Facial Expression Grounded Conversational Dialogue Generation

Bernd Huber[1,2], Daniel McDuff[2]
[1]Harvard University, Cambridge, MA, USA
[2]Microsoft Research, Redmond, WA, USA
bhb@seas.harvard.edu, damcduff@microsoft.com

***Abstract*— We present a novel conversational language model that is grounded with information about facial expressions. To our knowledge this is the first in-depth examination of grounding natural language models with facial cues. We train a neural language model that uses automatically detected facial action unit intensity information in images alongside text to generate conversational dialogue. We evaluate our model on a large and very challenging unconstrained real-world dataset from social media (Twitter), featuring 450,000 conversations with associated facial expressions. Systematic linguistic and crowdsourced analyses reveal the properties of our models: The facial expression grounding strengthens the sentiment of the resulting dialogue such that it is consistent with the valence of the facial expressions. Furthermore, the automatically generated conversational responses are rated as equivalent to the human gold-standard responses on relevance and emotion dimensions.**

I. INTRODUCTION

Conversational agents (CA) are becoming increasingly common. These agents can help people perform simple tasks such as sourcing a weather report, setting an alarm, or playing music with increased ease and efficiency. For such an agent to be truly useful and easy to engage with, CA need to be natural to interact with and respond to contextual information, such as non-verbal cues [6], [15], [23]. In fact, the main aim of certain new dialogue agents is to be able to be a chat companion and engage in open-ended dialogue with a user (e.g. *XiaoIce* [41].) However, this type of unstructured dialogue is challenging to generate and evaluate. While Natural Language Processing (NLP) continues to improve it is often found that resulting models have similar drawbacks. The range of possible responses to a question in an open-ended conversation can be quite large and models trained exclusively on text tend to provide rather non-committal responses such as "I don't know".

In real-world settings conversations are almost always associated with visual imagery. Contextual information from this imagery could include objects or people but also facial expressions. Affective information plays a particularly important role in many social interactions. Non-verbal expressions are critical for social functioning [34]. In conversations, facial expressions may serve as affective context that informs interlocutors about the emotion of the situation/conversation. Then why not include such contextual information in the dialogue generation process?

The relationship between conversation and facial expression, which includes non-verbal cues in natural language dialogue generation, is a challenging modeling and common sense reasoning problem. If a machine is to learn how to combine non-verbal cues and language understanding it will need a large amount of data about how humans behave.

978-1-5386-2335-0/18/$31.00 ©2018 IEEE

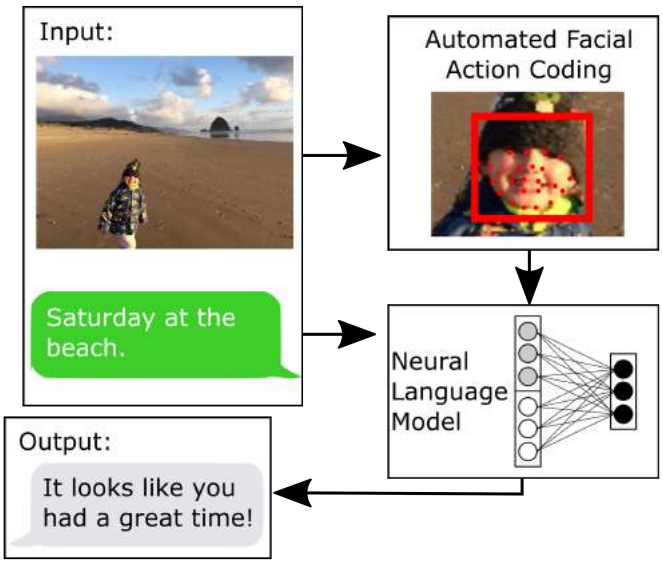

Fig. 1. We present a novel facial expression grounded conversational dialogue generation system. Our model leverages automated facial coding and textual context to generate dialogue that is closer to the sentiment in the associated images.

Affective crowdsourcing [31] has been demonstrated as a highly effective way of collecting large volumes of affective data "in-the-wild". Social media platforms provide access to large-scale conversational data between individuals. Many of these conversations (28% on Twitter [30]) are accompanied by images. Of these about 60% include at least one human face. Natural data posted on social media is unconstrained in terms of topics and the use of grammar and vocabulary. In addition, imagery (and facial expressions) are highly varied and contain lighting and pose challenges. However, this data has high ecological validity, and therefore represents an extensive and valuable source of training examples from which machines can learn conversational skills.

To computationally model textual context and facial expressions we need to choose a taxonomy for coding facial behavior. The facial action coding system (FACS) [12] is the most widely used and comprehensive taxonomy for such purposes. FACS provides a natural set of features for

modeling as each action can be clearly described. Manual FACS coding is laborious and requires trained coders. Automated FACS coding provides a highly scalable alternative to manual annotation, albeit at the loss of some precision when compared to expert human coders.

Prior work on understanding the complex meaning of facial expressions has often been constrained by the limited scalability of research methods that involve data collection in lab-based settings. Complementing lab-based analyses, large analyses of unconstrained ecologically valid data sources using automated techniques have been conducted. Research of this kind has revealed cultural [26], age [25] and gender [26] differences in facial responses. We leverage large-scale social media data to build a joint generative model of facial actions and language. Using this model we are able to explore how facial actions are related to language sentiment in associated conversations and find which words/topics are related to which facial actions.

