That is a good looking car !: Visual Aspect based Sentiment Controlled Personalized Response Generation

Anonymous ACL submission

Abstract

In a conversational system, generating utterances that communicate consistent and relevant preferences is vital for more personalized conversations. In this paper, we propose a task of generating utterances grounded on some assigned aspect-preferences-profile. These aspect-preference profiles consist of a list of aspect-sentiment tuples, denoting the preference of the speaker for some aspect in the form of sentiment ("positive" or "negative"). Since no prior dataset containing such profiles is available, we enhance Image-Chat data by assigning these profiles to each user in a conversation. The conversations in this dataset are based on an image, therefore the aspects are present in images as well as dialogue history. We build a BERT and ResNet-based encoder-decoder model with a memory network to store preference-profile. Through our experiments, we show that our model can generate responses that convey the sentiment of relevant aspects in accordance with the assigned profile. Both automatic and manual evaluations show the effectiveness of our model and dataset[1]. Our proposed system when using these profiles achieves a BLEU-1 score of 15.93 on this new task, which is an improvement of 2.92 points from the baseline experiment that does not use aspect-preference profiles.

1 Introduction

Multimodal conversational systems that can perceive the world visually and are able to converse with humans about its perceptions is an important next step towards the development of conversation systems. Since vision plays a major role in forming the world view in humans, it is only natural to model visual knowledge into conversation systems. Traditional chit-chat systems are primarily text based, with some control or grounding factors like knowledge (Dinan et al., 2019), empathy (Rashkin et al., 2019a), persona (Zhang et al., 2018), etc. In contrast to such traditional systems, Image-Chat dataset (Shuster et al., 2020) aims at building systems that can converse around a given image and the conversation style is also grounded on the persona type of the speaker. Persona of a speaker is another crucial aspect in open-domain chit-chats, and can have many dimensions. Persona can range from psychological classes of the speaker categorized as OCEAN (Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism) (Wiggins, 1996), persona profile consisting of factual statements about the user (Zhang et al., 2018), to style of speaking as present in Image-Chat.

Figure 1: An example of the proposed task. $S1$ and $S2$ are two speakers and the first row describes their aspect-preference profile. The rows after that contain the utterances by $S1$ and $S2$, respectively.

Another previously unexplored dimension of persona for conversational systems can be based on the aspect preferences of the system. The preferences or desires of a person develop around the age of 2 (Wellman and Woolley, 1990), much earlier than development of beliefs (i.e. being aware of the preferences of others). Therefore, for a chit-chat system to have a distinct persona, it is essential for it to express its preferences. This is usually

[1]The codes and datasets will be made available

expressed in the form of aspect-level-sentiments in a conversation. These pre-existing preferences are often aroused by vision (Gardner et al., 2003), leading to conversation about aspects in the visual modality, while expressing the aspect-sentiment in accordance with the pre-existing preference-profile of the user. In order to model these psychological properties, we need to build a system that can, (i). be visually aware and able to recognize aspects in an image, (ii). map the visual and textual (dialogue history) aspects with the appropriate aspects in the assigned aspect-preference profile, (iii). generate utterances about the aspects in the images that convey sentiments polarity in accordance with the assigned polarity from the assigned profile.

In this paper, we create a system based on transformers (Vaswani et al., 2017), where we initialize the encoder with the weights of pre-trained BERT (Kenton and Toutanova, 2019). A sentiment embedding is trained and memory network is used to store the aspects for the preference. A combination of sentiment embeddings and the aspects in the memory network represents the aspect-sentiment profile. Image representation is taken from the Resnet (He et al., 2016). Through our experiments, we show that our system is able to map the visual aspects with the preference-profile, and generate utterances that reflect the desired sentiments about the aspects in the image. Since no previous dataset is tailored to deal with this task, we enhance the Image-Chat dataset by assigning preference-profiles to it. This creation and assignment of profiles is done in an automated manner and no manual intervention is needed. Each preference-profile consists of a list of tuples consisting of aspects and sentiment. The sentiment in the tuple denotes the preference ('positive' or 'negative') of the speaker for the respective aspect. We train our model on this new dataset, to achieve our goal of generating utterances that can correctly express their preferences for the aspects present in the visual modality.

The main contributions of current work are as follows: (i). we propose a new type of persona profile consisting of aspect-preferences, (ii). we formulate a new task of generating utterances with profile consistent sentiments for the aspects present in the visual modality, and (iii). we propose a novel system that is capable of generating utterances grounded on the image, while expressing profile consistent sentiments in a conversation.

