Emotional Dialogue Generation using Image-Grounded Language Models

Bernd Huber[1,2], **Daniel McDuff**[2], **Chris Brockett**[2], **Michel Galley**[2], **and Bill Dolan**[2]
[1]Harvard University, Cambridge, MA, USA
[2]Microsoft Research, Redmond, WA, USA
bhb@seas.harvard.edu, {damcduff,chris.brockett,mgalley,billdol}@microsoft.com

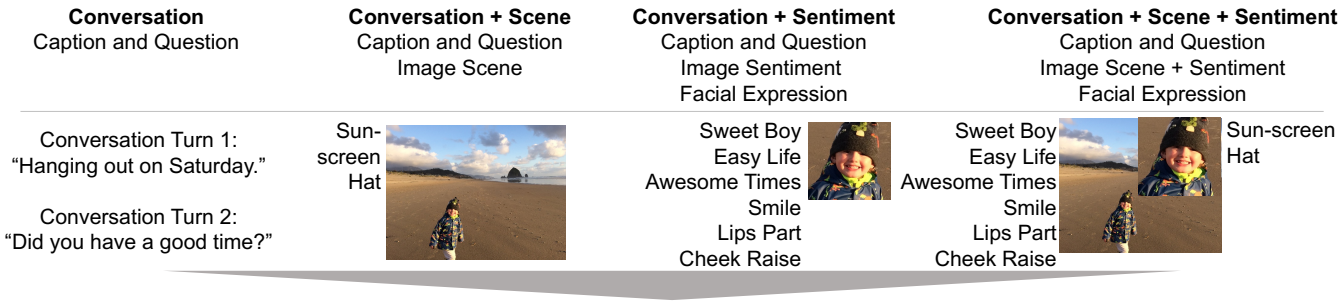

Figure 1: We present the first image-grounded dialogue model that combines scene and sentiment recognition with a natural language model. We analyze how image content (including objects, scenes and facial expressions) influences generated dialogue. We show that specific features can be used to tune the dialogue qualities. The system was trained and tested on one million real social media conversations.

ABSTRACT

Computer-based conversational agents are becoming ubiquitous. However, for these systems to be engaging and valuable to the user, they must be able to express emotion, in addition to providing informative responses. Humans rely on much more than language during conversations; visual information is key to providing context. We present the first example of an image-grounded conversational agent using visual sentiment, facial expression and scene features. We show that key qualities of the generated dialogue can be manipulated by the features used for training the agent. We evaluate our model on a large and very challenging real-world dataset of conversations from social media (Twitter). The image-grounding leads to significantly more informative, emotional and specific responses, and the exact qualities can be tuned depending on the image features used. Furthermore, our model improves the objective quality of dialogue responses when evaluated on standard natural language metrics.

ACM Classification Keywords

H.5.2 User Interfaces: Natural Language

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

© 2018 ACM. ISBN 978-1-4503-5620-6/18/04…$15.00
DOI: https://doi.org/10.1145/3173574.3173851

Author Keywords

Dialogue, conversation, emotion, computer vision, conversational agents

INTRODUCTION

Computer-based conversational agents (CA) are becoming a ubiquitous presence in our everyday lives. These assistants are useful for various tasks such as information retrieval (e.g., health information access [7]), information management (e.g., calendar organization [9]), and entertainment. Recent progress in the field of natural language processing (NLP) has enabled new functionalities for CA, such as generating logical responses to questions in constrained settings (e.g., website customer support [10]) or providing a companion to chat with (e.g., *XiaoIce* [48]). However, for a CA to become truly valuable to the user, it must be natural to interact with, and generalize to a broad range of contexts. But how can such capabilities be integrated into a CA?

Previous research suggests that CA need to be informative and empathetic to be engaging in conversations [8, 5, 6]. This research furthermore suggests that engaging conversations include visual cues (e.g., a photo or a video shared in a conversation between humans). On Twitter, for example, almost one third (28%) of posts are accompanied by an image (statistic from June 2015) [32]. Information contained in these images is often integral to the conversation. Figure 2 shows an example of a caption and question that could be associated with either of the two images. The appropriate response to the question would be very different depending on which of these

Figure 2: An example of how appropriate responses to a question may differ based on a tagged image. This example illustrates that by providing different image sentiment and image content, different responses may be expected and highlights the meaningful nature of image-grounded conversations.

images appeared with them. Thus, it is natural to suppose that a CA would be more effective if this information were part of its underlying conversational model.

But what is it in an image that provides this additional information? Contextual information could include objects and people but also facial expressions and the collective sentiment of a scene or situation. Affective information plays a particularly important role in many social interactions and non-verbal expressions are critical for social functioning [38]. Thus, information about the facial expressions of individuals within a scene can be very important for creating systems that allow for meaningful and rich conversational interactions. To this end, we present the first example of image-grounded dialogue generation using image sentiment, facial expressions and image scene features, as illustrated in Figure 1.

Mostafazadeh et al. [33] presented an approach for using low-level image inputs (i.e., pixels) in order to ground a conversational model. However, these convolutional neural network (CNN) features are high dimensional and difficult to interpret. In recent years, significant advances in computer vision have led to marked improvement in the state-of-the-art in object detection, scene understanding, and facial and body analysis [24]. We leverage these methods to extract high-level contextual features from images associated with conversations and use them to ground the language model of a CA. This approach surfaces relationships between images and conversations, that could not be seen otherwise.

In this paper we focus on the task of generating responses to questions. We restrict our analyses to this task because question answering allows for more objective quality evaluation (compared to open ended conversations). Specifically, we present a model for generating responses to questions using text, image scene and sentiment understanding. Our aim was to train a language model that produces dialogue that is informative, emotional, interesting, specific and logical. Hence, the main contributions of this work are:

1. To present the first image-grounded dialogue generation that combines scene and sentiment understanding with a natural language model.
2. To evaluate this model on highly-naturalistic real-world social media conversations using a rigorous human evaluation scheme.
3. To systematically analyze the relationship between image properties and conversational qualities. This analysis shows that both image content and sentiment play important roles in generating the best responses.