In this paper we focus on the task of generating responses to one conversational turn that includes a facial image (see Figure 1). Specifically, we present a model for generating responses that can leverage facial expression understanding. Our aim was to train a language model that produces dialogue that is more emotionally relevant to the original posts and images. Thus, the main contributions of this paper are to: 1) present the first facial expression grounded conversational dialogue generation system, 2) evaluate this model on a very large, real-world dataset of text and images with facial expressions from social media, 3) characterize the performance of an open source facial action coding tool [4] on social media and web images, 4) analyze the impact of the facial features on generated dialogue.

In the remainder of this paper, we present our conversational language model with image grounding and describe the social media dialogue dataset. We then evaluate our model using human judgments and linguistic analyses, and discuss the implications for CA.

II. RELATED WORK

Our work builds on research in affective computing, computer vision (specifically automated facial coding) and natural language processing. To our knowledge this is the first system that models facial expressions in natural language conversation generation.

A. Dialogue Generation

Deep neural networks (DNN) have proven to be very successful for open-ended dialogue generation. These networks commonly model conversations as a problem of predicting the next sentence (response), given the previous conversation (context). The context may consist of one or multiple turns. A widely adopted DNN approach to this problem is a sequence-to-sequence architecture (seq2seq) [36], [40], [22].

Combined modeling of imagery and language can take a number of forms. Image captioning aims to generate a salient textual description of an image. Recent work has effectively combined these models with attention networks such that the salient objects/features in the images/videos can also be highlighted [42]. Furthermore, previous work has shown how to effectively encode image and text together for sentence prediction tasks [20].

In AI research, there has been increasing interest in CA that allow for multi-modal user input, such as the combination of language and visual information [9], [15]. Various tasks have evolved around the combination of visual imagery and language. Visual Question Answering (VQA) focuses on generating answers to questions about images. This relies on the question being answerable from the contents of the image alone. A different approach to VQA is to generate the questions themselves [32]. We define Image-Grounded Conversation (IGC) as using images as contextual information in conversations. This task is different from Visual Question Answering (VQA), which aims at generating answers directly about contents of the image. Thus, we do not feel VQA is a good baseline to compare against. IGC creates an interesting paradigm where the questions could be related to the image but not specifically about the image content. In this paper, we focus on dialogue, requiring the model to generate responses using the caption of the image and the image as context.

B. Affective Computing

The role of facial expressions as non-verbal cues in social interactions has been studied in-depth in psychology. With the rise of automated methods for coding facial expressions the affective computing community has been able to increase the repeatability and scalability of this research and at the same time reveal new scientific insights. Specifically relevant to our work is the design of multi-modal systems that model language and non-verbal cues together.

Prior work has typically focused on the fusion of language and non-verbal cues for the automated recognition of affective states [7], [8], [10], [33]. In these models the features can be fused early, creating a joint representation of the linguistic and visual information, or late, using a form of voting on the predicted affect from each modality. Meta-analyses have shown that multi-modal models consistently outperform unimodal models in affect detection tasks [11]. For a relevant recent survey of multi-modal machine learning methods see Morency et al. [2]. This body of research clearly demonstrates a link between language and facial expressions.

C. Automated Facial Coding

Due to the laborious nature of manually annotating FACS it has become a necessity to leverage automated methods for any scalable application of facial coding. Over the past 20 years these methods have improved considerably. While the focus of our work is not a new model for facial action unit detection, it is helpful to provide some background on the state-of-the-art. For a recent comprehensive review see [38].

Results from the FERA challenges have shown consistent improvement in facial coding methods. The most recent focusing on performance across a broader range of head

poses [39]. Most methods proposed recently rely on convolutional neural network (CNN) models [18], [14] that perform well given a sufficiently large number of training images. These networks do not need prescribed feature descriptors but rather learn both a feature representation and classifier. Beyond the architecture of the models themselves, research has shown that combining multiple datasets [3] and leveraging large volumes of "crowd sourced" data from many different people can significantly improve accuracy [35], [28].

III. DIALOGUE MODEL

We propose a neural language generation model that combines textual inputs and facial action features to generate natural responses to dialogue. We compare the face-grounded model (text-image) to a model trained on text alone (text-only). Figure 2 shows the architecture of each model.

Textual Model (Text-Only): This model maps an input sequence (one or multiple turns) to an output sequence (Seq2Seq model ([37]) using an encoder and a decoder recurrent neural network (RNN). The initial recurrent state is the 500-dimensional encoding of the textual context. GRU stands for gated recurrent units.

Textual and Facial Expression Model (Text-Image): The textual feature vector is obtained using an RNN. The vector is then concatenated to the action unit feature vector and fed into a fully connected (FC) feed forward neural network. The result being a single 500-dimensional vector encoding both facial action and textual context, which then serves as the initial recurrent state of the decoder RNN.

Training of the models and the analysis of the output dialogue is described in Sections V and VI.