2 Related Work

Open domain chat-bots have become increasingly popular lately, leading to a release in several datasets. Dinan et al. (2019) proposed a chit-chat dataset where the topic of the conversation is grounded to a paragraph extracted from wikipedia. Rashkin et al. (2019b) proposed EMPATHETIC-DIALOGUES dataset, which is another interesting corpora in chit-chat domain. This dataset is created by giving the speakers an emotion label (like 'afraid', 'proud' etc.) and the speaker is asked to write a paragraph about a situation when they felt that way. Then the speaker is asked to converse with another speaker describing them the story. In this way the built corpora, is grounded to a given 'situation' and an emotion 'label'. In PERSONA-CHAT dataset (Zhang et al., 2018) persona profile in the form of statements about the speaker is associated to each speaker. An open domain conversation then takes place between the speakers while being grounded on their assigned persona. In this paper we propose a new type of 'persona-profile' called 'aspect-preference-profile', associated with the user. In this profile tuples of aspect and their preference (given by 'positive' or 'negative' sentiment) is stored for a speaker. The speakers are given an image and a conversation is built around it. In the conversation the speaker's should express their sentiment about aspects according to the given profile. It is to be noted that these aspects can be present in either images or dialogue history. We modify the Image-Chat dataset given by Shuster et al. (2020) to serve our purpose. This dataset consists of conversations around some image grounded on one of the 215 styles (like 'honest', 'hateful' etc.). A prior work by Firdaus et al. (2021) proposed aspect controlled response generation, where an aspect is given as input and a response is generated containing the aspect term. In contrast to this work, we do not provide any aspect-term for response generation, rather we provide an 'aspect-preference-profile' of the speaker. The aim is to make the user express sentiment in accordance with their preference if and when they give their opinion about some aspect.

3 Methodology

3.1 Problem Definition

The task requires as input an image I, aspect-preference-profile AP_i of the speaker S_i, and

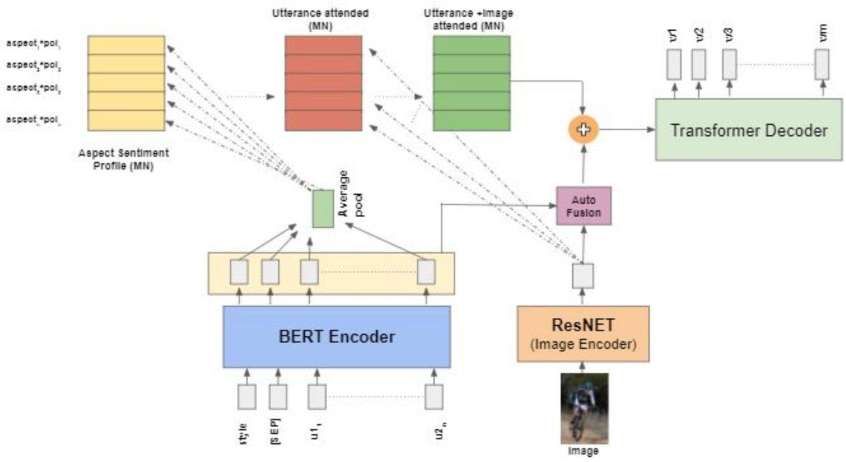

Figure 2: Model architecture of the proposed system. Here $pol_i \in \{positive, negative\}$ is the sentiment embedding for the aspect term $aspect_i$.

utterance history of the conversation $C = \{S_1U_1, S_2U_1, S_1U_2, ..., S_2U_{l-1}\}$. Here, $AP_i = \{(a_1, p_1), (a_2, p_2), ..., (a_n, p_n)\}$, where a_k is an 'aspect', $p_k \in \{\text{"}positive\text{"}, \text{"}negative\text{"}\}$ is the sentiment polarity associated with the aspect and S_jU_m is the m^{th} utterance by j^{th} speaker ($j \in \{1, 2\}$). Given the pre-requisite input, the task is to generate an utterance S_1U_l that is consistent with the conversation history C and expresses appropriate sentiment-polarity for the aspects in AP_1 that are also present in the visual modality, i.e. I.

An example of the task is shown in Figure 1. The image in the figure consists of a person playing a guitar. For speaker $S1$, the aspects most appropriate with respect to the image in AP_1, are 'music' and 'band', both of whose sentiment-polarity are 'positive' (here, $C = \{\phi\}$). The utterance S_1U_1 generated for S_1, thus expresses the 'positive' sentiment for the aspect 'band'. For speaker $S2$, the most appropriate aspect in AP_2 with respect to both $C = \{S_1U_1\}$ and I is 'music' with 'negative' sentiment. This is properly conveyed in the utterance S_2U_1 for the speaker $S2$.

The overall architecture consists of (c.f. Figure 2): (i). Utterance and persona-style encoder, (ii). Image encoder, (iii). Aspect-preference memory, (iv). Aspect selection module, (v). Multi-modal fusion mechanism, and (vi). Dialogue decoder.

