Incorporating Affect into Dialog Generation

by

Zana Buçinca

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Computer Science and Engineering

April, 2019

Incorporating Affect into Dialog Generation

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Zana Buçinca

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Tevfik Metin Sezgin, Koç University

_____

Assoc. Prof. Engin Erzin, Koç University

_____

Asst. Prof. Reyyan Yeniterzi, Özyeğin University

Date: _____

Për nanën dhe babin

To mom and dad

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, despite few studies which perceive affect as constitutive aspect of response generation, most of the existing dialog systems produce identical responses to a given inquiry, irrespective of affective information. In this thesis, we introduce *AffectON*, an approach for integrating affect into dialog generation. For generating language in a targeted affect, our approach leverages a probabilistic language model, an affective space and word embeddings. *AffectON* is language model agnostic since the probabilistic distribution can originate from any probabilistic language model (i.e. sequence-to-sequence models, neural language models, n-grams). Hence, it can be employed for both affective dialog and affective language generation. We experiment with affective dialog generation and conduct subjective and objective evaluations. For the subjective part of the evaluation, we design a designated user interface for rating and provide recommendations for the design of such interfaces. The results, both subjective and objective show that our approach is successful in pulling the generated language toward the targeted affect, with little sacrifice in syntactic coherence.

ÖZETÇE

İfade ediciliğinden ötürü doğal dil, insanlar arasında açık ve örtülü duygusal durumun aktarımı için çok önemlidir. Konuşmacıların duygusal durumuna ve konuşmanın içeriğine bağlı olarak, aynı sorunun (örneğin: Nasılsınız?), farklı duygu içerikli cevapları olabilir. Duygunun cevap oluşumunda yapılandırıcı bir özellik olarak kullanıldığı bir takım çalışmalar mevcut olsa da, diyalog sistemlerinin çoğu, duygusal durumdan bağımsız bir şekilde verilen bir soruya aynı cevabı üretmektedir. Bu tezde, duyguyu diyalog oluşturmada entegre eden bir yaklaşım olan, *AffectON*-u, tanıtıyoruz. Hedeflenen bir duyguda dil üretmek için yaklaşımımız, olasılıksal bir dil modelini, bir duygu uzayını ve kelime yerleştirmelerini kullanmaktadır. *AffectON*'un kullandığı olasılıksal dağılım herhangi bir olasılıksal dil modelinden kaynaklanabileceğinden (örneğin, diziden diziye modeller, sinir dili modelleri, n-gramlar) yaklaşımımız dil modeli agnostiğidir. Bu nedenle, hem duygusal diyalog hem de duygusal dil üretimi için uygundur. Bu çalışmada, duygusal diyalog oluşturma ile deneyler yapıp, öznel ve nesnel değerlendirmeler gerçekleştirdik. Değerlendirmenin öznel kısmında, derecelendirme yapılabilmesi için bir kullanıcı arayüzü tasarlayıp, bu tür arayüzlerin tasarımı için öneriler sunduk. Hem öznel hem de nesnel değerlendirme sonuçları, yaklaşımımızın oluşturulan dili hedeflenen duyguya doğru çekmede başarılı olduğunu ve sözdizimsel tutarlılık konusunda çok az fedakarlıkta bulunduğunu göstermektedir.

ACKNOWLEDGMENTS

I am grateful to my advisors Prof. Metin Sezgin, Prof. Yucel Yemez and Prof. Engin Erzin, without whom this thesis would not have existed. They gave me the independence to explore a topic of my choosing, yet steered me in the right direction every time it was necessary. I feel tremendously lucky to have gained insights from three different perspectives for every research problem I have encountered. Most importantly, I am forever indebted to my professors for their understanding, support, compassion and patience they showed me throughout my thesis journey. If I am to become someone's advisor one day, I am lucky to have the paragon of advisors as role models.

At Koc, I was fortunate to be a member of two amazing labs: Intelligent User Interfaces (IUI) and Multimedia Vision & Graphics Lab (MVGL). I thank my teammate, Berker, for his patience, the knowledge he shared and all the fun times we had in the midst of work. For the entertaining lunch breaks and the interesting discussions, I thank the IUI past and current team members: Kurmanbek, Dogancan, Ozan, Alican, Ferhat, Elif, Alpay, Piril and Zeynep. For all the fun we had at MVGL, at times working past midnight, I thank: Ogun, Arda, Nusrah, Gonca, Tugtekin, Serkan, Rizwan, Tugce and all the neighbor lab members who always dropped by to say hi.

My deepest thanks go to my family for having brought me up with love and the highest human values, the importance of which I comprehended only after living kilometers apart from them. I thank my eternal cheerleader, my mother, for implanting in me the love of books, the thirst for knowledge and the desire to always challenge my limits. I thank my peace radiator, my father, for teaching me to question every piece of information I receive and for teaching me the importance of patience and

perseverance in any inquiry. I wholeheartedly thank my dearest sisters for always being a phone call away and for paving the way for me to be here today. Last but not least, I am grateful to my partner, who chose to travel four hours daily just to be by my side. His unconditional love lightened up my days and gave me the strength to run through the finish line.

I also thank Prof. Reyyan Yeniterzi for being in my thesis committee.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

Word2Vec - Word to vector

VAD - Valence Arousal Dominance

Chapter 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 even their affective states. Yet, we are surrounded by computers that until two decades ago could interpret commands and convey outputs solely in an "unnatural" language - code. This gulf between verbal human-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. In this context, while the existing systems are increasingly aspiring for more naturalistic interfaces, they still lag behind on the incumbent ability of affect generation.

Affect is defined as a set of observable manifestations of a subjectively experienced emotion [2]. 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 [3], [4], [5], or vocally by speech features [6], [7], [8]. 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 [9], 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 [10]. The three principal dimensions of affect are considered *valence* (pleasant - unpleasant), *arousal* (active - passive), and *dominance* (dominant - submissive) [11]. 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 [12], the text should also belong to the targeted affect. If in addition to these requirements, the generated language 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.

In this thesis, 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 by leveraging a probabilistic language model, an affective space and a word embeddings. 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 *log likelihood*. For subjective evaluation, we design a designated 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, an affective and a word embedding space. To the best of our knowledge, we introduce the first language model agnostic affective language generation approach, which can be employed by any probabilistic language model.

- We design a user interface for subjective evaluation of affective text generation systems and provide design recommendations for such interfaces.

