AffectON: Incorporating Affect Into Dialog Generation

Zana Buçinca, Yücel Yemez, Engin Erzin, Metin Sezgin

Abstract—Due to its expressivity, natural language is paramount for explicit and implicit affective state communication among humans. The same linguistic inquiry (e.g. *How are you?*) might induce responses with different affects depending on the affective state of the conversational partner(s) and the context of the conversation. Yet, most dialog systems do not consider affect as constitutive aspect of response generation. In this paper, we introduce *AffectON*, an approach for generating affective responses during inference. For generating language in a targeted affect, our approach leverages a probabilistic language model and an affective space. *AffectON* is language model agnostic, since it can work with probabilities generated by any language model (e.g., sequence-to-sequence models, neural language models, n-grams). Hence, it can be employed for both affective dialog and affective language generation. We experimented with affective dialog generation and evaluated the generated text objectively and subjectively. For the subjective part of the evaluation, we designed a custom user interface for rating and provided recommendations for the design of such interfaces. The results, both subjective and objective demonstrate that our approach is successful in pulling the generated language toward the targeted affect, with little sacrifice in syntactic coherence.

Index Terms—affective computing, affective dialog generation.

1 Introduction

CONSIDERED as the foundation of civilization, language is the most sophisticated and intuitive means of communication amongst humans. With their utterances, along with exchanging information and expressing ideas, people explicitly or implicitly indicate their affective states. Yet, we are surrounded by computers that for a long time could interpret commands and convey outputs solely in an "unnatural" language - code. This gulf between verbal human and computer communication has been vastly bridged by advances in natural language processing, resulting in emerging interfaces such as conversational agents and dialog systems. Nevertheless, as Picard [1] states, for genuinely natural and intelligent human-computer interaction, computers should have and express emotions.

Previous research shows that emotionally adaptive (vs. non-adaptive) conversational agents are rated more positively and perceived more helpful and trustworthy by users [2], [3], which in turn contributes to more efficient human-computer teams. Moreover, evidence suggests that affective conversational agents can be helpful in an array of applications. They assist students in learning efficiently [4], alleviate mental health conditions (e.g. depression [5], dealing with bullying [6]), show empathy when delivering emotional information [7], provide medical care humanely [8] and improve the overall emotional well-being [9]. In this context, while the existing systems are increasingly aspiring for more naturalistic and humane interfaces, they still lag behind on the incumbent ability of affect generation that prior research has shown to have positive impact on users.

Affect is defined as a set of observable manifestations of a subjectively experienced emotion [10]. Being an intrinsic aspect of humans, it is exhibited through visual, vocal and verbal cues. To illustrate, people reveal their joy by their happy facial and vocal expressions, but also through their rich-in-positive-words utterances. A plethora of studies have delved into synthesizing affect in computers (i.e. agents or robots) visually by facial expressions and body language [11], [12], [13], or vocally by speech features [14], [15], [16]. On the other hand, not as much effort has been spent on the equally vital task of affective language generation.

Following Munezero et al.'s definition [17], in this study, affective language generation is deemed different from sentimental or emotional language generation. Although in the literature, there persists a lack of lucid distinction among affect, emotion, feeling and sentiment terms, as they are frequently used interchangeably, affect is unanimously considered as the term that subsumes emotion, feeling and sentiment [18]. The three principal dimensions of affect consist *valence* (pleasant - unpleasant), *arousal* (active - passive), and *dominance* (dominant - submissive) [19]. Evidently, comprised of three dimensions, affect is a continuous multifaceted phenomenon, in contrast to sentiment which is gauged solely by polarity (*positive – negative*), or emotion which has discrete categories (e.g. *happy, sad, angry, afraid, disgust*).

Generating affective language adds further complexity to the already challenging task of natural language generation. Besides meeting the natural language requirements in terms of semantics, syntax and morphology [20], the text should also convey the targeted affect. If the generated language, in addition to these requirements, is also conditioned on another sentence, the task transforms into affective dialog generation. As a result, affective language generation can be regarded as a subproblem of affective dialog generation.

- *Koç University, Istanbul 34450, Turkey.*
E-mail: {zbucinca16, yyemez, eerzin, mtsezgin}@ku.edu.tr

Manuscript received October 21, 2019;

In this paper, we present *AffectON*, an approach for affective dialog generation, which can also be applied to the affective language generation setting. *AffectON* is a language model agnostic algorithm which generates natural language in a targeted affect during inference by leveraging a probabilistic language model and an affective space. Since affective dialog generation subsumes affective language generation, the focus of the paper is affective dialog generation. Though *AffectON* is compatible with any probabilistic language model, in this study we experiment with a sequence-to-sequence neural language model for affective dialog generation. We evaluate the generated responses objectively, in terms of *valence* and *perplexity*. For subjective evaluation, we design a custom user interface for rating the generated text in terms of *valence, arousal, dominance, syntax*, and *appropriateness*. The results indicate that our approach is successful in generating responses close to the targeted affect, while preserving the syntactic and semantic requirements.

Our contributions can be summarized as follows:

- We present an approach for affective dialog and language generation by leveraging a language model and an affective space. To the best of our knowledge, we introduce the first language model agnostic affective language generation approach, which can be employed by any underlying language model.
- We design a protocol and a user interface for subjective evaluation of affective text generation systems and provide design recommendations for such interfaces.

2 Related Work

The main lines of research that tackle affective language generation fall under the categories of textual style transfer and affective dialog systems. The relevant related work in these areas is discussed below.

2.1 Textual Style Transfer

Style transfer is regarded as the task of generating semantic content in a targeted style. Recent advances in deep learning have led to a surge of research on style transfer via deep neural nets. While the task has enticed chiefly the computer vision community, with multiple works achieving remarkable results in style transfer on images [21], [22], [23], far fewer studies exist in language style transfer. Akin to image style transfer, language style transfer rephrases a style-free content with the desired stylistic attributes. For instance, Jhamtani et. al [24] transfer phrases from standard English to Shakespearean English. They utilize parallel corpora and an encoder-decoder network for training.

Though Tikhonov and Yamshchikov argue that sentiment is not a stylistic attribute of language [25], many studies consider sentiment to be so, hence manipulate with it accordingly. For instance, Hu et. al [26] aim to generate controllable sentences by learning disentangled latent representations. They leverage variational auto-encoders, while constraining generation based on specified attributes, such as sentiment or tense. Shen et al. [27] also generate sentences in a targeted sentiment by separating semantic and stylistic content of the text. They train an encoder that extracts style from the source text, yielding a style-free latent representation, which is then fed to a decoder together with a targeted style. Wang et. al [28] mix multiple generators for sentimental text generation. Each generator is responsible for learning to generate sentences with a single sentiment, but they all go through one discriminator.

By modifying sequence-to-sequence architectures, Fu et al. [29] propose two models for style transfer. In the first one, they encode the source sentence into a latent representation, by removing the stylistic information. Afterwards, they feed the content to two separate decoders which endow the content with predefined styles (sentiments). The other model utilizes a single decoder, which on top of content is conditioned on style also. Recently, pre-trained large transformer based models have been used to condition the sentiment of the generated text [30]. Our work is similar to the studies discussed here as it also transfers the affect of the sentences. Nonetheless, we are interested in the whole affect of the sentence, not only the sentiment, which is confined to *positive* or *negative*. In contrast to these studies, we do not regard affect as a stylistic attribute of the text. In addition, our approach is post-hoc, as we do not update model weights when transferring the affect of the sentence.

2.2 Affective Dialog Systems

Research on affective dialog system can be broadly classified into rule-based and data-driven approaches. In rule-based approaches, hand-crafted rules are embedded into the dialog systems to make them more affective. These approaches are specific to the domain and cannot be employed in open domain settings. For instance, Mahamood et al. [7] generate affective text for systems providing information to parents of neonatal infants.

Advances in deep learning gave impetus to data-driven dialog modeling. Vinyals et al. [31] showed that relative success in data-driven conversation modeling could be achieved with end-to-end training of sequence-to-sequence frameworks. Nonetheless, these approaches entail shortcomings such as producing short, dull and emotionless answers. To obtain affectively richer responses, a couple of studies have incorporated emotion into dialog generation. Zhou et. al [32] introduce a modified sequence-to-sequence architecture that produces answers, conditioned on one of eight emotion categories (angry, disgust, fear, like, happy, sad, surprise, other). Their architecture is comprised of an external and internal memory. The external memory is responsible for deciding whether to choose an emotional or non-emotional word during decoding. Whereas, the internal memory unit controls the amount of the emotion expressed during the decoding. Another study [33] that conditions responses on emotion categories, experiments with the ways of concatenating emotion tokens with the latent representation of the source utterance. Although, the results did not show a huge difference among the ways of concatenation, they observed that for some emotion categories (i.e. *fear*) concatenating emotion token prior to the context vector yielded marginally better results. The aforementioned studies are relatively successful in generating emotionally charged responses. However, the treatment of emotions as discrete categories makes these approaches impractical for conversation modelling, since the conversation would have abrupt, and unnatural changes in terms of emotion. Meanwhile, following the work of [34], we regard affect as a continuous, multidimension phenomenon while generating responses. This in turn enables smooth transitions among affective states.