IV. DATA

A. Social Media Conversations

We sourced a total of 450,000 conversations on Twitter. The definition of a conversation for our purposes is at least two tweets from different Twitter accounts, and an image with only one face associated with it. Our crawling revealed that approximately 60% of multi-turn conversations with images on Twitter had one face in the image. We performed experiments training on the 400,000 samples (with face images) and testing on 50,000 samples (with face images). The conversations were not filtered in any other way and represent highly naturalistic social media dialogue. From the images associated with the tweets we extracted facial action unit measures for 17 facial actions as described below.

B. Automated Facial Coding

We used facial coding software to extract the facial actions of the faces within the images [4]. The software provides an intensity score for 17 facial actions based on FACS. Numbers, names and examples of these facial actions are provided in Figure 3.

Validating OpenFace on Web Images Given the challenging nature of automatic detection of facial actions in real-world images, it is important we characterize the performance of facial coding algorithms. We collected an independent set of 10,000 images from social media and Internet search. These closely mirror the types of images found in Twitter posts (e.g. images of celebrities, selfies and personal photographs.)

TABLE I

F1-SCORES FROM THE OPENFACE [4] FACIAL AU DETECTION ON OUR WEB IMAGE DATASET. WE REPORT THE STATIC CLASSIFIER RESULTS AS WE APPLY THE ALGORITHMS TO TWITTER IMAGES IN OUR DIALOGUE ANALYSIS. DISFA [24], BP4D [43] AND SEMAINE [29] RESULTS TAKEN FROM [4].

AU	Action	DISFA, BP4D Semaine [4]	10,000 Web Images	
		F1	N	F1
1	Inner Brow Raise	.27	4,354	.47
2	Outer Brow Raise	.02	3,605	.48
4	Brow Furrow	.66	4,721	.55
5	Eye Widen	.55	2,538	.39
6	Cheek Raise	.41	4,858	.56
7	Lid Tighten	.75	5,942	.59
9	Nose Wrinkle	.23	1,755	.34
10	Upper Lip Raise	.68	4,202	.52
12	Lip Corner Pull	.87	2,667	.50
14	Dimpler	.38	1,663	.26
15	Lip Depressor	.05	1,337	.22
17	Chin Raise	.32	1,930	.32
20	Lip Stretch	.30	1,654	.25
23	Lip Press	.36	1,045	.16
25	Lips Part	.85	9,037	.74
26	Jaw Drop	.53	3,711	.43
43	Eyes Closed	.31	NA	NA

The images were coded by FACS trained experts. We will refer to this dataset as the Internet Labeled Image AU dataset (ILIAD). Table I shows the F1-score of the OpenFace static facial action classifiers on data from the DISFA [24], BP4D [43] and Semaine [29] datasets (scores taken from [4]) and ILIAD. We did not have the ability to code intensity values on the Internet image dataset and therefore evaluate the detection performance only. These results show that the OpenFace AU detectors work acceptably well across most of the action units. For AU 2 and AU 15 in particular the results on the web images are better than on the other datasets. However, we should be careful about drawing conclusions from the AUs with weak performance. In theory our neural language model should learn which action unit observations provide a reliable signal and which contain noise, but this validation provides extra evidence to help us interpret the outputs of the model.

V. TRAINING

In our analysis we trained two models. The text-only model and the text-image model. The models were trained with 400,000 conversations that were all social media posts with associated images. The text-only model only received the text portion of the conversation as input, the text-image model received the 17 action unit intensity values in addition to the text as input. We trained our models using a Tensorflow implementation on Titan X Graphics Processing Units. The

