Affective Neural Response Generation

Nabiha Asghar[†], Pascal Poupart[†], Jesse Hoey[†], Xin Jiang[‡], Lili Mou[†]
[†]Cheriton School of Computer Science, University of Waterloo, Canada
{nasghar,ppoupart,jhoey}@uwaterloo.ca, doublepower.mou@gmail.com
[‡]Noah's Ark Lab, Huawei Technologies, Hong Kong
jiang.xin@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 emotionally rich responses that are more interesting and natural.

Introduction

Human-computer dialogue systems have wide applications ranging from restaurant booking (Wen et al. 2015) to emotional virtual agents (Malhotra et al. 2015). Inspired by the recent success of deep neural networks in natural language processing (NLP) tasks such as language modeling (Mikolov et al. 2010), machine translation (Sutskever, Vinyals, and Le 2014), and text summarization (Rush, Chopra, and Weston 2015), the artificial intelligence (AI) community is aggressively exploring the paradigm of neural dialogue generation.

In a neural network-based dialogue system, discrete words are mapped to real-valued vectors, known as *embeddings*, capturing abstract meanings of words (Mikolov et al. 2013); 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. Latest advances in this direction have demonstrated its efficacy for both task-oriented dialogue systems (Wen et al. 2015) and open-domain response generation (Shang, Lu, and Li 2015; Li et al. 2016b; Serban et al. 2017).

While most of the existing neural conversation models generate syntactically well-formed responses, they are prone to being off-context, short, dull, or vague. Latest efforts to address these issues include diverse decoding (Li, Monroe, and Jurafsky 2016; Vijayakumar et al. 2016), diversity-promoting objective functions (Li et al. 2016a), adversarial learning (Li et al. 2017), latent variable modeling for diversity (Cao and Clark 2017), human-in-the-loop reinforcement learning (Li et al. 2016b), online active learning (Asghar et al. 2017), latent intention modeling (Wen et al. 2017), and content-introducing approaches (Mou et al. 2016; Xing et al. 2017). These advancements are promising, but we are still far from our goal of building autonomous neural agents that can consistently carry out interesting human-like conversations.

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, phrases, or emotions. For instance, the attention mechanism in a sequence-to-sequence (Seq2Seq) model can learn syntactic alignment of words within the generated sequences (Bahdanau, Cho, and Bengio 2015). Similarly, neural word embedding models like Word2Vec learn word vectors by context, and can preserve low-level word semantics (e.g., "king"−"male"≈"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 using a cognitively engineered word-level affective dictionary (Warriner, Kuperman, and Brysbaert 2013), where affectively similar constructs are close to one other. In this way, the ensuing neural model is aware of words' emotional features.
2. We propose to augment the standard cross-entropy loss with affective objectives, so that our neural models are explicitly 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.

Related Work

Affectively cognizant virtual agents are generating interest both in the academia (Malhotra et al. 2015) and the industry,[1] due to their ability to provide emotional companionship to humans. Endowing text-based dialogue generation systems with emotions is also an active area of research. 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 (Pittermann, Pittermann, and Minker 2010; Callejas, Griol, and López-Cózar 2011).

Despite these, our work is mostly related to two very recent studies:

- Affect Language Model (Ghosh et al. 2017, Affect-LM) is an LSTM-RNN language model which leverages the Linguistic Inquiry and Word Count (Pennebaker, Francis, and Booth 2001, LIWC) text analysis program for affective feature extraction through keyword spotting. Affect-LM considers binary affective features, namely *positive emotion*, *angry*, *sad*, *anxious*, and *negative emotion*; at prediction time, Affect-LM generates sentences conditioned on the input affect features and a learned parameter of affect strength.
  Our work differs from Affect-LM in that we consider affective dialogue systems instead of merely language models, and we have explored more affective aspects including training and decoding.
- Emotional Chatting Machine (Zhou et al. 2017, ECM) is a Seq2Seq model. It takes as input a prompt and the desired emotion of the response, and then produces a response. It has eight emotion categories, namely *anger*, *disgust*, *fear*, *happiness*, *like*, *sadness*, *surprise*, and *other*. Additionally, ECM contains an internal memory and an external memory. The internal memory models the change of the internal emotion state of the decoder, and therefore encodes how much an emotion has already been expressed. The external memory decides whether to choose an emotional or generic (non-emotional) word at a given step during decoding.
  Our approach does not require the input of desired emotion as in Zhou et al. (2017), 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 affective embeddings.

