JAIST Repository
https://dspace.jaist.ac.jp/

Title	非タスク型ダイアログシステムのための多モーダルユーザー印象認識に関する研究
Author(s)	魏, 文青
Citation	
Issue Date	2025-03
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/19926
Rights	
Description	Supervisor: 岡田 将吾, 先端科学技術研究科, 博士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Abstract

With the continuous development of human-computer interaction technology, the application of dialogue systems in various fields is becoming increasingly widespread. Among them, non-task-oriented dialogue systems have shown great potential in fields such as chatbots and open-domain dialogue systems. While improving the quality of non-task-oriented dialogue systems is crucial for enhancing user interactions. A high-quality dialogue system must not only understand user intent but also generate accurate and natural responses. This necessitates a robust evaluation framework to assess the system's capabilities. Additionally, evaluation serves as the foundation for system improvement and optimization. Through evaluation of the dialogue system, weaknesses in the system can be identified, making it easier to fine-tune models, data, or algorithms to improve overall performance.

With the rise of multimodal dialogue systems, the demand for their evaluation has also increased. However, existing evaluation methods for dialogue systems often focus solely on text-to-text interactions, neglecting the importance of multimodal data in dialogue systems. In contrast, unimodal systems typically rely only on language content or speech intonation, which may lead to neglect or misinterpretation of users' emotions. Multimodal information, such as speech intonation, facial expressions, and body movements, can better capture users' emotional changes. Therefore, utilizing multimodal information to evaluate multimodal dialogue systems is crucial.

Moreover, text-based evaluation metrics, such as BLEU and ROUGE, are insufficient for assessing multimodal dialogue systems. At the same time, existing multimodal databases face limitations in data collection, particularly in collecting speech and image data, resulting in incomplete and limited evaluation methods. Motivated by these challenges, this research aims to address data collection issues in the evaluation of multimodal non-task-oriented dialogue systems and propose innovative evaluation methods. By collecting, organizing, and utilizing multimodal data, we aim to evaluate dialogue system performance more comprehensively and accurately, thereby enhancing user experience and impressions. Therefore, this research has significant theoretical and practical implications and will make important contributions to the development of the field of multimodal dialogue system evaluation.

Above all, to establish an automated, robust, and accurate model for evaluating multimodal dialogue systems. Firstly, we introduce a method for identifying user satisfaction at the dialogue level, filling a gap in previous research. We use a method based on multimodal modeling, which comprehensively considers various information such as text, speech, and images to evaluate dialogue system performance more comprehensively. Then, we utilize deep learning models to comprehensively analyze user satisfaction at the dialogue level and user impressions at the exchange level, enhancing the accuracy and reliability of the evaluation methods. Through experimental evaluation, we confirm the effectiveness and feasibility of the proposed methods in the field of multimodal dialogue system evaluation, providing new insights and methods for further research in this area.

The user impression can be analyzed and evaluated at two levels: the exchange level and the dialogue level. These two levels are closely interconnected but differ in focus, making them well-suited for capturing information at different hierarchical levels. While the relationship between user impressions at the dialogue level and user sentiments at the exchange level is secondly explored, which proposes a multi-task learning model that comprehensively considers information from both levels. By analyzing the relationship between 18 dialogue labels and user sentiment during dialogue exchanges and utilizing multi-task learning models, we successfully achieve accurate identification of user impressions at the dialogue level, bringing new insights and methods to the field of dialogue system evaluation.

Lastly, we address the issue of existing methods neglecting the influence of users' personal information which included age, gender, and personality on their impressions by proposing a model based on adversarial learning. By reversing the gradient direction during training, our network learns adversarial features that remain consistent across different users' personal information domains, effectively mitigating the influence of users' personal information and making the model applicable to evaluate non-task-oriented dialogue systems. Through experimental validation, we confirm the effectiveness and feasibility of the proposed method, providing new insights and methods for further research in the field of dialogue system evaluation.

In conclusion, this study proposes novel approaches to addressing the evaluation challenges encountered by multimodal non-task-oriented dialogue systems. The proposed methods improve the accuracy and comprehensiveness of dialogue system evaluation, offering valuable insights for enhancing user experience and satisfaction in various applications.

Keywords: Dialogue system, Multimodal, Evaluation, User impression, User traits adaptation.