論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Broadening the Context: User-specific Information and Visual Scenes for Conversation Tasks
著者(和文)	FIKRI ABDURRISYAD
Author(English)	Abdurrisyad Fikri
出典(和文)	学位:博士（工学）, 学位授与機関:東京工業大学, 報告番号:甲第11757号, 授与年月日:2022年3月26日, 学位の種別:課程博士, 審査員:奥村 学,熊澤 逸夫,中山 実,篠崎 隆宏,船越 孝太郎
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11757号, Conferred date:2022/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Doctoral Thesis

Broadening the Context: User-specific Information and Visual Scenes for Conversation Tasks

Abdurrisyad Fikri

Department of Information and Communications Engineering
School of Engineering
Tokyo Institute of Technology

Supervisor: Manabu Okumura

November 19, 2021

Abstract

In real-world conversations involving humans, a conversation context is often dynamic, not only depending on the conversation history (previous utterances) but also by being influenced by other factors such as speakers' personalities, surrounding objects, etc. Moreover, the same message can be responded to in various manners by different persons.

Producing human-like responses is challenging. While the responses are expected to be correct or relevant, diversity or engagingness is also often considered as a human-like quality. We argue that incorporating additional information could help the model capture the conversation context and improve the conversation tasks' performance. Following this postulate, we have two aims in our study: (1) to drive generated responses to resemble a real person's style, and (2) to provide a better conversation context through a multi-modal (visual and language) approach. To realize these aims, we experimented on the models for each aim: (1) response generation models with user-specific information, and (2) response classification models augmented with conversation scene images.

Our experiment results show that our first model can produce more resembling responses to the actual users than the baseline, and using conversation scene images can improve the response classification task performance in the second work.

Contents

Abstract **i**

List of Tables **v**

List of Figures **vi**

1 Introduction **1**
1.1 Background 1
1.2 Objectives 3
1.3 Field of Study 4
1.4 Contributions 5
1.5 Outline of Thesis 6

2 Related Work **7**
2.1 Response Generation Tasks 7
2.2 Visual and Language Conversation Tasks 10
2.2.1 Image-based Conversation Datasets 10
2.2.2 Vision and Language Models 11

3 User-specific Response Generation **13**
3.1 Task Definition 13
3.2 Datasets 13
3.2.1 Conversation Dataset 13
3.2.2 User-info Dataset 14
3.3 Method 15
3.3.1 Encoder-Decoder Model 15
3.3.2 User-specific Information 16
3.3.3 Attentional Conversation Model 18
3.4 Experiment 19
3.4.1 Data 19
3.4.2 Implementation 20
3.4.3 Baseline and Comparison Models 21
3.4.4 Evaluation Setup 22
3.5 Results and Discussion 25

3.5.1 Automatic and Human Evaluation 25
3.5.2 Discussion 28

4 Response Classification with Visual Scene Dataset 31
4.1 Scenario and Task Definition 31
4.1.1 Scenario Description 31
4.1.2 Task Definition 31
4.2 Dataset 32
4.2.1 Dataset Construction 32
4.2.2 Dataset Analysis 33
4.3 Experiment 35
4.3.1 Data 35
4.3.2 Transformer-based Baselines and Error Analysis 36
4.3.3 Conversation Secene Image Augmented Model 39
4.4 Result and Discussion 41

5 Conclusion and Future Work 46
5.1 Conclusion 46
5.2 Future Work 47

Acknowledgement 48

References 49

List of Publication 60

List of Tables

1 Existing dialogue datasets associated with visual information. 10

2 Comparison models on the effectiveness of the proposed user-specific information features. 22

3 Automatic (BLEU, perplexity, distinct) and Human evaluation results for *fluency* and *relevance*, presented as raw score percentages. UNK + Info does not have validation set hence no perplexity for this model. Our UNK + Info model with unseen users gains 26.5% more for fluency and 36.5% more for relevance compared to the baseline. 25

4 Results on *fluency* and *relevance* criteria of four alternative models. From table on the left (a), we can learn that the vanilla Seq2Seq model achieved the highest acceptable response on *fluency*. However, from table on the right (b), in average, UNK + Info achieved the highest scores on both criteria. . . . 27

5 Table on the left (a) shows that UNK + Info achieved better style compared to the baselines with user-specific information. In the table on the right (b), we further compared the outputs from UNK + Info and DialoGPT in the preferable measure. We argue that the low agreement among the evaluators (0.2068 on Fleiss' Kappa) shows the similar quality between the two response alternatives. 28

6 Examples of responses from different users generated by our model, using known users and their user-info respectively, and its variant model, using unseen users and the same user-infos. 29

7 Quality assurance for image-conversation in terms of utterance-response *relevance*, *object* and *speakers* appropriateness, and *naturalness*. 35

8 Results of our models with different input settings. *Original images* were inputted as *scene images* by default. S, R, and O denote *speaker*, *respondent*, and *topic object* cropped images, respectively. Difference between a pair of * marked scores was not statistically significant. Differences between all other pairs (from our model) were significant (with T-test, $p < 0.05$). . . 42

List of Figures

1 Woman's response (picking girl up) can be fully understood only by seeing whole scene. Blue boxes marked with S and R denote a *speaker* and a *respondent*, respectively; *Topic object* is marked with yellow box. 3

2 Overview of our neural conversation model with attention to user-specific information. We use two-layer LSTM for both encoder and decoder. The attention layer attends to source hidden states $\bar{h}_{s'}$ and user-info embeddings P_i for user i. User embeddings u_i are concatenated to decoder input at every step. 19

3 Example of a user's Twitter bio and sample tweets used in style evaluation. We censored any mentions of other accounts. 24

4 Distribution of conversation topics in VCSD. Each conversation is represented by GloVe vector and colored according to its cluster, i.e., Food, Fashion, Sport, and Animal. 33

5 LXMERT failed to predict this pair as *correct*, while BERT and our model correctly predicted it. 37

6 Examples of visual conversation types: image-referring response (top) and scene-understanding (bottom). 38

7 (a) Overview of our model architecture being used in experiments. (b) We applied attention over visual features. 39

8 Cropped object image (yellow box) provides more information for understanding the conversation context. 43

9 Example case when both BERT and our model failed to predict the positive response correctly. 44

1 Introduction

1.1 Background

In recent years, dialogue or conversation tasks have been gaining a lot of attention (Li et al., 2016b; Zhou et al., 2017b; Gao et al., 2019; Zhang et al., 2020a; Zheng et al., 2020). In favor of a large amount of textual data available on the Internet, fully data-driven neural network models are now dominating the dialogue tasks. Starting with simple end-to-end sequence-to-sequence models, that simply generate responses given the contexts (Ritter et al., 2011; Sordoni et al., 2015a; Sutskever et al., 2014), some studies have introduced various approaches to integrate more information and features in an attempt to improve the quality of the responses.

While mimicking actual human conversations is the ultimate goal, as it is proven to be difficult, studies on conversation tasks often need to focus on limited aspects. Also, in real-world conversations involving humans, conversation context is often dynamic, not only depending on the previous utterances but also by being influenced by other factors such as speakers' personalities, surrounding objects, etc. One particular problem that is repeatedly being addressed is that neural network models tend to generate safe and general responses, e.g., "I don't know" or "I'm OK" (Sordoni et al., 2015b; Li et al., 2016b). While this type of response is not wrong, actual conversations involving humans would produce more interesting and non-monotonous responses. Moreover, the same utterance might elicit various answers from different people. Therefore, we argue that incorporating more information to improve the model's ability to "understand" the conversation context is necessary.

In conversation tasks, using single-modality (text-only) is the most straightforward method. However, according to recent studies on multi-modal tasks (Mostafazadeh et al., 2017; Das et al., 2017; Murahari et al., 2019), adding corresponding images grounded to the conversation context has proven useful in improving model performance. However, in existing studies on multi-modal conversation tasks, it is common to use topic/object images as visual clues. We argue that to understand the conversation context better, a visual

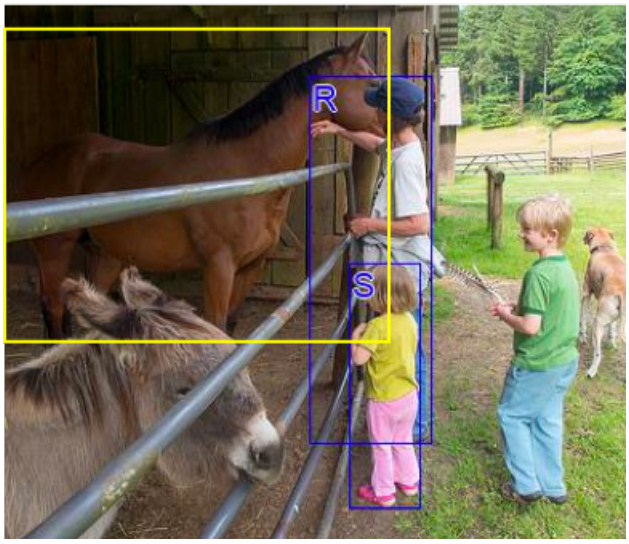
clue that includes the conversation scenes might be more useful than only the object.

In order to satisfy our intention to improve the quality of automated responses and also to investigate the effectiveness of visual conversation scenes in conversation task performances, we have two aims in our study: (1) to produce more diverse and human-resembling responses, and (2) to improve conversation task performance through better context with visual conversation scenes.

For the first aim, following several studies on user-specific features or persona for conversation tasks (Li et al., 2016b; Bak and Oh, 2019; Li et al., 2020), we experimented on capturing the user-level characteristic to drive a response generation model to generate "stylized responses" that resemble the ones from an intended user. We define the "stylized response" as a response that contains frequently used words or characters from the user. To conduct this experiment, we crawled Twitter as our source to construct the dataset. In addition to the conversation text, in the format of tweets and replies, we also collected the usernames as the user-identity features and individual non-reply tweets from each retrieved username, that was converted to user-specific information later.

For the second aim, while other studies on visual and language models for dialogue tasks mainly focused on topic object images as a visual clue, as we previously mentioned, we intend to explore the potential of scenic images in conversational settings. For instance, in Figure 1, since we can see the conversation scene, we can understand why the woman needs to pick up the little girl because the horse is taller than her. Also, in some real-world cases, for example, a conversation agent (robot) sometimes witnesses an interaction between humans in front of it; and such a situation requires the capability to understand a conversation from a third-person view.

To investigate the impact of visual conversation scenes on conversation tasks, we constructed a new third-person-view visual conversation scene dataset (VCSD). This proposed dataset provides scenic images corresponding to conversation utterances; each pair of data consists of (1) an original image, (2) the first utterance and its response, and (3) the corresponding speaker, respondent, and topic object annotations in the conversation image. Given the



Girl: Can I pet it?
Woman: Yes, let me pick you up so you can pet it.

Figure 1: Woman's response (picking girl up) can be fully understood only by seeing whole scene. Blue boxes marked with S and R denote a *speaker* and a *respondent*, respectively; *Topic object* is marked with yellow box.

popularity of transformer-based models, and to confirm our hypothesis that visual clues positively improve the task performance, we first experimented with a BERT (Devlin et al., 2019) (text-only) model as a baseline. We then proceeded to use our proposed dataset with the existing visual-and-language model and our proposed model. In our proposed model, we treated different focused part(s) of the images as additional image features, as opposed to other models which treated the visual input as regional features (Das et al., 2017; Tan and Bansal, 2019). By treating focused part(s) of the images as separate image features, we argue that it would increase the model's capacity to learn and associate the visual information with textual context better.