Chapter 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 on 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 computer vision community, with multiple works achieving remarkable results in style transfer on images [13, 14, 15], 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 [16] 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 [17], a lot of studies consider sentiment to be so, hence manipulate with it accordingly. For instance, Hu et. al [18] 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. [19], also generate sentences in a targeted sentiment by separating semantic and stylistic information 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 [20] 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. [21] 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. Our work is similar to the studies discussed here as it also transfers the affect of the sentences. Nonetheless, we are interested on the whole affect of the sentence, not only the sentiment, which is considered as *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. Rule-based approaches have been present for nearly a decade. In these works, 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. [22] generate affective text for systems providing information to parents of neonatal infants.

However, the advances in deep learning gave impetus to data-driven dialog modeling. Vinyals et al. [23] 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 [24] 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 [25] that conditions responses on emotion categories, experiments with the ways of concatenating emotion token 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 [26], we regard affect as a continuous, multidimensional phenomenon while generating responses, which in turn enables smooth transitions among affective states.

Asghar et al. [27] also 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 conversating. Instead, our responses and their affective direction develop naturally depended on multiple factors such as: the context of the conversation, our current

affective state and the relationship with the conversating partner. Thus, we argue that for naturalistic conversations the responses should be conditioned on affective information, predicted from the aforementioned factors (see future work).

Chapter 3

BACKGROUND

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

3.1 LSTM 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:

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

Surpassing n-gram models and achieving state-of-the-art results, the eminent Long Short-Term Memory (LSTM) architectures, are the current models of choice for language modeling. LSTM is a special type of Recurrent Neural Networks (RNNs), unfettered by the vanishing gradient problem typical RNNs suffer from [28]. Just like RNNs, LSTMs can handle variable size sequences and preserve information from predecessor words, offering an optimal structure for language modeling.

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 [29]. Here, we are particularly interested in their usage in conversation modeling. The idea is to encode one (or more) sentences to a vector and then to decode this vector to the response for the encoded utterance. For conversation modeling, sequence-to-sequence architectures employ LSTMs or their variants 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 :

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

Although many issues still persist, such as dull responses, lack of personality, and lack of consistency [30], when trained with dialog corpora (e.g. movie subtitles) these models generate conversations plausible to some extent. In this study we utilize sequence-to-sequence models for affective dialog generation.

3.2 Word Embeddings - word2vec

Constructing a semantic and syntactic vector space of language, where each word is represented by a vector, is another crucial task with myriad of applications in language analysis, classification and generation. The naïve approach is assembling one-hot vectors of vocabulary size. In these vectors, only one of the elements, namely

the index of the word we want to represent, will be 1 and the other elements will be 0. Hence, a representation of a single word yields a sparse vector of size $1xV$, where V is commonly very large. In addition to space complexity, this representation does not provide any information in terms of the semantic meanings of words. Words that appear in similar context can have indices far from one-another in the vocabulary, and vice-versa. Therefore, efforts on learning better word representations date back to 1980s [31].

Studies in this area generally utilize learning approaches for representing words as dense vectors. Especially, the success of work of Mikolov et al.'s [32] is widely considered as a milestone in the word embedding task. Mikolov et al. train a shallow neural network by learning to predict the target context from a word (skip-gram model) and the target word from the context (CBOW model). Therefore, words appearing in similar context have similar representations. As a consequence, they are close to each-other in word embedding (word2vec) space. These powerful vector representations learn semantic information that can be easily accessed through simple linear operations. For example, the algebraic operation vec("Paris") - vec("France") + vec("Italy") is closest to vec("Rome"). In addition to semantics, these representations hold also syntactic information. For instance, by using the analogies just as in the country-capital case, we can extract the plural of words. The algebraic operation of vec("women") - vec("woman") + vec("child") is closest to vec("children") in word2vec space.

A common example of the relation between countries and their capitals projected in two dimensional space of word2vec embeddings is depicted in Figure 3.1.

3.3 Valence, Arousal, Dominance (VAD) Word Representation

Word2vec embeddings carry semantics and syntactic information about the relation among the words in language. As discussed in section 3.2, words that appear in similar contexts are close to one-another in word2vec space. Nonetheless, this representation does not distinguish words according to the affect they induce. For example, words

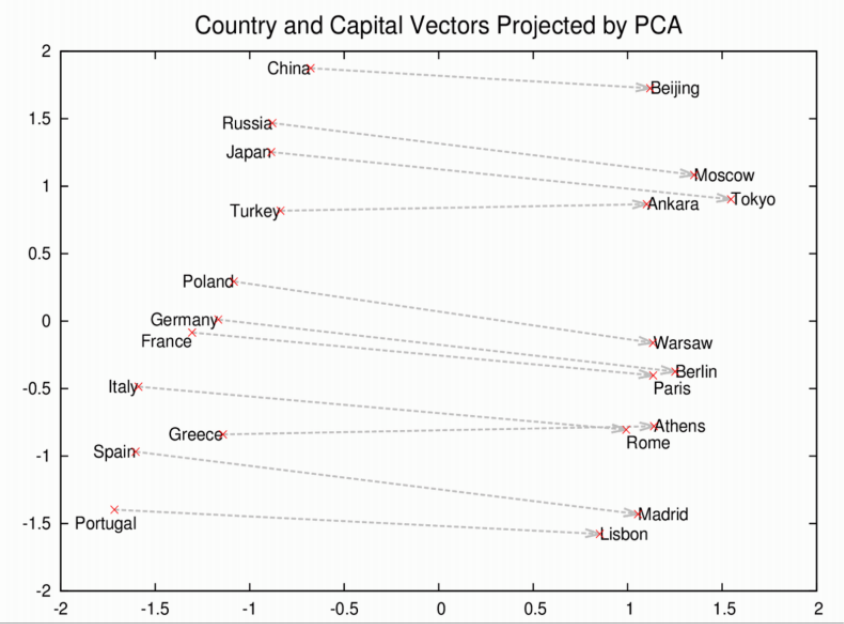
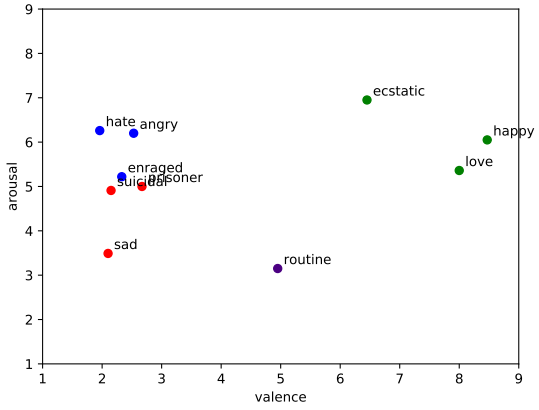


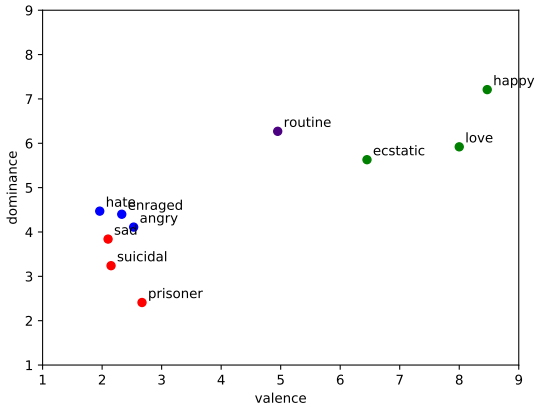
Figure 3.1: Countries and their corresponding capitals' vectors projected by PCA

with contrast affective meanings, such as *good* and *bad*, or *terrible* and *wonderful* appear nearby in word2vec space since they occur in similar context (e.g. "the weather was terrible" and "the weather was wonderful"). Akin to word2vec space, an affective space, where words similar in affect are proximate, would be valuable for analyzing and comparing words based on their underlying affect.