Background

Word Embeddings

In NLP, word embeddings map words (or tokens) to real-valued vectors of fixed dimensionality. In recent years, neural network-based embedding learning has gained tremendous popularity, e.g., Word2Vec (Mikolov et al. 2013). They have been shown to boost the accuracy of computational models on various NLP tasks.

Typically, word embeddings 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"). Therefore, we enhance traditional embeddings with a cognitively engineered dictionary, explicitly representing word affective features from several perspectives.

[1] https://www.ald.softbankrobotics.com/en/robots/pepper

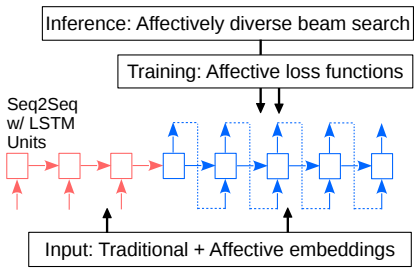


Figure 1: Overview of our approach, which is built upon a traditional Seq2Seq model. We propose three affective strategies for the input, training, and inference of Seq2Seq based on a cognitively engineered dictionary with Valence, Arousal, and Dominance (VAD) scores.

Seq2Seq Model

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 (Sutskever, Vinyals, and Le 2014). It consists of an encoder and a decoder, both of which are RNNs (typically with LSTM units). In the context of NLP, 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 $X = x_1, \cdots, x_m$ and $Y = y_1, \cdots, y_n$ are sequences of words/tokens, Seq2Seq models are typically trained with cross entropy loss (XENT):

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

where θ denotes model parameters. At prediction time, the model generates a response Y to a prompt X by computing

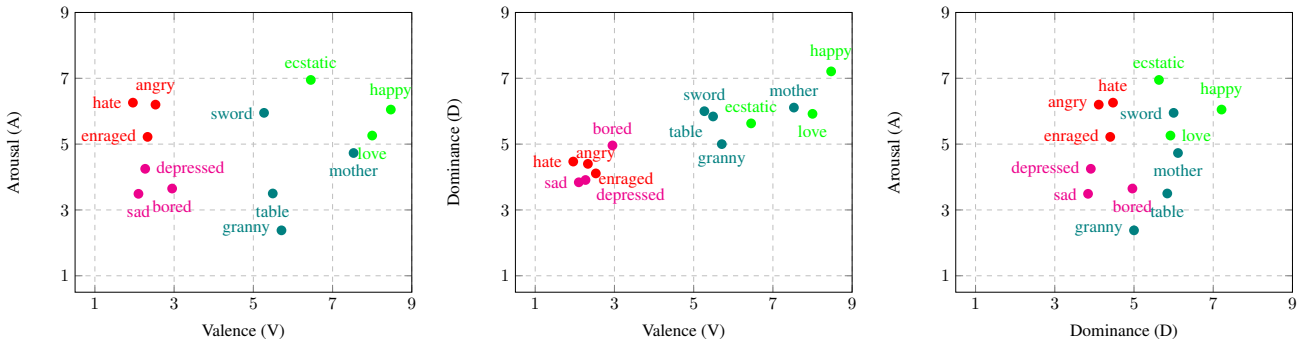
$$\arg\max_Y \{\log p(Y|X)\} \tag{2}$$

either greedily or by variants of beam search.

The Proposed Affective Approaches

In this section, we propose affective neural dialogue generation, which augments traditional neural conversation models with emotional cognizance.