1.2 Objectives

Our research objectives are as follows:

- Proposing a model/framework and learning strategy that can pro-

duce responses that resemble human-produced ones: By utilizing user-specific information, that consists of usernames and the set of users' own frequently used words, we propose a response-generation framework that can produce stylized responses that resemble the corresponding users.

- Proposing a multi-modal conversation dataset: Visual Conversation Scenes Dataset (VCSD). To our knowledge, this dataset is one of the first attempts to address the impact of visual conversation scenes in conversation tasks. We selected images from the Visual Genome[1] dataset, that contains scenes where a conversation is considered possible, and then annotated the images with conversation utterances. We believe this dataset would be fruitful for future research in bridging visual and language conversation tasks. To propose a strategy for capturing both language and visual contexts, especially using scenic images, we propose an approach for using various parts of one scenic image to optimize the model's ability to associate the corresponding textual context to the visual one.

1.3 Field of Study

A language is a primary tool for human communications in daily life, involving speaking, listening, reading, and writing. Languages that are commonly used and naturally developed by humans are called natural languages. The study of natural languages from the computational perspective that focuses on the interactions between computers and human languages is called Natural Language Processing (NLP).

One of the commonly studied tasks in NLP is dialogue systems or conversational agents. Please note that the terms 'dialogue' and 'conversation' are often used interchangeably. This task focuses on the interaction between computers and humans in a conversational setting. Generally, the computer agents would learn how to respond to a given utterance by humans. These conversational agents can employ one or more input modalities, such as text, speech (audio), graphics (still images or videos), etc.

[1] https://visualgenome.org

In our study, we address the possibility of using specific user-level features to enrich the system's ability to respond to the given utterance. In conversation tasks, it is common to have a pair of previous utterance(s) as the context and a response utterance as the expected output. However, using only previous utterance(s) might be insufficient to generate engaging responses as humans do. Therefore, we believe that adding additional features besides the conversation utterances would improve the system's performance.

The recent advance of computer vision fields, such as image recognition and object detection, also impacts the NLP fields, including the dialogue/conversation tasks. Several studies show that combining information extracted from images (still or videos) corresponding to the conversation text utterances could improve the performance for the tasks. In our study, we explore the potential of using a new type of image to improve the conversation performance.

1.4 Contributions

Our study offers contributions to the NLP field as follows:

- We propose a response-generation model/framework that can produce stylized responses driven by user-specific information. Using two types of user-specific features, our proposed model can handle unseen/unknown users, i.e., users that do not exist in the training data, and generate more resembling responses to the actual users than the baselines.
- We introduce a new multi-modal conversation dataset. To our knowledge, our third-person-view dataset is one of the first attempts to address the impact of visual conversation scenes in conversation tasks.
- We investigate and analyze the effectiveness of using a visual scene and its components to improve the conversation task performance. Our analysis offers a new approach for using scenic images, as opposed to commonly used topic images, as a visual clue. We also propose a neural network model that achieves better results than the baselines.
- We propose new visual conversation types that can help navigate future work on cross-modal conversation tasks.

1.5 Outline of Thesis

This thesis presents frameworks and strategies to improve the performance of conversation tasks, especially in user-level resemblance and multi-modality (visual and language) settings. This thesis also introduces a new multi-modal dataset that contains conversation scenic images and its focused parts (speaker, respondent, and topic object) and the corresponding conversation utterances. We outline our thesis as follows:

Chapter 2 describes the work related to two sub-tasks in this study, response generation tasks (in general and the varieties) and visual and language (multi-modal) conversation tasks.

Chapter 3 describes our framework and strategy to improve response generation tasks using user-specific information. The chapter starts with the task definition, dataset construction, method, and experiments and their results.

Chapter 4 describes our framework and strategy for improving conversation tasks with the visual scene dataset. This chapter starts with the scenario description and task definition, which is essential to understand the strategy on the dataset construction. We then continue to describe the experiments and the results.

Chapter 5 describes our conclusion from our research and its sub-tasks. We also highlight the main contributions of our research and provide insights for future work.

2 Related Work

2.1 Response Generation Tasks

Research on neural conversation models has been progressing particularly following the adaptation of sequence-to-sequence encoder-decoder framework of Sutskever et al. (2014). Some earlier studies are using previous utterances to predict the next utterance (Vinyals and Le, 2015; Sordoni et al., 2015a; Shang et al., 2015). However, as it has been acknowledged that one input can elicit various responses (Shang et al., 2015; Li et al., 2016a), using only previous utterances as the conversation context might not be sufficient to produce human-like natural responses. To address this issue, several approaches have been proposed: either by incorporating some specific mechanisms to alter the way responses are generated or by integrating additional information aside from the conversation utterances.

One of the early attempts to modify the response to follow some specific "styles" using a specific mechanism is proposed by Miyazaki et al. (2015). Using specific sentence-end expression on a Japanese corpus to determine the personal attributes, such as gender and age, their proposed model converted the conversation sentences to match the intended characteristic style. Another early attempt to convert dialogue sentences to express more *individuality* was proposed by Mizukami et al. (2015). Based on the statistical translation model, they combined paraphrasing and statistical language model techniques to transform the sentences to include the characteristic of the target. While Miyazaki et al. (2015) defined the specific attributes, a study by Zhou et al. (2017b) assumed that there exists a latent feature that acts as a responding mechanism that could be learned by the neural model. Using this feature, conditioned on the output of the input encoder, the model can generate diverse responses given the same utterance conditioned on different mechanisms. We argue that this work can be regarded as a foundation that we can associate some set of words to some specific features. Another example of this approach is the work by Zhang et al. (2020a), which attempts to diversify responses while maintaining the consistency of the persona of the users/speakers. Instead of using an additional specific feature to identify

the speakers, they extracted the persona or conversation topic from previous utterances/conversation history.

We can classify the strategy for integrating additional information into three categories: (1) adding abstract information, e.g., user-identification features, (2) incorporating character classifications, and (3) adding readable text information. In the first category, Li et al. (2016b) proposed a *persona-based* encoder-decoder model, which utilized a feature called *speaker embeddings* that is based on the user's identity. The *speaker embeddings* were learned together in the decoding phase with the intention of associating each speaker with some specific words. Correspondingly, some "nearby" speakers might have similar words. Similar to Li et al. (2016b), Wu et al. (2020) takes similar input information, while using a different variational generation approach to generate diverse responses. They utilized two regularization terms to guide the model to pay more attention on user's hidden information and user's language preference. Another variational generation approach was also proposed by Bak and Oh (2019). Using identity features from both speaker and respondent, they adopted variational hierarchical RNN architecture to personalized the responses. A study from Sato et al. (2017) also added some specific discrete variables to represent conversational situations such as involved users or conversation's timing. They treat these variables as additional input to the model.

In the second category, Li et al. (2020) introduced an approach to construct human-level attributes from movie character tropes and use them in a response selection task. They learned the language styles of the movie characters associated with a number of traits, and then retrieved the suitable response associated with the same traits with the target character. A similar attempt to infer persona information in a form of movie character tropes has been also previously proposed by Chu et al. (2018). Using attentive memory network, they classify the character trope given the dialogue snippets.

In the third category, Wang et al. (2017) proposed a way to steer the output style using some small sets of sentences called *scenting datasets*, where each set is based on one figure or persona. These sets consist of spoken-like sentences, e.g., speeches, movie subtitles, song lyrics, etc. While there is evidence that this method could produce a different style for each *scenting*

dataset, since a model is trained to focus only on a style, it is not capable to produce multiple styles at once. A similar approach was done by Zhang et al. (2018) and Mazaré et al. (2018). Using a set of introductory sentences as the user's profile/persona information, they attempted to let the model learn and "memorize" the way a user speaks.

While Wang et al. (2017) and Zhang et al. (2018) are utilizing parallel conversation-styled texts to guide the model, Gao et al. (2019) proposed a structured shared latent space, that enables the model to bridge between a conversation modelling and non-parallel style transfer using non-conversation texts. The work of Zheng et al. (2020) is also an attempt to capture the styles of unpaired non-conversational texts. Using Transformers-based encoder-decoder model (Vaswani et al., 2017), initialized with pre-trained GPT weights (Radford et al., 2019), they introduced an inverse dialogue module that aims to produce a pseudo input utterance from a once-stylized response. This pseudo input utterance then would be fed to the dialogue module to help maintain the conversation context while having the response stylized.

With the increasing popularity of Transformer-based models, Zhang et al. (2020b) extended the text generation model GPT-2 (Radford et al., 2019) to address the challenge of neural response generation. Inheriting the architecture of GPT-2, DialoGPT models a multi-turn dialogue session as a long text and frames the generation task as language modeling. Although not using user-specific information or mechanisms to capture user-specific response style, since this model is pretrained with a huge number of data and contains multi-layer self-attentive mechanisms, it is claimed to be able to capture the conversation style.

Our goal is similar to Mizukami et al. (2015) that we aim to express more individuality rather than pre-defined personality characteristics. In terms of input information, our work is similar to Li et al. (2016b) in using user-specific features to compute embedddings to help our model associate words to a specific user. However, we also incorporate a parallel readable text as style information for the corresponding user, which is similar to Zeng et al. (2019) that used user's description and numeric feature as additional information. Using this approach, our model can deal with multiple user-

Dataset	Setting	#Images	#Dialogues
VisDial (Das et al., 2017)	Series of QAs for content of image	120K	120K
GuessWhat?! (de Vries et al., 2017)	Series of QAs for object in image	66K	150K
MMD (Saha et al., 2018)	Dialogue between shopper and agent	4M	150K
IGC (Mostafazadeh et al., 2017)	Natural conversation about image	4K	4K
Image-Chat (Shuster et al., 2020)	Natural conversation about image	202K	202K
PhotoBook (Haber et al., 2019)	Conversation in image identification game	-	2.5K
VFD (Kamezawa et al., 2020)	Natural conversation from first-person view	35K	308K
VCSD (ours)	Natural conversation of people in image	8K	22K

Table 1: Existing dialogue datasets associated with visual information.

styles hence producing diverse responses, even for unseen users in the training dataset.

2.2 Visual and Language Conversation Tasks

2.2.1 Image-based Conversation Datasets

There are several situations where conversations are coupled with visual information, and many tasks and datasets have been proposed to better deal with these situations by using intelligent systems or robots. In these situations, the relationship between a speaker and an image can be classified into three cases: the speaker cannot see the image, the speaker can see the image, and the speaker is inside the image.

Visual Dialog (Das et al., 2017) and GuessWhat?! (de Vries et al., 2017) can be classified into a task setting where a person who cannot see an image repeatedly asks questions about the image to the person who can see it. These tasks can be seen as visual question-answering Antol et al. (2015) with multiple turns.

In the IGC (Mostafazadeh et al., 2017) and Image-Chat (Shuster et al., 2020) datasets, both speakers can see an image and talk about their impressions of the image. Huber et al. (2018) collected 1 million conversations associated with images from Twitter. They analyzed the influence of scene sentiment and facial expression features on conversation response generation tasks. The MMD dataset (Saha et al., 2018) simulates a process in which a shopper tries to find a product to buy with help from an agent. In the dialogue, the shopper and the agent can see multiple product images from which to select. PhotoBook (Haber et al., 2019) is a dataset for studying

how people achieve common ground in a dialogue using a multi-round online image-identification game online.

VFD (Kamezawa et al., 2020) assumes a "first-person view" setting where an agent (a person or a robot) sees a visual scene provided as an image and talks to a person within the image.

Our dataset differs from the previous studies above in that it considers a situation where two people having a conversation exist in the images. This setting is useful for developing conversation agents that can understand a conversation in a visual context from a third-person view. The characteristic properties of the datasets described in this section are summarized in Table 1.

