]} \times V$  is the set of all possible extensions based on the beams stored at time  $t - 1$ . DBS aims to overcome the diversity problem in BS by incorporating diversity among candidate outputs. It divides the top- $B$  beams into  $G$  groups (each group containing  $B' = G/B$  beams) and adds to traditional BS (Eq. 5) a dissimilarity term  $\Delta(Y_{[t]}^1, \dots, Y_{[t]}^{g-1})[y_t]$  which measures the dissimilarity between group  $g$  and previous groups  $1, \dots, g - 1$  if token  $y_t$  is selected to extend any beam in group  $g$ . This is given by

$$Y_{[t]}^g = \arg \max_{\substack{\mathbf{y}_{1,[t]}^g, \dots, \mathbf{y}_{B',[t]}^g \\ \in Y_{[t-1]}^g \times V}} \sum_{b=1}^{B'} \sum_{i=1}^t \log p(y_{b,i}^g | \mathbf{y}_{b,[i-1]}^g, X) + \lambda_g \Delta(Y_{[t]}^1, \dots, Y_{[t]}^{g-1})[y_{b,t}^g] \quad (6)$$

subject to  $\mathbf{y}_{i,[t]}^g \neq \mathbf{y}_{j,[t]}^g$ , where  $\lambda_g \geq 0$  is a hyperparameter controlling the diversity strength. Intuitively, DBS modifies the probability in BS by adding a dissimilar term between a particular sample (i.e.,  $y_{b,1}^g \dots y_{b,t}^g$ ) and samples in

other groups (i.e.,  $Y_{[t]}^1, \dots, Y_{[t]}^{g-1}$ ). We refer readers to [22] for the details of DBS. Here, we focus on the dissimilarity metric that can incorporate affective aspects into the decoding phase.

**Affectively Diverse Beam Search (ADBS).** The dissimilarity metric for DBS can take many forms as used in [22]: Hamming diversity that penalizes tokens based on the number of times they are selected in the previous groups, n-gram diversity that discourages repetition of n-grams between groups, and neural-embedding diversity that penalizes words with similar embeddings across groups. Among these, the neural-embedding diversity metric is the most relevant to us. When used with Word2Vec embeddings, this metric discourages semantically similar words (e.g., synonyms) to be selected across different groups.

To decode affectively diverse samples, we propose to inject affective dissimilarity across the beam groups based on affective word embeddings. This can be done either at the word level or sentence level. We formalize these notions below.

- *Word-Level Diversity for ADBS (WL-ADBS).* We define the word-level affect dissimilarity metric  $\Delta_W$  to be

$$\Delta_W(Y_{[t]}^1, \dots, Y_{[t]}^{g-1})[y_{b,t}^g] = - \sum_{j=1}^{g-1} \sum_{c=1}^{B'} \text{sim}(W2AV(y_{b,t}^g), W2AV(y_{c,t}^j)) \quad (7)$$

where  $\text{sim}(\cdot)$  denotes a similarity measure between two vectors. In our experiments, we use the cosine similarity function.  $y_{b,t}^g$  denotes the token under consideration at the current time step  $t$  for beam  $b$  in group  $g$ , and  $y_{c,t}^j$  denotes the token chosen for beam  $c$  in a previous group  $j$  at time  $t$ .

Intuitively, this metric computes the cosine similarity of group  $g$ 's beam  $b$  with all the beams generated in groups  $1, \dots, g-1$ . The metric operates at the word level, ensuring that the word affect at time  $t$  is diversified across groups.

- *Sentence-Level Diversity for ADBS (SL-ADBS).* The word-level metric  $\Delta_W$  in Eq. 7 does not take into account the overall sentence affect for each group. We propose an alternative sentence-level affect diversity metric, given by

$$\Delta_S(Y_{[t]}^1, \dots, Y_{[t]}^{g-1})[y_{b,t}^g] = - \sum_{j=1}^{g-1} \sum_{c=1}^{B'} \text{sim}(\Psi(\mathbf{y}_{b,[t]}^g), \Psi(\mathbf{y}_{c,[t]}^j)) \quad (8)$$

$$\text{where} \quad \Psi(\mathbf{y}_{i,[t]}^k) = \sum_{w \in \mathbf{y}_{i,[t]}^k} W2AV(w) \quad (9)$$

Here,  $\mathbf{y}_{i,[t]}^k$  for  $k \leq g$  is the  $i$ th beam in the  $k$ th group stored at time  $t$ ;  $\mathbf{y}_{b,[t]}^g$  is the concatenation of  $\mathbf{y}_{b,[t-1]}^g$  and  $y_{b,t}^g$ . Intuitively, this metric computes the *cumulative dissimilarity* (given by the function  $\Psi(\cdot)$ ) between the current beam and all the previously generated beams in other groups. This bag-of-affective-words approach is simple but works well in practice, as will be shown later.