Figure 1 delineates an overall picture of our approach. We leverage a cognitively engineered dictionary, based on which we propose three strategies for affective dialogue generation,



(a) V-A plot. (b) V-D plot. (c) D-A plot.

Figure 2: The relationship between several examples of adjectives, nouns, and verbs on the 3-dimensional VAD scale.

namely affective word embeddings as input, affective training objectives, and affectively diverse beam search. As will be shown in the experiments, these affective strategies can be combined to further improve Seq2Seq dialogue systems.

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 (Warriner, Kuperman, and Brysbaert 2013).[2]

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, or the degree of arousal evoked, by a stimulus), and Dominance (D, the degree of power/control exerted by a stimulus). Sociologists hypothesize that the VAD space (also known as the EPA[3] space) structures the semantic relations of linguistic concepts across languages and cultures; it captures almost 70% of the variance in affective meanings of concepts (Osgood 1952). VAD ratings have been previously used in sentiment analysis and empathetic tutors, among other affective computing applications (Robison, McQuiggan, and Lester 2009; Alhothali and Hoey 2015). To the best of our knowledge, we are the first to introduce VAD to dialogue generation.

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 (e.g., *pedophile*), neutral (e.g., *tablecloth*) and very positive (e.g., *happiness*), respectively. This axis is traditionally used on its own in most sentiment analysis techniques. Similarly, A $\simeq 1, 5$ and 9 corresponds to a word having very low (e.g., *dull*), moderate (e.g., *watchdog*), and very high (e.g., *insanity*) emotional intensity, respectively. Finally, D $\simeq 1, 5$ and 9 corresponds to a word/stimulus that is very powerless (e.g., *dementia*), neutral (e.g., *waterfall*) and very powerful (e.g., *paradise*), respectively. The VAD ratings of each word were collected through a survey in Warriner, Kuperman, and Brysbaert (2013) 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 Figure 2. For instance, the VAD vectors of the words *ecstatic* and *bored* are $[6.45, 6.95, 5.63]$ and $[2.95, 3.65, 4.96]$, respectively. This means that an average human rates the feeling of being *bored* as more unpleasant (V), less intense (A), and slightly weaker (D), compared with the feeling of being *ecstatic*. Similarly, the VAD vectors for the nouns *mother* and *granny* are $[7.53, 4.73, 6.11]$ and $[5.71, 2.38, 5.00]$, respectively. Thus, mothers are perceived to be more pleasant (V) and more powerful (D) than grannies, and evoke more intense emotions (A). From Figure 2a, we also see some clusters $\{$*angry, hate, enraged*$\}$ and $\{$*depressed, sad, bored*$\}$ that are slightly apart on the A axis. Also, the cluster $\{$*sword, table, granny*$\}$ is fairly neutral on the V axis, compared with the cluster $\{$*happy, mother, love*$\}$ in Figure 2b.

For words missing in this dictionary, such as stop words and proper nouns, we set the VAD vector to be the neutral vector $\vec{\eta} = [5, 1, 5]$, 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:

$$\texttt{W2AV}(w) = \begin{cases} \text{VAD}(l(w)), & \text{if } l(w) \in dict \\ \vec{\eta} = [5, 1, 5], & \text{otherwise} \end{cases} \quad (3)$$

where $l(w)$ is the lemmatization of the word w.

In this way, words with similar emotional connotations 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.

Affective Loss Functions

Equipped with affective vectors, we further design affective training loss functions to explicitly train an affect-aware

[2] Available at `http://crr.ugent.be/archives/1003`

[3] EPA refers to the affective dimensions *Evaluation*, *Potency*, and *Activity*, which are congruent to *Valence*, *Arousal*, and *Dominance*, respectively.