Asghar et al. [35] utilize a continuous affective space and sequence-to-sequence framework for affective response generation. To boost the performance, they utilize affective embeddings during encoding, augment the loss function with an affective objective and conduct affectively diverse beam search. Their approach does not require a target affect, instead the affective direction is guided by the loss function. They experimented with three distinct strategies for the loss function. The first strategy generates responses in the same affective direction with the source sentence. The second one generates responses in the opposite affective direction with the source sentence, and the final strategy tries to generate non-neutral responses irrespective of the direction. While these strategies might be useful in certain scenarios, we find them unnatural since humans do not have a pre-decided affective direction strategy when conversing. Instead, our responses and their affective direction develop naturally depended on multiple factors such as: the context of the conversation, our current mood and the relationship with the conversational partner. Thus, we argue that for naturalistic conversations the responses should be conditioned on affective information. Hence, in contrast to this study our approach generates responses conditioned on target affects. Moreover, we do not include affective information while training the seqeuence to sequence model, but utilize the learned weights with external affective information for response generation.

More recently and also closely related to our work, Colombo et al. [36] propose multiple approaches to generating emotional responses. They experiment with representation of emotional content as both distribution of probabilities on a vector of discrete emotions and a continuous affective state. Among their proposed approaches, Word Level Explicit model seeks to learn generation of affective sentences by adaptively sampling affective words. The main difference from our work is that we conduct the affective mapping during inference and not training. Because we change the affect during inference, our approach is robust and can be easily incorporated into any existing language model. Further, in terms of implementation their proposed model is a GRU-based model, whereas we use a transformer based model.

3 Background

Prior to elaborating on the proposed approach, we will briefly describe the three principal units our method employs.

3.1 Language Models

Language modeling is a core natural language processing task, applications of which span across multiple areas. A language model captures the distribution of a sequence of words in natural language. Given a sequence of words, it assigns a probability to the sequence. Essentially, we expect a language model to assign higher probabilities to sequences of words it has encountered in its training set. By exploiting the chain rule of probability, the problem of calculating the joint probability of a sequence N words $w_1, w_2, ..., w_N$ is formulated as:

$$P(w_1, w_2, ..., w_N) = \prod_{t=1}^{N} P(w_t | w_1, w_2, ..., w_{t-1}). \quad (1)$$

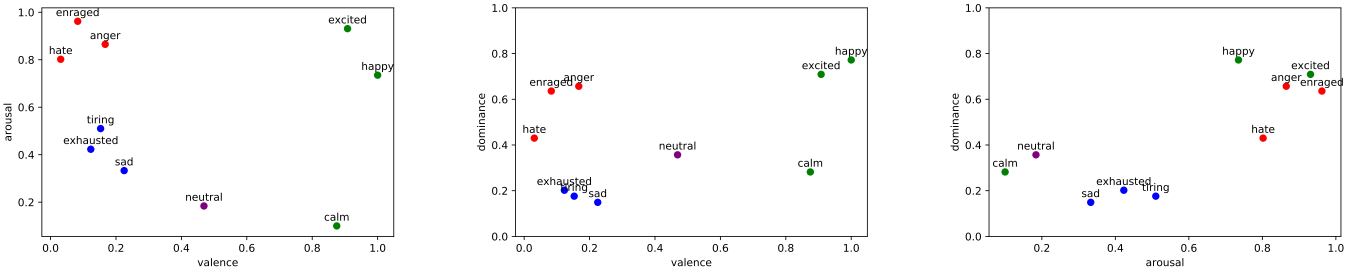
Common neural network architectures used for language modeling include the eminent Long Short-Term Memory (LSTM) architectures. LSTMs are a special type of Recurrent Neural Networks (RNNs) unfettered by the vanishing gradient problem typical RNNs suffer from [38]. Just like RNNs, LSTMs process one token at a time and can handle variable size sequences and preserve information from predecessor words, offering a robust structure for language modeling. Recently, transformer models have changed the landscape in natural language processing as they have outperformed prior models in language modeling and a vast number of other end tasks [39], [40]. In contrast to LSTMs, these architectures generally do not employ any recurrent networks. Instead, they use the attention-mechanism at each step to decide which parts of the input sequence are relevant.

3.1.1 Sequence-to-sequence Models

Sequence-to-sequence models are encoder-decoder neural networks with genesis from machine translation. Specifically, these architectures are utilized to encode a sentence into one language and decode it into another. Nonetheless, apart from translation sequence-to-sequence architectures have found applications in many multimodal and unimodal learning tasks [41]. Here, we are particularly interested in their usage in conversation modeling. The idea is to encode one (or more) sentences to an embedding and then to decode this embedding to the response for the encoded utterance. For conversation modeling, sequence-to-sequence architectures employ LSTMs or transformers for both the encoder and decoder. In this manner, they provide a medium for response generation of a given utterance. Sequence-to-sequence models build upon language models, as they are conditioned on the meaning of the encoded utterance, in addition to the predecessor words of the response currently being generated. Consequently, the problem sequence-to-sequence models tackle is finding the probability of a response $Y = y_1, y_2, ..., y_N$, given the utterance X:

$$P(Y|X) = \prod_{t=1}^{N} P(y_t | y_1, y_2, ..., y_{t-1}, X). \quad (2)$$

In this study we utilize a large transformer model fine-tuned on dialog corpus as a sequence-to-sequence model. Note that our approach can be implemented with any underlying language model. We chose to experiment with Generative Pre-Training (GPT) model [40] as one of the large and high performing transformer models. The pace by which new and superior language models are being released, highlights the importance of our approach being language model-agnostic and requiring no further training.



(a) Valence - Arousal

(b) Valence - Dominance

(c) Arousal - Dominance

Fig. 1. Sample words in affective space [37]
.

3.2 Affective Space - VAD

Language models capture contextual meaning of the words, but do not distinguish words according to the affect they elicit. For example, contrasting affective words, such as *good* and *bad*, or *terrible* and *wonderful* appear in similar contexts (e.g. ”the weather was terrible” and ”the weather was wonderful”) and are given similar probabilities under the language model. An affective space, where words similar in affect are proximate, would be valuable for analyzing and comparing words according to their underlying affect.

ANEW [34] is one of the earliest lexicons to consider affective meaning of words in three dimensions. In this study, humans rated 1034 words on a scale from 1 to 9 in terms of valence, arousal and dominance. This affective view accounts *valence* to range from unpleasant to pleasant, *arousal* to range from calm to excited and *dominance* from submissive to dominant. Following the same approach, Warriner et al. [42] provide an extended version of ANEW lexicon with 13,915 lemmas rated in valence, arousal and dominance (VAD) dimensions.

More recently, Mohammad [37] introduced *NRC Valence, Arousal, and Dominance (VAD) Lexicon*, which contains 20,007 VAD values for lemmas rated on a continuous scale from 0 (low) to 1 (high). These ratings were obtained by using *Best-Worst Scaling*, as an approach for overcoming the shortcomings of *Likert* scales (e.g. biases towards the middle of the scale) that were used for obtaining the ratings in prior work. We utilize this lexicon as an affective space. Figure 1 depicts sample words in VAD space.

4 Proposed Method: AffectON

In this section, we describe the proposed approach for text generation in a target affect. Specifically, we elaborate on affective response generation with a sequence-to-sequence model. However, our approach can be employed by any probabilistic language model.

