Journal of Theoretical and Applied Information Technology

30th November 2023. Vol.101. No 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MULTIMODAL LEARNING CONVERSATIONAL DIALOGUE SYSTEM: METHODS AND OBSTACLES

MUHAMMAD FIKRI HASANI[1,a] , GALIH DEA PRATAMA[2,b], ERNA FRANSISCA ANGELA SIHOTANG[3,c], FRANZ ADETA JUNIOR[4,d]

[1]Computer Science Program, Computer Science Department, School of Computer Science, Bina Nusantara University, West Jakarta, Jakarta, Indonesia

[2]Game Application and Technology Program, Computer Science Department, School of Computer Science, Bina Nusantara University, West Jakarta, Jakarta, Indonesia

[3]Statistics Department, School of Computer Science, Bina Nusantara University, West Jakarta, Jakarta, Indonesia

[4]Cyber Security Program, Computer Science Department, School of Computer Science, Bina Nusantara University, West Jakarta, Jakarta, Indonesia

E-mail: [a]muhammad.fikri003@binus.ac.id, [b]galih.pratama001@binus.ac.id, [c]erna.sihotang@binus.ac.id, [d]franz.junior@binus.ac.id

ABSTRACT

Humans converse with each other as one of the means to interact socially. But conversation not only served as media of communication, but also opened several occupancies like moderators, customer services, master-of-ceremony, or teachers. Dialogue system is a computer program that supports spoken, text-based, or multimodal conversational interactions with humans with many implementations recently and only work in single modality, such as text. However, human understanding is not limited into single data domain, but it needs the collective data domain information to understand the whole surroundings which is called as multimodality in the field of computer science and implemented further through the concept of artificial intelligence called multimodal learning. Multimodal learning has been subjected in research since years ago to increase artificial intelligence model result, such as enhancement of speech recognition through mouth image and facial expression recognition based on facial and landmark textures. This paper will provide reference in integration of multimodal learning in dialogue system, which will be useful to negate obstacles present in future research.

Keywords: *Multimodal learning, Dialogue system, Systematic literature review, Chatbot, Deep learning*

1. INTRODUCTION

Instinctively, humans are social beings, where conversations are done to interact with one another. Conversation is joint activity in which two or more participants use linguistic forms and nonverbal signals to communicate interactively [1]. Conversation itself is not only served as media of communication, but also opened several occupancies that involve around conversational proficiency, such as moderators, customer services, MC, lecturers, mentors, or teachers.

Recently, with the advance of technology, there exists a concept that tries to simulate and model on how human is doing conversation, namely dialogue systems or conversational agents. Dialogue system or conversational agent is a computer program that supports spoken, text-based, or multimodal conversational interactions with humans [2]. The application of dialogue system can usually in the form of applications, such as Google Assistant from Google, Siri from Apple, Cortana, and XiaoIce from Microsoft [3], or a bot in some instant message application. Initially, these dialogue systems only work in one data domain. For example, chatbot only understands input based on textual data. [4] only understands input from audio/speech data, and [5] only understand input based on textual data. However, human understand the information not only based on one data domain, but they combined multiple data domain to understand their surroundings. This phenomenon in computer science is called multimodality, and the concept of artificial intelligence that used multiple data domain for input source is called multimodal learning [6].

Multimodal learning has been conducted for some years ago, with focus to increase artificial intelligence model result. In [5], the authors combined features from audio and visual to enhance

speech recognition with cropped mouth image. In [7], the authors tried to increase facial expression recognition that combine facial textures and landmark features as multimodal approaches, and [8] which tries to take certain video segments through text query that combines text and image modalities, and many more research about multimodal learning in artificial intelligence.

From our findings, the purpose of various research involving chatbot is to enhance the understanding of chatbot to user conversation or dialogue using numerous techniques, currently using only single data domain. Despite that, the usage of single data domain is still limited, particularly when the user wanted to clarify the context by using multiple data domain, such as providing pictures, audio, and even video due to its benefits, especially in customer service area.