Seq2Seq conversation model. The philosophy of manipulating loss function is similar to Li et al. (2016a), 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 the time step i, the loss is computed by

$$L^i_{\texttt{DMIN}}(\theta) = -(1-\lambda)\log p(y_i|y_1,\cdots,y_{i-1},X) + \lambda\,\hat{p}(y_i)\left\|\sum_{j=1}^{|X|}\frac{\texttt{W2AV}(x_j)}{|X|} - \sum_{k=1}^{i}\frac{\texttt{W2AV}(y_k)}{i}\right\|_2 \quad (4)$$

where $\|\cdot\|_2$ denotes ℓ_2-norm. The first term is the standard XENT loss as in Equation 1. The sum $\sum_j \frac{\texttt{W2AV}(x_j)}{|X|}$ is the average affect vector of the source sentence, whereas $\sum_k \frac{\texttt{W2AV}(y_k)}{i}$ is the average affect vector of the target subsentence 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 $\hat{p}(y_i)$. λ 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, from the perspective of training an open-domain conversational agent whose main purpose is to entertain, it is interesting to generate responses that are *not* too affectively aligned with the prompts. Therefore, we design an objective function $L_{\texttt{DMAX}}$ that *maximizes* the dissonance by flipping the sign in the second term in Equation 4. (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 vague, dull, or emotionally neutral, while the model could choose the appropriate sentiment. 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

$$L^i_{\texttt{AC}}(\theta) = -(1-\lambda)\log p(y_i|y_1,\cdots,y_{i-1},X) - \lambda\,\hat{p}(y_i)\left\|\texttt{W2AV}(y_i) - \vec{\eta}\right\|_2 \quad (5)$$

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

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 widely 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.*" "*This is great!*" and "*This is so great!*" that lack semantic and/or syntactic diversity (Gimpel et al. 2013).

Diverse beam search (DBS) (Vijayakumar et al. 2016) 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.

In the following, we give an overview of BS and DBS, followed by a description of our proposed affective variants of DBS.

Beam Search (BS). 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 and let X be the input sequence. Ideally, decoding of an entire sequence $\mathbf{y}^*$ is given by

$$\mathbf{y}^* = y_1^*,\cdots,y_T^* = \arg\max_{y_1,\cdots,y_T}\left[\sum_{t\in T}\log p(y_t|y_{t-1},\cdots,y_1,X)\right] \quad (6)$$

where T is the length. BS approximates Equation 6 by computing and storing only the top-B high scoring (sub)sequences (called beams) at each time step. Let $\mathbf{y}_{i,[t-1]}$ denote the ith beam stored at time $t-1$, and let $Y_{[t-1]} = \{\mathbf{y}_{1,[t-1]},\cdots,\mathbf{y}_{B,[t-1]}\}$ denote the set of beams stored by BS at time $t-1$. Then at time t, the BS objective is given by:

$$Y_{[t]} = y^{1^*}_{1..t},\cdots,y^{B^*}_{1..t} = \arg\max_{\substack{\mathbf{y}_{1,[t]},\cdots,\mathbf{y}_{B,[t]}\\ \in Y_{[t-1]}\times V}}\sum_{b=1}^{B}\sum_{i=1}^{t}\log p(y_{b,i}|\mathbf{y}_{b,[i-1]},X) \quad (7)$$

subject to $\mathbf{y}_{i,[t]} \neq \mathbf{y}_{j,[t]}$

where $Y_{[t-1]}\times V$ is the set of all possible extensions (i.e., V) based on the beams stored at time $t-1$ (i.e., $Y_{[t-1]}$).

Diverse Beam Search (DBS). DBS aims to overcome the diversity problem in BS by incorporating diversity among candidate outputs. DBS divides the top-B beams into G groups (each group containing $B' = G/B$ beams) and incorporates diversity between these groups by maximizing the standard likelihood term as well as a dissimilarity metric among the groups.