4.1 Problem Formulation

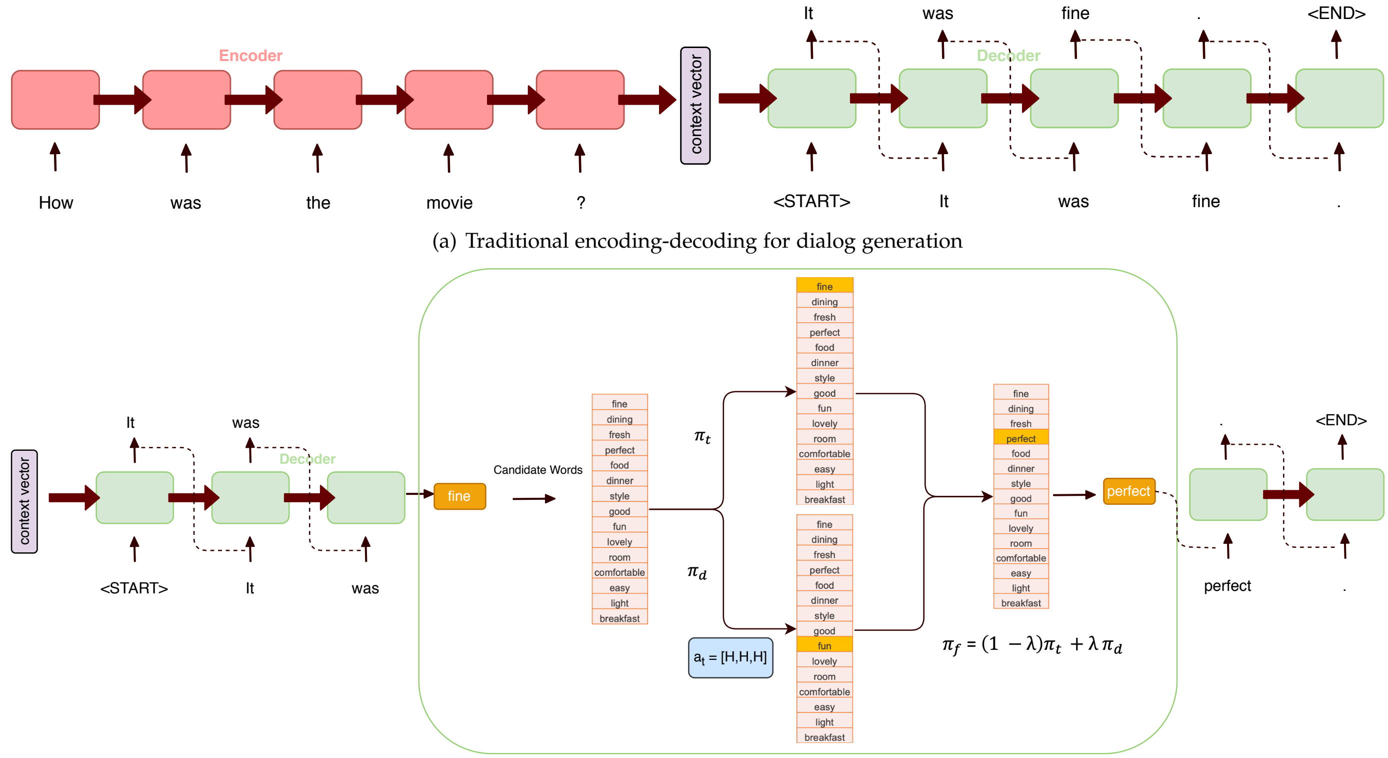
Let s_t be a sequence of words $w_1, w_2, ..., w_t$ that is semantically meaningful, syntactically correct and corresponds to the target affect a_t. If we choose the next word, w_{t+1} such that it also belongs to the target affect a_t (or neutral affect a_n) and preserves the semantic and syntactic validness of the sequence, we will have a longer sequence s_{t+1} $=w_1, w_2, ..., w_t, w_{t+1}$ that also holds the required premises (i.e., is a syntactically and semantically valid sequence in the targeted affect). The response generation scenario introduces an additional constraint, since the generated reply, in addition to being semantically meaningful, syntactically correct and in the targeted affect, should also be a plausible response for the given source utterance. Now, the problem transforms to the following: given a source utterance X, the aim is to generate a response Y for it, in the targeted affect a_t.

4.2 Candidate Words

We use a sequence-to-sequence model comprising an encoder and a decoder, which can be either trained from scratch or fine-tuned on a dialog corpus (see section 5.2 for implementation details). The sequence-to-sequence model encodes the source utterance X into a latent representation, and then decodes it to obtain the response Y. Thus, the model has learned not only the underlying language distribution, but also appropriate response generation.

To generate these responses in the targeted affect a_t, we encode the source utterance X into a latent representation h. We intervene with the decoding part. Figure 2 depicts the modification of the affect of a generated word with AffectON. At any given step t of the decoding, the model's output passes through a softmax function to yield the probability distribution of words $\pi_t \in \mathbb{R}^{\mathrm{V}}$, where V is the size of the vocabulary. The *argmax* of π_t is essentially the index of the word y'_t, which sequence-to-sequence model assigns highest probability to become the next word in the response Y. To continue, we can either use this predicted word y'_t or we can use the word y_t from the ground truth response coming from the corpus. Here, we continue with the ground truth response. We proceed by extracting the lemma form of the ground truth word $l_t = lemma(y_t)$ (e.g. $loved$ becomes $love$). This is done since the affective lexicon is constructed solely from lemmas. We check whether the lemma at step t, l_t exists in the affective lexicon. If the lemma is not in the affective lexicon, we do not possess any information on its affect, hence the original form of the lemma, y'_t is appended to the response Y without any further modification.

In the case when the lemma exists in the affective lexicon, our aim is to find a word that is contextually similar to y'_t, but is closer to the target affect a_t than l_t . To accomplish that, we pick k words which the model assigns the highest



(a) Traditional encoding-decoding for dialog generation

(b) Decoding with AffectON

Fig. 2. Traditional encoding-decoding (a) and the intervention of AffectON in decoding (b) are shown. Subfigure (b) depicts an example run through a simplified pipeline of AffectON with a sequence-to-sequence model. The highlighted words in the lists of words represent the words with the highest probability in that particular probability distribution.

probabilities as candidate words at step t from the probability distribution over all words in the vocabulary $\pi_t \in \mathbb{R}^V$. This reduced vector $\pi_t \in \mathbb{R}^k$ of probabilities is then put through a *softmax* to normalize the probabilities.

4.3 Incorporating the Affect of Candidate Words

To extract affective information for the candidate words, initially we obtain their lemma form. For each candidate lemma l, we compute the distance of its $V_l A_l D_l$ from the target affect's $a_t = V_t A_t D_t$, as Euclidean distance: $d = \sqrt{(V_l - V_t)^2 + (A_l - A_t)^2 + (D_l - D_t)^2}$, to obtain a vector of distances $d = [d_1, d_2, ..., d_k]$. To find a probability distribution over the distances, the vector of distances goes through a *softmax* function to yield $\pi_d \in \mathbb{R}^k$. Note that since a smaller distance from the target affect is more desirable, we take the negative values of the distances prior the *softmax* step. Finally, we calculate the fused probability $\pi_f \in \mathbb{R}^k$ for each candidate word, by taking into account the probability of the language model and the distance from the target affect as a probability:

$$\pi_f = (1 - \lambda)\pi_t + \lambda\pi_d. \tag{3}$$

The parameter λ in the equation, is tuned based on the priority we give the affect as opposed to the language model. In other words, it can be regarded as the *affect strength* we aim for. If $\lambda = 1$, then the generated utterances may be closer to the target affect, but will lose in syntactic correctness and/or semantic meaningfulness. On the other hand, if $\lambda = 0$ the generated utterances will aim for preservation of the latter requirements, without any concern about being in the target affect. In other words, it will be the case of plain sequence-to-sequence decoding. Finally, the candidate word with the highest probability π_f is picked as the next word of the response. We feed this word to the decoder, to continue the same described process for step $t + 1$. Note, that if the original word was not an affective one to begin with, it is not modified, instead it is immediately fed to the decoder.