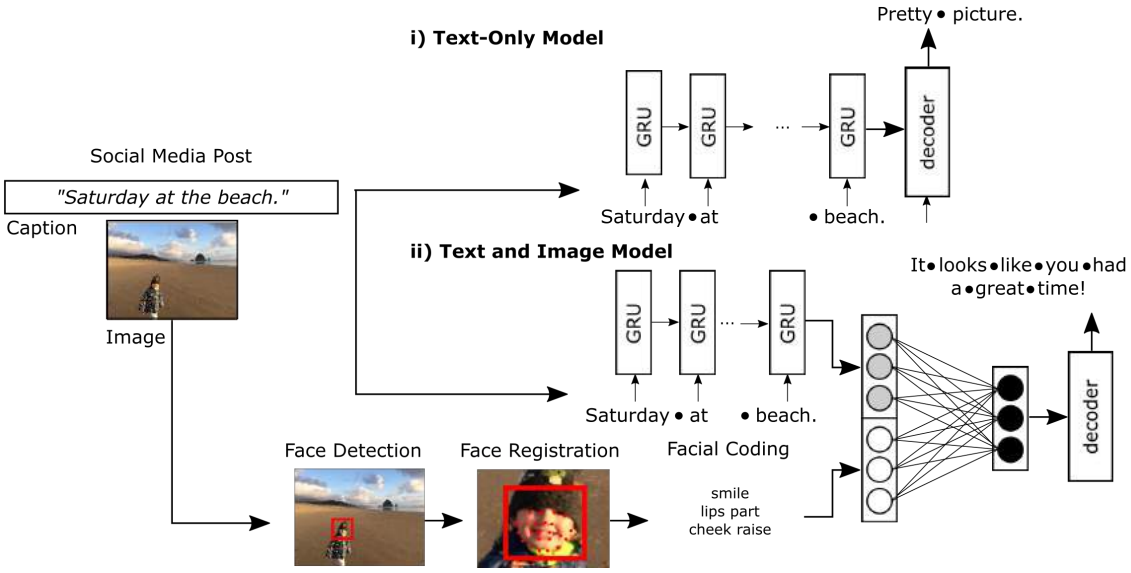

Fig. 2. The neural models that we used to generate dialogue. (i) The textual model captures the previous turn of a conversation and outputs the next turn in that conversation. It makes use of gated recurrent unit (GRU) cells that capture the time and context information in the text. (ii) The model that integrates the facial action coding. Faces are detected in the images and automated FACS coding performed. The architecture appends the FACS features with the textual information, which then ows into the decoder architecture.

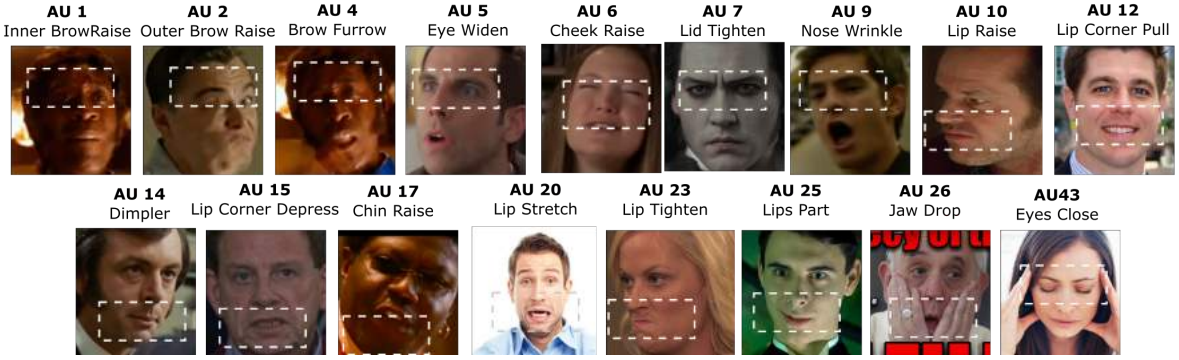

Fig. 3. Examples of the 17 facial actions we used to ground our dialogue model. The images for this figure are taken from the web image dataset (ILIAD) used for validation of the performance of the automated facial coding.

models were each trained for 500 epochs which took approximately 48 hours. The size of the 500-dimensional vector encoding of the textual and visual inputs was optimized during extensive prior training of dialogue systems. We trained all sets of weights using stochastic gradient descent with an exponentially decaying learning rate. We used early stopping and dropout to prevent overfitting. The perplexity of the models during training can be seen in Figure 4. It is clear that the models converge successfully. We did not observe over-fitting at any point in the training.

VI. RESULTS

A. Linguistic Analysis

We performed a rigorous computational analysis of the full 50,000 responses in our test set. Specifically we analyzed the impact of face features on the output of the model. Our target was a highly interpretable analysis, considering the large space of possible responses to a social media post. Therefore, we focus on relevant measures of:(1) language

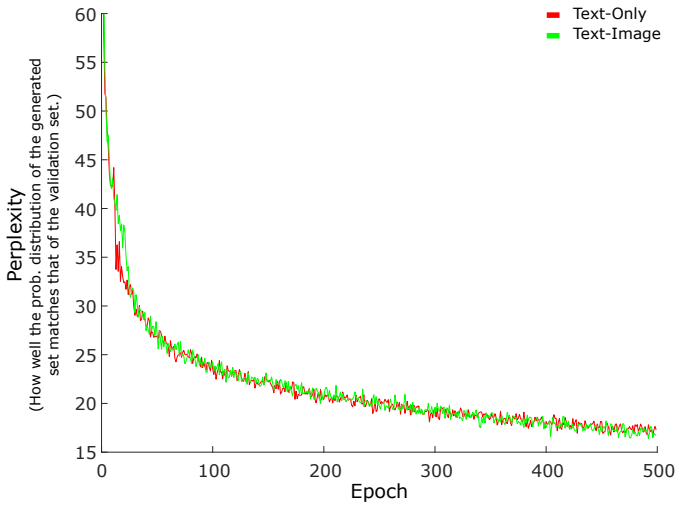

Fig. 4. Perplexity of the models for each training epoch (epoch = pass through the data). The models were each trained for 500 epochs.

sentiment, and (2) topic modeling. The following paragraph describes the methods we used and the results.

As a basis for the analyses, we used the 50,000 conversation samples held out for our test-set. We evaluated the impact of a facial action by selecting the 3,000 conversations in which the action was most intense. We computed sentiment and topic scores for the computer generated responses for those conversations. We repeated this for each action and compared the values and distributions. We also repeated the analysis for each of the models.