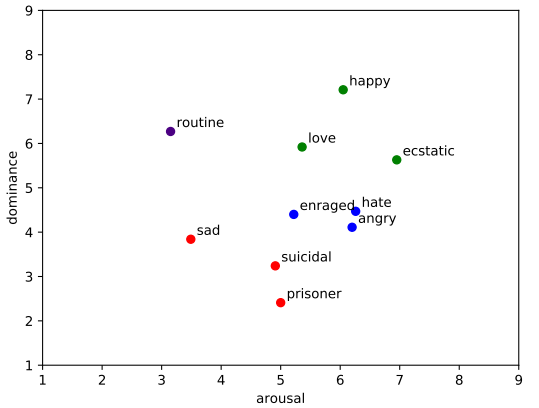
ANEW [26] is one of the earliest lexicons to consider affective meaning of words in three dimensions. In this study, humans rated 1034 words on a scale from 1 to 9 in terms of valence, arousal and dominance. This affective view accounts *valence* to range from unpleasant to pleasant, *arousal* to range from calm to excited and *dominance* from submissive to dominant. Following the same approach, Warriner et al. [33] provide an extended version of ANEW lexicon with 13,915 lemmas rated in valence, arousal and dominance (VAD) dimensions. We leverage this extended lexicon to construct our desired three-dimensional affective space. Figure 3.2 depicts sample words in VAD space.



(a) Valence - Arousal



(b) Valence - Dominance



(c) Arousal- Dominance

Figure 3.2: Sample words and their relationship in VAD space.

Chapter 4

METHOD: *AFFECTON*

In this section, we describe the proposed approach for text generation in a target affect, also shown in Algorithm 1. Specifically, we elaborate on affective response generation with a sequence-to-sequence model. However, our approach can be employed by any probabilistic language model.

The main idea lies behind the inductive step assumption that given a sequence of words $w_1, w_2, ..., w_t$ that is semantically meaningful, syntactically correct and corresponds to the targeted affect a_t, if we choose the next word, w_{t+1} such that it also belongs to the targeted affect a_t (or neutral affect a_n) and preserves the semantic and syntactic validness of the sequence, we will have a longer sequence $w_1, w_2, ..., w_t, w_{t+1}$ that also holds the required premises (i.e. is a syntactically and semantically valid sequence in the targeted affect). The response generation scenario introduces a further 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 becomes as follows: given a source utterance X, the aim is to generate a response Y for it, in the targeted affect a_t.

Initially, we train a sequence-to-sequence model comprising an encoder and a decoder on a dialog corpus (see section 4.1 for training 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 traditionally encode the source utterance X into a latent representation h, however we intervene with the decoding part. At any given step t of the decoding, the LSTM's output passes through

Algorithm 1 AffectON algorithm

Input: X, a_t ▷ X is input utterance, a_t is target affect
$Y \leftarrow \{\}$ ▷ set response to empty set
$t \leftarrow 0$
$y_t \leftarrow SOS$
$h \leftarrow \psi_e(X)$ ▷ encode first utterance
while $y_t \neq EOS$ **do**
    $h \leftarrow \psi_d(h, y_t)$ ▷ decode latent representation by feeding current picked word as input
    $t \leftarrow t + 1$
    $\pi_t \leftarrow softmax(f(h))$ ▷ 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
        $c_{W2V} \leftarrow kNN_{W2V}(y_t)$ ▷ candidate words from word embeddings
        $\pi_t \leftarrow \pi_t[c_{W2V}]$ ▷ probabilities of candidate words
        $l_{W2V} \leftarrow lemma(c_{W2V})$
        $d \leftarrow euclidean(l_{W2V}, a_t)$ ▷ 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
return Y ▷ return response

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 movie corpus. Here, we continue with the predicted word, however the ground truth responses can be used in an offline setting. 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 find the k-Nearest Neighbors of the word y_t in the word embedding space by utilizing cosine similarity where $k = 20$. The k-Nearest Neighbor words, we call the *candidate* words c_{W2V}, are words that are utilized in similar contexts with the word y_t, however might have different affects. We pick only the probabilities of these k candidate words from the probability distribution over all words in the vocabulary $\pi_t \in \mathbb{R}^{\mathrm{V}}$ yielded from the LSTM's output, and form a reduced vector $\pi_t \in \mathbb{R}^{\mathrm{k}}$.

To extract affective information for the candidate words, initially we obtain their lemma form l_{W2V}. 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 $\pi_d \in \mathbb{R}^{\mathrm{k}}$. 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 $\pi_f \in \mathbb{R}^{\mathrm{k}}$ for each candidate word, by taking into account the probability of the language model and the distance from the target affect as a probability:

$$\pi_f = (1 - \lambda)\pi_t + \lambda\pi_d \tag{4.1}$$

The parameter λ in the equation, is tuned according to 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 will 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. In our experimentation, we give priority to the language model, but we also allow the affect to have impact on the next lemma choice decision ($\lambda = 0.4$). 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.

4.1 Implementation Details

Dataset. To train the sequence-to-sequence model, we leveraged Cornell Movie Subtitle Corpus [34]. This corpus is a large collection of fictional conversation extracted from movie scripts, of a total 304, 713 utterances exchanged among characters.

Sequence-to-sequence. The sequence-to-sequence with attention [35] model was implemented in Pytorch. We utilized stochastic gradient descent with gradient clipping to prevent exploding parameters [36]. Our vocabulary is comprised of 20,000 most frequent words in the movie corpus. Words out of the vocabulary are marked with a designated unknown token. Our hidden layer is set 512 for both the encoder and the decoder. The encoder consists of two layers of bidirectional LSTM cells, whereas the decoder is composed of 4 layers of LSTM cells.