Concretely, DBS adds to traditional BS (Equation 7) a dissimilarity term $\Delta(Y^1_{[t]}, \cdots, Y^{g-1}_{[t]})[y_t]$ which measures the dissimilarity between group g and previous groups $1, \cdots, g-1$ if token y_t is selected to extend any beam in group g. This is given by

$$Y^g_{[t]} = \underset{\substack{\mathbf{y}^g_{1,[t]}, \cdots, \mathbf{y}^g_{B',[t]} \\ \in Y^g_{[t-1]} \times V}}{\arg\max} \Bigg[\sum_{b=1}^{B'} \sum_{i=1}^{t} \log p(y^g_{b,i} | \mathbf{y}^g_{b,[i-1]}, X) + \lambda_g \Delta(Y^1_{[t]}, \cdots, Y^{g-1}_{[t]})[y^g_{b,t}] \Bigg] \quad (8)$$

$$\text{subject to} \qquad \mathbf{y}^g_{i,[t]} \neq \mathbf{y}^g_{j,[t]}$$

where $\lambda_g \geq 0$ is a hyperparameter controlling the diversity strength. Intuitively, DBS modifies the probability in BS as a general scoring function by adding a dissimilar term between a particular sample (i.e., $y^g_{b,1} \cdots y^g_{b,t}$) and samples in other groups (i.e., $Y^1_{[t]}, \cdots, Y^{g-1}_{[t]}$). We refer readers to Vijayakumar et al. (2016) 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 Vijayakumar et al. (2016): 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 Δ_W to be

$$\Delta_W(Y^1_{[t]}, \cdots, Y^{g-1}_{[t]})[y^g_{b,t}] = -\sum_{j=1}^{g-1} \sum_{c=1}^{B'} \text{sim}\left(\texttt{W2AV}(y^g_{b,t}), \texttt{W2AV}(y^j_{c,t})\right) \quad (9)$$

where $\text{sim}(\cdot)$ denotes a similarity measure between two vectors. In our experiments, we use the cosine similarity function. $y^g_{b,t}$ denotes the token under consideration at the current time step t for beam b in group g, and $y^j_{c,t}$ 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 current beam b with all the beams generated in groups $1, \cdots, g-1$. The metric operates at the word level, ensuring that the word affect at time t is diversified across the groups.

• *Sentence-Level Diversity for ADBS (SL-ADBS).* The word-level metric Δ_W in Equation 9 does not take into account the overall sentence affect for each group up to time t. Thus, as an alternative, we propose a sentence-level affect diversity metric, given by

$$\Delta_S(Y^1_{[t]}, \cdots, Y^{g-1}_{[t]})[y^g_{b,t}] = -\sum_{j=1}^{g-1} \sum_{c=1}^{B'} \text{sim}\left(\Psi(\mathbf{y}^g_{b,[t]}), \Psi(\mathbf{y}^j_{c,[t]})\right) \quad (10)$$

$$\text{where} \qquad \Psi(\mathbf{y}^k_{i,[t]}) = \sum_{w \in \mathbf{y}^k_{i,[t]}} \texttt{W2AV}(w) \quad (11)$$

Here, $\mathbf{y}^k_{i,[t]}$ for $k \leq g$ is the ith beam in the kth group stored at time t; $\mathbf{y}^g_{b,[t]}$ is the concatenation of $\mathbf{y}^g_{b,[t-1]}$ and $y^g_{b,t}$.

The intuition is that, to capture sentence level affect, 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 in the experimental results.

It should be also noticed that several other beam search-based diverse decoding techniques have been proposed in recent years, including DivMBest (Gimpel et al. 2013), MMI objective (Li et al. 2016a), and segment-by-segment reranking (Shao et al. 2017). All of them use the notion of a *diversity term* within BS; therefore our affect-injecting technique can be potentially used with these algorithms.

Experiments

Dataset and Implementation Details

We evaluated our approach on the Cornell Movie Dialogs Corpus[4] (Danescu-Niculescu-Mizil and Lee 2011). It contains 300k utterance-response pairs, and we kept a vocabulary size of 12,000. All our model variants used a single-layer LSTM encoder and a single-layer LSTM decoder, each layer containing 1024 cells. We used Adam (Kingma and Ba 2015) for optimization with a batch size of 64 and other default hyperparameters. Listed as follows are detailed settings for each model.