3.2 Utterance and Persona-Style Encoder

The Image-Chat dataset associates with each utterance, a style class. This class acts as a control factor in determining the generation style of the responses. In total, there are 215 distinct style types. Since these styles are semantically related (like sweet, happy, eloquent, fickle, frivolous etc.) and not completely orthogonal, we use their semantic embeddings to represent them, instead of using one-hot encoding. We use BERT to encode the utterance history and the style control for the response. As input to the BERT model, we prepare a sequence of the form $SEQ_{ip} = \text{"}ST[SEP][S1]U1[S2]U2\text{"}$. Here, ST is the style type represented by its class name, $U1$ and $U2$ are the previous utterance by the speaker $S1$ and $S2$, respectively. To demarcate these segments, $[SEP]$, $[S1]$ and $[S2]$ are used as special tokens (unused BERT tokens are used). This input sequence SEQ_{ip} is passed as input to the BERT encoder and a hidden representation $H = [h_1, h_2, ..., h_k]$ is obtained.

3.3 Image Encoder

Our model utilizes pre-trained image features obtained from Resnet152 (He et al., 2016), which is a residual network with 152 layers. It is trained on the ImageNet dataset (Russakovsky et al., 2015) to classify images among 1,000 classes. We use the implementation provided in the torch-vision project (Marcel and Rodriguez, 2010). The extracted features I_r for an image I, has 2,048 dimensions. These features are then compressed to the size of hidden representation H by passing it through a trainable linear layer and a representation I_c is obtained.

3.4 Aspect-Preference Memory

The aspect-sentiment-persona profile of a speaker S_i can be represented as a set of tuples AP_i as discussed in Section 3.1. To store these aspects of the profile, we make use of an external memory network M and train $v_{positive}$ and $v_{negative}$ sentiment representations in an embedding matrix. We obtain the fastText embedding for each a_i in the profile and multiply it with the sentiment embedding v_{p_i} associated with it. We obtain the sentiment enriched aspect-embedding m_i for each aspect term a_i. The final memory-network M storing these embeddings would be of the form $M = [m_1, m_2, ..., m_n]$. It is to be noted that M would store the profile of the speaker whose utterance is to be generated.

3.5 Aspect-Selection Module

The aspect-preference profile contains the speakers sentiment with respect to multiple aspects. These aspects may not always be relevant to the given image or the ongoing conversation. Selection of contextually relevant aspects from the memory network needs to be done in order to correctly express sentiment. In order to make this selection, we utilize multi-hop attention mechanism Tran and Niedereée (2018). The attention mechanism works on a query q and an input sequence $IP = [ip(1), ip(2), ..., ip(m)]$. For each k in K hop attention, the following steps are executed:

$$s_t^{(k)} = tanh(W_q^{(k)} ip(t)) \odot tanh(W_g^{(k)} g^{(k-1)}) \quad (1)$$

$$\alpha^{(k)} = softmax(w_s^{(k)^T} s_t^{(k)}) \quad (2)$$

$$o_q^{(k)} = \sum_t \alpha_t^{(k)} ip(t) \quad (3)$$

Here, $W_q^{(k)}$, $W_g^{(k)}$ and $w_s^{(k)}$ are the trainable parameters, and m is a separate memory vector for guiding the next attention step. It is recursively updated using the following equation:

$$g_q^{(k)} = g_q^{(k-1)} + o_q^k \odot q \quad (4)$$

The initial value of vector $g^{(0)}$ is defined based on the context vector $o_q^{(0)}$, given by the equation 5:

$$o_q^{(0)} = \frac{1}{l} \sum_t h_q(t) \odot q \quad (5)$$

The representation $o_q^{(k)}$ is the final attended and summed representation of IP. At each k the representation $o_q^{(k)}$ is added to each step of the encoded representation H and H' is obtained.

The aspect selection is done using both the image representation I_c and the utterance history H as query. For computing attention on the aspect-preference memory M with respect to the image representation I_c, we set $IP = M$ and $q = I_c$ in the attention mechanism. Computing attention on M with respect to the utterance history H requires pooling of the representation (since it is a sequence). We obtain H_m by mean-pooling H and set $q = H_m$ to attend to the memory M by setting $IP = M$. H' is aspect-sentiment enriched hidden state after the attention based selection steps.

3.6 Multi-modal Fusion mechanism

Since our utterance decoder would work on the encoded representations of both, text and image, it is important to obtain a representation that fuses these modalities effectively. We use the auto-fusion mechanism proposed by Sahu and Vechtomova (2019) for this purpose. In this method, the unimodal representations (H' and I_c in our case) are first concatenated to obtain Z_m. These concatenated representations are then passed through a transformation layer to obtain an *autofused* latent vector H''. We then try to reconstruct the originally concatenated vector from the *autofused* latent vector and obtain the representation $\hat{Z_m}$. This is done by training the transformation layers to minimize the Euclidean distance between the original and reconstructed concatenated vector. This process also ensures that the learned vector does not contain arbitrary signals from the input concatenated latent vector. Training the model for the downstream task of response generation further incentivizes the layers to fuse the modality information without losing essential cues. The Euclidean distance between Z_m and $\hat{Z_m}$ is minimized by minimizing the mean-squared-error (J_m) as shown by equation 6.

$$J_m = ||Z_m - \hat{Z_m}|| \quad (6)$$

3.7 Dialogue Decoder

The representation H'' is the final multi-modal encoded representation, that contains the dialogue history, aspect-preference, style and image representation. Our decoder works on this representation to produce the target response $Y_{target} = \{y_1, y_2, ..., y_n\}$. Our decoder consists of d layers of stacked transformer decoder that work on H'', and is trained to reduce the log-likelihood L of generating the target response sequence using equation 7.

$$L = -\sum_1^n log(y_t^{target} | y_{<t}^{target}, X) \quad (7)$$

Here, $X = \{I, SEQ_{ip}, AP_i\}$, and n is the target sequence length.