2.2.2 Vision and Language Models

With the advancement of deep learning, the performance of image recognition and language processing has been greatly improved, and methods for integrating images and texts have been actively researched. Since the success of BERT (Devlin et al., 2019), pre-trained vision and language models have been quickly becoming popular due to their improved performance on a wide variety of vision and language tasks. Pre-trained vision and language models can be classified into two architecture types. In single-stream architecture models such as UNITER (Chen et al., 2020), a single transformer module takes both visual and textual inputs. In two-stream architecture models such as LXMERT (Tan and Bansal, 2019) and ViLBERT (Lu et al., 2019), one transformer module takes textual inputs, and another module takes visual inputs.

While many vision and language studies on dialogue tasks utilize still images and texts as the inputs, Hori et al. (2019) introduced a dialogue task which uses audio and visual features to emulate a scene as the conversation topic. Following their work, Alamri et al. (2019) introduced an audio-visual scene-aware dialogue (AVSD) dataset for the scene-aware dialogue task. In this scene-aware dialogue task, a conversation agent is tasked to answer questions about the scene in a short video, which means the video is treated as the conversation topic and the speakers are not involved in the scene. In contrast

with the scene-aware dialogue task's setting, in our setting the conversation is situated between two speakers within the image. Therefore, instead of using the visual scene only as a the conversation topic from outsider's point of view, our setting treats the conversation as being a part of the visual scene. Thus, unlike scene-aware dialogue task that uses any type of scenes (in video format), our setting required us to exclusively use *conversation scene* images. We argue that our proposed setting is more challenging and can be useful to study the ability of the model to understand the conversation scenes better.

3 User-specific Response Generation

3.1 Task Definition

In this study, we based our dialogue module on encoder-decoder model using end-to-end sequence-to-sequence approach. Given the input sequence of words $X = (x_1, x_2, ..., x_{n_X})$, the model aims to produce output sequence $Y = (y_1, y_2, ..., y_{n_Y})$ as a generated response. Since we aim to improve the output by using user-specific information by producing user-level stylized response, let $I = (i_1, i_2, ..., i_{n_I})$ denotes the set of users in dataset. Please note that, we defined "stylized response" as a response that contained frequently used words or characters from the corresponding user. For each i, we extract the user-identity feature and other user-specific features as the means for the model to learn the user-level response "style", i.e., user's frequently used words or characters.

3.2 Datasets

Since our target is to incorporate and emphasize the user-level speaking styles of actual human, i.e., the respondents, we need user identifications to be attributed to the conversation sentences. Therefore, we constructed our datasets from tweets, collected using Twitter API. We constructed two types of datasets: *conversation dataset* and *user-info dataset*.

3.2.1 Conversation Dataset

As the name implied, this dataset is designated to train the model to generate a conversation response given an input utterance. We crawled the conversations from Twitter, and retrieved only the tweets that satisfy following conditions: (1) the tweets have to be reply tweets, i.e., responses to other tweets, (2) the corresponding users have had engaged in the conversations with a minimum of three turns. We paired each reply with the tweet that it is a response of, as *response* and *input utterance*, respectively. Then, we used the respondents' usernames as the user identification attribute, hence *user embeddings*. Note that, in order to associate the user embeddings with

their respective sentences, we only constructed user embeddings from this conversation dataset.

To improve the data quality, we pre-processed the retrieved tweets to remove some noises, such as tweets with non-ASCII characters, duplication, and non-English tweets. To simplify the conversations, we also removed URLs, hashtags, and mentions from the tweets. Then, we filtered to only include sentence with maximum 15 words. The final conversation dataset consists of 233,293 pairs of input utterances and responses associated with 207 human users. We limited the vocabulary tokens to 35,000 words with average lengths of utterances and responses are 8.35 and 8.3 tokens respectively.

3.2.2 User-info Dataset

We constructed this user-info dataset from individual tweets, i.e., non-reply tweets, from the same users in the *conversation dataset*. There are two designated purposes for this dataset: (1) to be a complementary style information for available users in the *conversation dataset*, and (2) as a mean to handle unseen users, i.e., non-existing users in the training phase. The usage of this dataset will be explained more later in Section 3.3.3.

As the numbers of available single tweets from each user are greatly varied, 135 tweets in average with 395 tweets as the highest, we intend to normalize and extract the important words only. To extract the selected words, we compared several scoring methods: word frequency (selecting the most frequent words), TF-IDF Sparck Jones (1988), and pointwise mutual information (PMI) Church and Hanks (1990). We treated all the words in the retrieved tweets as the input of every method, then extracted top N words for each method. Hence, we have three sets of N words initially. We then deployed these sets in a separated preliminary experiments and evaluated the results qualitatively. We asked two fluent English speakers to compare the quality of generated responses using the three sets of words. We provided three sets of samples of the generated responses and asked them to evaluate the *fluency* and *relevance* to the input message. Based on their evaluation, we decided to choose TF-IDF as the scoring method to extract the important

words for our *user-info dataset*. We limited the maximum number of each user's words to 50 and regard this set of words as user's style characteristic. From all users in the *conversation dataset*, about 15% of users don't have this set of words.

While we collected the sentences from the users in *conversation dataset*, we can also use this dataset independent of those corresponding users in a "mix-and-match" setting. For example, we can use *user-info dataset* from *user A* and use it to enhance the style of *user B*. More details will be explained in Section 3.3.3.

3.3 Method

As described in Section 2.1, there are various approaches to integrate "styles" into the generated or selected responses. To accommodate our intention to use user-embeddings (user identification features) and user-info embeddings (user's frequent choice of words) mentioned in Section 3.2, we need our model to take these features into account in generating stylized-response. Considering the effectiveness of RNN and sequence-to-sequence approach in conversation tasks (Li et al., 2016b; Sato et al., 2017; Bak and Oh, 2019), we decided to follow Li et al. (2016b) and Zeng et al. (2019) to use RNN for our architecture and to use attention mechanism as our strategy to guide the response generation.

3.3.1 Encoder-Decoder Model

For our encoder-decoder model we adopt the Long Short-term Memory (Hochreiter and Schmidhuber, 1997) for the encoder to compute the representation of the input sequence. Given the input sequence $X = (x_1, x_2, ..., x_{n_X})$, the model aims to produce output sequence $Y = (y_1, y_2, ..., y_{n_Y})$ as a generated response. Then, we keep the encoder hidden state $\bar{h}_s$ to feed into the decoder unit. We use attention-mechanism (Luong et al., 2015) in the decoding phase to pay attention to the input utterance and also user-specific style information. In general, the decoding process can be interpreted through the

following equations:

$$
\begin{aligned}
p(y_t|y_{<t}, x) &= \text{softmax}(W_s\tilde{h}_t), && (1)\\
\tilde{h}_t &= \tanh(W_c[c_t; h_t]), && (2)\\
h_t &= LSTM(y_{t-1}, h_{t-1}), && (3)\\
c_t &= \sum_{s=1}^{S} a_t(s)\bar{h}_s, && (4)\\
a_t(s) &= \text{softmax}(h_t^\top W_a \bar{h}_s), && (5)\\
&= \frac{\exp(h_t^\top W_a \bar{h}_s)}{\sum_{s'} \exp(h_t^\top W_a \bar{h}_{s'})}
\end{aligned}
$$

With y_t denotes the response token at the time step t, the objective is to learn its probability distribution given the preceding response tokens $y_{<t}$ and previous utterance x. The model measures the alignment scores $a_t(s)$ between the source hidden states $\bar{h}_s$ and the current encoder output h_t. Then, we calculate the current context c_t by weight averaging the alignment scores $a_t(s)$ and the source hidden states $\bar{h}_s$. Please note that we also calculate the alignment scores for the user-specific style information that will be explained in the next section. Then, we concatenate all contexts and apply $tanh$ function to obtain $\tilde{h}_t$ to feed into $softmax$ function to produce the response token distribution; W_a, W_c, and W_a denote decoder learned parameters.

Following Luong et al.2015, we also apply input-feeding approach as an attempt to make the model capture the previous alignment. Input-feeding is done by concatenating $\tilde{h}_t$ to the decoder input a the next time step. For both the encoder and the decoder, we employ two-layer LSTM architectures.

3.3.2 User-specific Information

As our intention is to capture the characteristics of the users, i.e., the persons who respond to the given utterance, we argue that using user related information would be useful to achieve that objective. Different with Li et al. (2020) which classified the movie characters into a number of associated traits, we wanted the model to learn the style on the user-level. Following the approach from persona-based model by Li et al. (2016b), we aim to

pair conversation utterances with corresponding users-identity features and train them together. However, as this approach can only accommodate users present in the training data, we intend to overcome this limitation by adding a small dataset for each user to serve as another characteristic feature.

We defined two types of user-specific information: *user embeddings* and *user-info embeddings*. *User embedings* are derived from usernames in the training data, while *user-info embeddings* are derived from separate collections of the words that frequently used by the corresponding users in the conversation dataset.

Following the setup described in Section 3.3.1, let I_{train} denote the set of users (respondents) in the training data, K_{word} the dimension of word embeddings, and K_{user} the dimension of user embeddings. We convert each input sequence X to embeddings with size K_{word}. Then, we define a user identity embedding u_i with size K_{user} for each user $i \in I_{train}$. The user embedding u_i is shared to all conversations involving user i.

The second type of user-specific information involves a collection of users' selected words. In order to capture the characteristic, especially the speaking style, of each user, we argue that we need to define a feature or a set of information that can capture it. As the "speaking style" is an abstract concept, we define our user-specific style as a set of selected words that can represent each user's behaviour in common/daily conversation.

Let I denotes the set of users. Note that I_{train} is a subset of I. For each user in I, several sentences can be collected. From this collection of sentences, we then extract N words to represent the characteristics of the user. As described in Section 3.2.2, we extracted the selected words using TF-IDF method. Each of these words was then converted to an embedding of dimension K_{word}, hence *user-info embeddings*. Please note that as we extracted the selected words for each user in I, we have the corresponding style information not only for the user in I_{train} but also for all users in I if available. Therefore, our *user-info embeddings* also can be used for users that not found in the training data.

3.3.3 Attentional Conversation Model

Our attentional LSTM model takes three features as input: *input word embeddings*, *user embeddings*, *user-info embeddings*; The *input word embeddings* are used in encoding phase, while *user embeddings* and *user-info embeddings* are used in decoding phase. Since our model also incorporates the input-feeding approach, the input for the decoding phase is the concatenation of the output of the previous decoding time step y_{t-1}, user embedding u_i, and input-feeding $\tilde{h}_{t-1}$. The user-info embeddings will be used later as the input for additional attention mechanism. Hence, the decoding process can be described as follows:

$$h_t = LSTM([y_{t-1}; u_i; \tilde{h}_{t-1}], h_{t-1}). \tag{6}$$

The user-info embeddings are constructed from the collection of top N ranked words uttered by intended users, where the users are not necessarily present in the training data. Using the same embeddings as input word embeddings, we compose $P_i = \{p_{i_1}, ..., p_{i_N}\}$ $(\forall k,\ p_{i_k} \in \mathbb{R}^{K_{word}})$ as user-info embeddings for user i.