- For the baseline L_{XENT} loss, we used 1024-dimensional Word2Vec embeddings as input and trained the Seq2Seq model for 50 epochs by using Equation 1.
- For the affective embeddings as input, we used 1027-dimensional vectors, each a concatenation of 1024-D Word2Vec and 3-D W2AV embeddings. Training was also done for 50 epochs.

[4]The dataset is available at https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html

Model	Syntactic Coherence	Natural	Emotional Approp.
Word embed.	1.48	0.69	0.41
Word + Affective embeddings	**1.71**	**1.05**	**1.01**

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

Model	Syntactic Coherence	Natural	Emotional Approp.
L_{XENT}	1.48	0.69	0.41
L_{DMIN}	**1.75**	0.83	0.56
L_{DMAX}	1.74	0.85	0.58
L_{AC}	1.71	**0.95**	**0.71**

Table 2: Effect of affective loss functions.

- For affective loss functions (L_{AC}, L_{DMIN}, and L_{DMAX}), we trained the models in two phases. In the first phase, each model was trained using L_{XENT} loss for 40 epochs. In the second phase, each model was fine-tuned using the affective loss functions for 10 epochs. This two-phase approach was adopted because we observed inferior results (syntactical errors in particular) for single-phase training with the full loss functions for 50 epochs.
- The ADBS decoding was deployed at test time (both word-level and sentence-level metrics, Δ_W and Δ_S in Equations 9 and 10, respectively). We set $G = B$ for simplicity, that is, each group contains a single beam. Therefore, diversification among groups in our case was equivalent to diversification among all the beams.
- The λ hyperparameters for L_{DMIN}, L_{DMAX}, and L_{AC} were manually tuned through validation and set to 0.5, 0.4, and 0.5, respectively. For affectively diverse BS, λ was set to 0.7 (Equation 8).

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, Liu et al. (2016) show that these metrics have weak or no correlation with human judgments. It is also unclear how to evaluate affective aspects by automated metrics. Therefore, in this work, we recruited human judges to evaluate our models, following several previous studies (Shang, Lu, and Li 2015; Mou et al. 2016).

To evaluate the quality of the generated responses, we had 5 workers to evaluate 100 test samples for each model variant in terms of *syntactic coherence* (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. We also evaluated inter-annotator consistency by Fleiss' κ score (1971), and obtained a κ score of 0.447, interpreted as "moderate agreement" among the judges.[5]

Model	Syntactic Diversity	Affective Diversity	No. of Emotionally Approp. Responses
BS	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 3: Effect of affectively diverse decoding. H-DBS refers to Hamming-based DBS used in Vijayakumar et al. (2016). WL-ADBS and SL-ADBS are the proposed word-level and sentence-level affectively diverse beam search, respectively.

The evaluation of diversity was conducted separately (i.e., the results in 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 number of beams that were emotionally appropriate to the prompt. The scores obtained for each question were averaged. Moreover, we had three annotators in this experiment (fewer than the previous one), as it required more annotations (3 responses for every test sample). The Fleiss' κ score for this protocol was 0.471, also signifying "moderate agreement" between the judges.

In the following, we first evaluate the performance of three affective strategies individually, namely affective word embeddings as input, affective loss functions, and affectively diverse decoding. Then we 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 three additional features, largely improve the Seq2Seq model. The improvement is consistent in terms of all three 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, and that 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).

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

[5] https://en.wikipedia.org/wiki/Fleiss%27_kappa

Model	Syntactic Coherence	Naturalness	Emotional Appropriateness
Traditional Seq2Seq	1.48	0.69	0.41
Seq2Seq + Affective Embeddings	1.71	1.05	1.01
Seq2Seq + Affective Embeddings & Loss	**1.76**	1.03	1.07
Seq2Seq + Affective Embeddings & Loss & Decoding	1.69	**1.09**	**1.10**

Table 4: Combining different affective strategies.