5 Experiments

5.1 Model

As large pre-trained language models are being deployed at a rapid pace, it would be useful to generate affective text by leveraging these models' generation powers either by fine-tuning them or intervening during inference. Being language model-agnostic, our proposed approach allows the utilization of any underlying model for affective text generation, thus is useful in keeping up with the pace of these developments. We experimented with affective dialog generation by utilizing an instance of such large pre-trained model fine tuned on dialog corpora. We used an off-the-shelf model [43] which won automatic metric track of the dialog competition ConvAI2 [1] at *NeurIPS2018*. The generative model is a multi-layer Transformer encoder based on Generative Pretrained Transformer [40]. The decoder

[1] http://convai.io/

Algorithm 1 AffectON algorithm

```
Input: $X, a_t$                                  ▷ $X$ is input utterance, $a_t$ is target affect
$Y' \leftarrow \{\}$                             ▷ set response to empty set
$t \leftarrow 0$
$h \leftarrow \psi_e(X)$                         ▷ encode the utterance
while $y'_t \neq EOS$ do                         ▷ while the generated word is not End Of Sentence token
    $t \leftarrow t + 1$
    $\pi_t \leftarrow softmax(h)$                ▷ probability distribution over the vocabulary at step $t$
    $y'_t \leftarrow argmax(\pi_t)$
    $l_t \leftarrow lemma(y'_t)$                 ▷ get the lemma form of the word
    if $l_t$ not in affective lexicon then
        $Y \leftarrow Y \cup y'_t$               ▷ add word to response if not affective
    else
        $c \leftarrow top(\pi_t, k)$             ▷ top-k words with highest probabilities
        $\pi_t \leftarrow softmax(\pi_t[c])$     ▷ probabilities of candidate words
        $l \leftarrow lemma(c)$                  ▷ lemmas of candidate words
        $d \leftarrow euclidean(l, a_t)$         ▷ distances of candidate words from target affect
        $\pi_d \leftarrow softmax(-d)$           ▷ probability distribution over the distances
        $\pi_f \leftarrow (1 - \lambda)\pi_t + \lambda\pi_d$   ▷ aggregated scores of candidate words
        $y''_t \leftarrow argmax(\pi_f)$         ▷ candidate with highest score
        $Y \leftarrow Y \cup y''_t$              ▷ add word to response
        $h \leftarrow \psi_d(h, y''_t)$          ▷ feed the current picked word to the decoder
return $Y$                                       ▷ return response
```

is comprised of a 12-layer decoder only transformer with masked self-attention heads (768 dimensional states and 12 attention heads). This model is fine-tuned on a dialog dataset – PERSONA-CHAT [44]. PERSONA-CHAT is a crowd-sourced dataset, where each speaker was asked to converse with another person conditioned on a few sentences that defined their personality. We refer the readers to the original paper [43] for more details on training of the model, as it falls outside the scope of our work.

While we use the aforementioned model in the generation of the affective dialog, we do not condition the text on any predefined personality.

5.2 Implementation Details

We conducted experiments with four different affective targets. Marking significant points on the affective space, affective targets with *Valence Arousal Dominance* values of *0.0–0.5–0.0*, *0.0–0.0–0.0*, *0.5–0.0–0.5*, and *1.0–1.0–1.0* were selected for experimentation. The affective target *0.0–0.5–0.0*, which we refer as Low-Medium-Low (LML), represents sentences that are negative, elicit moderate arousal and are low in dominance, such as sentences including words akin to *stress*, *death*, *torture*. Similarly, the affective target *0.0–0.0–0.0* (LLL), represents sentences that are negative, low in dominance, but elicit no arousal, such as sentences with words *sad*, *exhausted*, *tiring*. Whereas, *0.5–0.0–0.5* (MLM) affective target is considered the neutral affect, as neutral sentences belong to the middle of the scale in terms of valence and dominance, but induce no arousal (e.g. words close to *0.5–0.0–0.5*: *routine*, *curtain*, *textbook*) [42]. The *1.0–1.0–1.0*, (HHH for High-High-High), affective target is the other extremum, as it signifies sentences that are exceptionally positive, prompt high arousal, and are also dominant, such as those that include words *happiness*, *fun*, *excitement*.

For each experiment, we mapped Cornell Movie Subtitle Corpus [45] to one of the affective targets. This corpus is a large collection of fictional conversations extracted from movie scripts. It contains a total of 304,713 utterances exchanged among characters. We experimented with five λ values (1.0, 0.7, 0.5, 0.3, 0.0) for each of the affective targets. Note that $\lambda = 0$ results in the same output for all of the affective targets, since during decoding no weight is given to affect. We experimented with $k = 20$, $k = 30$ and $k = 50$, as the number of candidate words.

6 EVALUATION

We perform subjective and objective evaluation on our proposed approach.

6.1 Subjective Evaluation

For the subjective part of the evaluation, the movie corpus was mapped into *HHH*, *MLM*, *LML* affective targets by utilizing the approach proposed in section 4 (with $\lambda = 0.5$ and $k = 30$, as we observed it to be optimal from objective evaluation). We chose these three targets for subjective evaluation, as the objective results showed them to be distinctive from one another in terms of VADER score. Also we did not collect subjective evaluations for target *LLL*, because from objective evaluations we noticed that sentences mapped into *LLL* changed less than those mapped into *HHH* and *LML*. We were interested to obtain human ratings on how well our approach has mapped sentences that have changed more substantially.

Since not all of the utterances change affect during the mapping, for human evaluation we picked the utterances that were modified the most by calculating n-grams between the mappings and the original utterances. These utterances are most likely to be in the targeted affect, but also to be worse off in terms of perplexity. Other than this criteria, the utterances shown were randomly selected from the three mapped corpora. To avoid any kind of rater bias, we also shuffled the utterances originating from the different corpora when they were presented to the raters.

6.1.1 Task & Procedure

Each rater rated a total of 21 pairs of utterances plus two "golden" pairs that were used to filter the ratings high in quality. A *pair* refers to an utterance from the original corpus and its mapped version in one of the affective targets. Utterances were evaluated on a scale from 0 to 100 regarding each affective dimension (i.e. *valence, arousal, dominance*). In addition, they were scored on a scale from 0 to 2 for *syntactic coherence* (Is the utterance grammatically correct?) and *appropriateness* (Is this utterance a plausible reply for the preceding utterance?). In the *appropriateness* part of the evaluation, participants were shown also the preceding dialog utterance, and were asked whether the modified utterance was an appropriate response to the preceding one. Human raters evaluated a total of 273 pairs of utterances, of those 180 distinct distinct pairs of utterances, for the three targets (60 per target).

6.1.2 Design and Analysis

The study was a within-subject design. Each participant was presented both with the original utterance and its mapped version in either of the affective targets. We used one-way ANOVA for analyses.

6.1.3 Participants

We recruited participants online via Amazon Mechanical Turk. Participation was limited to adults residing in the United States. We excluded the data of participants based on their response to two golden questions. These two utterances had opposite valence (e.g. "she was very happy!" and "she was very sad!"). Participants that rated these two utterances as having similar valence were removed from analysis. 13 out of 20 participants who completed the task were retained for final analysis. Each participant was paid 3$ per task (10$/hour). Participants were introduced with concepts of valence, arousal and dominance through a series of explanations and examples prior to proceeding with the task.

6.1.4 User Interface

The interface design is regarded crucial to and highly associated with the quality of ratings. Evidently, an inadequate user interface design leads to bias in rating tasks [46]. We designed a custom web-based interface for the rating task.

Participants rated the utterances, in terms of valence, arousal, dominance, syntactic coherence and appropriateness. Two approaches can be taken in this rating scenario, namely utterance based or dimension based. In the first approach, participants score an utterance in all dimensions (VAD, syntactic coherence, appropriateness) then proceed to the other utterance to score it. The second approach is for participants to score initially all of the utterances in a single dimension, then proceed to the other dimension to score the utterances. A pilot study we conducted indicated that participants had a harder time with the first approach, since changing the mode of the evaluation in addition to increasing the time per rating, also led to confusion among the meaning of dimensions. Hence, we asked participants to score the all utterances in each dimension sequentially.

Figure 3 depicts the user-friendly interface designed for the study. In this example, the rater was asked to score the utterances in terms of valence. Just as shown in the figure, for the dimension that was being scored (i.e. valence) we provided the definition, the manikin adopted from [47] and the words associated with the extrema of the scale, as an aide-mémoire for the raters. Because our affective lexicon has scores on a scale from 0-1, we use a scale from 0-100, as it easier to grasp than decimals, and display it as a slider. Raters were asked to adjust the slider where they thought suitable for the utterance in question. For scoring syntactic coherence and appropriateness, a Likert scale (0-2) like question was presented to raters.

6.2 Objective Evaluation

Evaluation of affective language generation systems entails two aspects, namely assessment of how good the generated language is and whether the generated language is in the targeted affect. Below, we discuss the objective metrics used for the two parts of the evaluation.