In the rest of this paper we describe how we built an image-grounded CA, the image analysis we conducted, human and machine-based evaluation of the resulting CA, and we present how features can be tuned to influence dialogue properties. Finally, we discuss implications for the design of CA.

RELATED WORK

Our work builds on research in both HCI and artificial intelligence (AI). Related work includes the design of social and affective agents, the integration of visual information with CA, and a survey on evaluating CA.

Designing Conversational Agents

Conversational agents often do not meet users' expectations [27]. Luger et al. previously showed that there is a large gulf between peoples' expectations about the capabilities of CA, and what such systems can actually deliver. The work reveals multiple design challenges arising from this gulf between user expectation and experience, such as how a CA may reveal its current state, or how one might design system feedback and clearly communicate the goal of the system.

Similarly, researchers have studied the *perceived intelligence* of agents in-depth [8]. The work by Cassell et al. studies how multi-modal interactions affect the experience of using a CA. According to Cassell, two main factors that determine perceived intelligence are how the system interface represents its functionality, and how knowledge is communicated to the user. Whilst the focus of Cassell's work has been multi-modal representations of intelligence (physical gestures in addition to voice), the central concept of the users' need to 'locate intelligence', and thereby the need to represent intelligence to the user, is an important concept in CA research.

By adding visual context into natural dialogue generation, we aim to make the dialogue more engaging, emotional and specific, without simply referencing the visual context itself (i.e. in contrast to visual question answering, the task of answering questions about objects or concepts, directly derivable from an image). This extends the modalities of interaction with the agent and may therefore be a fundamentally new step towards improving user experience of CA.

Visual Conversational Agents

In AI research, there has been increasing interest in CA that allow for multi-modal user input. Of most relevance here, various tasks have evolved around the combination of language and visual information [13, 29]. Visual captioning uses imagery as input to a machine learning model to generate a caption describing the picture [51]. For visual captioning, the objective is that the caption captures information about the objects/scene in the image. Visual Question Answering

(VQA) extends visual captioning to a more open-ended question answering paradigm. In VQA, questions are constrained to be answerable directly from the image [2]. An alternative to answering questions about an image is generating questions about the image [34]. In this case, natural and engaging questions are desired. Dialogue generation differs from VQA as it does not require the dialogue to be directly referencing contents of the image, as for example, in [11, 12], but the image rather serves as additional context to the conversation. For example, a conversation spurred by an image may reference related concepts but not the scene itself.

Rich sentiment and emotion information can be gleaned from both text and image analysis of social media [50]. We tackle the novel problem of answering questions posed by a user in a natural way, combining textual and visual context, where the questions may not generally be about the contents of the image itself. We use and extend the data-driven paradigm of conversation generation [39, 44, 40, 41, 46, 25] in which neural models are built, typically from social media data. The most closely related work to this approach [28, 43]demonstrates that image sentiment can be used for generating relevant sentiment in image captioning tasks. The generated language from these models contains emotional adjective-noun pairs. In image-grounded conversation, our goal is different, in that we want to generate social, engaging conversations that are grounded in the image.

Evaluating Conversational Agents

There is not a clear consensus on what the objective function of CA should be, especially when comparing AI and HCI literature. In AI research, a good dialogue model is supposed to generate conversations as close to the dialogue that a human would produce as possible (with the Turing test being the ultimate evaluation method). A large bulk of HCI research, in contrast, suggests that an agent behaving as a human is not necessarily the main criterion when it comes to creating engaging conversations [42]. HCI literature seems to focus more on specific behavioral traits. Gratch et al. [18] show that in order to create rapport, agents should provide more positive, emotional feedback from time to time. Walker et al. [47] show that the level of informativeness is an important measure for engaging dialogues. Measures of attitudes and perceptions have been applied in [5, 20] to evaluate a CA. Bickmore and Cassell [5] had users complete a standard questionnaire on trust after interaction with the agent to this end. These methodologies target a more specific kind of goal, rather than generally being human-like. In our work, we follow a similar approach by measuring specifically the emotionality and informativeness of an agent.

IMAGE-GROUNDED 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. The previous conversation may consist of one or multiple turns. A widely adopted DNN approach to this problem is a sequence-to-sequence architecture (*seq2seq*) [41, 46, 25]. These models have been very effective in various tasks such as dialogue generation and language translation. To integrate visual information into these models, Mostafazadeh et al. [33] proposed a modified *seq2seq* structure that uses visual input together with textual input for conversational language generation. We extend their architecture by adding a layer of higher-level image understanding to the network. We learned all the weights in our model using stochastic gradient descent with an exponentially decaying learning rate. We used early stopping and dropout to prevent overfitting, in a similar manner to Mostafazadeh et al. [34].

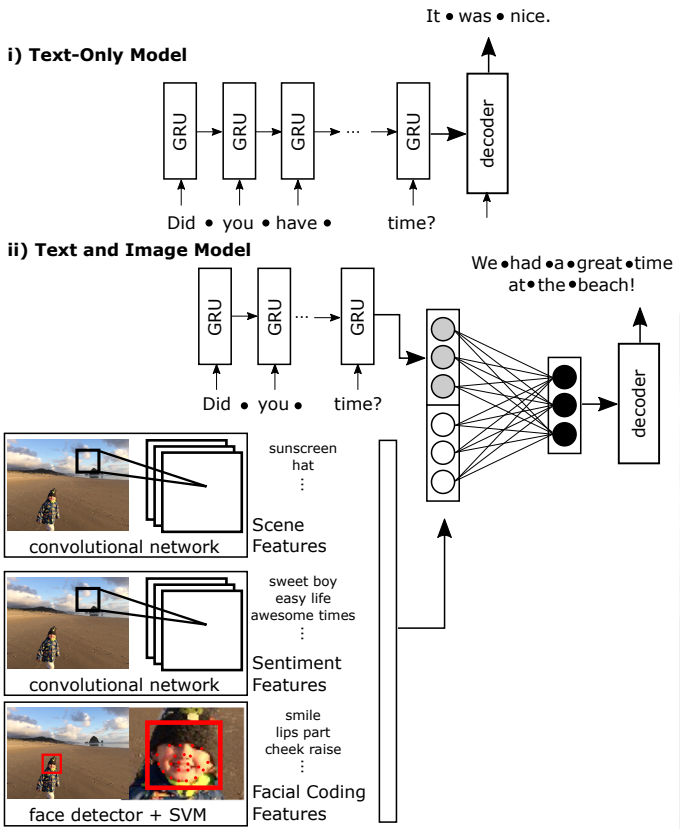

Figure 3: The two different deep learning models that we used to generate dialogue. (i) The text-only model captures one or multiple turn conversations and outputs the next turn in that conversation. It makes use of GRU-cells that capture the temporal and contextual information in the text. (ii) The model that integrates the image features appends the image with the textual information, which then flows into the decoder architecture.