Experiment	*Word Overlap*			*Semantic Similarity*				*Profile Consistency*	
	BLEU-1	***BLEU-2***	***Rouge_L***	***Ave.***	***Gre.***	***Ext.***	***SkTS***	***ASim***	***ASenti***
Style+AP+Att	**15.93**	**5.6**	**0.157**	**0.82**	**0.65**	**0.50**	**0.52**	**0.54**	**74%**
AP+Att	15.01†	5.3†	0.151†	0.80	0.63	0.41†	0.51	0.47†	66%†
Style+AP	14.37†	4.7†	0.144†	0.79†	0.61†	0.39†	0.48†	0.42†	64% †
Style	13.01†	4.2†	0.129†	0.79†	0.60†	0.39†	0.47†	0.36†	58%†

Table 1: Automatic evaluation results obtained on experiments using different combination of *Style*, *Utterance History (Hist)*, *Aspect Preference Memory (AP)* and *Aspect Selection Attention module (Att)*. Transformer model with BERT as encoder (Section 3) is used for all the experiments. The results using *AP* show improvement over the baseline using only *Style*. The results marked by "†" are significantly worse than the results of the experiment *"Style+Hist+AP+Att"* in t-test with $p < 0.05$ level.

Experiment	*Fluency (F)*	*Aspect Relevance (AR)*	
		IAR	***DAR***
Style+AP+Attn	**2.84**	**1.98**	**2.24**
AP+Attn	2.77	1.82	2.08
Style+AP	2.71	1.73	1.96
Style	2.61	1.68	1.92

Table 2: Manual evaluation results measuring *Fluency (F)*, *Image Aspect Relevance (IAR)* and *Dialogue Aspect Relevance (DAR)*.

4 Dataset Creation

In this section, we discuss corpus creation process and the models build for the purpose of building the dataset[2].

4.1 Dataset

For our task we enhance the Image-Chat dataset (Shuster et al., 2020) with aspect-preference profiles for the speakers. The aspect-preference for a speaker should reflect in their utterances in the form of sentiments. Therefore, we cannot assign arbitrary sentiments to the aspects mentioned in the utterances. Manually looking for aspects in utterances and putting these aspects with correct sentiment in the preference-profile is a time consuming and expensive task. Fortunately, aspect extraction and aspect-sentiment classification tasks have been well explored and have several publicly available datasets. We use datasets from SemEval 2014, 2015 and 2016 (ABSA task) to train BERT based aspect extraction and aspect-sentiment classification systems. We only consider positive and negative polarities for our experiments. We use our trained models to extract aspects and their sentiments from the dialogues in the Image-Chat dataset. For a speaker in the conversation, the aspects and sentiments extracted from their utterances are kept in the preference-profile. We limit the number of aspect-sentiment pairs in the profile to 15. If the extracted aspects from the speaker utterance do not complete the profile, the rest of the aspect-sentiment slots in the profile are filled by randomly selecting aspects and assigning them random sentiments. These random aspects act as distractors and forces the model to learn how to ignore the irrelevant aspect and focus on only the aspects that are relevant to the image and the conversation history. The speaker's profile remain same throughout the conversation. Some conversations in the Image-Chat dataset do not contain aspect term in any of the utterances, such conversations are removed from the dataset. Even if one utterance containing aspect-sentiment pair is present in the conversation, the conversation is kept in the dataset. The detailed statistics of the dataset is given in the appendix B.

4.2 Aspect Based Sentiment Analysis

We train a pipeline of BERT-based models for aspect extraction and aspect level polarity classification. We utilize the ABSA SemEval dataset (Pontiki et al., 2014, 2015, 2016) for this purpose. These trained models are used to extract aspects and detect their sentiment polarities from the utterances of Image-Chat dataset[3]. We pose Aspect term extraction task as a sequence classification problem with BERT using the IOB2 format, where I, O and B denote Intermediate, Outside and Beginning. (Sang and Veenstra, 1999). This BERT model was fed the whole sentence as the input segment and it obtained an F1-score of 0.8012 (evaluation carried out similar to Sang and Buchholz (2000)). The sentiment polarity prediction task is posed as a sentence-pair classification problem for the BERT model, where the sentence is provided as the first segment and the aspect-term as the second segment at the input. The model trained in this manner, obtained an F1-score of 0.9080 for the positive polarity and 0.8239 for the negative polarity on the ABSA SemEval dataset.