Word2Vec. Word embedding models are generally trained in billions of words,

which require days of training even in modern-day GPUs. However, there exist publicly available pre-trained models. We have used Mikolov et al.'s model [37], also referred as Google's model, which has been trained on 100 billion words from Google News dataset. The feature vector is 300 dimensional.

Chapter 5

EVALUATION

We perform subjective and objective evaluation on our proposed approach.

5.1 Subjective

5.1.1 Experimental Setup

For the subjective part of the evaluation, we conducted experiments with three different affective targets. Marking significant points on the affective space, affective targets with *151*, *999* and *515* VAD values were selected for experimentation. The affective target *151* represents sentences that are negative, elicit moderate arousal and are low in dominance, such as sentences including words akin to *stress*, *death*, *torture*. Whereas, *515* 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 *515*: *routine*, *curtain*, *textbook*). The *999* 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 involve words *happiness*, *fun*, *excitement*.

The whole movie corpus was mapped into the aforementioned three affective targets by utilizing the approach proposed in section 4. Since not all of the utterances change affect during the mapping, we picked the utterances that were modified for human evaluation. 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.

Human raters evaluated a total of 225 pairs of utterances for the three targets,

75 pairs per target. 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 1 to 9 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.

5.1.2 Raters

We conduct the study with 12 participants. All of the participants were carefully selected based on their English-speaking capabilities (having a TOEFL iBT score higher than 110, 96th percentile). The female to male ratio of the raters is 7:5, they come from diverse cultural backgrounds (5 nationalities), and fall into 22-31 age group (mean 26.5). Initially, the participants went through a training to prepare them for the task, including rating a number of golden sentences to get accustomed. Hence, we regard the ratings provided by them of high quality.

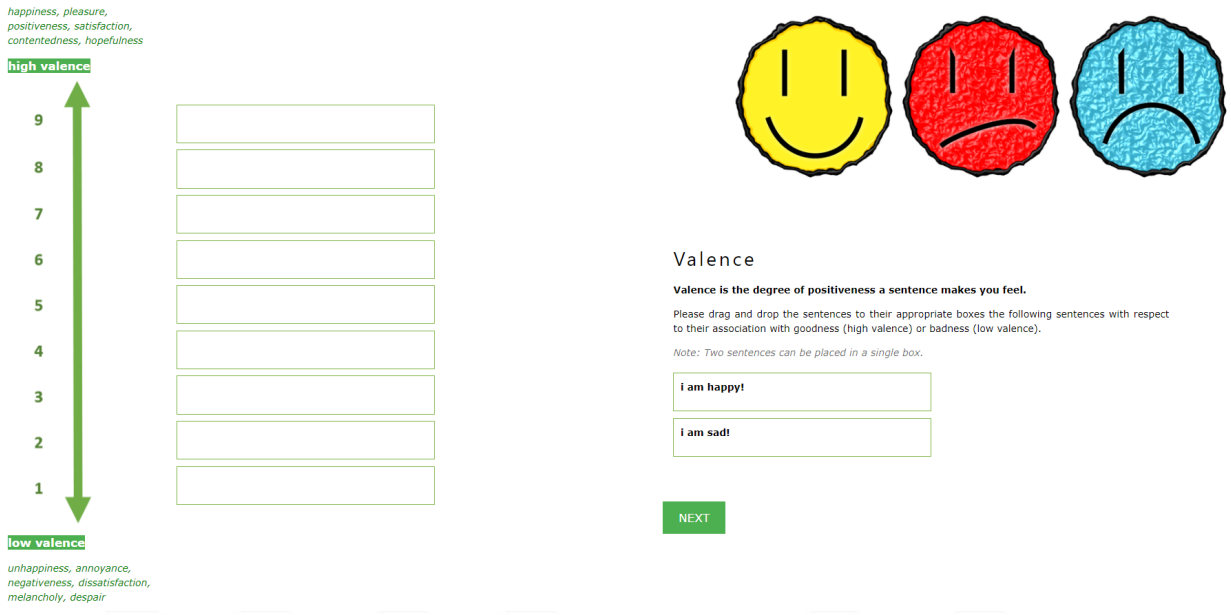
5.1.3 User Interface

The interface design is regarded crucial to and highly associated with the quality of ratings, as an inadequate design might lead to bias in rating tasks [38]. Consequently, instead of employing crowdsourcing platforms with generic interfaces (e.g. Amazon Mechanical Turk), we deliberately designed a customized web-based interface for the rating task. Another reason for not leveraging these platforms with anonymous workers, was that we preferred to train our raters on the study's concepts (e.g. valence, arousal, dominance) and the task itself to ensure qualitative ratings.

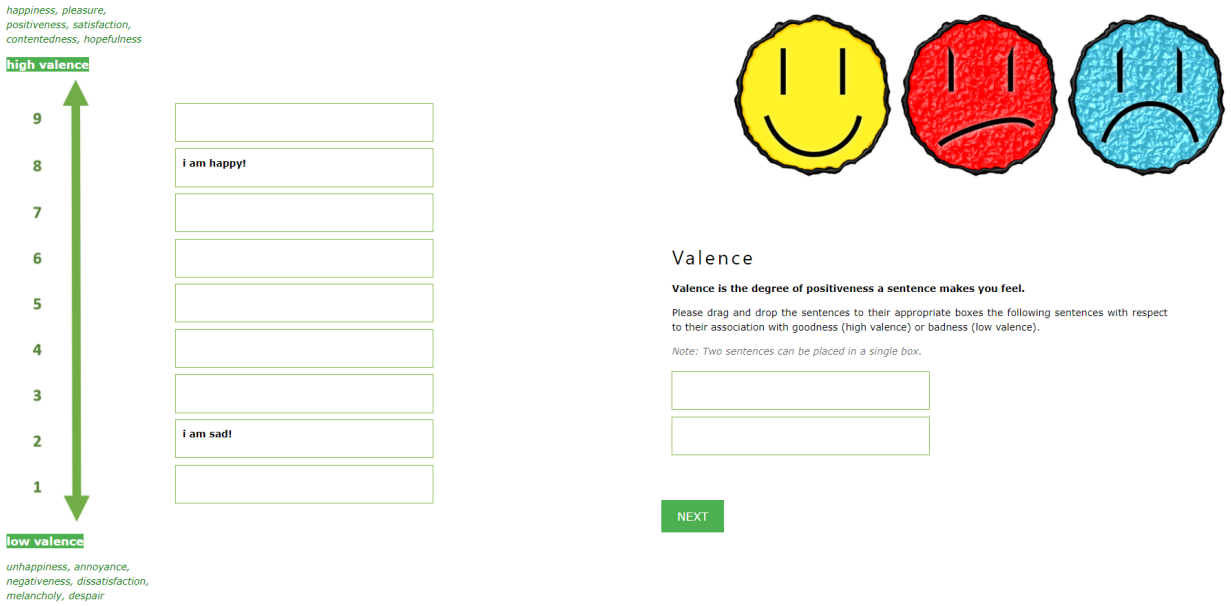
Having undergone the training, participants were ushered to a website for the rating task. The participants rated the utterances, in terms of valence, arousal, dominance, syntactic coherence and appropriateness. Two approaches can be taken

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 chose the second approach for the rating task.