Therefore, this paper conveys information from numerous references, which can be beneficial for future research with interest in dialogue system field. This will create many other possibilities in research, such as the integration of multimodal learning in dialogue system and its further use alongside the obstacles that might present during its implementation.

2. METHOD

This research uses systematic literature review (SLR) methodology based on multiple references from international publications. Reference-searching activity involves two main parts consist of research questions and inclusion-exclusion rules. Research questions are used to define the goals needed to be done in the research. To support those, inclusion-exclusion rules definition is done to filter past research for better output.

2.1 Research Question

Research questions are identified as basis for this publication, which is found through discussions by contributing authors. They also serve to filter the referenced papers, especially for the Related Works section. Based on the discussion, the research questions are as follows:

- How multimodal dialogue system is implemented in various cases?
- Which methods of multimodal learning dialogue system that are possible to be implemented?
- What challenges are being faced by the implementation of multimodal dialogue system?

2.2 Inclusion-Exclusion Rules

To produce better output in this research, certain rules regarding the selection of references need to be defined. In realization of that, inclusion-exclusion rules are made to select which papers are eligible to be referenced. Based on what this research discussed, these are inclusion-exclusion rules used as follows:

- Publication from international conference and journal published in recent six years. In this case, the earliest paper should be published in 2017.
- Publication indexed in Google Scholar, IEEE, or Elsevier.
- Publication that specifically discusses dialogue system and/or multimodal learning dialogue system.

3. RESULTS AND DISCUSSION

This section will explain the results and discussion in this research. This section has four parts, starting with Related Works that shows the publications used for evaluation. It is then continued with Quantitative Result, Qualitative Result, and Discussion.

3.1 Related Works

This literature review involves deep analysis regarding papers spanning from 2017 until 2021. To enhance the quality in analysis of multimodal learning conversational dialogue system, reference-filtering is done strictly. Referenced papers are taken from numerous conferences and journal across the world, with the summary can be seen on Table 1.

Table 1. List of publication for analysis

Publication	Methods
[9]	Reinforcement Learning with Q-Learning Algorithm.
[10]	Knowledge-aware Multimodal Dialogue (KMD) collaborated with HRED network.
[11]	Bidirectional Recurrent Neural Network (Bi-RNN) with Gated Recurrent Units (GRU) to encode textual data, VGG-16 for visual data encoding, and standard RNN GRU to generate response.
[12]	Two-way RNN, enhanced with attention-based Convolutional Neural Network (CNN) and processed further with multimodal Factorized Bilinear Pooling (MFB)

Publication	Methods
	module.
[13]	Pre-trained BERT for textual data, Mel-Frequency Cepstral Coefficient (MFCC) for audio data, and pre-trained Residual Network (ResNet) for visual data.
[14]	RNN to encode text-based utterance, ResNet to extract visual features, Multi-Layer Perceptron (MLP) with cross-entropy loss function to optimize network, and RNN to decode responses.
[15]	Multimodal Graph Convolutional Network (GCN).
[16]	Deep Averaging Network (DAN) to match utterances, and mini-Xception model based on CNN to detect emotions.
[17]	DST model to generate response, and pre-trained GPT-2 language model to form prediction and more accurate output.
[18]	Sequence-to-sequence model based on Transformer network to generate multimodal dialogue response and translate text to image.

Currently, the literature review related to this research focuses only on single aspect, either multimodal learning [19,20] or chatbot technology [21–23]. From the former literatures, there is still no research which discusses the usage of chatbot technology alongside multimodal learning. Therefore, this paper will present various technical aspects not shown by any of these papers, due to implementation of multimodal learning in chatbot applications might be different than multimodal learning for other purposes.

3.2 Quantitative Result

The quantitative aspect shown in the section will consist of three parts; publication source, publication year of release, and modalities used in the publications. There are only two sources of publication used here, such as international conference and journal. In this publication, conference proceedings are used more than the journal ones, as seen on Figure 1.