A schematic of our approach can be seen in Figure 4. In detail, the text and text plus image models can be compared as follows.

Text-Only Model: The input is a text caption and question that is mapped to an output response. This model maps the input sequences to an output sequence (*seq2seq* model [45]) using an encoder and a decoder recurrent neural network (RNN). The initial recurrent state is the 500-dimensional encoding of the textual context.

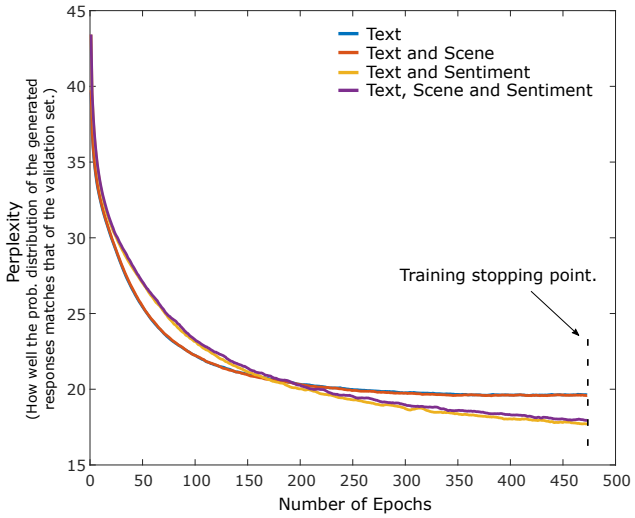

Figure 4: Convergence of four different models, measured in perplexity (lower is better). Each model was trained for 475 epochs (i.e., passes through given dataset), which takes about two days to train on a state-of-the-art computer architecture.

Figure 5: Example images of the 17 facial actions classified in the images. The relevant region of the face is highlighted by the white box.

Text and Image Model: As with the text-only model the textual feature vector is obtained using a recurrent neural network (RNN). The vector is then concatenated to the image feature vector and fed into a fully connected feed forward neural network. The results being a single 500-dimensional vector encoding both visual and textual context, which then serves as the initial recurrent state of the decoder RNN. We experimented with different combinations of scene, sentiment and facial coding image features which are described in the following section.

IMAGE UNDERSTANDING

This section describes the set of features we extracted from the images. We chose to extract a set of features that are interpretable and give a rich description of the content. These features allow for a deeper understanding of the relationship between image features and the generated dialogue.

Scene Understanding

To generate scene understanding features, we used a convolutional neural network (CNN) based scene recognition classifier trained on a scene-centric database called *Places*. The *Places* database features over seven million training images of scenes [52]. The classifier outputs probabilities for 1,183 scene features. To avoid over-fitting, we selected the 50 scene features with the highest probabilities across the training set (i.e., the most likely scenes to occur in our training set). The resulting features are shown in Table 1. These reflect the type of content that frequently occurs in social media posts, such as clothing (e.g., hats, ties, sunglasses), everyday objects (e.g., TVs, phones) and places (e.g., museums). Figure 8 provides examples of the images from some of these categories.

Scene Sentiment

We used a CNN-based sentiment recognition classifier trained on the Multi-lingual Visual Sentiment Ontology (MVSO) [21] dataset to extract further information about the scenes. Given an image, the MVSO model provides probabilities for 4,800 adjective-noun pairs, which have shown to be highly correlated with the overall sentiment in the image. In addition, each of these noun pairs is associated with a sentiment score from -1 (negative valence) to $+1$ (positive valence). To avoid overfitting we selected the 50 scene sentiment features with highest cumulative probabilities across the training set. The resulting features are shown in Table 1. Again, these reflect the type of content that frequently occurs in social media posts. The adjective noun pairs describe people, physical appearance and life stages. Figure 8 provides examples of the images from some of these categories.

Facial Coding

The facial action coding system (FACS) [14] is the most widely used and comprehensive taxonomy for coding facial actions. We chose AUs over face emotion classifiers as a representation of the face as they provide a more objective, comprehensive and fine-grained description. We did not use higher-level emotion categories as that may have limited the behaviors we were able to capture (e.g., to the common set of six "basic" emotions.) We used facial coding software to extract the facial actions of the faces within the images [3]. The classifier extracts appearance-based information from the face region-of-interest and a Support Vector Machine (SVM) classifier provides a probability score for 17 facial actions based on FACS. Figure 5 shows example images of the different facial actions. The actions can be associated with emotional valence based on psychology studies [22]. Using this basis we assigned lip corner pull (smile) and cheek raise actions with positive

Table 1: Scene and sentiment features extracted from the images. The sentiment features are a set of adjective-noun pairs that are mapped to specific sentiments in the MVSO corpus. The face features are the facial action units from the facial action coding system (FACS). Each feature represents the probability of the specific label being present in a given image.