Fig 5.1 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 and the words associated with the extrema of the scale, as an aide-mmoire for the raters. On the right part of the screen a pair of utterances is displayed. Following the work of [26] we use a scale from 1-9 and display it on the left part of the screen. Raters were asked to drag-and-drop the utterances on the scale on the left where they thought suitable. The two utterances could be placed on the same box (have the same score), if necessary. The rater would proceed by clicking the next button, until they scored all of the utterances.



(a) User prompted with a pair of utterances (the original one and its modified counterpart)



(b) User rates the utterances

Figure 5.1: User interface

5.2 Objective

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.

5.2.1 Language Generation

Generative models in general suffer from lack of established objective evaluation metrics [39]. For natural language generation, the most widely reported score is the *log likelihood* score the model assigns to the generated language. Thus, higher *log likelihood* scores indicate better language generation. However, there are a number of caveats to mind while comparing different models in terms of *log likelihood*, since vocabulary size and out-of-vocabulary (OOV) rate should be in the same range for fair assessments. To illustrate, in language modeling, the words in the dataset that fall outside the vocabulary are replaced with a special *unknown* token. Now, consider a dummy language model that has only the *unknown* token in its vocabulary. In this scenario, the language model will assign high *log likelihood* scores to every sentence, since they are all made of only from the *unknown* token. However, we know that the sentences are not meaningful. Still, in our scenario *log likelihood* provides an objective gauge for the assessment of our affectively generated sentences, considering the comparable scores originating from the same model.

Another broadly established metric in evaluation of conversational models with genesis from machine translation is BLEU (Bilingual Evaluation Understudy) score [40]. BLEU is a metric that evaluates how good is a generated sentence compared to a reference sentence. 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 a ground truth affective sentences. Therefore, we do not report BLEU score.

5.2.2 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 [41], [42]. 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)[42]. VADER is a rule-based system for sentiment 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 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.

Chapter 6

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 responses and their modified versions in different targets are provided in Table 6.

6.1 Subjective

The results of the subjective evaluation of utterances in terms of valence, arousal and dominance are shown in Fig. 6.1. Whereas, Fig.6.2. 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. For the statistical significance between the original and the modified utterances *Mann-Whitney U test* was used.

The results indicate that our approach was successful in pulling the mean of valence, arousal and dominance of the utterances towards the targeted affects. Particularly, *valence* shows statistically significant change ($p << 5e-2$) for every targeted affect. Since the selected original utterances are relatively high in valence prior to being modified (mean: 6.85), as expected the largest difference ($p << 1.5e-22$) is seen when the target valence is $V_t = 1$ in the target affect $a_t = [1, 5, 1]$.

Dominance also, exhibits statistically significant difference between the target affect $a_t = [1, 5, 1]$ and the original utterances. While the utterances are effectively pulled toward $D_t = 1$ and show a statistically significant difference from the original utterances ($p << 3.41e-6$), the mean is not in the vicinity of $D_t = 1$. A similar oc-

Table 6.1: Sample responses from the movie corpus and their affectively modified counterparts.

original utterance	It's been a real pleasure, Eve.
target VAD - 515	It's been a great experience, Eve.
target VAD - 999	It's been a wonderful delight, Eve.
target VAD - 151	It's been a little disappointment, Eve.
target VAD - 111	It's been a little sadness, Eve.
original utterance	It's Will, the really funny good looking guy you met at the bar.
target VAD - 515	It's Will, the really comfortable nice looking guy you met at the bar.
target VAD - 999	It's Will, the really hilarious strong looking guy you met at the bar.
target VAD - 151	It's Will, the really stupid tough looking guy you met at the bar.
target VAD - 111	It's Will, the really boring tough looking guy you met at the bar.
original utterance	Dang, you are right, Mr. Rothstein, I am so sorry.
target VAD - 515	Dang, you are good, Mr. Rothstein, I am so stupid.
target VAD - 999	Dang, you are good, Mr. Rothstein, I am so happy.
target VAD - 151	Dang, you are wrong, Mr. Rothstein, I am so angry.
target VAD - 111	Dang, you are wrong, Mr. Rothstein, I am so disappointed.

currence is also noticed for target dominance $D_t = 9$ (target affect $a_t = [9, 9, 9]$). This is due to the fact that there are very few words on the extrema of the scale in terms of dominance, as depicted in Figure 6.3.c. Consequently, though our approach changes the dominance of the utterances toward the desired direction, it cannot reach an extreme target as there exist no plausible words there. In the target affect $a_t = [5, 1, 5]$, this is not the case, as a significant portion of the words fall into the proximity of $D_t = 5$ (see Figure 6.3.c.). Hence, our approach has a wide range of choices, and is successful in reaching the target dominance subsequently.

In contrast to valence and dominance, in word arousal distribution, a substantial amount of words is accumulated in lower parts of the scale (Figure 6.3). This lack of words high in arousal is perceived in the results of extreme target $a_t = [9, 9, 9]$, with $A_t = 9$. Again, the original utterances are pulled toward $A_t = 9$, but the extremum is unachievable. For the target affect $a_t = [1, 5, 1]$, the utterances are pulled close to $A_t = 5$, since there are abundant words in the proximity. While we would expect a lower arousal for the target affect $a_t = [5, 1, 5]$, results indicate a relatively small decrease, though statistically significant ($p << 7.43e - 3$).

As expected, the modified utterances were not as syntactically coherent as their original counterparts. However, their mean is above 1.5 on a scale from 0 (not grammatically correct) to 2 (grammatically correct) compared to 1.8 of original utterances, which indicates that our model perturbs the grammar of the utterances only slightly. Especially, when the target affect is close to the original affect of the utterances, as it is the case with $a_t = [5, 1, 5]$, we observe even smaller detriment in terms of syntactic coherence.