Figure 1. Statistic of publication based on the sources

Year of release is also used in consideration for choosing the proper publications. For this publication, the earliest publication is released in 2017 and the latest is in 2022, which statistics can be seen on Figure 2.

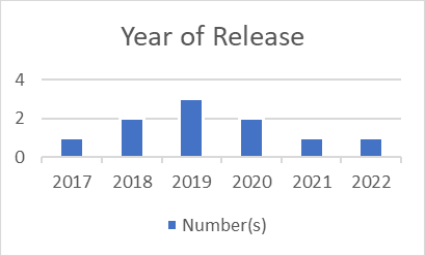

Figure 2. Statistic of publication based on year of release

In each of the publications, there is more than one modality used for the experiment, hence it is called multimodal. Text is the commonly used modality, where it can fit in many cases provided. Audio and image are also mostly used and in tandem with any other modalities like video and motion. Modalities such as video and motion are the least used ones in the publications cited, that is featured on Figure 3.

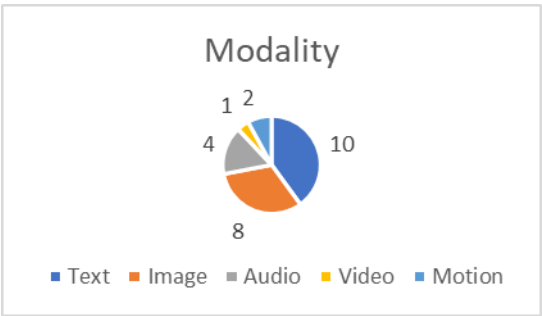

Figure 3. Statistic of publication based on modalities

The multimodal dialogue system can be implemented case-by-case in any domain. The common implementation of the system can be

found in fashion domain, where it can predict what kind of fashion suits the need of the user. The implementation is also put in retail domain with the same purpose, mainly in fashion and furniture domain. There are also many domains put for the implementation, which can be seen in Figure 4.

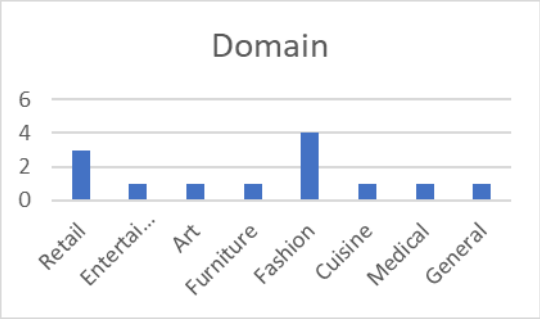

Figure 4. Statistic of publication based on domain

3.3 Qualitative Result

In this section, each of the publications used in related works will be discussed further. This will cover what the publication aimed from the use of multimodal learning, advantage of the usage and the disadvantages present in the said research.

[9] focuses on the implementation of multimodal system for Human-Robot interaction which combines task-based and chatbot-style dialogue inside. The system will then be used for entertainment purposes where it enables user to interact and gives responses accordingly through text, audio, and motion modalities for shopping mall domain.

There are three main components used in the system, with the first one is an altered *Program AB* which implemented the chatbot to suit shopping mall domain. Then there comes second component called *BURLAP*, a Reinforcement Learning framework that uses standard *Q-Learning* algorithm [24] to train agent using a simulated user. The third component used in the system is *IrisTK* [25] which integrates the subsystems and handles speech recognition and speech synthesis.

The system is then evaluated in comparison with the chat-based and task-based version of the system. The hybrid system managed to get higher average reward and it can produce longer dialogues with real users, which made the system more engaging.

[10] implements knowledge-aware multimodal dialogue system which aims to provide special consideration to the semantics and domain knowledge does not present in any text-based dialogue systems. The dialogue system is used in fashion domain with the usage of text and image modalities to give more interactive responses.

The system used knowledge-aware multimodal dialogue (KMD) which is collaborated with HRED network. There are three main things in the system, such as Taxonomy-based Visual Semantic Learning, Domain Knowledge embedding, and End-to-end Reinforcement Learning [26]. The system is then experimented with various dataset of fashion domain.