Test		Features
Scene	[52]	analog clock, band aid, bath towel, bathing cap, binder, book jacket, bow tie, carton, cash machine, cassette, cellphone, comic book, cowboy hat, digital clock, drumstick, envelope, hair slide, hair spray, hand blower, harmonica, ice lolly, iPod, laptop, lighter, lipstick, lotion, mask, menu, modem, monitor, museum, neck brace, notebook, packet, paper towel, pill bottle, plunger, remote control, rubber eraser, ruler, screen, sunglass, sunglasses, sunscreen, syringe, t-shirt, television, toilet tissue, web site, wig
Sentiment	[21]	amazing girls amazing people, awesome times, bipolar disorder, blessed life, broken hearts, changing lives, chronic pain, comic sans, creative advertising, creative agency, creative cloud, creative direction, creative director, easy life, eternal life, fast company, fit girls, fit life, funny food, funny jokes, funny quotes, funny stuff, good quality, great business, handsome men, healthy chocolate, hot site, interactive media, late dinner, light rain, low price, magic cards, medical practice, open education, personal injury, personal trainers, professional portfolio, real men, real music, real talk, sexy lips, short sale, sparkling heart, special offers, sweet boy, teen pregnancy, true friends, visual identity, wise words
	[3]	inner brow raise, outer brow raise, brow furrow, eye widen, cheek raiser, nose wrinkle, lip raise, lip corner pull, dimpler, lip corner depressor, chin raise, lip stretch, lip tighten, lips part, jaw drop, lip suck, eyes closed

valence and inner brow raise, brow furrow, eye widen, eye tighten, lip depressor, lip tighten and lip stretch actions with negative valence. The valence is included as a feature in our model.

DATA

One million conversations (image, textual context, question, response tweet threads) were mined from the Twitter fire hose. The only criterion for the conversations was that they featured an image and caption, followed by a question and a response to the question. The data is otherwise unconstrained and represents a very challenging language modeling research problem, as well as a large variety of content; Twitter users do not always use standard grammar or spellings, and frequently use colloquial language. We believe positive results on such a dataset bodes well for a broad range of language datasets. Prior work supports the use of Twitter as the source of millions of natural conversations for conversation modeling (e.g., [33]).

To get a sense for the data, we looked at the type of questions people asked of their social network in a subset of 3,000 conversations. Question types, i.e., question intentions, were similar to the question categories found by Morris et al. [31], being mostly social, opinion-based, or factual knowledge. Topics were mostly appearance-related or entertainment-related. Generally, compared to conversations without images, image-related question may be more related to the content in the image. Using our face feature extractor, we detected that about 27% of the images were close-up photographs of faces, which may explain the large amount of appearance related discussions.

RESULTS

We trained one version of the text-only and three versions of the text and image model to compare the impact of different types of text and image features on the generated dialogue. Each model was trained for 475 epochs (i.e. passes through given dataset), which takes about two days to train on a state-of-the-art computer architecture.

1. **Text**: The text-only model was trained using captions, questions and responses. Thus, there was no additional information from the images.

2. **Text + Image Scene**: The text and image model was trained with text from the conversations (captions, questions and responses) and the additional image scene features (N = 50).

3. **Text + Image Sentiment**: The text and image model was trained with text from the conversations and the additional image sentiment features (facial expressions and scene sentiment) (N = 68).

4. **Text + Image Scene and Sentiment**: The text and image model was trained with text from the conversations and both the image scene and sentiment features (N = 118).

We used both human annotation and automatic methods to evaluate the responses generated by our models, which will be presented in the following sections.

Human Judgment

Human judgments of the computer generated dialogue responses were performed using a crowdsourcing task. Demographics of the crowdworkers were restricted to English language speakers in the UK and USA. Workers were paid 10 cents (USD) per task and were limited to completing a maximum of 15 tasks. The worker recruitment platform we used provided a work quality assurance mechanism by regularly assigning quality assessment tasks to workers, and filtering workers that do not pass quality standards. This pool of workers therefore is assumed to provide judgments of a reasonably high quality levels.

In the task, the crowdworkers were presented with one conversation and the corresponding image. Each sample consisted of

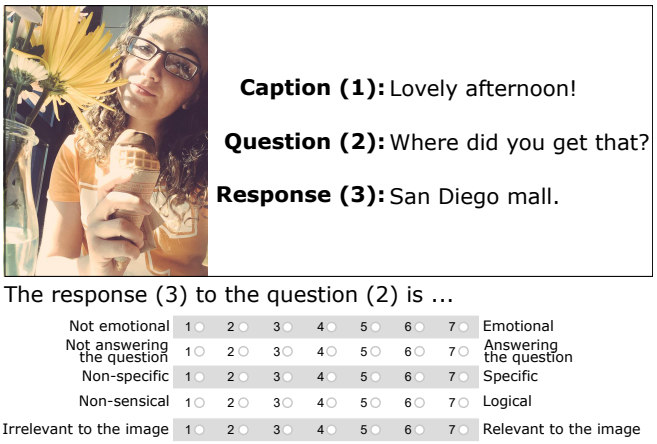

Figure 6: The human annotation task to rate the generated responses. For one judgment, crowdworkers were presented with one conversation and the corresponding image. The response was generated from one of the four dialogue models, which was assigned randomly. Crowdworkers were asked to rate the response based on how emotional, informative, specific, logical and relevant it was.

the image, caption and question from Twitter followed by the computer generated response from one of the models. Crowdworkers were asked to evaluate the quality of the response. Specifically, they were asked: "The response to the question is:" and responded on five seven-point Likert scales. The end points were:

i) Non-emotional - Emotional

ii) Not Answering the Question - Answering the Question

iii) Non-specific - Specific

iv) Nonsensical - Logical

v) Irrelevant to the Image - Relevant to the Image

Figure 6 shows the design of the human rating task for one of the conversations.

The related work by Ghosh et al [16], presents a language generation model that can be grounded with explicit emotion labels. The paper evaluates this model on Amazon Mechanical Turk, asking workers to rate emotionality and grammatical correctness of a generated sentence. In contrast, in image-grounded conversations, the mapping between the input image and corresponding conversation may not be as linear as the mapping of one emotion label for each sentiment. While the final goal of an emotional agent might be congruence, to disentangle the relationship between the image information and the generated language, we chose to evaluate emotion on an absolute scale.

It would be infeasible to manually annotate the full 112,000 conversations that were used for testing the model. Therefore, we selected a subset for manual annotation. The selection criteria were: 1) the image had at least one face (and thus all features could be computed), 2) the gold response was not simply yes or no, 3) the question was not appearance related, rhetorical question (e.g., "Why are you so cute?" or "Can I please look like you?"). These questions were excluded because answers to those were either heavily biased towards one answer (e.g., yes/no questions are usually phrased to be answered with "Yes", appearance related rhetorical questions are typically answered with expressions of gratitude, e.g., "Awww thanks babe"). We also selected face-only conversations as a subset because we wanted to make full use of the feature set, and because the face-related discussions are more specific discussions, typically around humans (as compared to more factual posts such as a website screenshot). Using this selection scheme, we otherwise randomly selected 200 conversations for human rating. We assigned each conversation to ten independent crowdworkers. Since every conversation was rated for each of the four feature combinations, we collected a total number of 8,000 judgments.