The model is trained to attend not only to the input source, i.e., the hidden states of the encoder, but also to the user-info embeddings. Therefore, since this model uses two contexts, we need to adjust Equation 3 to

$$\tilde{h}_t = \tanh(W_c[c_t^{(X)}; c_t^{(P)}; h_t]), \tag{7}$$

where we define $c_t^{(X)}$ as the context for input source and $c_t^{(P)}$ as the context for user-info embeddings. This proposed model is illustrated in Figure 2. We can see in the Figure that the user-info embeddings that co-attended are fed directly to the attention module. By doing this, we can inject any user's *user-info embeddings* as the pair of the *user embeddings* in "plug-and-play" mechanism, even for the non-existing user in the training data.

In terms of using co-attention beside the main attention on the input source, our model is similar with Zeng et al. (2019) which applied co-attention to the user's self-description sentence. However our model is different on what to attend to and how we treat the user-level features. We argue that adding

attention to a user-related word collection would be more effective than to a self-description sentence which sometimes has no correlation to the identity of the user. In terms of user-identifying feature, we use *user embeddings* as the feature and concatenate these embeddings to the decoder inputs in every time step. In comparison, they used embeddings from user's numeric feature, such as age and gender, and applied gated memory module on the embeddings before combine them together with other decoder inputs. We postulate that including the *user embeddings* in each time step of decoding phase would let the model learn easier of the user-level styling.

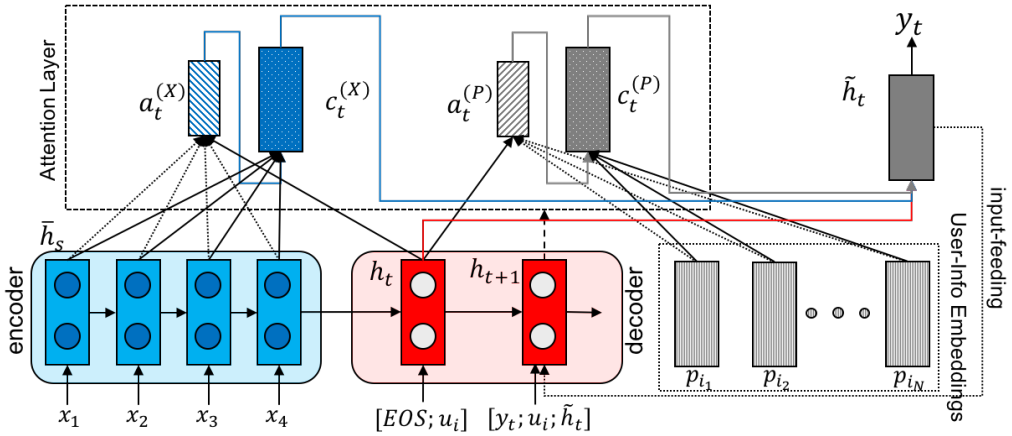


Figure 2: Overview of our neural conversation model with attention to user-specific information. We use two-layer LSTM for both encoder and decoder. The attention layer attends to source hidden states $\bar{h}_{s'}$ and user-info embeddings P_i for user i. User embeddings u_i are concatenated to decoder input at every step.

3.4 Experiment

3.4.1 Data

We used both *conversation dataset* and *user-info dataset* as described in Section 3.2. For validation, we used 1000 pairs of utterances and responses of 6 users from *conversation dataset* alongside the *user-info dataset* with the average of 47 tokens for each user. For test phase, we extracted different set of tweets sentences from 12 random users in *conversation dataset* resulting

in 100 pairs of utterances and responses. To test the learned model against unseen users, we simply replaced the users to the new users, not existing in the training data, which then will be concerted to unknown (UNK) token.

3.4.2 Implementation

We implemented our model using PyTorch[2] framework. Both our encoder and decoder employed two-layer LSTM. Some hyper-parameter details are as follows:

- Each LSTM layer contains 300 hidden units.
- Embedding size is set to 300.
- Network parameters are initialized with uniform distribution $[-0.05, 0.05]$.
- Training batch size is set to 128.
- Learning rate for the encoder is set to 0.0001, multiplied by 2.5 for the decoder.
- Dropout rate is set to 0.1.
- Vocabulary size is 35,000.

We trained the model by using the Adam optimizer(Kingma and Ba, 2014a) with different learning rates between the encoder and decoder. We conducted several procedures to determine the training stop condition. We observed the decrease in loss $H_{y'}(y) := -\sum_i y'_i \log(y_i)$. We set the epoch to 50, then we set the early stop condition to when the loss difference is less than 7% for a long period. As we were cautious to rely on numbers only, we also asked two fluent English speakers to evaluate the trained model quality by evaluating a small set of generated responses. Finally, we stopped training at the 47th epoch. We also limited the maximum length of an utterance to 15 words per sentence. The training was run on a single Titan X GPU for about three days.

[2]https://pytorch.org/

The input utterance and user-info embeddings were initialized with GloVe embeddings (Pennington et al., 2014a). We replaced the words not in the vocabulary with UNK tokens. The same treatment was applied to unseen users in user embeddings. We set the UNK token embedding to a vector of all zeroes at the initial stage. To select the token prediction, we applied the greedy approach.

3.4.3 Baseline and Comparison Models

We adopted the persona-based model of Li et al. (2016b) to serve as the benchmark for our model. Their model utilizes user-identification attribute, named *speaker embeddings*, in the decoding phase to let the model associate words that often appear together with some specific users. Therefore, user-specific associated words can be regarded as user's style.

In terms of using user embeddings in the decoding phase, our model are similar to their model. However, as mentioned in Section 3.2, user embeddings cannot cover unseen users. Our model overcomes that issue by using user-info embeddings. The decoder input of both ours and their model can be represented by Equation 6. Since the baseline model does not have user-info embeddings, our model's attentional hidden $\tilde{h}_t$ is different from theirs. The attentional hidden of the baseline model would be the same as 3, while our model's is represented by Equation 7.

We prepare a variant of our proposed model, using unseen (UNK) users for user embeddings. The rationale for this setting is to investigate whether our model could generate better responses against our baseline's handicap. We only used this variant in the evaluation phase. We also compare our model with the standard sequence-to-sequence (seq2seq) model, without user embeddings and user-info embeddings. This baseline can also work as an ablation study to verify whether additional user-specific features can actually improve the generation in general, particularly in real-world type conversation. The summary of the baseline comparison models is shown in Table 2.

As the transformer-based (Vaswani et al., 2017) models are gaining popularity, we argue that it would be useful to compare our model's outputs to

Models	User-specific Features	
	user embeddings	user-info embeddings
User + Info (Main)	✓	✓
UNK + Info	UNK token	✓
UserOnly (Baseline)	✓	-
Seq2Seq	-	-

Table 2: Comparison models on the effectiveness of the proposed user-specific information features.

the transformer-based model's. For this evaluation, we adopted DialoGPT (Zhang et al., 2020b) as our comparison model. DialoGPT is a large neural response generation model that extends GPT-2 (Radford et al., 2019) on the conversation data from Reddit. We fine-tuned DialoGPT on our conversation dataset and generated responses using the same set of conversation pairs used in the previous stages. Please note that the DialoGPT model does not utilize user-specific information features.

3.4.4 Evaluation Setup

Compared with other generation tasks that have finite groundtruth, e.g., machine translation, a conversation task can have multiple correct outputs given the same input. Therefore, an evaluation metric that can measure an output prediction against multiple references is preferable, for example, BLEU(Papineni et al., 2002). Originally designed to evaluate the machine translation task, BLEU measures the overlapping words or n-grams between the output and references.

Despite the common usage of BLEU in evaluating conversation tasks (Sordoni et al., 2015a; Li et al., 2016b; Zheng et al., 2020), according to Liu et al. (2016), BLEU score is lowly correlated with human judgments of dialogue systems. To overcome this issue, several studies have proposed new metrics to evaluate open-domain dialogue systems(Lowe et al., 2017; Ghazarian et al., 2019; Yuma et al., 2020). However, these new metrics require learning models to measure the scores. Therefore, we decided to still use BLEU and perplexity as the means to compare the models performance. We also included ***Distinct*** (*dist-1* and *dist-2*) scores to represent the diversity

on the generated responses, calculated by averaging the amount of n-gram by the total generated tokens (Li et al., 2016b).

To complement the automatic evaluation metrics, we also provide human evaluation on our models' outputs. We hired human evaluators from Amazon Mechanical Turk service to evaluate the quality of our generated responses. We asked the evaluators to evaluate the responses using the following criteria:

- ***fluency*** or ***naturalness***: Whether the response could be produced by a (English speaking) human.
- ***relevance*** or ***adequacy***: Whether the response could be accepted as a suitable answer or contains useful information regarding the input utterance.
- ***style***: Whether the response could be produced by the same person if some profile information was provided.

Aside from measuring the ***style***, we also wanted to ascertain that the model is able to produce fluent sentences while maintaining relevance to the conversation context. Hence, the ***fluency*** and ***relevance*** criteria.

Since evaluating ***style*** is supposed to be more difficult compared to other criteria, we conducted the evaluation in two stages: 1) assessing ***fluency*** and ***relevance*** at one time, 2) evaluating ***style*** similarity. For these stages, we prepared a separate set to produce the model-generated responses. We randomly picked 12 users from the conversation dataset and retrieved tweets that they replied to as input utterances. As we retrieved 5 to 10 tweets from each user, 100 tweets were collected. Each pair of an input utterance and its response was then evaluated by 10 judges.

In the first stage, we provided the human evaluators with pairs of input utterances and the respective response from each baseline model. Since we have four baseline models, each input utterance is paired with four response alternatives. We used a three-point Likert scale labeled {bad, enough, good}, which later converted to {-1, 0, +1} respectively, and asked the evaluators to score every response alternative in terms of ***fluency*** and ***relevance***.

For the second stage, since we wanted to investigate the effect of user-specific information, we only evaluate models that utilize user-specific fea-

ture(s), i.e., we omitted vanilla Seq2Seq in this stage. We provided the evaluators with Twitter user bio, i.e., a user's short biography or profile information that commonly contains keywords, and some sample tweets from the respective users. We asked the evaluators to compare the style of generated responses to the sample tweets of each respective user. We also required the evaluators to take into account the user's bio when judging the responses. An example of the provided user's information is shown in Figure 3. As we need to see which model can generate responses with more resemblance to the user-specific style, the evaluators were required to rank the response alternatives from each model. We used 5-score ranks with 1 as the most resembling style and 5 as the least resembling one; tie scores were permitted.

Figure 3: Example of a user's Twitter bio and sample tweets used in style evaluation. We censored any mentions of other accounts.

To evaluate and compare the generated responses from both our models and the fine-tuned DialoGPT, we recruited three volunteers, who evaluated

Models	Fluency (%)			Relevance (%)			BLEU-1	Perplexity	Distinct	
	bad	enough	good	bad	enough	good			dist-1	dist-2
UserOnly	19.5	27.3	53.2	51.8	25.2	23.0	0.0705	294.14	0.047	0.0941
Seq2Seq	8.2	25.8	66.0	40.1	29.4	30.5	0.0625	**280.8**	0.0735	0.1456
User + Info	17.5	26.4	56.1	44.9	28.2	26.9	0.069	291.15	**0.0745**	**0.165**
UNK + Info	9.0	23.7	**67.3**	37.4	31.2	**31.4**	**0.0754**	-	0.066	0.1378

Table 3: Automatic (BLEU, perplexity, distinct) and Human evaluation results for *fluency* and *relevance*, presented as raw score percentages. UNK + Info does not have validation set hence no perplexity for this model. Our UNK + Info model with unseen users gains 26.5% more for fluency and 36.5% more for relevance compared to the baseline.

the same 100 pairs of conversation. We asked the evaluators to compare two response alternatives from our model and DialogGPT and to choose which one is preferable given the input utterance. We randomly assigned the two response alternatives as *Response 1* or *Response 2* to help the evaluators stay objective. Then, we asked the evaluators to choose from {Response 1, Response 2, Neither} options.