[2]The implementation details for all the experiments are given in appendix A.1

[3]The quality of the extracted aspect-polarities is discussed in appendix C

	***Aspect-Preference:*** (shopping, positive), (geneology, negative), (stage, negative), (muses, negative), (chess, negative), (flat, positive), (center, positive), (military, negative), (goods, negative), (lands, positive), (washing, negative), (baked, negative), (mindset, negative), (car, positive), (lightning, negative) **Style:** shy
	Output: i would love to drive that car .
	Aspect-Preference: (shopping, positive), (geneology, negative), (stage, negative), (muses, negative), (chess, negative), (flat, positive), (center, positive), (military, negative), (goods, negative), (lands, positive), (washing, negative), (baked, negative), (mindset, negative), (car, negative), (lightning, negative) ***Style:*** shy
	Output: i would never ride that car .
	***Aspect-Preference:*** (jersey, positive), (bee-hive, positive), (peons, positive), (cheers, negative), (building, positive), (love, positive), (road, negative), (bait, positive), (little, negative), (lemon, positive), (striped, positive), (os, negative), (visual, negative), (arts, negative), (varmit, positive) ***Style:*** passionate
	Output: this building is so beautiful , i love the architecture .
	Aspect-Preference: (jersey, positive), (bee-hive, positive), (peons, positive), (cheers, negative), (building, negative), (love, positive), (road, negative), (bait, positive), (little, negative), (lemon, positive), (striped, positive), (os, negative), (visual, negative), (arts, negative), (varmit, positive) ***Style:*** passionate
	Output: this house is so sad .
	***Aspect-Preference:*** (shirt, negative), (dress, negative), (guy, negative), (sea, positive), (sweat, negative), (virtue, positive), (shop, negative), (waiter, negative), (white, negative), (chinese, negative), (horse, positive), (arena, positive), (land, positive), (veteran, negative), (baseball, positive) ***Style:*** abrasive (annoying , irritating)
	Output: that shirt is so ugly , i hate it .
	Aspect-Preference: (shirt, positive), (dress, positive), (guy, negative), (sea, positive), (sweat, negative), (virtue, positive), (shop, negative), (waiter, negative), (white, negative), (chinese, negative), (horse, positive), (arena, positive), (land, positive), (veteran, negative), (baseball, positive) ***Style:*** abrasive (annoying , irritating)
	Output: this guy is so annoying

Table 3: Some interesting examples showing the effect of changing profile on generated utterance on with same image as input. In all these examples the first utterance is generated, showing the ability of the model to map relevant aspects in preference-profile to the image.

5 Evaluation Metrics, Results and Analysis

5.1 Evaluation Metrics

We report the results of our experiments for both automatic and human evaluation metrics. In automatic evaluation, we use **BLEU-1**, **BLEU-2** and **Rouge-L** to measure word overlap between the generated response and the gold response. The higher their value the more the overlap. To measure semantic relevance between the generated and gold response, we utilize the embedding based evaluation metrics. More specifically, we use the embeddings of bag-of-words to represent both the generated and ground-truth response, and calculate their **Average similarity (Ave.)**, **Greedy similarity (Gre.)**, and **Extrema similarity (Ext.)**. Apart from embeddings for bag-of-words, we obtain the sentence vector representation (Skip-Thought) for both the generated and gold response, and compute cosine similarity between them to obtain the **Skip-Thought-similarity (SkTS)**[4]. Along with the aforementioned automatic evaluation metrics, we also need to compute the consistency of our outputs with respect to the aspect preference-profile. In order to measure this, we introduce two more automatic evaluation scores to compute **Aspect similarity (ASim)** and **Aspect-sentiment match (ASenti)**. Aspect similarity computes the average cosine similarity between *fastText* word embeddings of the aspects present in aspect-preference-profile and the predicted utterance. We compute the embedding

[4]we use nlg-eval to compute these scores https://github.com/Maluuba/nlg-eval

	***Aspect-Preference (S2) :*** (rattlesnakes, negative), (cars, negative), (lor, negative), (detail, negative), (soil, positive), (fields, positive), (pattern, negative), (poultry, positive), (coal, positive), (demographic, positive), (architectural, negative), (cupcakes, negative), (designs, negative), (swirl, negative), (rotting, positive) ***Style (S2):*** solemn ***S1 (utterance):*** a sacred place .
	Output (S2 [style]): i don't know what that is. ***Output (S2 [style + AP]) :*** it is a terrible building . ***Gold:*** would be a lot more sacred if it werent for the cars around. disgusting to see how industrialization is soiling history and faith.
	***Aspect-Preference (S2):*** (life, positive), (yellow, negative), (groundskeeper, positive), (cream, negative), (trash, positive), (clump, negative), (bodies, positive), (sodas, negative), (wire, negative), (office, positive), (compactor, positive), (whipped, negative), (rough, positive), (barefoot, negative), (students, positive) ***Style (S2):*** impersonal ***S1 (utterance):*** those kids look so uninterested, are schools even trying to engage students anymore?
	Output (S2 [style]): i don't think they are doing anything . ***Output (S2 [style+AP]):*** they are probably just having a good time . ***Gold:*** the kids are learning .