**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.	<b>1.71</b> ↑	<b>1.05</b> ↑	<b>1.01</b> ↑

**Table 2.** The effect of affective loss functions.

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

It should be also noticed that several other beam search-based diverse decoding techniques have been proposed in recent years, including DivMBest [7] and MMI objective [10]. All of them use the notion of a *diversity term* within BS; therefore our affect-injecting technique can be used with these algorithms.

## 5 Experiments

We evaluated our approach on the Cornell Movie Dialogs Corpus [4], which contains  $\sim 300\text{k}$  utterance-response pairs. All our model variants used a single-layer LSTM encoder and a single-layer LSTM decoder, each layer containing 1024 cells. We set the vocabulary size to 12,000 and used Adam [9] optimizer.

For the baseline  $L_{\text{XENT}}$  loss, we used 1024-D Word2Vec embeddings as input and trained the Seq2Seq model for 50 epochs using Eq. 1. For the affective embeddings as input, we used 1027-D vectors, each a concatenation of 1024-D Word2Vec and 3-D W2AV embeddings. Training was done for 50 epochs. For each of the affective loss functions ( $L_{\text{AC}}$ ,  $L_{\text{DMIN}}$ , and  $L_{\text{DMAX}}$ ), we trained the model using  $L_{\text{XENT}}$  loss for 40 epochs, followed by 10 epochs using the affective loss functions. The ADBS metrics  $\Delta_W$  and  $\Delta_S$  were deployed at test time with  $G = B$ .

### 5.1 Results

Recent work employs both automated metrics (e.g., BLEU, ROUGE, and METEOR) and human judgments to evaluate dialogue systems. While automated metrics enable high-throughput evaluation, they have weak or no correlation with human judgments [12]. It is also unclear how to evaluate affective aspects by automated metrics. Therefore, in this work, we recruited 3–5 human judges to evaluate our models, following several previous studies [15, 20].

To evaluate the quality of the generated responses, we used 5 workers to evaluate 100 test samples for each model variant in terms of *syntactic coherence*

**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	<b>1.51</b> ↑	1.25 ↑	1.30 ↑
SL-ADBS	1.45 ↑	<b>1.31</b> ↑	<b>1.33</b> ↑

**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	<b>1.76</b> ↓	1.03 ↓	1.07 ↑
Seq2Seq+Affective emb. & Loss & Decoding	1.69 ↓	<b>1.09</b> ↑	<b>1.10</b> ↓

(Does the response make grammatical sense?), *naturalness* (Could the response have been plausibly produced by a human?) and *emotional appropriateness* (Is the response emotionally suitable for the prompt?). For each axis, the judges were asked to assign each response an integer score of 0 (bad), 1 (satisfactory), or 2 (good). The scores were then averaged for each axis (Tables 1 and 2). We evaluated the inter-annotator consistency by Fleiss’  $\kappa$  score [5], and obtained a  $\kappa$  score of 0.447, interpreted as “moderate agreement” among the judges.<sup>3</sup> We also computed the statistical significance of the results using one-tailed Wilcoxon’s Signed Rank Test [25] with significance level set to 0.05. This is indicated in Tables 1 and 2 through arrows: a down-arrow indicates that the model performed equally well as the baseline, and an up-arrow indicates that the model performed significantly better than the baseline.

The evaluation of diversity was conducted separately (Table 3). In this experiment, an annotator was presented with top-three decoded responses and was asked to judge *syntactic diversity* (How syntactically diverse are the five responses?) and *emotional diversity* (How affectively diverse are the five responses?). The rating scale was 0, 1, 2, and 3 with labels bad, satisfactory, good, and very good, respectively. The annotator was also asked to state the

<sup>3</sup> <https://en.wikipedia.org/wiki/Fleiss%27.kappa>.

number of beams that were emotionally appropriate to the prompt. The scores obtained for each question were averaged. We used three annotators in this experiment (fewer than the previous one), as it required more annotations (3 responses for every test sample). The Fleiss’  $\kappa$  score for this protocol was 0.471, signifying “moderate agreement” between the judges. As before, Wilcoxon’s Signed Rank significance test was used to compare each model with the baseline (vanilla BS).

Next, we evaluate the performance of 3 affective strategies individually, namely affective word embeddings as input, affective loss functions, and affectively diverse decoding. We then show how these strategies can be integrated.