3.5 Results and Discussion

3.5.1 Automatic and Human Evaluation

The summary of the evaluation can be seen in Table 3. In automatic evaluation, we find that our model User + Info which being fed with seen users and user-info embeddings can achieve higher diversity on the generated words both shown by dist-1 and dist 2 scores. It is also interesting that our model which being fed with unseen users in the testing phase, UNK + Info, achieved highest BLEU score compared to other settings. To help understand the models' performance, we argue that qualitative evaluation is required.

For human evaluation, we first evaluated all models in Table 2 in terms of *fluency* and *relevance*. For both criteria, the evaluators needed to assign a score from three choices: bad, enough, and good. To compare the models, first we counted the number of each score label every model received, we call this scores as *raw scores*. According to these scores, for both criteria, UNK + Info (our model with unseen users) received the highest *good* score,

followed by the Seq2Seq model. The UNK + Info gains 26.5% more fluency point compared to the baseline. To calculate this gain, we simply compared the percentage obtained by UNK + Info (67.3%) against UserOnly (53.2%).

To determine the overall quality of the response generation, in terms of *fluency* and *relevance*, we also need to include the *enough* scores from both criteria. We combined *enough* and *good* scores as *acceptable*. According to the *acceptable* score, as shown in Table 4a, Seq2Seq model achieved the highest score in terms of *fluency*, gaining 0.87% more point compared to our UNK + Info model. As the gain is arguably small, we consider our model is good enough in terms of *fluency*.

In terms of *relevance*, we can see from Table 3 that all models received higher *bad* score compared to their *good* score. We argue that this occurrence implies that *relevance* is harder to achieve than *fluency*. Yet, our UNK + Info achieved the highest *acceptable* score in this criteria, gaining 36.5% more compared to the baseline.

We also calculated the average scores by converting the {bad, enough, good} scores to {-1, 0, 1} respectively for each model; the results are shown in Table 4b. According to these average scores, our UNK + Info tops in both *fluency* and *relevance* criteria. Using one-way ANOVA as the significance test, we confirmed that our main model (User + Info) is significantly better than the baseline (UserOnly) in the relevance criteria.

In the second stage, we measured the style similarity between the generated responses and the corresponding user's sample tweets. As in the first stage, each of the 100 input-response pairs was evaluated by 10 human evaluators, resulting in 1,000 samples, from which we removed some results that did not show consistency, e.g., results with identical responses with different rank scores. As described in Section 3.4.4, since our intent was to investigate the influence of user-specific information, we evaluated only three models in this stage. The results can be seen in Table 5a.

As we asked the evaluators to assign a rank score in scale of 1 to 5 (smaller number is better), according to the Table 5a, the average scores relatively positioned in the middle rank, i.e., around rank three. It suggests that, in general, all models only generate "good enough" responses in terms of style. Nevertheless, both our models achieved higher rank compared to baseline,

Models	**Fluency** (%) acceptable	**Relevance** (%) acceptable
UserOnly	80.5	48.2
seq2seq	**91.8**	59.9
User + Info	82.5	55.1
UNK + Info	91.0	**62.6**

(a) Acceptable or "Good enough" results with *good* and *enough* scores combined. seq2seq tops *fluency*, but our model with unseen users gets the highest *relevance* score.

Models	**Fluency**	**Relevance**
UserOnly	0.337 ± 0.06	-0.28 ± 0.06
seq2seq	0.578 ± 0.05	-0.09 ± 0.06
User + Info	0.386 ± 0.06	-0.18 ± 0.06
UNK + Info	**0.583 ± 0.05**	**-0.06 ± 0.06**

(b) Average scores for *fluency* and *relevance* criteria. For ***relevance***, our model achieved significantly better scores than the baseline (one-way ANOVA, $p < 0.05$).

Table 4: Results on *fluency* and *relevance* criteria of four alternative models. From table on the left (a), we can learn that the vanilla Seq2Seq model achieved the highest acceptable response on *fluency*. However, from table on the right (b), in average, UNK + Info achieved the highest scores on both criteria.

and we confirmed by Friedman Test that UNK + Info is significantly better than the baseline. Some examples generated by our models can be seen in Table 6.

As we noticed that our UNK + Info model performed better than other models, we further compared the responses generated by UNK + Info to the outputs of DialoGPT Zhang et al. (2020b). Following the setup described in Section 3.4.4, we recruited three volunteers who evaluated 100 pairs of conversation. The volunteers are asked to choose the preferred response from two alternatives generated by our model and DialoGPT. The results can be seen in Table 5b.

We can see that DialoGPT's outputs got selected more times at 40.67% than our UNK + Info model at 35.67%. Considering that DialoGPT's capacity and model size is bigger than our model, we argue that our achievement is decent. Additionally, we measured the inter-annotator agreement among the evaluators by calculating the Fleiss' Kappa Fleiss et al. (1971) of the results, resulting in 0.2068 for the Kappa score. This low score indicates a slight agreement among the evaluators Landis and Koch (1977). We argue that the slight agreement could indicate the similar quality between our model and DialoGPT, which makes the evaluation relatively difficult.

Models	Style Rank
UserOnly	3.37 ± 0.09
User + Info	3.29 ± 0.09
UNK + Info	**3.16 ± 0.09**

(a) Results of style evaluation. Smaller values are better. Our variant model was significantly better than the baseline (Friedman Test, $p < 0.05$).

Models	Preferred
UNK + Info	35.67%
DialoGPT	**40.67%**
Neither	23.67%

(b) Results of the comparison between UNK + Info and DialoGPT. The Fleiss' Kappa score of 0.2068 shows a slight agreement among the evaluators.

Table 5: Table on the left (a) shows that UNK + Info achieved better style compared to the baselines with user-specific information. In the table on the right (b), we further compared the outputs from UNK + Info and DialoGPT in the preferable measure. We argue that the low agreement among the evaluators (0.2068 on Fleiss' Kappa) shows the similar quality between the two response alternatives.

3.5.2 Discussion

Our main intention is to incorporate an individual user's characteristics to machine-generated responses. We specifically attempted to incorporate more user-specific information to emphasize the response style of different users. Therefore, we conducted the experiments to incorporate additional user-specific information, and the evaluation we performed has shown that the human evaluators recognized a better change in response style.

Furthermore, one aspect that distinguishes our model from others is the application of external resources in addition of user-specific features. Generally, if a conversation model was trained to pick up some user-specific or persona-specific traits, such features should be included in the training. However, our external resource (user-info dataset) can be used even without existing corresponding user in the training. Our work also can serves as an evidence of the usability of using external resources for response generation model with simple architecture.

An interesting finding is that the UNK + Info model, which being fed with unseen users, achieved better scores than our User + Info model on the style preference by human judges. Despite achieving lower scores on the distinct evaluation metric, which is considered to represent the diversity,

User	Input	Responses	
		User + Info	**UNK + Info**
User1	*i'm going to have a cardiac episode and my wife is not here*	*i love you*	*what if i do*
	hate to burst your bubble it snowed last night & this morning	*and now you get to the game*	*ahh ! ! ! i was so excited !*
	they cut so much out of the finale !	*i was thinking it was*	*it was a good season ! !*
User2	*if lebron loses this series does he leave and where does he go ?*	*he scored the house and got his back house*	*he was playing with his rb / 3*
	he was five for his last 50 you literally had to pound the middle of the zone three times	*are you saying that were no reason ?*	*he has a good draft ?*
	hey guess what ? the bills will draft a few football players this weekend	*lol you got that man ! ! ! ! ! !*	*you beat me ?*
User3	*this is such a mood and i can't even explain why*	*did you get braces or walmart ?*	*hello ! ! !*
	you can skip iron man 3	*do i get some rest ?*	*yeah i am*
	heard he cheated on her	*don't forget*	*she was a lesbian*

Table 6: Examples of responses from different users generated by our model, using known users and their user-info respectively, and its variant model, using unseen users and the same user-infos.

this UNK + Info model also achieved the highest BLEU score compared to others. Although these scores could imply that BLEU correlate negatively with response diversity, we would need more experiments and analysis to

achieve the conclusion. However, we argue that these scores imply that style preferability is not always aligned with the response diversity.

Another point to analyze is that using both user-specific information, i.e., seen *user embeddings* and *user-info embeddings*, might overwhelm the response generation that leads the outputs to be too "stylized". Through manual observation, we assume that a model with more injected information can be too focused on the styles and lose some relevance to the input utterance. However, the baseline, with less information, still received lower scores on automatic and human evaluation. We postulate that our architecture with co-attention to external information could manage to generate better response in terms of style without loosing the relevance and other expected quality.

We also evaluated the human preference between the generated responses from UNK + Info model and DialoGPT. Considering that DialoGPT is bigger than our model, it is not surprising that DialoGPT acquired the higher score in this evaluation. However, given the low agreement between the evaluators, we can assume that even for a big transformer-based model, the open-domain response generation task with real-world conversations is a difficult task. Therefore, we argue that our model's performance is acceptable.

In conclusion, a problem still persists in styling generated responses. Regardless of the results being better than the baseline, generating fluent and relevant responses with an expected style is challenging. Nevertheless, we expect that with the popularity of the transformers-based model and other big pre-trained methodology, incorporating a specific and intended style into the generated responses would be possible in the near future.

4 Response Classification with Visual Scene Dataset

4.1 Scenario and Task Definition

4.1.1 Scenario Description

In a real-world situation, a conversation agent (robot) sometimes witnesses an interaction between humans in front of it, so the capability to understand conversation from a third-person view is necessary. This agent needs to determine which humans are having a conversation and what they are talking about by evaluating their utterances and visual clues. Figure 1 illustrates a scene captured by the agent. Given the first utterance and the visual clues, we task the agent with predicting the correct response from available candidates, i.e., we implement this task as a multi-modal conversation response-selection task.

The minimal inputs for the robot are (1) an original image (a conversation scene), (2) an utterance made by a speaker, and (3) a response candidate. As a novel feature of our dataset, we can choose to input (4) a speaker or (5) a respondent from the conversation scene. We consider specifying the speaker and the respondent to be a natural setting for robots since it should be clear who made an utterance and who should respond to it. Another interesting possible input to the robot, though not realistic, is (6) a topic object. This setting is where the robot has prior knowledge of what the conversation's content is.

4.1.2 Task Definition

In this sub-task of our research, because our aim is to investigate the effectiveness of the visual clues on conversation tasks, we decided to simplify the task to be a response binary classification task. Let $\mathbf{x}$ and $\mathbf{r}$ be a pair of input and output of a conversation classification model, respectively. Then, the model would be required to predict whether $\mathbf{r}$ is the correct response for input $\mathbf{x}$ or not. Given the multi-modal setting, the variable $\mathbf{x}$ could consist of conversation utterance text sequence and accompanied by other modality

features; visual clues in this case.

4.2 Dataset

4.2.1 Dataset Construction

We constructed our dataset by annotating the images from the Visual Genome dataset (Krishna et al., 2017) with three stages of crowdsourcing. We defined an image as a conversation scene if it satisfies three criteria: (1) having two or more persons, (2) the speakers are relatively close to each other, and (3) there is at least one nearby object that can be used as a general conversation topic. We used Visual Genome annotations such as "person," "man," "woman," and "kid" for speakers and respondents, while topic objects could be any object, e.g., building, sky, or hat. We call this set of images *potential images*. The obtained 16,333 *potential images* were then annotated by crowd-workers via Amazon Mechanical Turk. The number of workers who participated in this study was 1539. We specified English-speaking countries and native fluency as the minimum requirements. Three stages of crowdsourcing were done by different groups of workers.