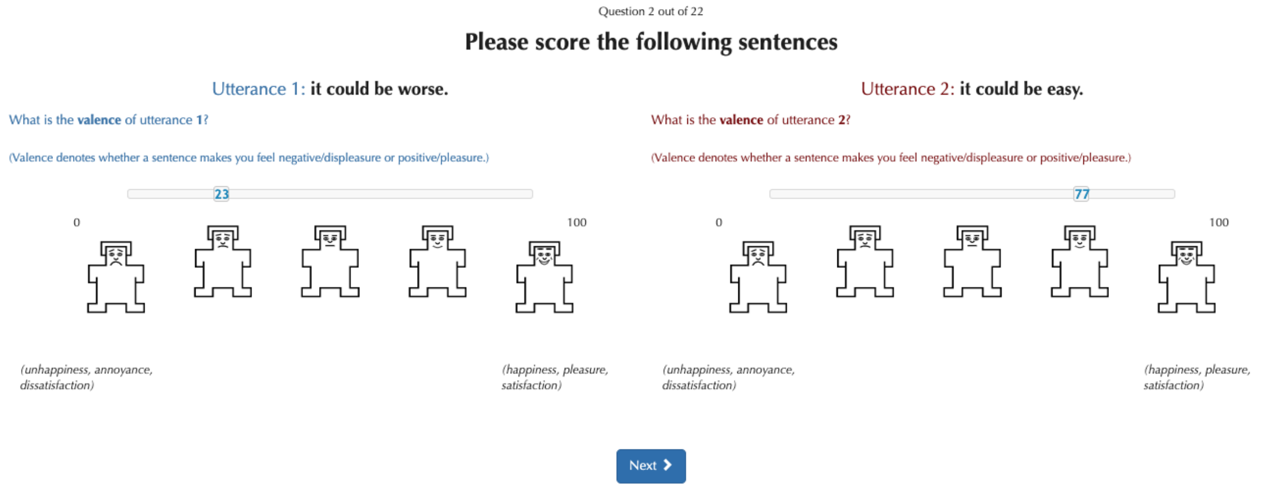
6.2.1 Language Generation

Generative models in general suffer from lack of established objective evaluation metrics [48]. For natural language generation, the most widely reported score is the *perplexity* of the model. Perplexity is a measurement of how well a probability distribution or probability model predicts a sample. A lower *perplexity* indicates that a sentence has higher probability under the language model. Following prior work [30], we evaluate our generated responses in terms of perplexity with another large language model – GPT-2 [49].

Another broadly established metric in evaluation of conversational models with genesis from machine translation is BLEU (Bilingual Evaluation Understudy) score [50]. It measures the lexical overlap between the generated responses and the reference ones. Though this metric can be used in cases where there is a reference, or ground truth sentence to be compared to, in our setting we do not have ground truth affective sentences. We report BLEU metric with the reference being the original corpus to indicate how different is the affective corpus from the original one. A higher BLEU score indicates fewer utterances were changed in the mapped corpus.

6.2.2 Sentiment analysis

While there is no tool available for automatic measurement of affect of a text in terms of VAD, a great deal of effort has been spent on developing automatic tools for sentiment analysis [51], [52]. Practically, the available tools measure only the valence or polarity of text and do not consider arousal and dominance dimensions. To measure the valence of our generated sentences, we use one of these tools: Valence Aware Dictionary for Sentiment Reasoning (VADER)[52]. VADER is a rule-based system for sentiment



Question 2 out of 22

Please score the following sentences

Utterance 1: it could be worse.

What is the valence of utterance 1?

(Valence denotes whether a sentence makes you feel negative/displeasure or positive/pleasure.)

23

Utterance 2: it could be easy.

What is the valence of utterance 2?

(Valence denotes whether a sentence makes you feel negative/displeasure or positive/pleasure.)

77

0
100
(unhappiness, annoyance, dissatisfaction)
(happiness, pleasure, satisfaction)
0
100
(unhappiness, annoyance, dissatisfaction)
(happiness, pleasure, satisfaction)
Next

Fig. 3. The designated user interface for subjective evaluation. Illustrated is a sample question where participants were requested to score the utterances in terms of *valence*. The interface looks similarly for *arousal* and *dominance* dimensions, with respective manikins adopted from [47].

analysis. In addition to utilizing a large lexicon of human valence ratings, VADER considers exclamation points, degree modifiers (e.g. *really* good), negation words (e.g. *not* bad), when computing the valence of a given sentence. VADER's compound score classification thresholds are set by the authors [52] at –0.05 and +0.05. If a sentence has a score higher than +0.05 it is considered positive in valence, if it has a score between -0.05 and +0.05 it is considered neutral and if it has a score lower than -0.05 it is considered negative.

7 Results & Discussion

In the following section we discuss the subjective and objective results of our evaluation. Overall, the results suggest that our approach is successful in pulling the generated language towards the targeted affect. Some sample inputs and their responses in different affective targets are provided in Table 3.

7.1 Subjective Results

The results of the subjective evaluation of utterances in terms of valence, arousal and dominance are shown in Figure 5 (a). Whereas, Figure 5 (b) depicts the subjective evaluations of utterances in terms of syntax and appropriateness as a response. For each of the target affects, the mean of the modified sentences and the mean of their original/unmodified counterparts is depicted. Additionally, mean of syntactic coherence and appropriateness are shown.

We measure the rater agreement for the 186 sentences rated by at least two raters with Pearson correlation for continuous variables (valence, arousal, dominance) and Cohen's κ for categorical variables (syntax and appropriateness). Our analysis indicate a very strong correlation for valence (Pearson $r = 0.97, p << 0.0001$), a moderate one for arousal (Pearson $r = 0.42, p << 0.0001$) and a weak one for dominance (Pearson $r = 0.16, p = 0.005$). This might be partly due to the fact that arousal and dominance are harder concepts to gauge in comparison to valence. For syntax and appropriateness, raters showed moderate (Cohen's κ = 0.52) and fair agreement (Cohen's κ = 0.38), respectively.

The results of subjective evaluations indicate that our approach was successful in pulling the mean of valence of the utterances towards the targeted affects. For each of the dimensions, we observed a significant change in *valence*. The valence of original utterances $(M_{o_{HHH}} = 42.47)$ was pulled towards positive affect in *HHH* mapping $(M_{HHH} = 82.09, F_{1,181} = 77.99, p << .0001)$. We observe that valence of the original utterances in *MLM* is negative, whereas valence of *LML* is positive. This is due to the fact that for subjective evaluation, we picked utterances that changed the most, by calculating distinct n-grams between the mappings and the original. These are the utterances that have most likely changed affect in the mapping. For *MLM*, we observe that the original utterances $(M_{o_{MLM}} = 26.27)$ pass 50, the neutral threshold $(M_{MLM} = 75.17, F_{1,181} = 176.10, p << .0001)$. While the result corresponds to the objective VADER score in Figure 4(b), for $\lambda = 0.5$, as discussed more thoroughly in section 7.2, this phenomenon might be a result of sentiment bias of the language model. Original utterances $(M_{o_{MLM}} = 71.38)$ were successfully pulled toward negative valence for affective target *MLM* $(M_{MLM} = 29.09, F_{1,181} = 112.10, p << .0001)$.

In contrast to valence, in the distribution of arousal, the words are mostly accumulated in the middle of the scale (Figure 6). This lack of words high in arousal is perceived in the results of extreme targets HHH. While arousal was changed significantly from original $(M_{o_{HHH}} = 49.68)$ towards higher values for target affect *HHH* $(M_{HHH} = 60.95, F_{1,181} = 8.73, p = .003)$, the change was not as perceivable as for valence. Significant change in the negative direction of arousal was also detected for *LML*. Original utterances $(M_{o_{LML}} = 56.28)$ shifted towards low arousal when mapped $(M_{LML} = 47.35, F_{1,181} = 5.37, p = 0.02)$. For affective target *MLM*, original utterances $(M_{o_{MLM}} = 54.62)$ moved toward the lower part of the arousal scale $(M_{MLM} = 51.39, F_{1,181} = 0.67, p = n.s.)$, although no significant effect was observed.

Dominance was pulled significantly in the respective targets for each target affect. For target *HHH*, original utterances ($M_{o_{HHH}} = 45.81$) are effectively pulled toward high dominance ($M_{HHH} = 64.94, F_{1,181} = 23.76, p << .0001$). A similar occurrence is also noticed for target dominance target affect MLM, where original sentences shift from ($M_{o_{MLM}} = 40.08$) to ($M_{MLM} = 56.15, F_{1,181} = 13.62, p = .0003$). For target *LML*, utterances shifted from ($M_{o_{LML}} = 56.07$) toward lower dominance ($M_{LML} = 39.65, F_{1,181} = 15.58, p << .0001$), as desired.