**Experiment #1: Affective word embeddings as input.** Table 1 compares Seq2Seq open-domain dialogue systems with and without the affective word embeddings. We see that the cognitively engineered affective embeddings, even with 3 additional features, improve the baseline Seq2Seq model significantly (as indicated by the up-arrows). The improvement is consistent in terms of all 3 evaluation aspects, and its effect is the most significant in emotional appropriateness. The results show that traditional word embeddings learned end-to-end during training are not sufficient to capture emotional aspects; using additional knowledge makes the model more aware of affect.

**Experiment #2: Affective loss functions.** We compare in Table 2 the proposed loss functions—namely minimizing affective dissonance (DMIN), maximizing affective dissonance (DMAX), and maximizing affective content (AC)—with traditional cross-entropy loss (XENT).

**Table 5.** Examples of the responses generated by the baseline and affective models.

Message	Model	Response
I wish I was a billionaire	Traditional Seq2Seq	Um, you know...
	$L_{DMIN}$	But you work for the casino?
	$L_{DMAX}$	You are personally responsible
	$L_{AC}$	It’s suicide!
	$L_{XENT} + \text{Aff. emb.}$	Don’t trouble yourself about it
	$L_{AC} + \text{Aff. emb.}$	Would it matter in the end
	$L_{AC} + \text{Aff. emb.} + \text{SL-ADBS}$	Would it change you?
I hate you	Traditional Seq2Seq	It was supposed to be
	$L_{DMIN}$	Aww c all guys do...
	$L_{DMAX}$	Keeps fairies away!
	$L_{AC}$	I haven’t done anything crazy
	$L_{XENT} + \text{Aff. emb.}$	Don’t say that!
	$L_{AC} + \text{Aff. emb.}$	I still love you!
	$L_{AC} + \text{Aff. emb.} + \text{SL-ADBS}$	I don’t want to fight you

As shown in Table 2, DMIN and DMAX yield similar results, both outperforming XENT. Moreover, AC generally outperforms DMIN and DMAX in terms of naturalness and appropriateness. The results imply that forcing the affect vector in either direction (towards or against the previous utterance) helps the model, but its performance is worse than AC. The mediocre performance of  $L_{\text{DMIN}}$  and  $L_{\text{DMAX}}$  could be explained by the fact that the relationship between a prompt and a response is not always as simple as minimum or maximum affective dissonance. It is usually much more subtle; therefore it makes more sense to model this relationship through established sociological interaction theories like the Affect Control Theory [8]. By contrast, the AC loss function encourages affective content without specifying the affect direction; it works well in practice and significantly out-performs the baseline XENT loss on all three axes.

Considering both Tables 1 and 2, we further notice that the affective loss function alone is not as effective as affective embeddings. This makes sense because the loss function does not explicitly provide additional knowledge to the neural network, but word embeddings do. However, as will be seen in Experiment #4, these affective aspects can be directly combined. Another interesting observation is the improved syntactic coherence of the affect-based models; we hypothesize that these models replace grammatically incorrect words with affectively suitable options that turn out to be more grammatically sound.

**Experiment #3: Affectively Diverse Decoding.** We now evaluate our affectively diverse decoding methods. Since evaluating diversity requires multiple decoded utterances for a test sample, we adopted a different evaluation setting as described before. Table 3 compares both word-level and sentence-level affectively diverse BS (WL-ADBS and SL-ADBS, respectively) with the original BS and Hamming-based DBS used in [22]. We see that WL-ADBS and SL-ADBS beat the baselines BS and Hamming-based DBS by a statistically significant margin on affective diversity as well as number of emotionally appropriate responses. SL-ADBS is slightly better than WL-ADBS as expected, since it takes into account the cumulative affect of sentences as opposed to individual words.

**Experiment #4: Putting them all together.** We show in Table 4 how the affective word embeddings, loss functions, and decoding methods perform when they are combined. Here, we chose the best variants in the previous individual tests: the loss function maximizing affective content ( $L_{\text{AC}}$ ) and the sentence level diversity measure (SL-ADBS). In this table, the statistical significance arrows denote the comparison of each row with the previous row, rather than with the baseline. As shown, the performance of our model generally increases when we gradually add new components to it, though some of the incremental improvements are statistically insignificant.

Note that our setting is different from ECM [26], the only other known emotion-based neural dialogue system to the best of our knowledge. ECM requires a desired affect category as input, which is unrealistic in applications. It also differs from our experimental setting (and our research goal), making direct comparison infeasible. However, our proposed affective approaches can be potentially integrated to ECM.