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 1.47 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. $a_t = [5, 1, 5]$). We observe

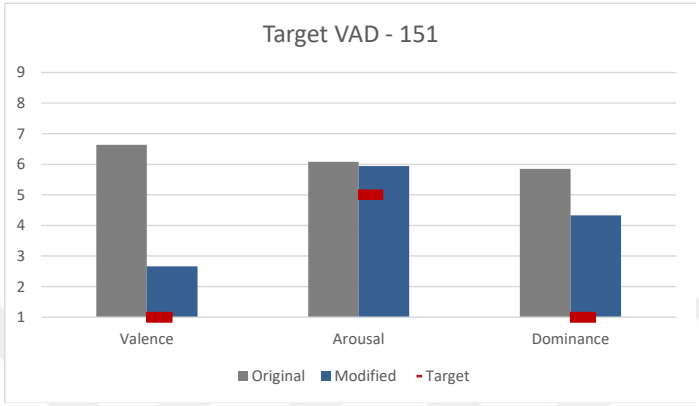
that the appropriateness score deteriorates when the difference between target and original affects increases, probably due to the fact that in the experimental setting we self-define the target affect, which might cause abrupt changes in the conversation. Nonetheless, the application of this approach is intended for settings, where the target affect is extracted from environmental factors (i.e. visual or speech cues). In this scenario, the generated responses would be more appropriate in terms of affect.

6.2 Objective

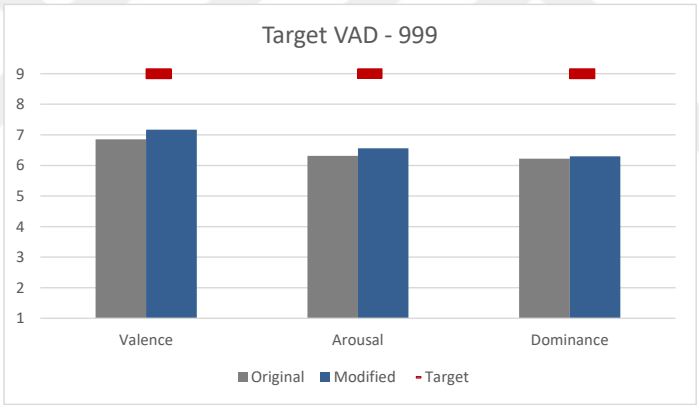
As a baseline for the objective evaluation, we utilize our approach without employing any language model. In other words, in the equation 4.1, λ is set to 1. Note that although no language model is leveraged in this case, implicitly word embeddings hold information regarding the context of the word, just as a language model does.

Each point in Fig. 6.4. depicts the average VADER score and log likelihood over all movie corpus (304, 713 utterances). Comparing the sequence-to-sequence model with the baseline, we observe that since the baseline has no language model constraint, it has a wider range of VADER scores (valences) than the sequence-to-sequence model. However, the baselines log likelihood scores are inferior to those of sequence-to-sequence models. Log likelihood scores imply that our approach performs the affect modification without a huge deterioration in language structure. Among those, again we observe that targets closer to original affect have better scores.

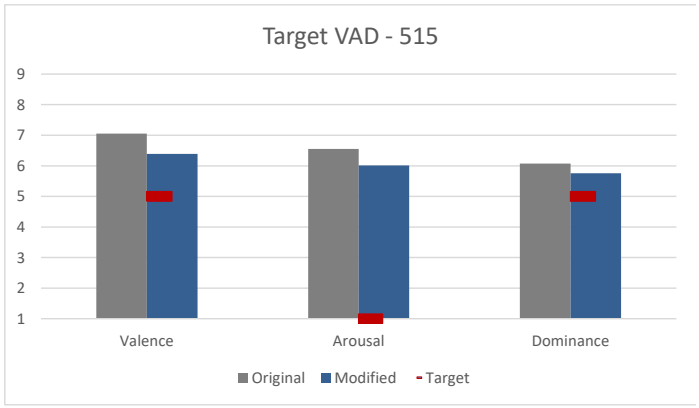
For the sequence-to-sequence case, as expected the original utterances and the modified utterances with the neutral target affect $a_t = [5, 1, 5]$ have the same average compound score 0.04, which indicates neutral valence. Nonetheless, this score indicates that the utterances are closer to the positive valence. This positiveness of the original sentences was also conspicuous in the subjective results. AffectON with sequence-to-sequence exceeds the valence thresholds of VADER compound scores (-0.05 and +0.05) for the affective targets $a_t = [1, 5, 1]$ and $a_t = [9, 9, 9]$, respectively. For the target affect $a_t = [1, 1, 1]$, although the VADER threshold -0.05 is not exceeded, the utterances are pulled toward the negative valence.



(a) Target VAD - 151



(b) Target VAD - 515



(c) Target VAD - 999

Figure 6.1: Human ratings of valence, arousal and dominance for affective targets 151, 515 and 999

