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
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報通信 系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 (Engineering) Doctor of
学生氏名： Student's Name	Abdurrisyad Fikri	指導教員（主）： Academic Supervisor(main)	奥村学
		指導教員（副）： Academic Supervisor(sub)	

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

In dialogue or conversation tasks, mimicking actual human conversations is the ultimate goal. However, since producing human-like qualities is challenging, studies on conversation tasks often need to focus on limited aspects.

One aspect of real-world conversations is, humans can produce interesting or engaging responses as opposed to general-sounding ones such as “I am OK” or “I don't know”. 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.

Another aspect is that conversation context is dynamic, not only depending on the conversation history but also being influenced by other factors such as speakers' personalities, surrounding objects, etc. Arguably, in conversation tasks, using single-modality (text-only) is the most straightforward method. However, according to recent studies on multi-modal tasks, adding corresponding images grounded to the conversation context has proven useful in improving model performance.

Thus, in our study, we have two aims to focus on: (1) to produce automated responses that resemble human-produced ones, and (2) to utilize multi-modal features to capture the conversation context better. We translate our two research aims into two sub-tasks: (1) user-specific response generation, and (2) response classification with a visual scene dataset.

In the first sub-task, we experimented on capturing the user-level characteristic to drive a response generation model to generate a “stylized response” that resembles the ones from intended users. We defined the “stylized response” as a response that contained frequently used words or characters from the users. We crawled Twitter as our source for the dataset. In addition to the conversations' 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 later converted to user-specific information. Using both features, we trained our model to pay attention to both conversation context (previous utterance) and user-specific information to generate a response that incorporates user-level response style.

In the second sub-task, we experimented on using scene images in a conversation setting as the visual clue. Unlike other studies on visual and language models for dialogue tasks, that mainly focused on topic object images, we intend to explore the potential of scenic images in conversation settings. We argue that, instead of seeing only the object of the conversation topic, we would understand the context better if we can see the speakers' condition also.

Moreover, 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.

For this task, 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 popularity of transformer-based models, and to confirm our hypothesis that visual clues positively improve the task performance, we first experimented with a BERT(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.

We also proposed a model that treats the focused annotated part(s) of the images as separate image features.

Our experiment result shows that our proposed model outperformed the baseline on a binary response classification task.
We also did an analysis on the conversation types and propose three types of visual conversation: 1) visual question-answering, (2) image-referring response, and (3) scene understanding.
Also, we believe our dataset would be useful for future research on visual and language conversation tasks.

Through these sub-tasks, we offer contributions to the NLP fields as follows:
(1)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, and generate more resembling responses to the actual users compared to the baselines.

(2)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.

(3)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 of using scenic images, as opposed to commonly used topic images, as the visual clue. We also proposed a neural network model that achieves better results compared to the baselines.

(4)We propose new visual conversation types that can help navigate future work on cross-modal conversation tasks.

To conclude, producing human-like qualities, such as engagingness and the ability to infer the context not only from utterances, in automated responses is challenging.
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.
Nevertheless, our results show that we can control the automated response, though in a limited scope, by using related information. Our results also suggest that using a multi-modal context and widening the focused area/adding more features can increase the model “understanding” of the conversation context.