Table 2: Human annotator agreement for the automatically generated responses. We had collected 10 annotations per task.

Task	Krippendorff Alpha (α) [23]	KappaQ (κ) [4]
Emotional	.704	.708
Informative	.902	.827
Specificity	.825	.706
Logic	.855	.763
Relevance	.826	.659

Table 3: Results of automatic linguistic analysis (a scalable complement to the human ratings.) The relative improvement in the dialogue *BLEU* and *word2vec* scores are shown as percentages (compared to the text-only model). These scores were computed on the whole test set which consisted of 112,000 image grounded conversations.

		BLEU score	*word2vec* score
	Text (Baseline)	4.47	.311
Text +	Scene	4.59 (2.57%)	.311 (.15%)
	Sentiment	4.74 (5.97%)	.315 (1.28%)
	Scene & Sentiment	4.73 (5.56%)	.319% (2.56%)

We use two measures for assessing the agreement between the coders. First, a weighted Krippendorff alpha (α) [1]. Second, a Kappa Q (κ), that is a generalization of Bennett et al.'s S score [4] proposed by Gwet [19]. Both are suitable for ordinal scales. Table 2 shows the α and κ for each question. The effective reliability of the annotators is moderate to high across all metrics. Given the highly naturalistic data and open nature of the responses the metrics provide confidence that there is agreement between the coders.

Figure 7 shows bar plots of the average score of conversational responses from the different models.

Linguistic Analysis

Although we believe that human ratings of responses are most insightful and meaningful, we conducted additional automatic evaluations to verify the linguistic qualities of our models. It

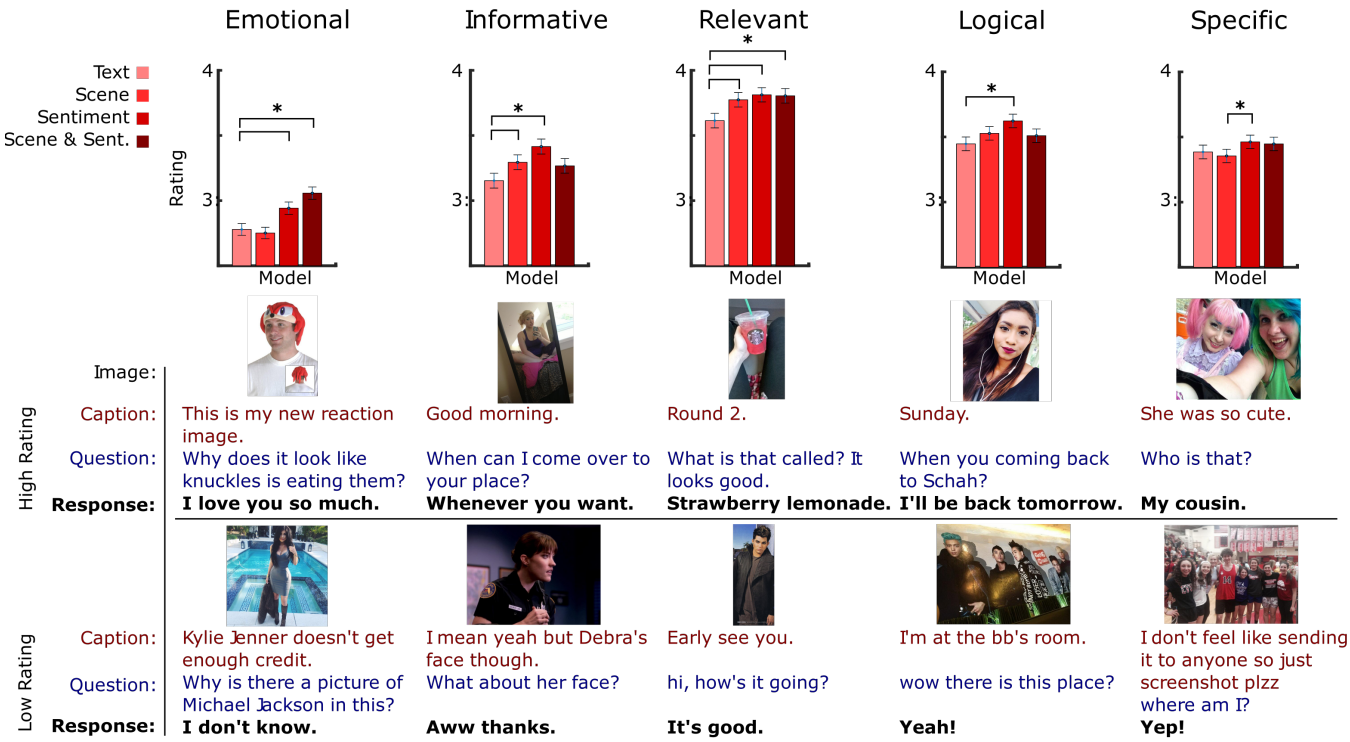

Figure 7: Human rating of responses for the four different models. The ratings show the mean Likert-scale ratings between 1 and 7. We show examples to illustrate responses with high and low ratings for each of the evaluation metrics, including the corresponding images. Example responses were taken from the model with the full scene+sentiment feature set.

is a non-trivial task to design automated metrics for evaluating open ended conversations (it is often hard to determine what a "correct" conversation looks like.) Several metrics are frequently used, the most common being the *BLEU* score [35], the gold standard measure in machine translation. For completeness we also used a *word2vec* score as a secondary metric for evaluation.

BLEU Score

The *BLEU* score [35] approximates the similarity between the actual responses from the Twitter conversations in the test set and the generated responses from our models. It is a precision-oriented metric that measures the amount of lexical overlap (phrases of 1-4 words) between system and reference responses, and that also incorporates a brevity penalty to penalize system outputs that are too short.