Sentiment Analysis: The sentiment analysis involved calculating the textual sentiment (from -1.0 (very negative) to 1.0 (very positive)) of the generated responses described above. We used the sentiment classifier presented by Hutto et al [17]. Bare in mind that the same conversations were analyzed for each model; however, naturally the generated response for each model are different.

Figure 5 (B) shows the average response sentiment (y-axis) over all samples, given the activation of one specific face action unit (x-axis). Pairwise comparisons of the average (mean) sentiment between text-only and text-image models (using a two-tailed Kolmogorov-Smirnov test) show that differences are significant for AU 6 ($p = 0.0236$), AU 7 ($p = 0.0399$), AU 12 ($p < 0.0001$), AU 17 ($p = 0.0034$), and AU 45 ($p < 0.0001$). Marginally significant differences also exist for AU 15 ($p = 0.0668$) and AU 26 ($p = 0.0783$).

To further frame this outcome we calculated the correlation between the generated response sentiment and the feature intensity value. We performed this analysis across the complete set of 50,000 conversations. Table II indicates whether the text sentiment of the responses was positively (pos), negatively (neg) or not correlated (N.S.) with the action unit intensity. The results were significant for all actions, with AUs 2, 5, 6, 7, 10, 12, 14 and 25 being positively correlated (i.e. the text sentiment increased in positive valence as these actions became more intense) and AUs 1, 4, 9, 15, 17, 20, 23, 26, 43 being negatively correlated (i.e. the text sentiment decreased in positive valence as these actions became more intense.) For context we also highlight how these facial actions might be associated with positive and negative emotional valence interpretations based on two sources from prior literature [19] and [13]. Our language results are broadly consistent with these valence associations. To illustrate how action unit intensity actually changes the language, we provide some qualitative examples of response pairs of low and high intensities of the action units. Table II shows pairs of response samples from the test corpus.

Topic Analysis: To get an overview of the generated topics associated with each facial action, we again analyzed the generated responses for each action unit separately. We used the 3,000 conversations with highest AU intensity for each action. We then ranked words in the generated responses based on term-frequency inverse-document frequency (TFIDF) scores. To quantify the topic difference between the text-only and text-image models, we use a latent Dirichlet allocation (LDA) topic modeling to compute the difference in topic from no face feature input to high feature intensity. Figure 5 (A) shows the Likelihood for every action unit separately. This gives us an idea for how much impact a facial action unit actually has on the topic of the output text. The lower the average topic likelihood, the more variance there is in the topics. We found that topics varied most with AU 5, 17 and 10 and varied least with AU 12. It is possible that responses to a smile (AU 12) are more homogeneous than responses to other actions that might be associated with multiple emotional states (e.g. AU 5 is associated with surprise, fear and anger.) Overall, the topic analysis is less interpretable than the sentiment analysis and thus it is hard to draw many conclusions about the nature of the effects.

B. Human Judgment

Human judgments of the computer generated dialogue responses were performed using a crowdsourcing task. As we only modeled English language dialogue in this study we recruited native English speakers from the US and the UK using a crowdsourcing platform.

It would be infeasible to manually annotate the full set of 50,000 conversations that were used for testing the model. Therefore, we randomly selected a subset of 350 for manual annotation. Those in which the face was largely obscured were removed leaving 312 for testing. In the task, the crowdworkers were given a dialogue sample, caption (C) and response (R), and a corresponding image (I). Each sample consisted of the image and caption from Twitter followed by the computer generated response from one of the models. We ran a similar process to evaluate the gold-standard (real) responses from Twitter. The workers were paid five cents (US) for coding one response. Each response was evaluated by seven crowdworkers.

Figure 6 shows the crowdsourcing task interface. Workers were asked to rate the response on four metrics: (1) the overall emotionality, (2) the general relevance to the image, (3) the similarity of the emotion in the response and the image emotion, and (4) the similarity of the emotion in the response and the caption emotion. Each question was evaluated on a seven-point Likert scale.

We use two measures for assessing the agreement between the coders. First, a weighted Krippendorf alpha (α) [1]. Second, a Kappa Q (κ), that is a generalization of Bennet et al.'s S score [5] proposed by Gwet [16]. Both are suitable for ordinal scales. Table III shows the α and κ for each question. The effective reliability of the annotators is moderate to high across all metrics. Given the highly unconstrained data and open nature of the responses the metrics provide confidence that there is sufficient agreement between the coders. As expected the agreement for the Gold (human) responses was slightly higher than that for the machine responses. The κ for the Gold responses for the four metrics (1-4) was 0.802, 0.761, 0.810 and 0.726 compared to 0.691, 0.704, 0.621 and 0.523 for the computer responses.