Table 4: Analysis of generated utterances having previous conversation history.

similarity between aspects, as aspects generated by the model may not exactly match with that in the aspect-preference-profile, but may still be semantically similar and therefore correct (e.g. "girl" and "lady"). To obtain the aspect-sentiment match we compute the percentage of instances where the sentiment of the aspect in the generated output matches that in the profile. We use the trained BERT based aspect-extraction and aspect-sentiment detection model (as discussed in Section 4.2) to obtain the aspects and their sentiments from the generated outputs. The results obtained on the automatic evaluation metrics is given in Table 1.

In manual evaluation, we compute **Fluency (F)** to measure the grammatical correctness or readability of the generated response. A generated utterance may contain aspects which may or may not be relevant to the given image or conversation history (even if they appear in the aspect-preference-profile). We need to measure the **Aspect Relevance (AR)** of the response generated by looking at the given image and conversation history. We divide the **Aspect Relevance** into two parts, *viz.* (i). *Image Aspect Relevance (IAR):* It measures whether the aspects in the generated utterance are relevant to the given image; (ii). *Dialogue Aspect Relevance (DAR):* It measures if the aspects generated are attuned to the aspects mentioned in the previous context of the dialogue. Three human experts with post-graduate qualifications were asked to rate 100 responses generated from the proposed model. These experts are the regular employees in our research group and have approximately 2 years of experience for the similar work. They were asked to give a score of 1/2/3 for bad/normal/good quality to rate both Fluency and Aspect Relevance. The results of manual evaluation are shown in the Table 2[5].

5.2 Analysis

The results obtained for both automatic and manual evaluations (Table 1 and Table 2 respectively), clearly show that best results are obtained when both *style* and *aspect-preference memory* is used in conjunction with *aspect selection attention module*. In automatic evaluation, it can be seen that removing *style* from the experiment results in marginal drop in all metrics. The most significant drop occurs is observed in *ASenti* ($\downarrow$ 8% points). A reason for this drop is that, many categories of *style* often co-relates with the sentiments expressed in the utterance. Since the *AP* are constructed using aspect-sentiment association extracted from these utterances, often the sentiment expressed for an aspect plays a major role in determining the style of the utterance. As an illustration, for a *style* of type "hateful", the hate is often expressed towards some aspects; which in turn results in the aspect having negative sentiment associated with it in the *AP*. Removing attention based selection mechanism leads to a big drop in BLEU-1 ($\downarrow$ 1.56 points) and BLEU-2 ($\downarrow$ 0.9 points). The drop in *ASim*

[5]The inter-annotator agreement using Krippendorff's alpha (Krippendorff, 2011) was found to be 0.87, 0.81 and 0.83 for *F*, *IAR* and *DAR* respectively

($\downarrow$ 0.12) is expected due to the lack of specialized aspect-sentiment selection mechanism, resulting in the drop in *ASenti* ($\downarrow$ 10% points) too. Using only style as the control parameter yields significantly lower word overlap scores ($\downarrow$ 2.92, $\downarrow$ 1.4 and $\downarrow$ 0.028 for BLEU-1, BLEU-2 and Rouge_L respectively). In terms of profile consistency too there is a huge drop in *ASim* ($\downarrow$ 0.18 points) and *ASenti* ($\downarrow$ 16% points). Experiment using only *style* under-performs considerably in terms of all the *semantic similarity* based metrics too.

The manual evaluation results further confirm the importance of every component of our experiment. It can be seen from Table 2 that using *style*, *aspect-preference-profile* with *attention based selection*, produces the best results in terms of both *fluency (F)* and *aspect relevance (AR)*. It is interesting to observe that *image aspect relevance (IAR)* is lower than *dialogue aspect relevance (DAR)* for all the results. This shows that correctly mapping aspects in image with those in *AP* is far difficult than doing such mapping from textual dialogue history.

Table 3 shows some example outputs of first utterance in the conversation, where the generated output is based on a given image and style, along with the assigned *AP* for the speaker. The first two examples show that changing the sentiment of the aspects 'car' and 'building' from positive to negative, produces the utterances that correctly reflect these changed sentiments. It is interesting to note that in the second output of the second example the aspect term 'house' is produced in the utterance. The aspect term in the *AP* closest to this is 'building'. The generated output often does not contain an exact term mentioned in *AP*, but produces an aspect similar to it (e.g. 'house' and 'building' are interchangeable in this case). The third example is a great instance where the relation between style and sentiment is captured. In the first part of the example, the output produced expresses negative sentiment towards the aspect 'shirt', which is consistent with the sentiment of similar aspects ('shirt' and 'dress') in the *AP*. When we flip the sentiment of these aspects in *AP* (from negative to positive). The output produces a response that expresses negative sentiment towards the aspect term 'guy', which is again consistent with *AP*. This happens because the style of generation was set as 'abrasive (annoying, irritating)'. This style of generation would mostly contain negative sentiments. Therefore changing the sentiment to positive, merely makes the model focus on the next most relevant aspect in *AP* with a negative sentiment.