BLEU has been extensively used in evaluation of conversation generation tasks [44, 49, 25]. Since conversation generation is an open ended problem, where plausible outputs are inherently diverse, scores in these tasks are typically much lower than those observed in machine translation. Note that even humans usually only attain a *BLEU* score of approximately 5% in this task [26]. While [26] suggest that *BLEU* correlates poorly with human judgment at the sentence-level, corpus-level scores are shown to correlate well with human judgments when measuring differences in system performance [37, 15, 17].

Word2vec Score

Recent methods for learning vector space representations of words, such as *GloVe* [36] and *word2vec* [30] have succeeded in capturing fine-grained semantic and syntactic regularities. Those vectors are pretrained on very large text corpora and can be downloaded and used out of the box. *Word2vec* vectors have the advantage, in comparison to *BLEU* scores, of capturing paraphrasing (e.g., "yes", "yeah" or "yup" would be very similar word vectors). We added a *word2vec* based similarity measure as a second automated evaluation metric. The similarity between two sentences is computed as the cosine distance between the average word vector for each sentence:

$$score_{w2v} = cos(gen, ref) \quad (1)$$

Where *gen* is the average word vector for the computer generated response and *ref* is the average word vector for the reference (Twitter) response.

In our evaluation, we used the Twitter responses as the reference responses, and computed scores for each of the four dialogue models. The results can be seen in Table 3. We find that adding more image information leads to better automated evaluation (both sentiment and scene related). The BLEU score for the text and image model (using scene and sentiment features) was significantly higher than the text-only model (t-test p=0.012). The *word2vec* score for the text and image

model (using scene and sentiment features) was significantly higher than the text-only model (t-test p<0.0001).

Image Impact Analysis

The previous analyses showed that adding image information to conversational agents can improve their performance, both on informational and emotional scales. However, we also wanted to understand *why* the agent generates better responses on these scales. We therefore analyzed the impact of every feature on the output, to find out which were the most important features that influenced the sentiment and content in the responses generated by an agent.

There are two common ways to analyze the importance (saliency) of one specific feature on the output of a neural network model. We can either analyze the output weights with respect to a specific feature, or we can analyze post-hoc how the generated responses change. The latter approach allows for more interpretable analyses. Thus we chose to analyze how the output of the CA changes when one specific feature is varied.

We analyzed the change in the generated responses on two metrics, a content-based measure, and a sentiment based measure. The content-based measure evaluates the average number of words that changed when changing one specific feature. The sentiment-based measure evaluates the average change in sentiment when changing one specific feature.

Figure 8 shows the most impactful features for both the sentiment-based and the content-based measures. The corresponding images are samples with the feature being more activated from left to right. We grouped the features by the feature category (scene, sentiment or facial action).

DISCUSSION

Performance

Our results show that human evaluators rate the responses from the dialogue model with image-grounding as significantly more emotional ($\chi^2_{7,N=4,000} = 16.55, p < .0001$). This suggests that the conversational agent learned a stronger relationship between the question and the response when having information about the image, and that this information increased the likelihood of the response being emotional. Figure 7 shows examples of sentences that were ranked highly emotional (e.g., "I love you so much!") and examples that were ranked highly non-emotional (e.g., "I don't know."). We observe that on some occasions the very emotional responses appeared less relevant to the question, compared to less emotional responses. However, overall the models with scene and sentiment features were also judged better across all other categories compared to the text-only model. This suggests that visual information is necessary when teaching CA to become more human-like, and emotive. We performed a second experiment using 200 images without faces. The human judgments of these were not significantly different across the models for any of the questions. This suggests that our model works most effectively on images with faces, perhaps because these images tend to feature more emotion and thus sentiment features are more informative.

Secondly, our results show that human evaluators rated the responses from the dialogue model with image-grounding as significantly more informative ($\chi^2_{7,N=4,000} = 5.58, p < .0181$) and relevant to the conversation ($\chi^2_{7,N=4,000} = 8.34, p < .0039$), when compared to a model that uses purely textual information. One way to interpret this is that responses become more committal. The range of possible responses to a question in an open-ended conversation can be quite large and text-only models tend to provide rather non-committal responses such as "I don't know" or "You know it". These types of responses have a higher likelihood of being *correct* regardless of the question. Our findings suggest that image grounding provides a way to reduce the range of possible answers, hence leading to more relevant and informative conversations. This aligns with the observations in [34]. Conversations with an image-grounded agent might ultimately become more engaging to users due to this property.

Figure 7 shows an example of a response that was rated very informative (question: "When can I come over?", response: "Whenever you want".) Although the image in this conversation does not provide a direct answer to the question, it may still provide some grounding that makes the answer more specific. As a further note, these types of examples justify our comparison with a baseline without any visual information, since there is not necessarily direct information in the image. This is also the reason why we did not compare our results with a VQA task, in which responses are designed to be about the contents of the image.

Feature Importance

Unlike previous work that used raw image features, using interpretable scene and sentiment features allowed us to analyze the impact of the image properties on the conversational responses in much greater detail. Figure 8 shows the features with most impact on the dialogue response sentiment and content. The corresponding images are samples of each class, the relevant class having higher probability from left to right. Image scene sentiment and facial expression features had most impact on the language sentiment. While the features that had the largest impact on the content in the language were image scene (content) related.

Figure 9 shows how varying the lip corner pull (smile) feature influences the dialogue generated my the model for the question: "Do you like it?". The impact of this feature on the response sentiment is quite intuitive. Bigger smiles lead to more positive responses from the model. It is interesting to observe how the response becomes more and more positive and enthusiastic, as the feature changes.

Designing Conversational Agents

Our work has important implications for the design of conversational agents. Just as for humans, it is not possible for an agent to be highly emotional, informative, specific and creative all the time. However, there is much room for improvement in generating natural dialogue.