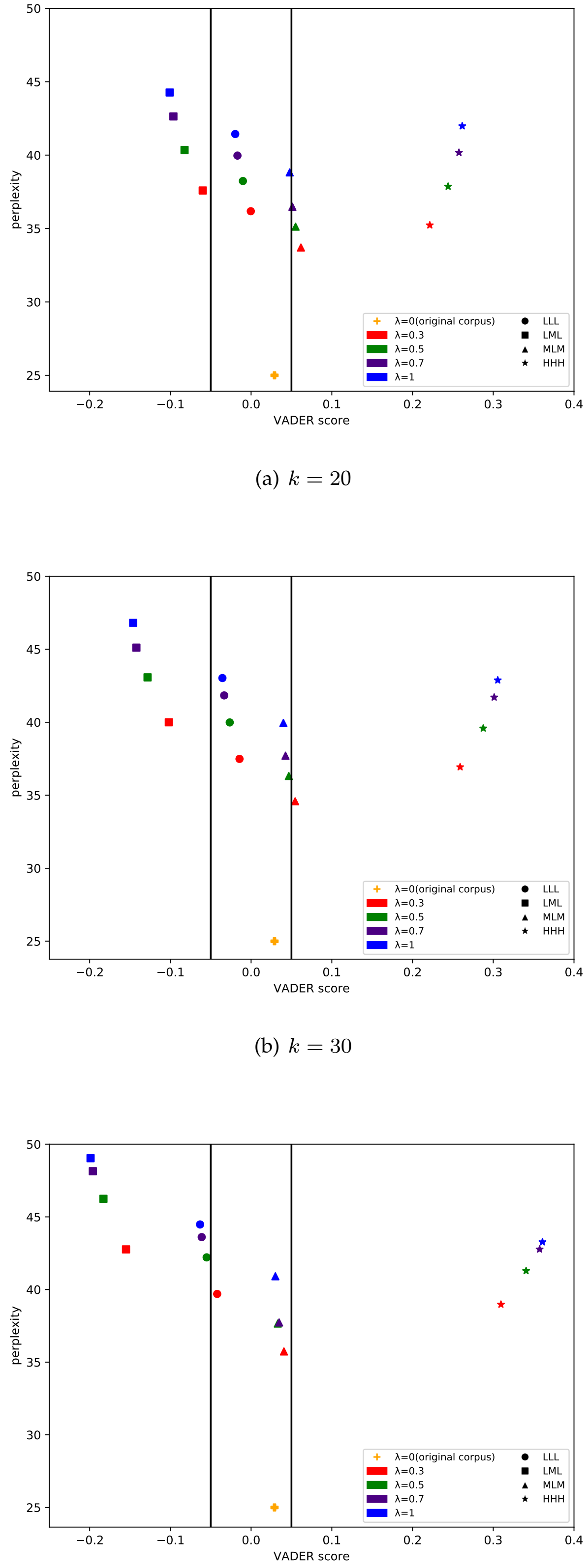
We did not observe any significant effect in terms of difference in syntactic coherence between the modified sentences and the original ones, for either any of the target affects: *HHH* ($M_{o_{HHH}} = 1.50, M_{HHH} = 1.51, F_{1,181} = 0.61, p = n.s.$), *MLM* ($M_{o_{MLM}} = 1.59, M_{MLM} = 1.58, F_{1,181} = 0.56, p = n.s.$), *LML* ($M_{o_{LML}} = 1.55, M_{LML} = 1.49, F_{1,181} = 0.62, p = n.s.$). However, the mean of both the modified and original utterances is close to 1.5 on a scale from 0 (not grammatically correct) to 2 (grammatically correct) which indicates that the original utterances, also, were not syntactically coherent to begin with. We observe that our model preserves the syntactic coherence. For example, original *MLM* utterances, have higher syntactic coherence and their modified counterparts also stay in the same range. As seen in the objective evaluation, affective target *LML* disturbs perplexity the most among other targets. This is also reflected in the subjective evaluations.

Since the dialog pairs originate from a movie subtitle corpus, they generally lack the naturalness of daily conversation. This unnaturalness is reflected on the *appropriateness* score, where even the original dialog pairs have a relatively low mean of 1.2 on a scale from 0 (inappropriate response) to 2 (appropriate response). As it is the case with syntactic coherence, the utterances are more appropriate when the original affect and the targeted affect are not very different (i.e. affective target is MLM). However, we do not observe significant difference in terms of appropriateness of original and modified utterances for any of the affective targets: *HHH* ($M_{o_{HHH}} = 1.20, M_{HHH} = 1.25, F_{1,181} = 0.24, p = n.s.$), *MLM* ($M_{o_{MLM}} = 1.25, M_{MLM} = 1.3, F_{1,181} = 0.33, p = n.s.$), *LML* ($M_{o_{LML}} = 1.28, M_{LML} = 1.21, F_{1,181} = 0.24, p = n.s.$)

Overall, our results indicate that our approach is mostly successful in pulling the generated language towards the target affect, without any significant loss in terms of syntactic coherence and appropriateness.

7.2 Objective Results

Figure 4 depicts the averages over the entire movie corpus (304, 713 utterances) for $k = 20$, $k = 30$ and $k = 50$. Comparing the results of different λ values for each k, we observe that the larger the λ value, the less importance is given to the language model constraint, hence the wider range of VADER scores (valences). However, as expected the perplexity worsens as λ increases. We observe that when λ is small (e.g. $\lambda = 0.3$), perplexity scores are better than for larger λ values, but the range of VADER scores is very narrow. Thus, some of the affective targets do not exceed their respective VADER score boundary due to the priority

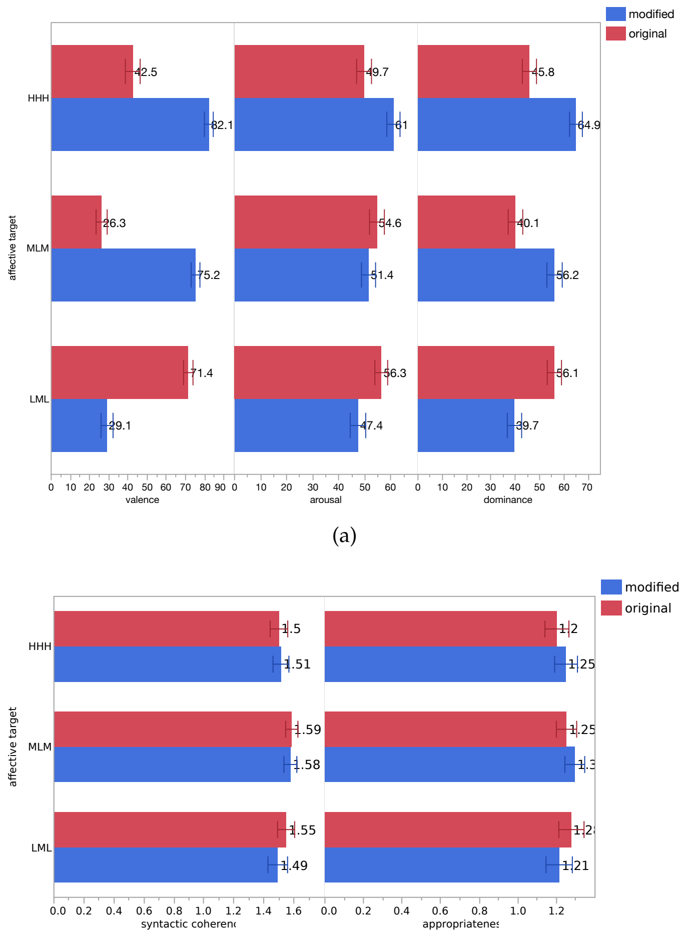


(a) $k = 20$

(b) $k = 30$

(c) $k = 50$

Fig. 4. Results of experiments with different number of candidate words k and λ values. Each point represents the average VADER score and the perplexity of the movie corpus mapped to a given target affect. The two vertical lines depict the VADER score boundaries for positiveness ($>= 0.05$) and negativeness ($=< -0.05$).



(a)

(b)

Fig. 5. Human ratings of (a) valence, arousal and dominance (b) syntactic coherence and appropriatenes for affective targets HHH, MLM and LML (blue) and their original counterparts (red). Error bars indicate one standard error.

that is given to language model. Perplexity scores imply that our approach performs the affect modification without a huge deterioration in language structure for smaller λ. Among modifications with the same λ values but different target affects, again we observe that targets closer to original affect have better scores.

Comparing the results of different k values, we notice that as we increase number of candidate words we see better mappings for valence. For $k = 50$, except when $\lambda = 0.3$, *LLL* falls in the negative valence portion, and the neutral affect *MLM* falls in the neutral zone, as desired. However, the trade-off is seen in terms of perplexity. Evidently, increasing the number of candidate words allows the algorithm to select a word from a larger pool of words, which might be closer to the target affect, yet has lower probability under the language model.