Table 4 shows example output utterances having some conversation history. The generated outputs are compared to the gold responses and the outputs generated using only style (ignoring the *AP*). In the first example it can be seen that output of our model expresses negative sentiment about the aspect 'building' (present in the image). Despite 'building' not being present in *AP*, our model focuses on the the most similar aspect to the image, i.e. 'architecture' (with sentiment 'negative' associated with it). Although worded very differently, the response manages to express similar sentiment as that in gold. The response is also relevant to both the dialogue and the image. In contrast the response generated using *style* only is very generic and not very relevant to the conversation. Similar phenomenon can be observed in the second example (Table 4), where the image and conversation context map to the sentiment of the aspect 'student' in the profile. An interesting observation here is that the aspect-term is not mentioned in output, instead a pronoun 'they' referring to the term 'kids' in the previous utterance is produced. Although the output is consistent with the profile, image and dialogue-history; such samples are missed while computing *ASim*, reducing the evaluation-score. The response generated by using only *style*, conveys negative sentiment to the target-aspect; contradictory to the sentiment in the gold response.

6 Conclusion

In this paper we propose a new task of controlling the output of a chat-bot by grounding it to an 'aspect-preference-profile'. This profile consists of a list of aspect-sentiment tuples. We obtain a dataset for this task by enhancing the Image-Chat data with such profiles. Since this corpora is made up of conversations around images, the aspects whose sentiment are controlled can be present in both visual and textual (dialogue history) modality. Next, we create a system using BERT, ResNet and Memory network based encoder-decoder model, that can produce responses around image and dialogue history, while still being grounded to an assigned 'aspect-preference-profile'.

Relationship between 'style' and 'aspect-sentiment' can be explored as an interesting case-study for future work.

7 Ethical Declaration

We use a freely available dataset under MIT license to create our new dataset. The dataset has been used only for academic purposes, in accordance with the license. The dataset created in this work will be made available only after filling and signing an agreement declaring that the data will be used only for research purposes. The annotation for manual evaluations was done by human experts, who are the regular employee of our research group. There are no other issues to declare.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. 2021. Aspect-aware response generation for multimodal dialogue system. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(2):1–33.

Judith M Gardner, Bernard Z Karmel, and Michael J Flory. 2003. Arousal modulation of neonatal visual attention: Implications for development.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Sébastien Marcel and Yann Rodriguez. 2010. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1485–1488.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019a. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019b. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Gaurav Sahu and Olga Vechtomova. 2019. Adaptive fusion techniques for multimodal data. *arXiv preprint arXiv:1911.03821*.

Erik F Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. *arXiv preprint cs/0009008*.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*, pages 173–179. The Association for Computer Linguistics.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2414–2429. Association for Computational Linguistics.

Nam Khanh Tran and Claudia Niedereée. 2018. Multihop attention networks for question answer matching. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 325–334.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Henry M Wellman and Jacqueline D Woolley. 1990. From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, 35(3):245–275.

Jerry S Wiggins. 1996. *The five-factor model of personality: Theoretical perspectives*. Guilford Press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

All the models were implemented using PyTorch (Paszke et al., 2017). The BERT model was implemented using the transformers library (Wolf et al., 2019). Models are trained with an initial learning rate of 1e-4 with a linear schedule and a warmup (Vaswani et al., 2017), using the Adam Optimizer (Kingma and Ba, 2015). Mini-batches of size 12 were used during training. For storing aspect representations on memory network, fastText embeddings (Bojanowski et al., 2017) were used. The models were each trained for 40 epochs on our modified Image-Chat dataset. K in k-hop attention was set to 3. The dimension of the hidden state H was 512, while the dimensions of aspect embeddings obtained from *fastText* was 300. The decoder consists of three stacked transformer ($d = 3$) decoder. The total number of parameters in the model was 199,126,175. The best model based on validation loss was saved, and with five runs for each experiment. The experiments were conducted on GeForce RTX 2080 Graphics Processing Unit (GPU) with a GPU memory of 11,019 MBs. On a batch size of 12, average time taken per epoch was 3 hours.

B Dataset Statistics

Table 5 shows the data statistics of our preference-profile enhances dataset. Conversations from Image-Chat data for which no aspect could be extracted, are removed. In total 64,911 unique aspects were extracted from utterances of Image-Chat dataset. Table 6 shows the data statistics of the the SemEval dataset on which our aspect-extraction and aspect-sentiment classification models were trained.

Split	***Train***	***Test***	***Valid***
Number of Images	163,940	7,467	3,725
Number of Dialogues	163,940	7,467	3,725
Number of Utterances	287,338	22,400	11,174

Table 5: Dataset statistics of the enhanced Image-Chat data. Conversations not containing any aspect-term is dropped.