It is helpful if we can design an agent to have the linguistic style of our choosing. For example, in some cases we desire

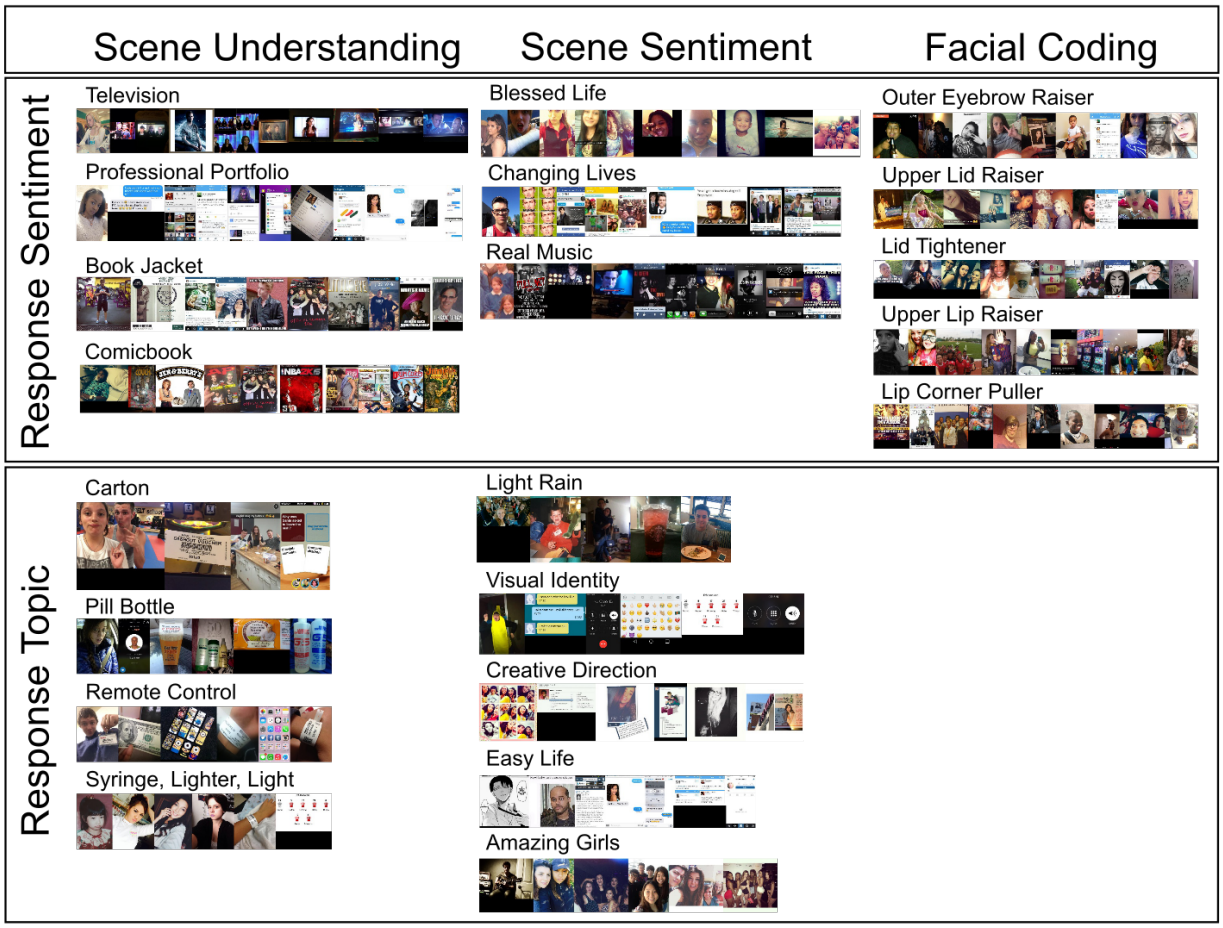

Figure 8: Analysis of the image features that have the most impact on the generated dialogue responses. Top) Features that influenced the sentiment of the dialogue responses the most. Bottom) Features that influenced the topic (term-frequency) of the language responses the most. We sample images on the continuum of each specific feature, from low probability (left) to high probability (right). For example, the feature for *Television* shows sample images with increasing probability of a TV being in the image, from left to right.

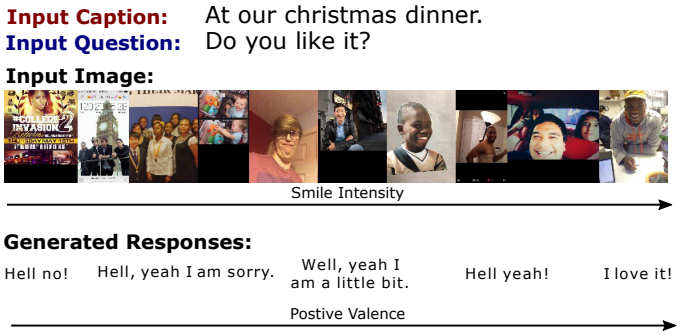

Figure 9: An example of how our text and image model changes its response to a question depending on the change of the lip corner pull (smile) feature. The feature changes from left to right from not activated at all to very activated. The responses are always generated from the same conversation, the only thing we changed is the image that was input into the model.

that an agent is more emotive and in other cases informativeness might be more important. Designing an emotional agent has been shown to be more engaging and useful in a variety of scenarios [8, 5]. Our results suggest that the type of image features used, whether more content or sentiment oriented, can influence the properties of the dialogue generated by the language system.

Future work is needed to further explore these properties. For example, how does one avoid potential biases that such an affective agent might learn. Little prior work on dialogue modeling has evaluated text across categories such as emotionality or empathy.

Implications for Social Media Analysis

Our analyses reveal interesting relationships between image content and question answering in social media data. For example, facial expressions seem to have a large impact on both arousal and valence in peoples' responses. There might be all kinds of relationships between how people respond

to questions and the facial expression shown in an image. Furthermore, an interesting future research direction would be to use a model such as ours as a way of revealing underlying patterns in social media interactions.

Repeatability and Applications

One strength of our work is that Twitter data is publicly accessible via the Twitter API, and can be obtained by researchers. We used the raw Twitter stream, so similar data is readily accessible, which allows for quickly prototyping a working model of this type. Our work can be reproduced with publicly available feature extraction toolkits. Dialogue can be generated in below one second on a regular laptop.

We imagine various application scenarios of image-grounded conversation models. One scenario would be bringing visual content into a conversational model for the visually impaired. Another would be visual grounding in a conversational agent aimed at having a social interaction with the user. Overall, visual grounding may enrich the social experience in conversations with computers.