The mean score for each question is shown in Table IV. The only significant difference ($p<0.01$) was in how emotional the responses were. The text-only and text-image models produced responses that were rated as significantly

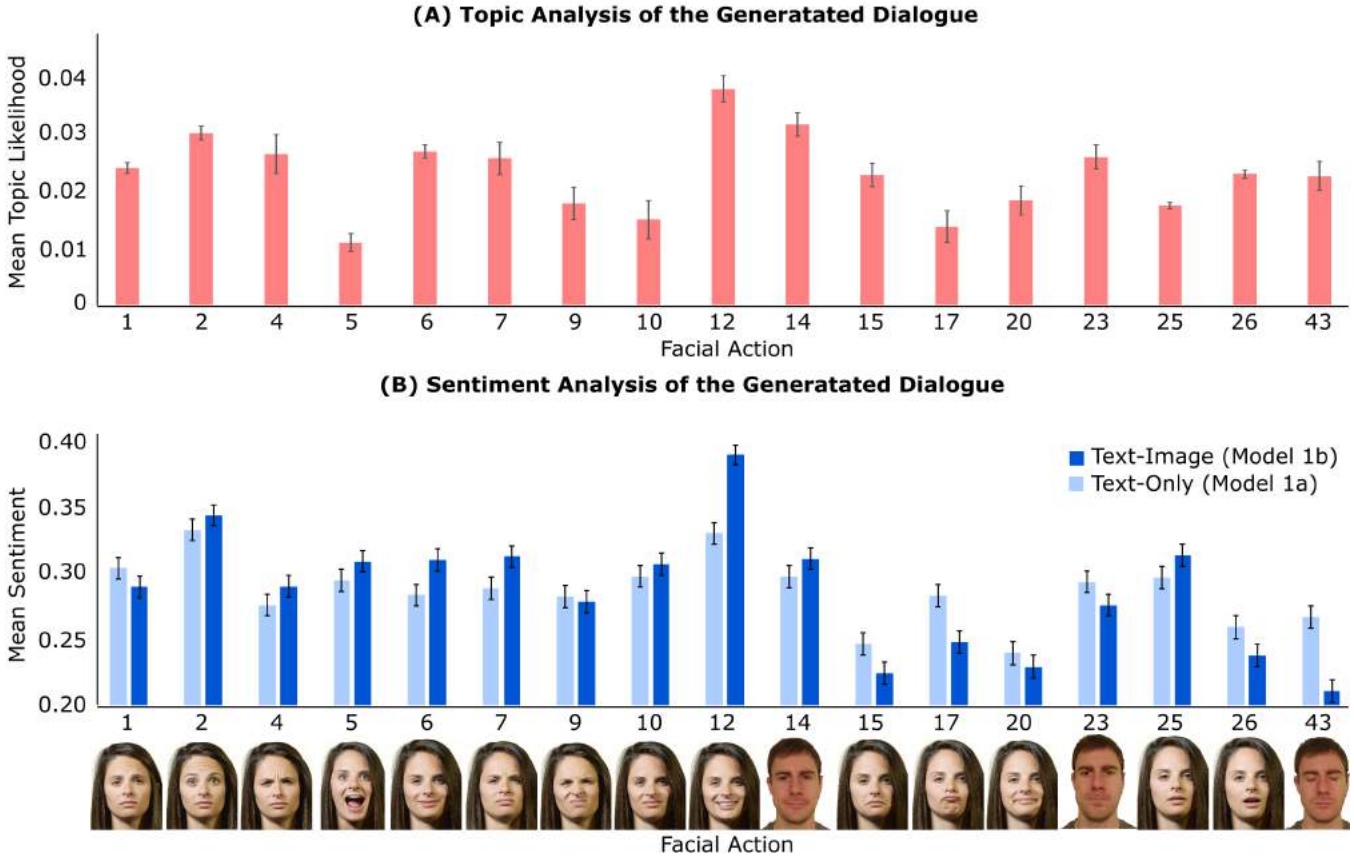

Fig. 5. (A) Average likelihood for topic model. The higher the value, the more similar the generated response between less intense and more intense facial actions. In other words, the lower the Likelihood, the more impact the feature has on the output. A likelihood value that is > 0.05 means that the topics are more-or-less the same. (B) Average response sentiment (y-axis), given the activation of one specific facial action unit (x-axis). Every feature sentiment value is computed on a set of 50,000 responses. Example images [27] of the corresponding actions are shown below the bars.

TABLE II

EXAMPLE PAIRS OF GENERATED RESPONSES WHEN EACH FACIAL ACTION IS LOW INTENSITY (LEFT) AND HIGH INTENSITY (RIGHT). THE LOW INTENSITY AU COLUMN CORRESPONDS TO THE TEXTUAL OUTPUT GIVEN A NEUTRAL FACE. VALENCE OF EACH FACIAL ACTION (VALENCE) AND CORRELATION OF THE ACTION INTENSITY WITH THE RESPONSE SENTIMENT (SENTIMENT CORR.) ARE SHOWN. ALL CORRELATIONS ARE SIGNIFICANT.