Split	***# Sentences***	***# Aspects***	***# Unique Aspects***
SemEval (Train + Valid)	2,242	4,016	1,437
SemEval (Test)	401	513	269

Table 6: Dataset statistics of SemEval dataset

C Outputs

Table 7 shows the some utterances from Image-Chat dataset, with extracted aspects and their sentiments, using the BERT models trained on SemEval ABSA datasets. It can be observed that despite being trained on reviews dataset, the models work well when extracting aspects and their sentiments from utterances too. Table 8 shows some more example outputs from our model, showing how using *AP* helps in expressing sentiments for an aspect in accordance with the preference.

Utterance	*Extracted Aspect*	*Aspect-Sentiment*
home sweet home	home	positive
its a house, so like it	house,	positive
how can you get any work done in such a disorganized office?	work	negative
is the street skewed or am i just kind of drunk?	street	negative
this ugly box of a building should be torn down and turned into tombstones.	box	negative
this ugly box of a building should be torn down and turned into tombstones.	building	negative
i can't wait to buy these shirts for my mom's birthday!	shirts	positive
she knows my jokster mentality	jokster	positive
these flowers look very expensive.	flowers	negative
expensive or not they look amazing to make a bouquet out of.	bouquet	negative
i would love to put some of these in a vase to set on a window sill.	vase	positive
you were in band before, when was that?	band	negative
these are the most disgusting candies. if you like them you should be ashamed of yourself.	candies.	negative
i love cockpit shots like this.	shots	positive
this band was good but a little too up-tempo for me.	band	positive
the best part about that band is their promotional art, i don't think they sound very good.	promotional art,	negative
oh no, did i feed my fly traps lately?	fly traps	negative
you will get used to it after the headache goes away.	headache	negative
i would love to have dinner as the sun sets with my loved on facing this statue.	dinner	positive
looking at those windows makes me want to throw rocks at them.	windows	negative
now if only a little bird would land on the beam so i can take a pretty picture.	bird	positive
the view always turns me on.	view	positive

Table 7: Random samples from Image-Chat dataset for which aspects are extracted and their sentiment are assigned. The BERT based models discussed in Section 4.1 is used for this purpose.

	***Aspect-Preference (S2) :*** (ceilings,’negative), (around, negative), (squares, positive), (celebrations, negative), (tour, negative), (hind, negative), (industrial, negative), (comission, positive), (wrap, negative), (safari, negative), (plants, negative), (topography, positive), (key, positive), (antenna, negative), (business, negative) ***Style (S2):*** fearful ***S1 (utterance):*** i wonder how many years a flower like that will bloom or if it will even bloom with different colors depending on the soil conditions.
	Output (S2): what if the plant dies ? ***Gold:*** i am scared this plants might give me a rash.
	***Aspect-Preference (S2):*** (cycle, positive), (brontosaurus, negative), (guitar, negative), (muss, negative), (sundays, negative), (brook, positive), (chemical, negative), (red, positive), (pandas, positive), (ointment, positive), (ponys, positive), (photowork, negative), (husk, negative), (air, negative), (mountain, positive) ***Style (S2):*** eloquent (well-spoken, expressive) ***S1 (utterance):*** this is the most amazing place i have ever laid my eyes on. lets go!
	Output (S2): i would enjoy the view of the mountain range. ***Gold:*** i agree , i think that walking along side that babbling brook and wonderful mountain would be great.
	***Aspect-Preference (S2):*** (bicycling, positive), (mixer, negative), (scene, positive), (glue, positive), (statue, positive), (factories, negative), (valve, positive), (showgirl, negative), (power, negative), (shirts, negative), (varooom, negative), (plaid, negative), (white, positive), (cocaine, negative), (washer, negative) ***Style (S1):*** complex
	Output (S1): this statue is so beautiful , i would love to see it in person . ***Gold:*** the white statue against a deep blue sky is every artist’s dream.
	***Aspect-Preference (S2):*** (migrants, negative), (lump, negative), (of, negative), (pine-cone, negative), (mollusks, negative), (orlando, negative), (of, negative), (proteins, negative), (style, negative), (music, negative), (pet, positive), (building, positive), (parrot, positive), (coal, negative), (eye, negative) ***Style (S1):*** meticulous (precise,thorough)
	Output (S1): this building is very well designed. ***Gold:*** the artistic skill and precision it took to build this building is amazing .
	***Aspect-Preference (S1) :*** (golbins, positive), (metal, negative), (plane, negative), (exotic, positive), (planes, negative), (rainbow, positive), (space, positive), (tutu, positive), (carpe, positive), (ad, positive), (payment, negative), (tulle, positive), (firm, negative), (password, positive), (stuff, negative) ***Style (S1):*** morbid ***S1 :*** i wonder how many planes that blue one shot down ***S2:*** the situation it was in was pretty horrific, i cannot imagine the anguish and pain.
	Output (S1): that plane will crash into the plane ***Gold:*** this plane has probably caused some death.

Table 8: Some output samples from our model using *AP*