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}+ Aff. Emb.	Don't trouble yourself about it.
	L_{AC}+ Aff. Emb.	Would it matter in the end.
	L_{AC}+ Aff. Emb. + 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}+ Aff. Emb.	Don't say that!
	L_{AC}+ Aff. Emb.	I still love you!
	L_{AC}+ Aff. Emb. + SL-ADBS	I don't want to fight you.

Table 5: Examples of the responses generated by the baseline (traditional Seq2Seq) and affective models.

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_{DMIN} and L_{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 (Hoey, Schröder, and Alhothali 2016). By contrast, the AC loss function encourages affective content without specifying the affect direction; it works well in practice.

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 itself does not provide additional knowledge to the neural network. However, as will be seen in Experiment #4, these affective aspects are "orthogonal" to each other, and can be directly combined.

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 Vijayakumar et al. (2016). We see that WL-ADBS and SL-ADBS beat the baselines BS and Hamming-based DBS by a fair 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_{AC}) and the sentence level diversity measure (SL-ADBS).

As shown, the performance of our model generally increases when we gradually add new components to it. This confirms that the three affective strategies can be directly combined for further improvement, and we achieve significantly better performance compared with the original Seq2Seq model, especially in terms of emotional appropriateness.

Note that our setting is different from the Emotional Chat Machine (ECM) (Zhou et al. 2017), 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 in addition to their manually specified emotion category.

Case study. We present several sample outputs of all the models in Table 5 to give readers a taste of how the responses differ. L_{XENT} responses are generic and non-committal, as expected. L_{DMIN} tries to match the affect of the word *billionaire* with *casino*, L_{DMAX} responds to *hate* with *fairies*, L_{AC} maximizes affective content of the responses with the words *suicide* and *crazy*. L_{XENT} with affective embeddings produces responses with more subtle affective connotations.

Conclusion and Future Work

In this work, we address the problem of affective neural dialogue generation, which is useful in applications like emotional conversation partners to humans. 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 the future, we would like to investigate affect-based attention mechanisms for neural conversational models. We would also like to explore affect-based personalization of neural dialogue systems using reinforcement learning.

Acknowledgments

We thank Marc-André Cournoyer for his helpful Github repositories on neural conversation models, and the annotators who helped us evaluate our models.

References

[Alhothali and Hoey 2015] Alhothali, A., and Hoey, J. 2015. Good news or bad news: Using affect control theory to analyze readers' reaction towards news articles. In *NAACL-HLT*, 1548–1558.

[Asghar et al. 2017] Asghar, N.; Poupart, P.; Jiang, X.; and Li, H. 2017. Deep active learning for dialogue generation. In *Proc. Conf. Lexical and Computational Semantics*, 78–83.

[Bahdanau, Cho, and Bengio 2015] Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

[Callejas, Griol, and López-Cózar 2011] Callejas, Z.; Griol, D.; and López-Cózar, R. 2011. Predicting user mental states in spoken dialogue systems. *EURASIP J. Advances in Signal Processing* 2011(1):6.

[Cao and Clark 2017] Cao, K., and Clark, S. 2017. Latent variable dialogue models and their diversity. In *EACL*, volume 2, 182–187.

[Danescu-Niculescu-Mizil and Lee 2011] Danescu-Niculescu-Mizil, C., and Lee, L. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proc. Workshop on Cognitive Modeling and Computational Linguistics*, 76–87.

[Fleiss 1971] Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382.

[Ghosh et al. 2017] Ghosh, S.; Chollet, M.; Laksana, E.; Morency, L.-P.; and Scherer, S. 2017. Affect-LM: A neural language model for customizable affective text generation. In *ACL*, 634–642.

[Gimpel et al. 2013] Gimpel, K.; Batra, D.; Dyer, C.; Shakhnarovich, G.; and Tech, V. 2013. A systematic exploration of diversity in machine translation. In *EMNLP*, 1100–1111.