AU	Valence [19], [13]	Sentiment Corr.	Caption	Low Intensity Action	High Intensity Action (see Fig. 5 (B) image examples)
1	Neg.	Neg.	hey y'all .	you 're so pretty !	i hate you
2	Neg.	Pos.	who wants to see a photo of me looking like mogli?	i'm so dead	you 're so cute !
4	Neg.	Neg.	biggest one went 30 " and 9lbs	your hair is so cute !	what the hell .
5	Neg.	Pos.	look at us beauties	that 's fake	oh my gosh . so cute .
6	Pos.	Pos.	snapped by my bish lil vinny	no it 's not a little bit	good luck bro
7	Pos.	Pos.	i looked at the problems and i was like what is this	damn	pretty color !
9	Neg.	Neg.	happy birthday to my boy	you 're so pretty	loser
10	Neg.	Pos.	subway date with my son	stop	oh my gosh i love you
12	Pos.	Pos.	kenzie at her finest moments	oh my gosh i look so weird	i love you
14	Neg.	Pos.	she looks great here	oh my gosh no	you 're beautiful
15	Neg.	Neg.	selfie cause i 'm single and wearing a vest.	i love your mom	lool i hate you
17	Neg.	Neg.	flex friday	he 's so cute	your looks sick
20	Neg.	Neg.	grow hair , grow .	yeah that 's you at all	stop stop i 'm ugly
23	Neg.	Neg.	come to mama	love you !	i'm so jealous !
25	Pos.	Pos.	got ta love life !	i was not the only one	the best player in the world !
26	Pos.	Neg.	h8 cleaning my room .	you're so cute	that 's so hot i 'm crying
43	N.A.	Neg.	can i drop this here ?	yes .	i hate you so much

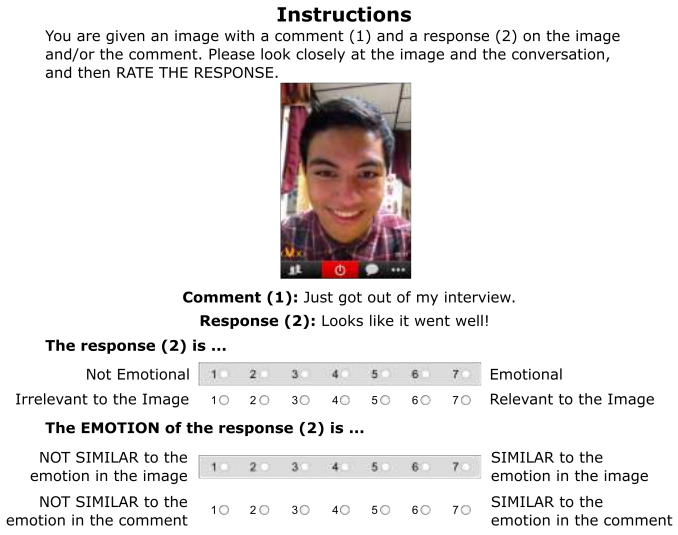

Fig. 6. A screenshot of the human judgment task used for evaluating the conversational responses.

TABLE III

HUMAN ANNOTATOR AGREEMENT FOR THE DIALOGUE RESPONSES. WE COLLECTED SEVEN ANNOTATIONS PER TASK.

Task	Krippendorf Alpha (α) [21]	KappaQ (κ) [5]
Emotional	.515	.691
Relevant	.685	.704
Similar to Emotion in Image	.564	.621
Similar to Emotion in Comment	.569	.523

more emotional than the gold-standard responses. There was no significant differences in the means for the other questions (i.e. the output of the models were as relevant to the image, as similar in emotion to the image and comment when compared to what humans actually wrote on Twitter). This shows that the neural language models do produce good quality responses.

VII. DISCUSSION

Incorporating facial action coding into a language model significantly influenced language sentiment. Overall, the conversation sentiment became more positive with the inclusion of facial actions. However, the variance also increased. On the feature-level we observe that more expressive and salient facial actions have strong impact on language sentiment. This observation holds for both positive face actions (e.g. AU 12 or 6) which made language sentiment more positive, and for negative face actions (e.g. AU 1 or 15) which made language sentiment more negative. These changes are consistent with what we might expect from prior analysis of the valence of the facial actions [19], [13]; however, we did not prescribe the valence of the facial actions at any point in the modeling. In a few cases the valence impact of actions on language sentiment was different from what might be excepted from prior literature on facial action unit valence [19], [13], e.g. AU 2 or 14. However, one could argue that both of these actions could have negative and positive valence.

AU 5, 10 and 17 appear to impact the topic generated by the language model the most. This shows that topic shift and sentiment shift might be two separate processes and be influenced by facial expressions in different ways. Topics can often be described in a positive or negative manner. Further analysis of how topics changes with the presence of different facial expressions would be very interesting future research.

We show interpretable examples for both the topic changes and the sentiment changes in the output text. Samples in Table II illustrates how the sentiment of the output changes when different actions are more or less intense. These examples show *how* language sentiment change is reflected in the generated language.

TABLE IV

MEAN RATING FOR EACH MODEL AND THE GOLD RESPONSES ON EACH OF THE FOUR QUESTIONS. RATINGS WERE ON A SCALE FROM 1 TO 7. THE NEURAL LANGUAGE MODEL RESPONSES WERE RATED THE SAME AS THE GOLD RESPONSES IN ALL CASES WITH THE EXCEPTION THAT THE MODEL RESPONSES WERE SLIGHTLY MORE EMOTIONAL.