Based on the experiments done in said publication, KMD gives the most satisfying and accurate result than any dialogue systems being experimented. While KMD only gives performance of almost 70% at average in text response, it excels better in image response experiment with performance reaching 97% compared to other dialogue system that topped only at 93% in the same testing.

[11] discusses the creation of baseline model for multimodal dialogue system in fashion domain. The publication also explains the hierarchical approach to text and image data, which then creates new dataset that can be used publicly. The research then continues with the usage of multimodal encoder-decoder.

The method starts with data collection that is helped by fashion expert to make sample conversations which consist of text and image. For the classification task, the system uses multimodal encoder-decoder. In text-only utterance, it is encoded through Bidirectional Recurrent Neural Network (Bi-RNN) model enhanced with Gated Recurrent Unit (GRU) cells [27]. In image-only utterance, it uses FC6 from VGGNet-16 [28] as encoder. For multimodal utterance, it concatenates the encoding process from two prior utterances. For decoding purpose, the standard RNN GRU cells are used to generate responses.

The experiment will then put two multimodal Hierarchical Encoder Decoder (HRED) to test, one with basic architecture and the other one with attention layer put in the network. Those models are put in test with unimodal baseline HRED for each task. From the result of experiment below, it can be inferred that multimodal HRED performs well in text-based task, while multimodal HRED with attention layer gives the best performance in all versions.

[12] continues with the focus of multimodal dialogue system that is guided by user attention. The system is made to give more appropriate response per user's requests, integrating it with text and image modalities present in retail and fashion domain. It will also be compared with other similar systems based on previous publications.

The system in the publication uses two-way Recurrent Neural Network (RNN) model which is applied in interaction between user and chatbot. To perform text-based task, the model is reinforced with attention mechanism based on Convolutional Neural Network (CNN) [24] that puts text as input and produces more attentive text feature. On the other hand, image-based task is processed with CNN model and continued with taxonomy-based combined tree that is given more weight on textual features. After that, the two features are processed deeper through multimodal Factorized Bilinear Pooling (MFB) [29] to produce utterance vector. The experiment will compare the proposed system with various similar systems, such as unimodal SEQ2SEQ [30], unimodal HRED, multimodal HRED with no additional layer, and knowledge-aware multimodal dialogue system (KMD). Through the experiment, the proposed system produces the best performance in every task. The system is also compared individually with other systems and scored based on four aspects such as fluency, relevance, logical consistency, and informativeness. From the individual comparison, the proposed system won in every aspect and only pales to unimodal HRED in informativeness aspect.

[13] also takes interest in similar research to identify sentiment from human behavior. To achieve that, the publication will focus on multimodal deep neural network for identifying opportunities when the agent should express positive or negative empathetic responses. It is used in medical segment and uses text, audio, and video (visual) modalities in the system, which is a fusion of Recurrent Neural Network (RNN).

To train and evaluate the system, the portion of Distress Analysis Interview Corpus – Wizard-of-Oz (DAIC-WOZ) [31] is used as the dataset and labels from Amazon Mechanical Turk (MTurk) [32] will be served to validate the wizard's empathetic responses from former dataset. The proposed system consists of three main parts in feature extraction, with each one focuses on one modality. The pre-trained BERT [33] is used to process textual features. On the other hand, audio features are processed through extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [34] and Mel-frequency cepstral coefficients (MFCC) [21], which then extracted using OpenSMILE [35]. The visual representation is experimented with two different feature sets, where OpenFace [36] is used to extract the intensity of facial action units based on FACS and face embedding from OpenFace which uses ResNet-50.

The experiment will then evaluate the proposed system on a dataset of 2185 instances of conversation excerpts from 186 participants. The result shows that the system produces moderate accuracy in MTurk and Wizard labels.