In the first stage of crowdsourcing, we filtered out images unsuitable as conversation scenes. Using Visual Genome's annotations, for each image, we marked two speaker/respondent objects with bounding boxes. Please note that one image can contain several "person" objects; hence, one image can have multiple potential speaker/respondent pairs. Then, we asked the crowd-workers to determine whether the pair of speakers appear to be talking. The crowd-workers needed to choose among five answers: *yes, likely, possibly, no*, and *invalid (no visible two speakers)*. We kept only the images labeled with *yes* and *likely*, amounting to 8,356 images.

In the second stage, we asked the crowd-workers to choose between one of the two annotated speakers as the first speaker and then select one object from the image as the general topic of conversation. Afterward, the crowd-workers were required to pose as the first speaker and write an appropriate first utterance. To increase the conversation diversity, we annotated each image four times. Therefore, the same image could have different first speakers or topics. We also asked the crowd-workers to label whether a sentence

was natural or not. Later, we dismissed the data labeled as unnatural.

The last annotation stage was to collect responses appropriate to the first speaker's utterances. The crowd-workers were required to pose as the second speaker and write a relevant response to the first utterance. While we asked the annotators to take into account the topic object, the response did not need to explicitly contain the topic word. The response could refer to the topic object implicitly. For example, if the topic object was "a building", it was possible to respond with "do you want to go there?". Annotators could also include a comment if the topic object was unclear, e.g., blurry, or the image quality was low. We removed the data that they regarded as unclear. By the end of this stage, we collected 22,331 pairs of *image-first utterance-response* data.

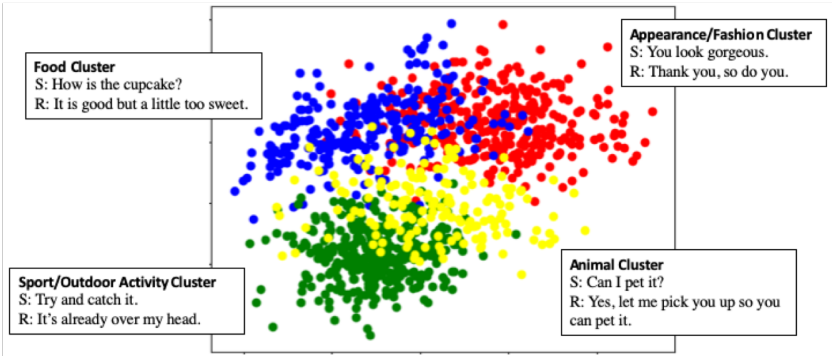


Figure 4: Distribution of conversation topics in VCSD. Each conversation is represented by GloVe vector and colored according to its cluster, i.e., Food, Fashion, Sport, and Animal.

4.2.2 Dataset Analysis

To help to investigate whether visual scenes are necessary to improve conversation tasks, we performed qualitative analysis on the utterance and response pairs. We sampled 100 pairs from our dataset and evaluated them, especially the utterance, with two questions: (1) "Is the image necessary to understand the context?", and (2) "Do we need context understanding to respond to the utterance?" In this analysis, we hid the corresponding images from the assessors. Our finding shows that 75% of a conversation needed to be accompanied

by visual information to be correctly understood. However, only 58% of the conversation utterances required context understanding to be responded to. For example, with the "Let's pick up the cones" utterance, we need to see the visual clue to understand the context, but this utterance can be responded to with a simple "Yes" or "Sure, no problem." This phenomenon indicates that the ability to respond does not necessarily corresponded to context understanding.

We also performed cluster analysis on the conversation text to understand the distribution of conversation topics. Using GloVe embedding (Pennington et al., 2014b), we converted words to 300-dimensional vectors and then averaged the word embeddings to obtain a vector representation of every pair in conversations. Then, we used Agglomerative Clustering (Rokach and Maimon, 2005; Pedregosa et al., 2011) to obtain 60 clusters of conversations. The clusters showed that our dataset has a diverse set of conversation topics. We selected four clusters, reduced the vector dimensions using principal component analysis (PCA), and visualized them in Figure 4. From these four clusters, i.e., *food*, *fashion*, *sport*, and *animal*, we can observe that a conversation from one cluster is sometimes similar to one in another cluster. For instance, conversations on the topic of *animal* might be related to those on the topic of *sport*, and vice versa. We also found that the conversation topics in our dataset cover topics indicated not only by "nouns" but also "verbs," e.g., *travel* or *transportation*, that can be indicated by the words "go," "come," or "drive." We can argue that this shows that the focus of natural human conversations is not necessarily about objects but often about actions regarding the objects.

To assure the quality of the collected data, we recruited four people from our research group to perform a quality assessment for 400 randomly sampled pairs; each person assessed 100 samples. We defined four assessment criteria: (1) the relevance of the utterance-response pair, (2) the reasonability of the topic object, (3) the reasonability of the speaker-respondent pair, and (4) the naturalness of the conversation given the visual scene.

The assessors were required to assign an assessment score on a 4-point scale, with 1 as the lowest score (clearly wrong), 2 (inappropriate), 3 (questionable), and 4 as the highest score (good). From the results shown in

Assessment Score	clearly wrong	inappr-opriate	questio-nable	good
Relevance	1%	1.25%	10.25%	87.5%
Object	2.25%	8%	26.25%	63.5%
Speakers	3.25%	4.75%	16.25%	75.75%
Naturalness	2.25%	7.25%	14.5%	76%

Table 7: Quality assurance for image-conversation in terms of utterance-response *relevance*, *object* and *speakers* appropriateness, and *naturalness*.

Table 7, we concluded that the amount of "noisy" data (about 1% to 3%) was acceptable.

4.3 Experiment

4.3.1 Data

We split our 22,331 pairs of data into training, validation, and test sets at a ratio of 70:15:15; 15,631 for the training set and 3,349 each for the validation and test sets. We used these data as positive instances. For negative instances, for each first utterance, we selected a random response from a set of four responses given the same image. However, as described in Section 4.2.1, we removed pairs of data that were considered unclear or completely wrong. This removal caused some images to have less than four response alternatives. Nevertheless, we made sure the negative samples are selected from the conversation using the same original images. In the rare cases where there is no alternative conversation with the same image, we selected other conversations with different images as the negative pairs. Some examples can be seen in Figure 6. We set the ratio of positive and negative instances to 1:1, resulting in over 30K instances for the training set and over 6K each for the validation and test sets.

Following the scenario described in Section 4.1.1, we needed a method to utilize the parts of the images that were focused on, i.e., the speakers and possibly the topic objects also. One approach is to use annotations for image object regions (bounding box positions), as used in Faster-RCNN (Ren et al., 2015) for object-detection tasks. Another approach is to use

cropped image parts together with the original images, as used in (Minh et al., 2018). By doing this, the model can learn both representations for an entire image and the parts focused on. Since our proposed dataset actually uses Visual Genome's annotations, we can use either approach to include partial information from images. However, using cropped images can increase the capacity of visual information and might be useful for model performance.

4.3.2 Transformer-based Baselines and Error Analysis

To confirm our hypothesis that multiple modalities, more specifically, using conversational scene images, can improve the conversation task, we first experimented with a single-modal (text only) model. We used BERT (Devlin et al., 2019) as the baseline for this task. We fine-tuned the BERT base model with first utterance - response pairs from our dataset with labels of 1 and 0 for positive and negative samples, respectively, i.e., as binary classification. While we expected BERT to achieve a moderate accuracy score on this task, according to our analysis in Section 4.2.2, wherein about 50% of cases would require context understanding, BERT exceeded our expectation by achieving an accuracy score of above 85%.

Then, we proceeded to experiment with LXMERT (Tan and Bansal, 2019) as a BERT counterpart for a cross-modal model. For this experiment, we trained LXMERT from scratch and did not use the pre-trained models. Since LXMERT accepts one image for each corresponding text, we used the original images. We trained the model using the same binary-classification task; the model learned to predict whether a pair of utterance-response and a visual scene belong to the same conversation. Since LXMERT expects the input images to be annotated by bounding boxes, we decided to use two variants of annotations: (1) speaker, respondent, and object bounding boxes, and (2) all Visual Genome's densed-annotation bounding boxes. While both settings produced considerably high scores, respectively 79.4% and 80.65%, they were still lower than BERT's achievement. We postulate that BERT's higher performance is due to its pretraining that utilizes a large text corpus. While LXMERT is also based on the BERT architecture, its textual component is not extensively pre-trained compared with BERT.

To better understand these phenomena, we did an error analysis of BERT's outputs through a manual qualitative observation of randomly sampled outputs. We discovered that, when there were enough details in the text to assume that a response follows the first-utterance, visual clues are not necessary to complete this task. For example, in Figure 5, BERT successfully predicted the label while LXMERT failed to do so. We suspected that LXMERT failed to associate the textual context of "warm" and "sweating" with the "cold" scene. This example shows that, in some cases, textual context would suffice to respond to an utterance correctly. Afterward, from about 15% of cases where BERT failed to predict the correct label, we randomly sampled 100 pairs of false negative cases. Then, we analyzed whether the visual information was actually required to correctly predict the label. We found out that 23% of the samples might not require visual clues to be classified correctly, thus they resulted in the complete failure of BERT classification.

Utterance: Are we warm enough?
Response: I'm sweating, you?

Figure 5: LXMERT failed to predict this pair as *correct*, while BERT and our model correctly predicted it.

For the rest of the 77%, we classified the conversation pairs into three types: (1) visual question-answering type, (2) image-referring response type, and (3) scene-understanding type. The *visual question-answering* type is pretty obvious; it involves cases where both the speaker and the respondent

are mainly asking and answering about the same specific object in an image. The *image-referring response* type is where the speaker does not explicitly refer to or mention an object, but the respondent does. We can see this as the opposite of the standard case where the speaker asks or talks about an object, and the respondent only needs to respond accordingly. The last one, the *scene-understanding* type, is more complicated to set a boundary to. As the name implies, this type requires contextual understanding both from the textual and its corresponding visual context. It might also require common-sense or specific assumptions on the conversation topic. Guided by these definitions, we observed that 35% of the data were of the first type, 11% were of the second type, and 31% were of the third type.

From the examples in Figure 6, we can see that, in the *image-referring response* type, the object "dog" in the response text needs to be referred to in the image. While it is similar, in the "scene-understanding" type, the textual context must be related to the scene. Considering these phenomena, we argue that we need an approach for extracting related and useful information from an image to complement the textual context.

Utterance: I am getting great pictures from the boat.
Positive: You want the dog in some of them?
Negative: Yes, we'll keep riding from here.

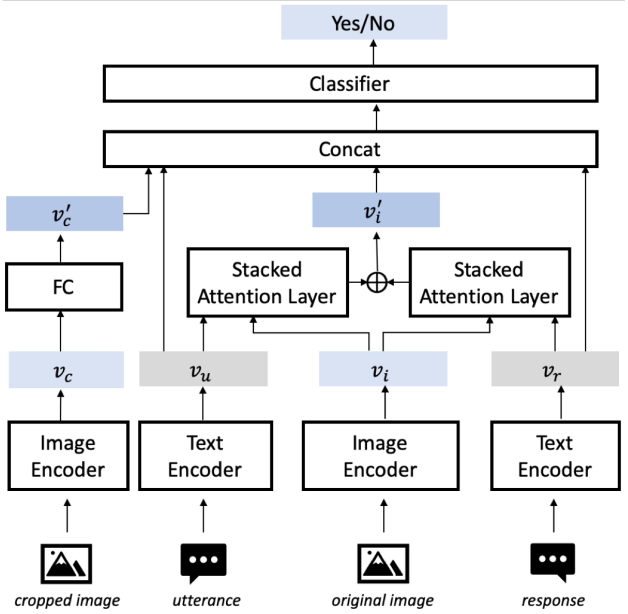
Utterance: Are you going to make it to the lift?
Positive: I'm going to jump out from this snow
Negative: Sure, someday I would love to climb the mountain.