Task	Gold	Text-Only	Text-Image
Emotional	4.37 (1.68)	4.81 (1.86)	4.83 (1.78)
Relevant	4.75 (1.74)	4.93 (2.01)	4.89 (1.93)
Similar to Emo. in Image	4.21 (1.66)	4.20 (1.90)	4.27 (1.91)
Similar to Emo. in Comment	4.03 (1.80)	4.08 (1.97)	4.08 (1.97)

VIII. CONCLUSIONS

We have presented the first example of a facial expression grounded language generation model. Combining facial actions and text to generate responses is a complex commonsense reasoning task. The results from our neural model show that grounding in facial actions helps strengthen and align the sentiment in computer generated dialogue.

We evaluated the model using large-scale linguistic analyses, crowdsourced human judgments and using illustrative qualitative examples. The human evaluation showed that our text-image model generates results similar in quality to human reference responses. The automated analysis showed that the model generates language that seems to approximate the emotion in the facial expression in the image and influences the response topic in a reasonable manner. Our work opens up a new avenue for generative conversational dialogue that leverages automated facial action coding. Future work will combine non-verbal cues with other scene information.

In addition, we presented a validation of a commonly used open source facial coding toolkit, OpenFace, on 10,000 images found on the Internet. The results showed that OpenFace provides accurate detectors for a majority of the facial actions. We deemed this validation necessary given the highly unconstrained nature of social media imagery.

REFERENCES

[1] J.-Y. Antoine, J. Villaneau, and A. Lefeuvre. Weighted krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *EACL 2014*, pages 10–p, 2014.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *arXiv preprint arXiv:1705.09406*, 2017.

[3] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–6. IEEE, 2015.

[4] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

[5] E. M. Bennett, R. Alpert, and A. Goldstein. Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3):303–308, 1954.

[6] T. Bickmore and J. Cassell. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 396–403. ACM, 2001.

[7] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM, 2004.

[8] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis. Multimodal emotion recognition from expressive faces, body gestures and speech. *Artificial intelligence and innovations 2007: From theory to applications*, pages 375–388, 2007.

[9] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[10] S. K. D'mello and A. Graesser. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20(2):147–187, 2010.

[11] S. K. D'mello and J. Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47(3):43, 2015.

[12] P. Ekman, W. V. Friesen, and J. Hager. *Facial action coding system: A technique for the measurement of facial movement*. Research Nexus, Salt Lake City, UT, 2002.

[13] W. V. Friesen and P. Ekman. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2(36):1, 1983.

[14] S. Ghosh, E. Laksana, S. Scherer, and L.-P. Morency. A multi-label convolutional neural network approach to cross-domain action unit detection. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 609–615. IEEE, 2015.

[15] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating rapport with virtual agents. In *International Workshop on Intelligent Virtual Agents*, pages 125–138. Springer, 2007.

[16] K. L. Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[17] 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.

[18] S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

[19] K. S. Kassam. *Assessment of emotional experience through facial expression*. Harvard University, 2010.

[20] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[21] K. Krippendorff. Computing krippendorff's alpha reliability. *Departmental papers (ASC)*, page 43, 2007.

[22] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *Proc. of NAACL-HLT*, 2016.

[23] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency. Its only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014.

[24] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[25] D. McDuff. Smiling from adolescence to old age: A large observational study. In *Affective Computing and Intelligent Interaction (ACII), 2017 International Conference on*. IEEE.

[26] D. McDuff, J. M. Girard, and R. El Kaliouby. Large-scale observational evidence of cross-cultural differences in facial behavior. *Journal of Nonverbal Behavior*, 41(1):1–19, 2017.

[27] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby. Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3723–3726. ACM, 2016.

[28] D. J. McDuff. *Crowdsourcing affective responses for predicting media effectiveness*. PhD thesis, Massachusetts Institute of Technology, 2014.

[29] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[30] M. R. Morris, A. Zolyomi, C. Yao, S. Bahram, J. P. Bigham, and S. K. Kane. With most of it being pictures now, i rarely use it: Understanding twitter's evolving accessibility to blind users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5506–5516. ACM, 2016.

[31] R. Morris, D. McDuff, and R. Calvo. Crowdsourcing techniques for affective computing. *The Oxford handbook of affective computing*, pages 384–394, 2014.

[32] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. 2016.

[33] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency. Utterance-level multimodal sentiment analysis. In *ACL (1)*, pages 973–982, 2013.

[34] R. E. Riggio. Social interaction skills and nonverbal behavior. *Applications of nonverbal behavioral theories and research*, pages 3–30, 1992.

[35] T. Senechal, D. McDuff, and R. Kaliouby. Facial action unit detection using active learning and an efficient non-linear kernel approximation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–18, 2015.

[36] L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. In *ACL-IJCNLP*, 2015.

[37] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[38] M. Valstar. Automatic facial expression analysis. In *Understanding Facial Expressions in Communication*, pages 143–172. Springer, 2015.

[39] M. F. Valstar, E. Sánchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang, L. Yin, and M. Pantic. Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. *arXiv preprint arXiv:1702.04174*, 2017.

[40] O. Vinyals and Q. Le. A neural conversational model. In *Proc. of ICML Deep Learning Workshop*, 2015.

[41] S. Weitz. Meet xiaoice, cortana's little sister. *Microsoft.[Online]. Available: https://blogs. bing. com/search/2014/09/05/meet-xiaoicecortanas-little-sister*, 2014.

[42] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[43] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.