[14] puts focus on representation of MAGIC, a multimodal dialogue system that uses adaptive decoder. The system has a function to produce proper response based on the context. The system is used for retail domain and centered between text and image modalities during usagae. MAGIC has various components, such as context encoder that implements RNN for textual-based encoding purposes and ResNet to extract visual features from certain products. To help the system understand the context deeper, Multi-Layer Perceptron network is used to predict probability distribution based on 15 intentions from context vector produced by context encoder. On the other hand, cross-entropy loss function is used to optimize the network.

To produce general response, the system uses simple RNN decoder based on general context vector found in multimodal dialogue dataset without basic knowledge. Other than that, MAGIC uses more knowledge-aware RNN decoder to fulfill its main function to produce proper response. After the encode-decode process, the system also embeds recommender which uses RNN model to combine visual features and other additional information.

The experiment phase will then put MAGIC and other baseline models into test. It will focus on best image selection and textual response generation. The proposed MAGIC is also compared side by side with the baselines to test the fluency, relevance, logical consistency, and informativeness factors. Through the experiment, MAGIC excels the baselines, where it can produce recall of 99% and surpasses the Bleu [37] and Nist [38] tests in textual response generation. When comparing the proposed system with each baseline, it wins in almost every testing factor, except when it's beaten by Hierarchical Encoder Decoder (HRED) in fluency factor and informativeness.

[15] also contributes to the usage of Graph Convolutional Network (GCN) to create aspect-guided multimodal dialogue system. The system aims to produce aspect-guided responses, which will provide more interactive and informative responses for better communication between the agent and the user. The research uses text and image modalities and focuses on cuisine domain.

The publication focuses on two things, creation of Multi-domain Multimodal Dialogue (MDMMD) dataset [11] and the implementation of GCN to fulfill multimodal dialogue system guided

with aspect. The model combines information from text and image features to produce responses relevant to the targeted aspect. It has four main parts, utterance encoder that uses bidirectional Gated Recurrent Units (Bi-GRU) [39], image encoder that uses pre-trained VGG-16, context encoder that is present as the final hidden representations, and decoder which uses another GRU for generating the response in a sequential manner.

Evaluation is done to the proposed system in two phases, automatic evaluation that compares the system with baseline models and human evaluation which tests fluency, relevance, aspect appropriateness, and domain consistency scores. In automatic evaluation, the proposed system excels every baseline with small margin. On the other hand, the human evaluation also bested every baseline in all four aspects of evaluation.

[16] discusses about multimodal dialogue system in form of chatbot used for experiential media system. The system is put to give more interaction between human and computer in theater production, which uses text, image, audio, and motion modalities.

The chatbot is built with various components and models, such as Natural Language Processing through Deep Averaging Network (DAN) [40] and mini-Xception model [41] based on Convolutional Neural Network (CNN). Those networks are enhanced with NVIDIA Jetson Nanos [42] to lower the latency of processing. The system is also integrated with various real mechanism in theater to make it more interactive for the audiences.

Evaluation of the system is done through user feedbacks, when it shows the reviews are somehow mixed. While some users found that the experience of storytelling is hindered, some others felt that the system gives brand-new experience of enjoying theater. The performance of the system is also measured through latency, which produces good results in most subsystems, such as chatbot, sentence matching, haiku poem, and emotions. The biggest latency present in crowd indexing subsystem, where it produces latency of almost 1.5 seconds.

[17] focuses on different purpose, where it is used as a new directive for virtual assistants, and it enables the better processing of multimodal inputs than baseline Natural Language Processing models. The publication is specifically directed to furniture and fashion retail domain that uses text, image, and audio modalities, creating Situated Multimodal Conversations (SIMMC) dataset and models.

The publication provides datasets totaling of ~13K human-human dialogues (~169K utterances) collected using a multimodal Wizard-of-Oz (WoZ) [43] setup on two retail domains: furniture grounded in a shared virtual environment that uses text and audio data and fashion grounded in an evolving set of images. SIMMC model consists of four main parts, such as Utterance and History Encoder, Multimodal Fusion, Action Predictor, and Response Generator. Specifically, Utterance and History Encoder has four main encoders like History-Agnostic Encoder (HAE), Hierarchical Recurrent Encoder (HRE) [44], Memory Network (MN) encoder [45], and Transformer-based HAE (T-HAE).