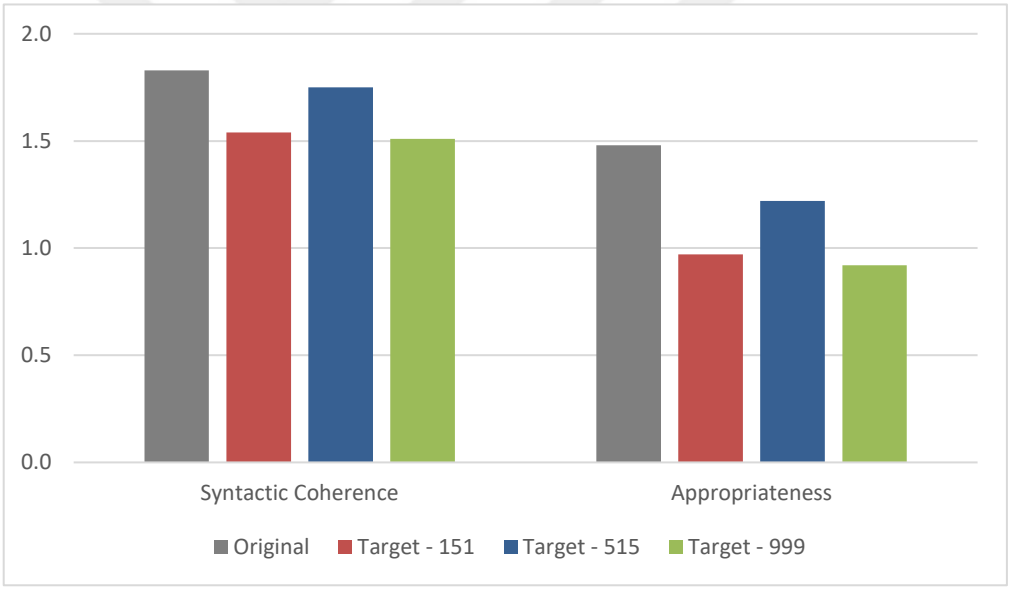
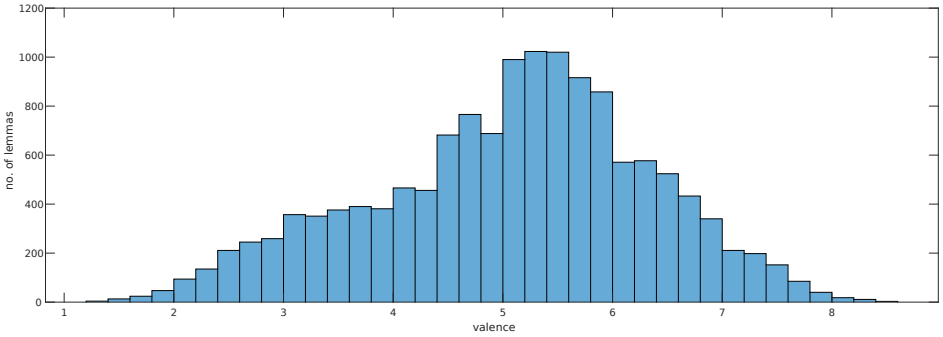
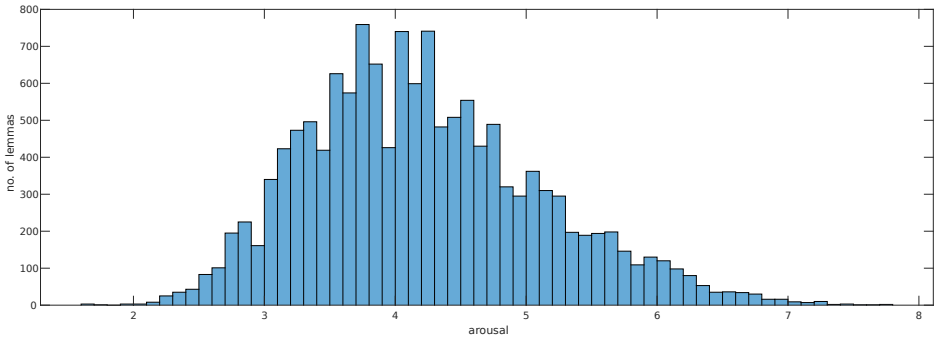


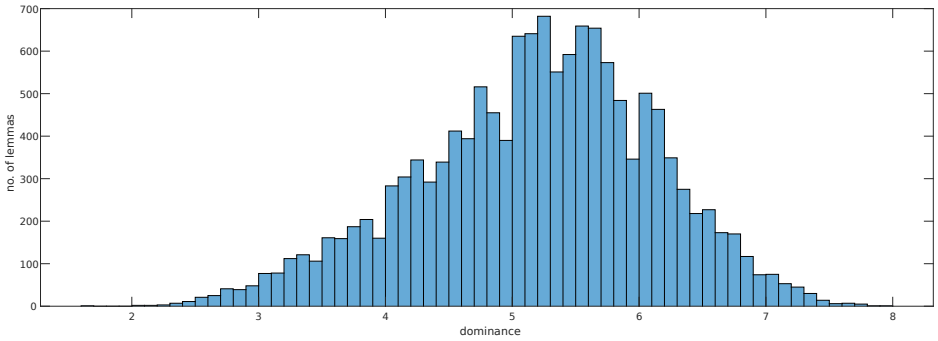
Figure 6.2: Human ratings of syntax and appropriateness for affective targets 151, 515 and 999



(a) Valence



(b) Arousal



(c) Dominance

Figure 6.3: Distribution of 13,915 lemmas in affective lexicon according to their VADs.

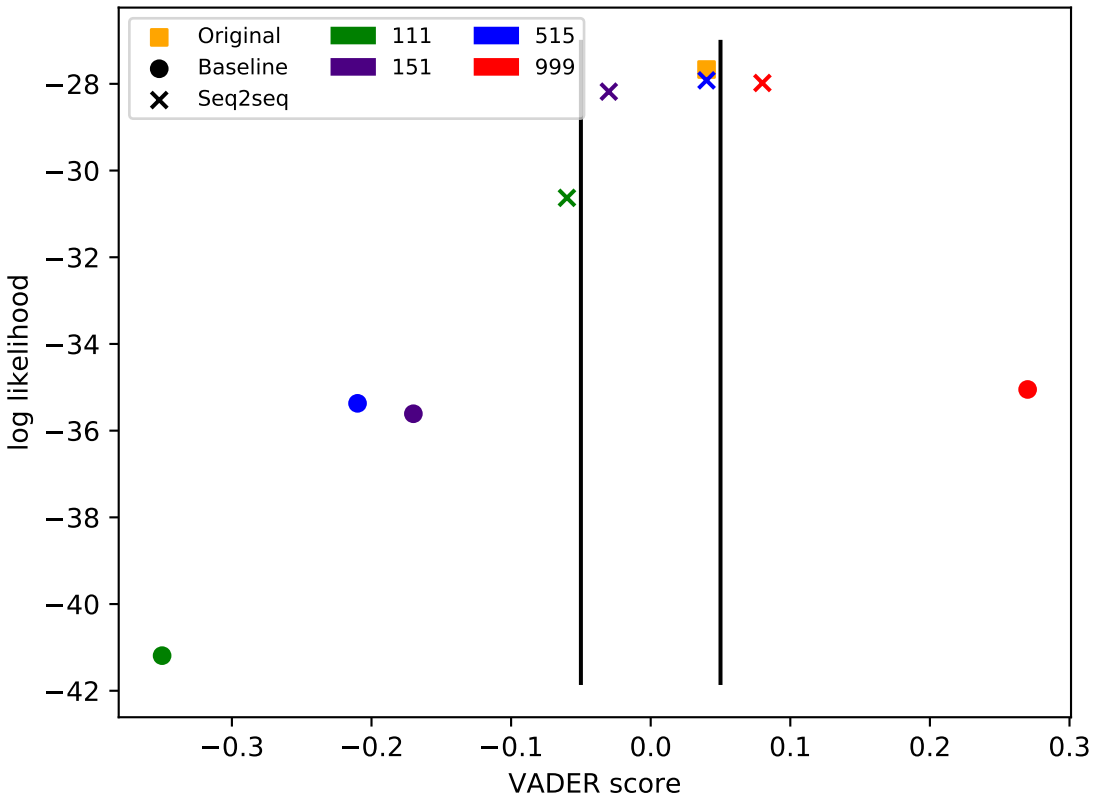


Figure 6.4: Objective Results. The two vertical lines depict the VADER score boundaries for positiveness ($>= 0.05$) and negativeness ($=< -0.05$).