CONCLUSIONS

We present the first example of an image-grounded conversational agent using visual sentiment, facial expression and scene features. We trained a novel CA on a large dataset of highly naturalistic conversations from Twitter. This model allowed for in-depth analyses of the relationships between such image information and the generated language.

Specifically, we analyzed and discussed the influence that image sentiment and image content have on the sentiment and content in the responses to the visually aided questions. Evaluation on an independent set of conversations showed that including image features increased how emotional, informative and relevant the generated dialogue was judged to be. We also found that visual sentiment and facial features in the images were the primary drivers of variations in sentiment in the generated responses. In addition, scene (content) features had more influence on the topic that was generated in the output. Our proposed model also significantly outperformed the baseline on automated linguistic metrics.

Grounding conversations using images is an exciting new research domain that could contribute to more natural and engaging CA. Finally, our work can benefit social media researchers as a means to discover novel insights in multimedia posts that combine imagery and dialogue.

LIMITATIONS

There are several limitations to the conclusions we can draw from this study. First, the system was trained on conversations from Twitter. These conversations do not always generalize to other natural language scenarios. Testing our system on other sources of data will be an important next step of our work.

Secondly, we excluded hashtags, emojis, and usernames from our model. While this makes the model more generalizable, there might be meaning in these additional signals that we did not capture. We believe that hashtags would provide valuable contextual information that might also help grounding a language model.

Finally, the overall performance of our model is still not perfect. This comes mainly from the fact that we used a dataset of one million conversations. Tens of millions of conversations would certainly improve the quality of the generated dialogue, a sample size frequently used for sequence-to-sequence models [41, 46]. However, for reasons of simplicity, we kept the number of conversations used in our training at one million.

REFERENCES

1. Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. 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*. 10–p.
2. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.
3. Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–10.
4. Edward M Bennett, R Alpert, and AC Goldstein. 1954. Communications through limited-response questioning. *Public Opinion Quarterly* 18, 3 (1954), 303–308.
5. Timothy Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 396–403.
6. Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 2 (2005), 293–327.
7. Timothy W Bickmore, Dina Utami, Robin Matsuyama, and Michael K Paasche-Orlow. 2016. Improving access to online health information with conversational agents: a randomized controlled experiment. *Journal of medical Internet research* 18, 1 (2016).
8. Justine Cassell, Tim Bickmore, Lee Campbell, Hannes Vilhjálmsson, and Hao Yan. 2000. Conversation as a system framework: Designing embodied conversational agents. *Embodied conversational agents* (2000), 29–63.
9. Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar. help: Designing a Workflow-Based Scheduling Agent with Humans in the Loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2382–2393.
10. Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. SuperAgent: A Customer Service Chatbot for E-commerce Websites. *Proceedings of ACL 2017, System Demonstrations* (2017), 97–102.

11. Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual Dialog. (2017). http://arxiv.org/abs/1611.08669

12. Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. In *International Conference on Computer Vision (ICCV)*. http://arxiv.org/abs/1703.06585

13. David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, and others. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1061–1068.

14. Paul Ekman, Wallace V Friesen, and John Hager. 2002. *Facial action coding system: A technique for the measurement of facial movement*. Research Nexus, Salt Lake City, UT.

15. Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proc. of ACL-IJCNLP*.

16. Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. *arXiv preprint arXiv:1704.06851* (2017).

17. Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate Evaluation of Segment-level Machine Translation Metrics. In *Proc. of NAACL*.

18. Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. 2007. Creating rapport with virtual agents. In *International Workshop on Intelligent Virtual Agents*. Springer, 125–138.

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

20. Kasper Hornbæk. 2006. Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies* 64, 2 (2006), 79–102.

21. Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. 2015. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 159–168.

22. Karim Sadik Kassam. 2010. *Assessment of emotional experience through facial expression*. Harvard University.

23. Klaus Krippendorff. 2007. Computing Krippendorff's alpha reliability. *Departmental papers (ASC)* (2007), 43.

24. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

25. Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proc. of NAACL-HLT*.

26. Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proc. of EMNLP*. https://aclweb.org/anthology/D16-1230

27. Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5286–5297.

28. Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2016. SentiCap: Generating Image Descriptions with Sentiments.. In *AAAI*. 3574–3580.

29. Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao, Oscar Romeo, Sushma Akoju, and Justine Cassell. 2016. Socially-Aware Animated Intelligent Personal Assistant Agent.. In *SIGDIAL Conference*. 224–227.

30. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

31. Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. 2010. What do people ask their social networks, and why?: a survey study of status message Q&A behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1739–1748.

32. Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P Bigham, and Shaun K Kane. 2016. 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*. ACM, 5506–5516.

33. Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*.

34. Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating Natural Questions About an Image. In *Proceedings of the Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, ACL âĂŹ16*.

35. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

36. Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

37. M. Przybocki, K. Peterson, and S. Bronsart. 2008. Official results of the NIST 2008 Metrics for MAchine TRanslation Challenge. In *MetricsMATR08 workshop*. `http://itl.nist.gov/iad/mig/tests/metricsmatr/2008/`

38. Ronald E Riggio. 1992. Social interaction skills and nonverbal behavior. *Applications of nonverbal behavioral theories and research* (1992), 3–30.

39. Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proc. of EMNLP*.

40. Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proc. of AAAI*.

41. Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *ACL-IJCNLP*.

42. Nicole Shechtman and Leonard M Horowitz. 2003. Media inequality in conversation: how people behave differently when interacting with computers and people. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 281–288.

43. Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. 2016. Image Captioning with Sentiment Terms via Weakly-Supervised Sentiment Dataset. In *BMVC*.

44. Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proc. of NAACL-HLT*.

45. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

46. Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. In *Proc. of ICML Deep Learning Workshop*.

47. Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 271–280.

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

49. Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. 2016. Conditional Generation and Snapshot Learning in Neural Dialogue Systems. In *EMNLP*. ACL, Austin, Texas, 2153–2162. `https://aclweb.org/anthology/D16-1233`

50. Quanzeng You. 2016. Sentiment and Emotion Analysis for Social Multimedia: Methodologies and Applications. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1445–1449.

51. Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4651–4659.

52. Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.