Figure 6: Examples of visual conversation types: image-referring response (top) and scene-understanding (bottom).

4.3.3 Conversation Secene Image Augmented Model

To complement the above investigation into the impact of visual scenes on conversation tasks, and to confirm whether we can improve the performance of transformer-based models, we continued our experiments by using the cropped part(s) of the conversation scene images to aid the conversation text in this task. To accommodate our need for feeding multiple images and their pairs of utterance-response corresponding to the model, we developed a neural network model with a simple architecture. Figure 7 shows the architecture of our model.

(a) Main architecture



(b) Stacked Attention

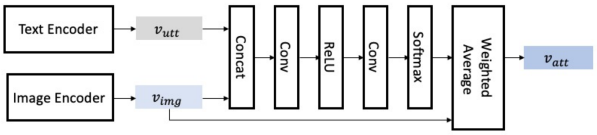


Figure 7: (a) Overview of our model architecture being used in experiments. (b) We applied attention over visual features.

We adapted the architecture for visual question-answering (VQA) tasks by Kazemi and Elqursh (2017) and combined it with the strategy of Minh

et al. (2018) for an addressee recognition task to optimize the extraction of visual information for the parts focused on. We modified the objective to predict whether a given pair is appropriate or not, i.e., as binary classification. The objective can be formulated as

$$a^* = \arg\max_{a \in [0,1]} p(a|\mathbf{x}, \mathbf{r}; \theta), \tag{8}$$

where $\mathbf{x}$ is multi-modal input representation including text utterance u, original image i, and cropped image area c, while $\mathbf{r}$ is a response candidate, we computed the probability of a, where the value is between 0 and 1, to get the classification $a*$. To calculate the $p(a|\mathbf{x}, \mathbf{r}; \theta)$, we first encode all the inputs and the response using separate neural encoders: $v_u = f_u(u), v_i = f_i(i), v_c = f_i(c)$ and $v_r = f_r(r)$. We represent the text encoder as f_u and the image encoder as f_i, respectively, to produce the vectors v_u,v_i, and v_c. We used the same (shared) image encoder for both the original image i and the cropped image c. Since a cropped image is a part of an original image, we assume it would be more useful for the encoder to be able to learn the mutual information between these visual clues. In contrast, we used a different text encoder f_r for response r. We used one-layer bi-directional LSTM (Hochreiter and Schmidhuber, 1997) for the text encoder and ResNet (He et al., 2016), pre-trained with the Places356 dataset Zhou et al. (2017a), for the image encoder.

For the original image vector v_i, a soft attention mechanism was used to align the information from image features and textual context, v_u and v_r. We adapted stacked attention used by Kazemi and Elqursh (2017) for this mechanism. This attention mechanism results in vectors v_i^u and v_i^r, which represent the alignment for the text utterance and response, respectively. Then, we combined both alignment vectors by element-wise addition:

$$v_i' = v_\alpha^u + v_\alpha^r. \tag{9}$$

Experiment Settings

We used Adam Kingma and Ba (2014b) as the optimizer for training and fixed the learning rate at 0.0001. The mini batch size was fixed at 20. We

defined an early-stopping condition, as the training convergence indicator, using a parameter called *patience*. We defined *patience* as the maximum consecutive iterations allowed when the average training loss is increasing greater than a threshold. For instance, if we set the *patience* to 10 and the threshold to 5%, the training should stop if the average loss is increasing above 5% (compared to the best average loss) for ten consecutive iterations. In our experiments, we fixed the *patience* at 5 and set the threshold to 5%. We trained our models on an Nvidia Titan RTX GPU with a 24-GB GPU-memory capacity. We limited the maximum training epoch to 50 epochs. Typically, the training converged in approximately 12 hours for our model. The dimensions of word embeddings and LSTM were set to 300 and 512, respectively, while we set the dimension of image features to 2048. As for input images, which consists of one original image and its three cropped parts, we pre-processed all images, original and cropped images, to have equal dimension of 256×256. Then, we applied data augmentation to the images during training, including random cropping, horizontal flipping, and normalization transformation.

4.4 Result and Discussion

A summary of the results can be seen in Table 8. Using the model described in Section 4.3.3, we achieved higher scores than BERT. We evaluated several input settings for combinations of cropped images. In line with our analysis on visual conversation types, feeding a cropped image of either the corresponding speaker/respondent/object seemed effective. Particularly, cropped object-images, which achieved an accuracy of 86.04%, got higher scores than BERT. Our analysis showed that this was related to the *visual question-answering* type, where BERT failed to capture the context from the provided text.

In comparison with LXMERT's results, our model with a cropped image as additional input feature also resulted in better scores. Please note that, as describe in Section 4.3.2, because LXMERT treats an image as a set of region features, we fed the image and focused region parts (speaker, respondent, and topic object) as the region features, i.e., not as cropped images. In our

Model	Input					Accuracy	Precision	Recall	F1
	Text	Scene Image	Focused-on Part						
			S	R	O				
BERT	✓	-	-	-	-	85.28	0.849	0.857	0.853
LXMERT	✓	✓	✓	✓	✓	79.4	0.78	0.82	0.8
LXMERT	✓	✓	-	-	-	80.65	0.79	0.83	0.81
Ours	✓	-	-	-	-	64.77	0.641	0.667	0.653
Ours	✓	✓	✓	-	-	82.88	0.806	0.867	0.835
Ours	✓	✓	-	✓	-	84.2	0.829	0.86	0.844
Ours	✓	✓	-	-	✓	86.04	0.843	0.884	0.863
Ours	✓	✓	✓	✓	-	88.8	0.871	0.909	0.89
Ours	✓	✓	✓	-	✓	**91.18***	0.894	**0.933**	**0.911**
Ours	✓	✓	-	✓	✓	90.07	0.889	0.914	0.901
Ours	✓	✓	✓	✓	✓	91.0*	**0.897**	0.924	**0.911**

Table 8: Results of our models with different input settings. *Original images* were inputted as *scene images* by default. S, R, and O denote *speaker*, *respondent*, and *topic object* cropped images, respectively. Difference between a pair of * marked scores was not statistically significant. Differences between all other pairs (from our model) were significant (with T-test, $p < 0.05$).

experiments, we have two variants of inputs for LXMERT: (1) Original image with focused-on Speaker, Respondent, and Object (SRO) regions, and (2) Original image without focused part, instead, we used all object annotations by Visual Genome which typically contains ten or more region features. The corresponding results of these input settings can be seen in Table 8. Also, in our experiments, we used same image features for both LXMERT and our model – we did not use Faster-RCNN to extract the image features for LXMERT.

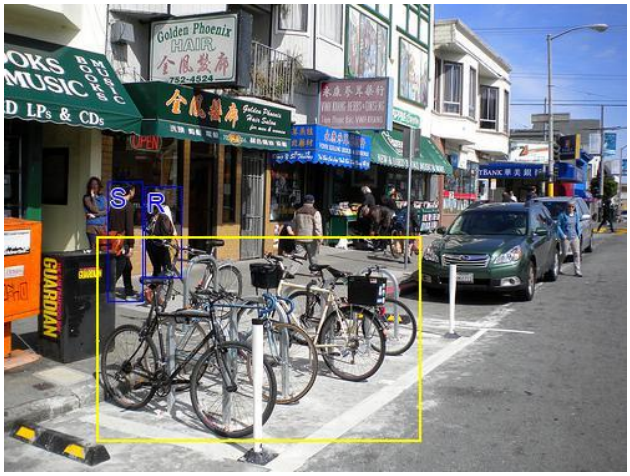
Our experiments with two cropped images showed more improvement in terms of accuracy and F1 score, particularly with *speaker* and *object* components. We saw this improvement as conforming to our visual conversation categories of the *image-referring response* type and *scene-understanding* type. In Figure 8, we can see one example where our proposed model correctly predicted the pair as correct, while BERT failed to do so. Using two cropped images, in particular, the *speaker* and *object*, can help the model to focus on the important parts. Furthermore, in some cases, *utterance* and *response* text might refer to different foci. For example, in both pictures in Figure 6, the topics of the utterance and response might seem unrelated. Therefore, having two related components could help the model to learn the relationship between each text pair and the visual information better. This setting

of using *speaker* and *object* components can be seen as a natural conversation scene, considering that, in actual conversations, the respondent occasionally needs to see the partner and/or the object of the topic to understand the context.

Utterance: Hey dude, look at that pretty girl watching us.
Positive: They are under age man.
Negative: You are going to bust your ass.

Figure 8: Cropped object image (yellow box) provides more information for understanding the conversation context.

Regarding BERT's performance, according to our findings, we postulate that when the responses could be perceived as semantically following the utterances, it would be enough for BERT to predict the label. As described in Section 4.3.2, in Figure 5 BERT predicted the label correctly while LXMERT, with visual information, failed to do so. However, our model with *speaker* and *object* (a sweater) cropped images could also correctly predict the label. We argue that since our model had a clearer object-related visual clue — a cropped image compared with the bounding boxes of LXMERT — it could associate the textual context of "warm" even though the scene was arguably "cold." Therefore, we argue that in order to capture a better context for the conversation, visual scene information is essential, particularly for the three visual conversation types: visual question-answering, image-referring response, and scene-understanding. Nevertheless, we found that in some



Utterance: Will our bikes be safe?.
Positive: Did you lock yours?
Negative: Do you have a key to unlock it.

Figure 9: Example case when both BERT and our model failed to predict the positive response correctly.

particular cases, for example, in Figure 9, even with the visual scene information, the model could not correct BERT's errors. This failure might have been caused by an inability to associate the visual clue of "bikes" with the response text.

Since the negative pairs were generated from conversations with the same images, we also analyzed sampled results from false-positive cases. We did this analysis to investigate whether the models' failure were caused by "naturally positive" conversations that made it difficult for the models to predict the negative label. While we found that some false-positive predictions were produced by negative pairs that can be considered a "natural positive conversation", especially if the input utterance was written in an open-question manner. However, as the models are supposed to consider not only the text but also the visual clue with different focused parts, e.g., different topic objects, such cases are very rare once visual contexts are taken into accounts. Since the number of false-positive answers is significantly smaller to the true-negative ones, for example, about 600 vs 2700 for LXMERT or 400 vs 2800 for our models, we argue that the negative samples worked as intended and

the failures were not mainly caused by false-positive cases.

5 Conclusion and Future Work

5.1 Conclusion

To improve the quality of automated responses in conversation/dialogue tasks, we arranged our research objective into two: (1) improving generated response quality with user-level styles (resemblance to the ones of intended users), and (2) improving the performance of conversation tasks using multi-modal (visual and language) contexts.

For the first objective, we proposed a framework and strategy to use user-specific information, consisting of usernames and user-info dataset (users' frequently-used words), to stylize the output response to resemble the ones of intended users. The evaluation through human judgment showed that the outputs of our model are better than the baseline overall, especially the variant with unseen users. Although our model has a simple architecture and is small in size, it could produce responses with acceptable quality. We believe that our experiments can serve as evidence that, even with a limited size, a simple and intuitive architecture can improve the response quality.