Chapter 7

LIMITATIONS AND FUTURE WORK

Despite its promises for affective language generation, this study has some limitations. First, the sequence-to-sequence model was trained on a movie corpus, which is not a representative dataset for naturalistic daily conversation. Experiments with much larger and more realistic conversational datasets would lead to a more robust sequence-to-sequence model. In such scenario, the model would have learned the underlying conversational distribution better and the word embeddings would not be necessary. Simply put, in our approach word embeddings augment the language model regarding the context of the words, which would be unnecessary with an ideal language model. Hence, one of our future research directions is experimenting with superior datasets.

Additionally, negation words (e.g. *not*) induce an issue, since we apply our approach word per 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 on our approach. A simple solution would be a rule-based consideration of these words during the affect modification.

In terms of application, we plan to incorporate *AffectON* to our robot's conversational model. The target affect for every response, will be extracted from environmental factors such as the human's affective state. The affective state will be computed by a separate audiovisual affect recognition module. In this area, there are many pathways to explore, such as choosing the optimal strategy for selecting the affective target. Another alluring task would be learning the affective target strategy from

multimodal datasets.

Chapter 8

CONCLUSION

In this paper, we presented AffectON, an approach for generating affective language. Particularly, we experimented with affective dialog generation. We leveraged sequence-to-sequence language models, a word2vec space and an affective space to generate affective responses. Both subjective and objective evaluations were conducted to assess our approach. We observe that AffectON can successfully pull text to a targeted affect, with a small sacrifice in terms syntax.

Appendix 1

APPENDIX

A.1 Detailed Results

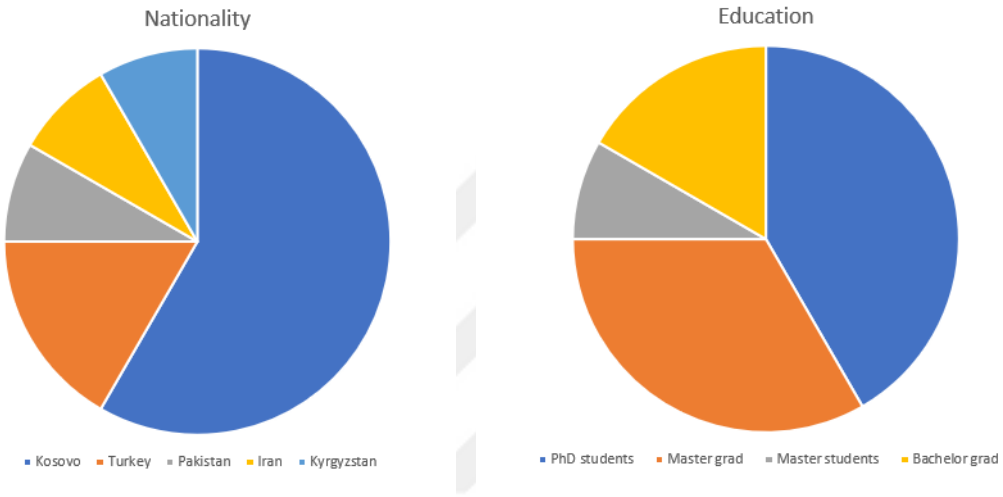
Table A.1: Objective Evaluation Results

	mean-VADER	**std - VADER**	**mean - LM score**	**std - LM score**
original	0.04	0.40	-27.67	29.75
		affective without LM (baseline)		
target VAD - 151	-0.35	0.57	-41.19	44.05
target VAD - 111	-0.17	0.42	-35.61	37.38
target VAD - 515	-0.21	0.47	-35.37	37.10
target VAD - 999	0.27	0.50	-35.05	36.68
		affective with sequence-to-sequence		
target VAD - 151	-0.06	0.40	-30.63	31.10
target VAD - 111	-0.03	0.38	-28.18	30.46
target VAD - 515	0.04	0.38	-27.92	30.11
target VAD - 999	0.08	0.39	-27.98	30.20

Table A.2: Subjective Evaluation Results

	target VAD - 151				
	mean-mapped	**std-mapped**	**mean-original**	**std-original**	**p-value**
valence	2.66	1.52	6.64	1.93	**1.53E-22**
arousal	5.94	1.84	6.08	1.55	4.50E-01
dominance	4.33	2.17	5.85	1.99	**3.41E-06**
syntactic coherence	1.54	0.64	1.82	0.36	**1.09E-03**
appropriateness	0.97	0.83	1.62	1.12	**3.26E-05**
	target VAD - 999				
valence	7.17	1.50	6.85	1.48	5.54E-02
arousal	6.56	1.67	6.31	1.55	1.21E-01
dominance	6.30	1.71	6.22	1.55	3.83E-01
syntactic coherence	1.51	0.72	1.83	0.46	**7.27E-04**
appropriateness	0.92	0.84	1.39	0.73	**1.90E-04**
	target VAD - 515				
valence	6.39	1.57	7.05	1.65	**1.42E-03**
arousal	6.01	1.61	6.55	1.65	**7.40E-03**
dominance	5.76	1.63	6.07	1.74	**4.15E-02**
syntactic coherence	1.75	0.52	1.85	0.40	8.76E-02
appropriateness	1.22	0.8	1.44	0.7	**5.00E-02**

A.2 Rater Demographics



(a) Nationality

(b) Education

(c) Gender

Figure A.1: Rater Demographics

BIBLIOGRAPHY

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

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

[4] 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*, pp. 2259–2266, IEEE, 2017.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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*, pp. 17–46, Springer, 1996.

[13] 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.

[14] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2018.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[16] H. Jhamtani, V. Gangal, E. Hovy, and E. Nyberg, "Shakespearizing modern language using copy-enriched sequence-to-sequence models," *arXiv preprint arXiv:1707.01161*, 2017.

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

[18] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," *arXiv preprint arXiv:1703.00955*, 2017.

[19] 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*, pp. 6830–6841, 2017.

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

[21] 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.

[22] 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*, pp. 12–21, 2011.

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

[24] 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.

[25] 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, pp. 49–54, 2018.

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

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

[28] 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.

[29] 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.

[30] S. Jiang and M. de Rijke, "Why are sequence-to-sequence models so dull? understanding the low-diversity problem of chatbots," *arXiv preprint arXiv:1809.01941*, 2018.

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, p. 533, 1986.

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[33] 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.

[34] 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.

[35] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[36] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, pp. 1310–1318, 2013.

[37] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[38] 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*, pp. 585–592, ACM, 2003.

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

[40] 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*, pp. 311–318, Association for Computational Linguistics, 2002.

[41] 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.

[42] 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.