Among these experiments, AffectON with $k = 30$ and $\lambda = 0.5$ manages to modify the affect noticeably, yet preserve the perplexity scores the most. We expected the original corpus ($\lambda = 0$) utterances and the modified utterances of $\lambda = 0.5$ with the neutral target affect $a_t = [M, L, M]$ to have the similar average VADER scores. However, in Figure 4(a) we see that λ (0.3, 0.5) values of the neutral affective target fall in the positive side of the VADER border. We observe a similar effect for the *LLL* affective target, for which the values fall in the neutral zone of the sentiment, instead of the negative. This is particularly interesting because it suggests that the underlying language model is biased towards positive sentiment, in line with evidence from prior work [53].

We also observe that mapping sentences to the neutral affet *MLM*, as expected, has the least negative impact on perplexity. We notice that target *LML* has the highest perplexity, although the distance between the original corpus and the mappings in terms of valence is smaller than that of mappings for affective target *HHH*. This again highlights the bias of language model towards the positive valence as it gives positive words higher probabilities.

Finally, to understand how distinct the mappings are from the original corpus we computed BLEU scores (computed using 1- to 4-grams) with the original corpus as the reference. BLEU scores, presented in Table 2, communicate a similar story as perplexity. As the number of candidate words increases, BLEU score decreases, indicating that the mappings become more different from the original corpus. We see a similar pattern with λ values, also. Higher λ values allow the model to pick words that are more affective, and in turn less similar to the original corpus, resulting in lower BLEU scores.

7.3 Correlation between Objective Metrics and Human Judgments

To understand how reliable VADER and perplexity are as objective measures for our task, we computed the correlations between these two metrics and human judgements of valence and syntactic coherence, respectively. In line with prior work that analyzes the correlation between objective metrics and human judgments for natural language generation [54], we compute these correlations on individual ratings of utterances and also their mean values. Correlations amongst mean values are computed because in natural language generation, metrics (e.g. perplexity, BLEU score) are more informative when reported on corpus level (i.e. averages) rather than per sentence [54]. However, the number of data points is reduced notably when computing correlations on mean values, thus significance cannot be shown.

We picked $k = 30$ and $\lambda = 0.5$ for the subjective evaluation, hence we compare mean values of VADER score and perplexity of that particular setting to mean human ratings. These scores are shown in Table 1. The results revealed a strong and highly significant correlation between 360 individual ratings of valence and VADER scores of the same utterances (Pearson $r = 0.73, p << 0.0001$). Note that these 506 utterances are the 253 rated utterances and their corresponding ground truth pairs. We observed a strong correlation between means of VADER and valence scores of affective targets, also (Pearson $r = 0.88$).

When measuring the correlation between perplexity and syntactic coherence on individual 506 ratings, we used one-way ANOVA, because syntactic coherence is a categorical variable for individual ratings. The results of the analysis were not significant ($F_{1,2} = 1.44, p = n.s.$). However, we

TABLE 1
Objective Measures and Human Judgments ($k = 30$, $\lambda = 0.5$)

	VADER	**valence**	**perplexity**	**syntactic coherence**
HHH	0.28	82.1	39	1.51
MLM	0.04	75.2	36	1.58
LML	-0.13	29.1	43	1.49

TABLE 2
BLEU scores for $k = 20$, $k = 30$ and $k = 50$ candidate words. Each number indicates the BLEU score of the mapped corpus.

$k = 20$	λ =1.0	λ =0.7	λ =0.5	λ =0.3
HHH	0.079	0.081	0.085	0.090
MLM	0.086	0.092	0.096	0.097
LML	0.086	0.089	0.093	0.096
LLL	0.086	0.089	0.094	0.097
$k = 30$	λ =1.0	λ =0.7	λ =0.5	λ =0.3
HHH	0.071	0.073	0.077	0.082
MLM	0.077	0.084	0.088	0.090
LML	0.074	0.078	0.082	0.087
LLL	0.076	0.079	0.084	0.088
$k = 50$	λ =1.0	λ =0.7	λ =0.5	λ =0.3
HHH	0.063	0.067	0.071	0.077
MLM	0.065	0.071	0.081	0.085
LML	0.062	0.070	0.074	0.078
LLL	0.069	0.071	0.075	0.081

noticed a strong correlation between the mean values of perplexities and syntactic coherence ratings per affective target (Pearson $r = -0.92$). Note that the correlation is negative because higher perplexity indicates a less probable sequence.

Overall, VADER seems to be a good predictor of human ratings of valence. We investigated the reasons behind not finding a significant effect between perplexity and syntactic coherence for individual ratings. Particularly, we looked into sequences where there was discrepancy between the two ratings. These included sequences that human rated as syntactically incoherent, while they had low perplexities under the language model. Examples included sequences such as: "He's a pretty man.", where other adjectives are more commonly used to describe the subject (e.g. *handsome*). Our observations suggest that perplexity is a relatively good gauge of the overall plausibility of sequences, especially when calculated on corpus level. However, subjective evaluation is still necessary to measure the quality of the generated language.

8 Conclusion

As conversational agents have become ubiquitous, it is paramount that our interaction with them is effortless and satisfying. This also means that these conversational agents need to be affectively cognizant and respond to our affective states appropriately. In this paper, we presented AffectON, an approach for generating affective language. Particularly, we experimented with affective dialog generation. We leveraged a sequence-to-sequence language model and an affective space to generate affective responses. We conducted subjective and objective evaluations to assess our approach. Our results indicate that AffectON can successfully pull text to a targeted affect, with a small sacrifice in terms of syntax. Our results also indicate that the objective metrics correlate reasonably well with subjective ratings. With the rapid pace of development of large language models, AffectON is promising as an approach that generates affective responses relatively well while intervening only during inference. Therefore, it can easily accommodate any underlying language model without the need for further training.

Despite its promises for affective language generation, this study has some limitations. For instance, negation words (e.g. *not*) are an issue, since we apply our approach word by word. In a positive target affect, for a instance, the original sentence *The movie wasn't bad.* might become *The movie wasn't good*. Though, the objective evaluation (VADER score) considers these negation words and the results show that our approach is successful in pulling the language toward the desired affect, in this context there is room for improvement.

While we present an approach for generating affective language, further research is necessary to understand affective dialog strategies. For example, when should the conversational agent match user's affective state and when it is beneficial to generate responses in a different affect? We should also look deeper into human-human interactions and understand which strategies work best for which types of situations, applications and people.

9 Acknowledgements

Authors M. Sezgin, Y. Yemez and E. Erzin thank ERA-Net CHIST-ERA (JOKER project) and The Scientific and Technological Research Council of Turkey (TUBITAK, grant number 113E324).

References

[1] R. W. Picard and J. Klein, "Computers that recognise and respond to user emotion: theoretical and practical implications," *Interacting with computers*, vol. 14, no. 2, pp. 141–169, 2002.

[2] B. Kühnlenz, S. Sosnowski, M. Buß, D. Wollherr, K. Kühnlenz, and M. Buss, "Increasing helpfulness towards a robot by emotional adaption to the user," *International Journal of Social Robotics*, vol. 5, no. 4, pp. 457–476, 2013.

[3] S. L. Lutfi, F. Fernández-Martínez, J. Lorenzo-Trueba, R. Barra-Chicote, and J. M. Montero, "I feel you: The design and evaluation of a domotic affect-sensitive spoken conversational agent," *Sensors*, vol. 13, no. 8, pp. 10 519–10 538, 2013.

[4] B. Kort, R. Reilly, and R. W. Picard, "An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion," in *Proceedings IEEE International Conference on Advanced Learning Technologies*. IEEE, 2001, pp. 43–46.

[5] S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Transactions on Affective Computing*, 2017.

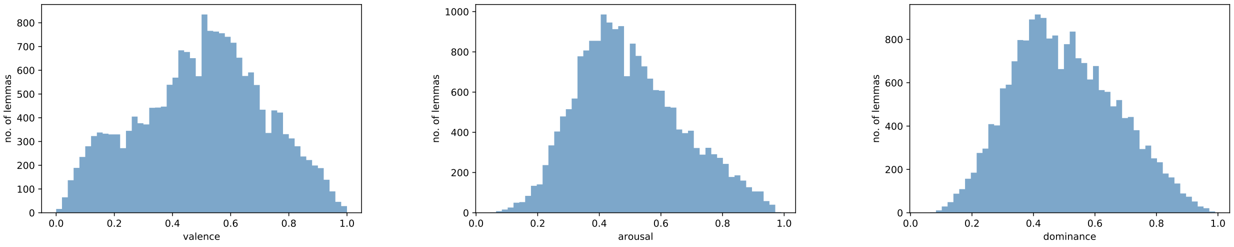
[6] C. Gordon, A. Leuski, G. Benn, E. Klassen, E. Fast, M. Liewer, A. Hartholt, and D. R. Traum, "PRIMER: an emotionally aware virtual agent," in *Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI 2019), Los Angeles, USA, March 20, 2019*, C. Trattner, D. Parra, and N. Riche, Eds., 2019.