Finally, we present several sample outputs of all models in Table 5 to give readers a taste of how the responses differ.

## 6 Conclusion

In this work, we advance the development of affectively cognizant neural encoder-decoder dialogue systems by three affective strategies. We embed linguistic concepts in an affective space with a cognitively engineered dictionary, propose several affect-based heuristic objective functions, and introduce affectively diverse decoding methods. In information retrieval tasks such as question-answering and dialogue systems, these techniques can help retain the users by interacting with them in a more empathetic and human way.

## References

1. Asghar, N., Poupart, P., Jiang, X., Li, H.: Deep active learning for dialogue generation. In: Proceedings of Joint Conference on Lexical and Computational Semantics, pp. 78–83 (2017)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR (2015)
3. Callejas, Z., Griol, D., López-Cózar, R.: Predicting user mental states in spoken dialogue systems. *EURASIP J. Adv. Signal Process.* **2011**(1), 6 (2011)
4. Danescu-Niculescu-Mizil, C., Lee, L.: Chameleons in imagined conversations: a new approach to understanding coordination of linguistic style in dialogs. In: Proceedings Workshop on Cognitive Modeling and Computational Linguistics, pp. 76–87 (2011)
5. Fleiss, J.L.: Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**(5), 378–382 (1971)
6. Ghosh, S., Chollet, M., Laksana, E., Morency, L.P., Scherer, S.: Affect-LM: a neural language model for customizable affective text generation. In: ACL (2017)
7. Gimpel, K., Batra, D., Dyer, C., Shakhnarovich, G., Tech, V.: A systematic exploration of diversity in machine translation. In: EMNLP, pp. 1100–1111 (2013)
8. Hoey, J., Schröder, T., Alhothali, A.: Affect control processes: intelligent affective interaction using a partially observable markov decision process. *Artif. Intell.* **230**, 134–172 (2016)
9. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
10. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. In: NAACL-HLT, pp. 110–119 (2016)
11. Li, J., Monroe, W., Ritter, A., Jurafsky, D.: Deep reinforcement learning for dialogue generation. In: EMNLP, pp. 1192–1202 (2016)
12. Liu, C.W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., Pineau, J.: How not to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. In: EMNLP, pp. 2122–2132 (2016)
13. Malhotra, A., Yu, L., Schröder, T., Hoey, J.: An exploratory study into the use of an emotionally aware cognitive assistant. In: AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments (2015)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)

15. Mou, L., Song, Y., Yan, R., Li, G., Zhang, L., Jin, Z.: Sequence to backward and forward sequences: a content-introducing approach to generative short-text conversation. In: COLING, pp. 3349–3358 (2016)
16. Osgood, C.E.: The nature and measurement of meaning. *Psychol. Bull.* **49**(3), 197–237 (1952)
17. Pennebaker, J.W., Francis, M.E., Booth, R.J.: *Linguistic Inquiry and Word Count*. Erlbaum Publishers, Mahwah (2001)
18. Pittermann, J., Pittermann, A., Minker, W.: Emotion recognition and adaptation in spoken dialogue systems. *Int. J. Speech Technol.* **13**(1), 49–60 (2010)
19. Robison, J., McQuiggan, S., Lester, J.: Evaluating the consequences of affective feedback in intelligent tutoring systems. In: *Proceedings of International Conference on Affective Computing and Intelligent Interaction*, pp. 1–6 (2009)
20. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation. In: *ACL*, pp. 1577–1586 (2015)
21. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: *NIPS*, pp. 3104–3112 (2014)
22. Vijayakumar, A.K., Cogswell, M., Selvaraju, R.R., Sun, Q., Lee, S., Crandall, D., Batra, D.: Diverse beam search: decoding diverse solutions from neural sequence models. arXiv preprint [arXiv:1610.02424](https://arxiv.org/abs/1610.02424) (2016)
23. Warriner, A.B., Kuperman, V., Brysbaert, M.: Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **45**(4), 1191–1207 (2013)
24. Wen, T.H., Gasic, M., Mrkšić, N., Su, P.H., Vandyke, D., Young, S.: Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In: *EMNLP*, pp. 1711–1721 (2015)
25. Wilcoxon, F.: Individual comparisons by ranking methods. *Biom. Bull.* **1**(6), 80–83 (1945)
26. Zhou, H., Huang, M., Zhang, T., Zhu, X., Liu, B.: Emotional chatting machine: emotional conversation generation with internal and external memory. arXiv preprint [arXiv:1704.01074](https://arxiv.org/abs/1704.01074) (2017)