For the second objective, we experimented on visual and language conversation tasks using specific scenic images as a visual clue. We presented a new multi-modal dataset containing pairs of conversation scene (third-person view) images, their corresponding focused part annotation, and conversation utterances. Using this dataset, we investigated the effectiveness of visual clues on a conversation task performance. Our findings show that visual information is necessary when context understanding is required. We also identified three types of visual conversations where visual conversation scenes are essential to understanding the conversation context: (1) *visual question-answering* type, (2) *image-referring response* type, and (3) *scene-understanding* type. On the basis of this analysis, we also proposed an approach for using scene components in the tasks, particularly the speaker, respondent, and topic object parts. Our model could significantly outperform the baseline models by partially coping with the first and second conversation types, obtaining 91% accuracy.

Regarding the simplicity of our second task, that is binary classification,

as mentioned in Section 4.1.2, we chose this task because our aim is to investigate the effectiveness of using still conversation scene images in conversation tasks. Nevertheless, we believe that we can apply the same strategy with more response candidates or labels to create more variation. Furthermore, it is also possible to use our dataset in, for example, a generation task. Since our conversation utterances are situated within the images rather than by outsiders observing the images, we postulate that we can learn a better conversation context in a more natural conversation scene.

5.2 Future Work

It still remains challenging to produce automated responses of human-like quality, such as engagingness and the ability to infer the context not only from utterances. Particularly in response generation tasks, it is difficult to properly incorporate and emphasize specific factors, such as style, to a response while maintaining the context or its relevance.

To improve the generation quality in general, especially in terms of fluency and engagingness, we believe that adopting pre-trained models, such as GPT-2, would help. However, incorporating specific response styles, especially at the user level, would be more complicated. We postulate that exploring the possibility of using graphs to represent the user-level information and to associate it with the utterances would be useful for future research.

As we proved that using scenic images in a conversation task improved the performance, we believe that more research in this field would be fruitful in the future. One possible future task is to improve the visual-conversation dataset by using images from movie scenes. We believe that increasing the data, especially the *image-referring response* and *scene-understanding* types, would improve the model's ability to "understand" the conversation context more. Another important future task is to explore more options in cross-modality processing. While our current method works for a simple task, improving alignment between visual and language information would be essential for visual and language research, including conversation tasks.

Acknowledgements

All praises to Allah, the Almighty and the Merciful, for giving me the strength to complete this thesis.

I would express my sincere gratitude to Prof. Manabu Okumura, Prof. Hiroya Takamura, and Dr. Hidetaka Kamigaito for their time, patience, and guidance throughout my research projects. Sincere thanks to Dr. Kotaro Funakoshi for the valuable comments and suggestions. I thank you for all of discussion, brainstorming sessions, and all of your valuable lessons during my graduate study period here in Tokyo Tech.

My special thanks to Dr. Nobuyuki Shimizu, Dr. Takashi Miyazaki, and Hiep Le for the countless discussion hours we had gone through during my internship period at Yahoo Japan Research Lab. I am forever indebted.

My gratitude and appreciation to all members of Okumura - Funakoshi Laboratory for the supports, especially Fujita, Boat, and Yukun for many days we had spent together in the lab room. My gratitude also for Ms. Iiyama for the all the help on the administrative works.

I also could not have finished this study without the endless support from my beloved wife and son, and all the love and prayer from my parents.

Lastly, thank you for MEXT for giving me the opportunity to study at Tokyo Institute of Technology through the scholarship.

References

H. Alamri, V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson, S. Lee, and D. Parikh. Audio-visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

J. Bak and A. Oh. Variational hierarchical user-based conversation model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1941–1950, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1202. URL https://www.aclweb.org/anthology/D19-1202.

Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

E. Chu, P. Vijayaraghavan, and D. Roy. Learning personas from dialogue with attentive memory networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2638–2646, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1284. URL https://www.aclweb.org/anthology/D18-1284.

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. URL https://www.aclweb.org/anthology/J90-1003.

A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

J. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

X. Gao, Y. Zhang, S. Lee, M. Galley, C. Brockett, J. Gao, and B. Dolan. Structuring latent spaces for stylized response generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1814–1823, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1190. URL https://www.aclweb.org/anthology/D19-1190.

S. Ghazarian, J. Wei, A. Galstyan, and N. Peng. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2310. URL https://www.aclweb.org/anthology/W19-2310.

J. Haber, T. Baumgärtner, E. Takmaz, L. Gelderloos, E. Bruni, and R. Fernández. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1184. URL https://www.aclweb.org/anthology/P19-1184.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. Lopes, A. Das, I. Essa, D. Batra, and D. Parikh. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2019. doi: 10.1109/ICASSP.2019.8682583. URL https://www.merl.com/publications/TR2019-016.

B. Huber, D. McDuff, C. Brockett, M. Galley, and B. Dolan. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173851. URL https://doi.org/10.1145/3173574.3173851.

H. Kamezawa, N. Nishida, N. Shimizu, T. Miyazaki, and H. Nakayama. A visually-grounded first-person dialogue dataset with verbal and non-verbal responses. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3299–3310, Online, Nov. 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.267.

V. Kazemi and A. Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *CoRR*, abs/1704.03162, 2017. URL http://arxiv.org/abs/1704.03162.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014a.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014b. URL http://arxiv.org/abs/1412.6980. cite

arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1): 32–73, May 2017. ISSN 0920-5691. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 1977.

A. W. Li, V. Jiang, S. Y. Feng, J. Sprague, W. Zhou, and J. Hoey. Aloha: Artificial learning of human attributes for dialogue agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8155–8163, Apr. 2020. doi: 10.1609/aaai.v34i05.6328. URL https://ojs.aaai.org/index.php/AAAI/article/view/6328.

J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119. Association for Computational Linguistics, 2016a. doi: 10.18653/v1/N16-1014. URL http://www.aclweb.org/anthology/N16-1014.

J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany, August 2016b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1094.

C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Lan-*

guage Processing, pages 2122–2132. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1230. URL http://www.aclweb.org/anthology/D16-1230.

R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1103. URL https://www.aclweb.org/anthology/P17-1103.

J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics, 2015. doi: 10.18653/v1/D15-1166. URL http://www.aclweb.org/anthology/D15-1166.

P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1298. URL https://www.aclweb.org/anthology/D18-1298.

T. L. Minh, N. Shimizu, T. Miyazaki, and K. Shinoda. Deep learning based multi-modal addressee recognition in visual scenes with utterances. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1546–1553. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/214. URL https://doi.org/10.24963/ijcai.2018/214.

C. Miyazaki, T. Hirano, R. Higashinaka, T. Makino, and Y. Matsuo. Automatic conversion of sentence-end expressions for utterance characterization of dialogue systems. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 307–314, Shanghai, China, Oct. 2015. URL https://www.aclweb.org/anthology/Y15-1035.

M. Mizukami, G. Neubig, S. Sakti, T. Toda, and S. Nakamura. *Linguistic Individuality Transformation for Spoken Language*, pages 129–143. Springer International Publishing, Cham, 2015. ISBN 978-3-319-19291-8. doi: 10.1007/978-3-319-19291-8_13. URL https://doi.org/10.1007/978-3-319-19291-8_13.

N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. Spithourakis, and L. Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan, Nov. 2017. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1047.

V. Murahari, P. Chattopadhyay, D. Batra, D. Parikh, and A. Das. Improving generative visual dialog by answering diverse questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1449–1454, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1152. URL https://www.aclweb.org/anthology/D19-1152.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel,

M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1162.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014b. URL http://www.aclweb.org/anthology/D14-1162.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593. Association for Computational Linguistics, 2011. URL http://www.aclweb.org/anthology/D11-1054.

L. Rokach and O. Maimon. *Clustering Methods*, pages 321–352. Springer US, Boston, MA, 2005. ISBN 978-0-387-25465-4. doi: 10.1007/0-387-25465-X_15. URL https://doi.org/10.1007/0-387-25465-X_15.

A. Saha, M. M. Khapra, and K. Sankaranarayanan. Towards building large scale multimodal domain-aware conversation systems. In S. A. McIlraith

and K. Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 696–704. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17104.

S. Sato, N. Yoshinaga, M. Toyoda, and M. Kitsuregawa. Modeling situations in neural chat bots. In *Proceedings of ACL 2017, Student Research Workshop*, pages 120–127, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-3020.

L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586. Association for Computational Linguistics, 2015. doi: 10.3115/v1/P15-1152. URL http://www.aclweb.org/anthology/P15-1152.

K. Shuster, S. Humeau, A. Bordes, and J. Weston. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.219. URL https://www.aclweb.org/anthology/2020.acl-main.219.

A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado, May–June 2015a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N15-1020.

A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado, May–June 2015b. Association for Computational Linguistics. doi: 10.3115/v1/N15-1020. URL https://www.aclweb.org/anthology/N15-1020.

K. Sparck Jones. Document retrieval systems. chapter A Statistical Interpretation of Term Specificity and Its Application in Retrieval, pages 132–142. Taylor Graham Publishing, London, UK, UK, 1988. ISBN 0-947568-21-2. URL http://dl.acm.org/citation.cfm?id=106765.106782.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

H. Tan and M. Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://www.aclweb.org/anthology/D19-1514.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

O. Vinyals and Q. V. Le. A neural conversation model. In *Proceedings of the 31th ICML Deep Learning Workshop*, Lille, France, 2015.

D. Wang, N. Jojic, C. Brockett, and E. Nyberg. Steering output style and topic in neural response generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2140–2150. Association for Computational Linguistics, 2017. URL http://aclweb.org/anthology/D17-1228.

B. Wu, M. Li, Z. Wang, Y. Chen, D. F. Wong, Q. Feng, J. Huang, and B. Wang. Guiding variational response generator to exploit persona. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 53–65, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.7. URL https://www.aclweb.org/anthology/2020.acl-main.7.

T. Yuma, N. Yoshinaga, and M. Toyoda. uBLEU: Uncertainty-aware automatic evaluation method for open-domain dialogue systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 199–206, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-srw.27. URL https://www.aclweb.org/anthology/2020.acl-srw.27.

W. Zeng, A. Abuduweili, L. Li, and P. Yang. Automatic generation of personalized comment based on user profile. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 229–235, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-2032. URL https://www.aclweb.org/anthology/P19-2032.

S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://www.aclweb.org/anthology/P18-1205.

Y. Zhang, X. Gao, S. Lee, C. Brockett, M. Galley, J. Gao, and B. Dolan. Consistent dialogue generation with self-supervised feature learning, 2020a.

Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.30. URL https://www.aclweb.org/anthology/2020.acl-demos.30.

Y. Zheng, Z. Chen, R. Zhang, S. Huang, X. Mao, and M. Huang. Stylized dialogue response generation using stylized unpaired texts, 2020.

B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017a.

G. Zhou, P. Luo, R. Cao, F. Lin, B. Chen, and Q. He. Mechanism-aware neural machine for dialogue response generation. In *AAAI*, 2017b.

List of Publication

Journal Papers

Abdurrisyad Fikri, Hiroya Takamura, Manabu Okumura. Stylistically User-specific Response Generation. In *Journal of Natural Language Processing*, Vol. 28 No. 4, 2021.

Abdurrisyad Fikri, Hiep V. Le, Takashi Miyazaki, Manabu Okumura, Nobuyuki Shimizu. Improving Conversation Task with Visual Scene Dataset. In *Journal of Natural Language Processing*, Vol. 29 No. 1, 2022.

Conference Papers

Abdurrisyad Fikri, Hiroya Takamura, Manabu Okumura. Stylistically User-specific Response Generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 89–98. Association for Computational Linguistics, 2018.