[7] S. Mahamood and E. Reiter, "Generating affective natural language for parents of neonatal infants," in *Proceedings of the 13th European Workshop on Natural Language Generation*, 2011, pp. 12–21.

[8] A. Malhotra, L. Yu, T. Schröder, and J. Hoey, "An exploratory study into the use of an emotionally aware cognitive assistant," in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. Citeseer, 2015.

TABLE 3
Sample inputs and their responses in different affective targets.

input	Good to see you!
target VAD - MLM	It's been a long day. How are you?
target VAD - HHH	It's been a wonderful pleasure chatting with you! How are you doing right now?
target VAD - LML	It's been a crazy day! How are you doing?
target VAD - LLL	It's been a little lonely here in the wilderness.
input	Maybe we could go somewhere and just eat?
target VAD - MLM	I'm not sure.
target VAD - HHH	I'm great at eating. I'm very good at it.
target VAD - LML	I'm not hungry. You can go.
target VAD - LLL	I'm not one for small talk. I'm a loner.
input	Can you help us carry this?
target VAD - MLM	I can be of use to you.
target VAD - HHH	I love to help people. I'm a fierce leader.
target VAD - LML	I need to leave. I don't have time.
target VAD - LLL	I suppose so. I'm not used to people helping me, though.



(a) Valence (b) Arousal (c) Dominance

Fig. 6. Distribution of 20,007 lemmas in the affective lexicon according to their VAD values [37].

[9] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, "Emma: An emotion-aware wellbeing chatbot," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 1–7.

[10] "Merriam-webster online," https://www.merriam-webster.com/dictionary/affect, accessed: 2019-01-27.

[11] E. André, M. Klesen, P. Gebhard, S. Allen, and T. Rist, "Integrating models of personality and emotions into lifelike characters," in *Affective interactions*. Springer, 2000, pp. 150–165.

[12] Y. Huang and S. M. Khan, "Dyadgan: Generating facial expressions in dyadic interactions," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 2259–2266.

[13] E. Hudlicka, "Virtual affective agents and therapeutic games," in *Artificial Intelligence in Behavioral and Mental Health Care*. Elsevier, 2016, pp. 81–115.

[14] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 157–183, 2003.

[15] C. G. Henton, "Method and apparatus for automatic generation of vocal emotion in a synthetic text-to-speech system," Jan. 12 1999, uS Patent 5,860,064.

[16] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *arXiv preprint arXiv:1806.09514*, 2018.

[17] M. D. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE transactions on affective computing*, vol. 5, no. 2, pp. 101–111, 2014.

[18] K. S. Fleckenstein, "Defining affect in relation to cognition: A response to susan mcleod," *Journal of Advanced Composition*, pp. 447–453, 1991.

[19] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[20] K. De Smedt, H. Horacek, and M. Zock, "Architectures for natural language generation: Problems and perspectives," in *Trends in Natural Language Generation An Artificial Intelligence Perspective*. Springer, 1996, pp. 17–46.

[21] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[22] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[23] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[24] H. Jhamtani, V. Gangal, E. Hovy, and E. Nyberg, "Shakespearizing modern language using copy-enriched sequence-to-sequence models," *EMNLP 2017*, vol. 6, p. 10, 2017.

[25] A. Tikhonov and I. P. Yamshchikov, "What is wrong with style transfer for texts?" *arXiv preprint arXiv:1808.04365*, 2018.

[26] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1587–1596.

[27] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, "Style transfer from non-parallel text by cross-alignment," in *Advances in Neural Information Processing Systems*, 2017, pp. 6830–6841.

[28] K. Wang and X. Wan, "Sentigan: Generating sentimental texts via mixture adversarial networks," in *IJCAI*, 2018.

[29] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan, "Style transfer in text: Exploration and evaluation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[30] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, "Plug and play language models: A simple approach to controlled text generation," in *International Conference on Learning Representations*, 2020.

[31] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[32] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[33] C. Huang, O. Zaiane, A. Trabelsi, and N. Dziri, "Automatic dialogue generation with expressed emotions," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, vol. 2, 2018, pp. 49–54.

[34] M. M. Bradley and P. J. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," Univ. of Florida, Tech. Rep., 1999.

[35] N. Asghar, P. Poupart, J. Hoey, X. Jiang, and L. Mou, "Affective neural response generation," in *European Conference on Information Retrieval*. Springer, 2018, pp. 154–166.

[36] P. Colombo, W. Witon, A. Modi, J. Kennedy, and M. Kapadia, "Affect-driven dialog generation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 3734–3743.

[37] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 174–184.

[38] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[40] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[41] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[42] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[43] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents," *arXiv preprint arXiv:1901.08149*, 2019.

[44] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2204–2213.

[45] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs." in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.

[46] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl, "Is seeing believing?: how recommender system interfaces affect users' opinions," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2003, pp. 585–592.

[47] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[48] Z. Xie, "Neural text generation: A practical guide," *arXiv preprint arXiv:1711.09534*, 2017.

[49] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners."

[50] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[51] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[52] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[53] K. Arnold, K. Chauncey, and K. Gajos, "Sentiment bias in predictive text recommendations results in biased writing," in *Proceedings of Graphics Interface 2018*, ser. GI 2018. Canadian Human-Computer Communications Society / Societe canadienne du dialogue humain-machine, 2018, pp. 33 – 40.

[54] E. Reiter and A. Belz, "An investigation into the validity of some metrics for automatically evaluating natural language generation systems," *Computational Linguistics*, vol. 35, no. 4, pp. 529–558, 2009.

Zana Buçinca is a PhD student at Harvard University. She earned her M.S. degree in Computer Science and Engineering from Koc Univeresity in 2019 and her B.S. degree in Computer Engineering from Izmir Institute of Technology in 2016. Her research interests include human-computer interaction, machine learning and affective computing.

Yŭcel Yemez received the BS degree from Middle East Technical University, Ankara, in 1989 and the MS and PhD degrees from Bogazici University, Istanbul, respectively, in 1992 and 1997, all in electrical engineering. From 1997 to 2000, he was a postdoctoral researcher in the Image and Signal Processing Department of Telecom Paris (ENST). Currently, he is an associate professor in the Computer Engineering Department, Koc University, Istanbul. His research interests include various fields of computer vision and graphics.

Engin Erzin (S'88-M'96-SM'06) received his Ph.D. degree, M.Sc. degree, and B.Sc. degree from the Bilkent University, Ankara, Turkey, in 1995, 1992 and 1990, respectively, all in Electrical Engineering. During 1995-1996, he was a postdoctoral fellow in Signal Compression Laboratory, University of California, Santa Barbara. He joined Lucent Technologies in September 1996, and he was with the Consumer Products for one year as a Member of Technical Staff of the Global Wireless Products Group. From 1997 to 2001, he was with the Speech and Audio Technology Group of the Network Wireless Systems. Since January 2001, he has been with the Electrical & Electronics Engineering and Computer Engineering Departments of Koc University, Istanbul, Turkey. Engin Erzin is currently a member of the IEEE Speech and Language Processing Technical Committee and Associate Editor for the IEEE Transactions on Multimedia, having previously served as Associate Editor of the IEEE Transactions on Audio, Speech & Language Processing (2010-2014). His research interests include speech-audio-visual signal processing, affective computing, human-computer interaction and machine learning.

Metin Sezgin received the graduate (summa cum laude) degree with Honors from Syracuse University, in 1999. He received the MS degree from the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, in 2001. He received the PhD degree in 2006 from Massachusetts Institute of Technology. He subsequently moved to University of Cambridge, and joined the Rainbow Group at the University of Cambridge Computer Laboratory as a postdoctoral research associate. He is currently an associate professor in the College of Engineering, Koc University, Istanbul. His research interests include intelligent human-computer interfaces, multimodal sensor fusion, and HCI applications of machine learning. He is particularly interested in applications of these technologies in building intelligent pen-based interfaces. His research has been supported by international and national grants including grants from DARPA (USA), and Turk Telekom. He is a recipient of the Career Award of the Scientific and Technological Research Council of Turkey.