[Hoey, Schröder, and Alhothali 2016] Hoey, J.; Schröder, T.; and Alhothali, A. 2016. Affect control processes: Intelligent affective interaction using a partially observable markov decision process. *Artificial Intelligence* 230:134–172.

[Kingma and Ba 2015] Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

[Li et al. 2016a] Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, 110–119.

[Li et al. 2016b] Li, J.; Monroe, W.; Ritter, A.; and Jurafsky, D. 2016b. Deep reinforcement learning for dialogue generation. In *EMNLP*, 1192–1202.

[Li et al. 2017] Li, J.; Monroe, W.; Shi, T.; Ritter, A.; and Jurafsky, D. 2017. Adversarial learning for neural dialogue generation. In *EMNLP*, 2147–2159.

[Li, Monroe, and Jurafsky 2016] Li, J.; Monroe, W.; and Jurafsky, D. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

[Liu et al. 2016] Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, 2122–2132.

[Malhotra et al. 2015] Malhotra, A.; Yu, L.; Schröder, T.; and Hoey, J. 2015. An exploratory study into the use of an emotionally aware cognitive assistant. In *AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments*.

[Mikolov et al. 2010] Mikolov, T.; Karafiát, M.; Burget, L.; Cernockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *INTERSPEECH*, 1045–1048.

[Mikolov et al. 2013] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, 3111–3119.

[Mou et al. 2016] Mou, L.; Song, Y.; Yan, R.; Li, G.; Zhang, L.; and Jin, Z. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *COLING*, 3349–3358.

[Osgood 1952] Osgood, C. E. 1952. The nature and measurement of meaning. *Psychological Bulletin* 49(3):197–237.

[Pennebaker, Francis, and Booth 2001] Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. *Linguistic Inquiry and Word Count*. Erlbaum Publishers.

[Pittermann, Pittermann, and Minker 2010] Pittermann, J.; Pittermann, A.; and Minker, W. 2010. Emotion recognition and adaptation in spoken dialogue systems. *Int. J. Speech Technology* 13(1):49–60.

[Robison, McQuiggan, and Lester 2009] Robison, J.; McQuiggan, S.; and Lester, J. 2009. Evaluating the consequences of affective feedback in intelligent tutoring systems. In *Proc. Int. Conf. Affective Comput. and Intell. Interaction and Workshops*, 1–6.

[Rush, Chopra, and Weston 2015] Rush, A. M.; Chopra, S.; and Weston, J. 2015. A neural attention model for abstractive sentence summarization. In *EMNLP*, 379–389.

[Serban et al. 2017] Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; and Bengio, Y. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, 3295–3301.

[Shang, Lu, and Li 2015] Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. In *ACL*, 1577–1586.

[Shao et al. 2017] Shao, L.; Gouws, S.; Britz, D.; Goldie, A.; Strope, B.; and Kurzwell, R. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. *arXiv preprint arXiv:1701.03185*.

[Sutskever, Vinyals, and Le 2014] Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*, 3104–3112.

[Vijayakumar et al. 2016] Vijayakumar, A. K.; Cogswell, M.; Selvaraju, R. R.; Sun, Q.; Lee, S.; Crandall, D.; and Batra, D. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

[Warriner, Kuperman, and Brysbaert 2013] Warriner, A. B.; Kuperman, V.; and Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45(4):1191–1207.

[Wen et al. 2015] Wen, T.-H.; Gasic, M.; Mrkšić, N.; Su, P.-H.; Vandyke, D.; and Young, S. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *EMNLP*, 1711–1721.

[Wen et al. 2017] Wen, T.-H.; Miao, Y.; Blunsom, P.; and Young, S. 2017. Latent intention dialogue models. In *ICML*, 3732–3741.

[Xing et al. 2017] Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W.-Y. 2017. Topic aware neural response generation. In *AAAI*, 3351–3357.

[Zhou et al. 2017] Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.