The evaluation phase pits the baseline models, such as TF-IDF and LSTM, with the proposed SIMMC models. There are two tasks done in the phase, first one called API Prediction and the second called Response Generation. API Prediction task is measured via accuracy, perplexity, and attribute accuracy, whereas Response Generation is measured through BLEU, recall@k (k=1,5,10), mean rank, and mean reciprocal rank (MRR) [46]. The HRE model excels in accuracy and perplexity during API Prediction using furniture dataset, while the same model won in all parameters of evaluation for fashion dataset. Furthermore, the HRE model beats every model in every measurement during Response Generation task, both in furniture and fashion dataset.

[18] discusses the importance of multimodality for an intelligent conversational agent, which also presents multimodal dialogue response generation model as the main product of the research. The modalities used in the research is text and image, and the model is used for general domain.

The model, which is called as Divter, consists of two important components such as generator of textual dialogue response and text-to-image translator. The textual dialogue response generator uses sequence-to-sequence model that has 24-layers Transformer with 1024 hidden size and 16 heads. On the other hand, text-to-image translator also uses sequence-to-sequence model with the same architecture as the former component, making both based on Transformer model.

Automatic metrics and human judgements are used as the basis of variants from proposed Divter model, comparing it with baseline models such as pre-trained BERT-base [33], T5-3B [47], SCAN [48], and S2S-TF. The automatic evaluation itself focuses on four aspects: (1) Image Intent Prediction that will predict whether an image should be

produced next for the context and uses F1 metric based on [49]; (2) Generation of Image Description that uses PPL [50], BLEU [37], Rouge [51], and F1 [52] as metric; (3) Quality of Image Generation that uses Frechet Inception Distance (FID) and Inception Score (IS) based on [53]; (4) Text Response Generation which uses the same metrics as (2). In (1), base Divter loses to T5-3B with difference of 2.7 F1 score. For (2), proposed Divter excels other baselines in every metrics. In evaluation (3), Divter bested every baseline, with Divter without joint learning scored best in FID metric and base Divter gives the highest score of IS metric. (4) shows that Divter gives stunning result in every metrics, especially the without joint learning variant and base one.

Meanwhile, the human evaluation is done in two phases, where the first phase compares Divter with SCAN and S2S-TF through four aspects, such as Context Coherence, Text Fluency, Image Quality, and Background Consistency. The second phase puts two versions of Divter (pure text and multimodal) in comparison to DialoGPT [54]. From the results, Divter bested the baselines, except in Image Quality evaluation. On the other hand, Divter with pure text loses to DialoGPT with slight differences, while multimodal Divter beats DialoGPT.

3.4 Discussion

Based on the quantitative and qualitative results, dialogue systems and especially multimodal learning dialogue systems can be implemented in many domains. Referred in Figure 4, we can see that dialogue system find it place in fashion more than the others. Mainly because people can use the utility of dialogue system for fashion, such as information of fashion product or transactions such as buying or checking orders. That's why the second domain position is retail, where the main business flow is much similar to fashion domain. Retail here refers to any kind of retail business, outside of fashion business.

Talking about modalities being used as shown on Figure 3, all the paper used text as modality while in tandem with other modalities. Second modality that being used is image, therefore it can be said that the most popular modality pair is text-image. This may result from the data source abundance. Text and image data are scattered around the internet or easily accessed for labelling or training. Figure 3 also able tell a story that the most popular research in multimodal dialogue system is enhancing the model using text and image-based multimodality, while the rarest modality pair is text-video and text-motion (video stream data). From here, we can say that the modality pair that have the most promising prospect to research is text-video because the scarcity of publication, therefore it is much easier to achieve novelty and state-of-the-art.

Artificial intelligence approach from authors all used deep learning (DL). It also can be inferred that the DL used are mainly found in natural language processing task, such as RNN-based (LSTM, GRU), Transformer, and encoder-decoder (Seq2Seq task), and CNN-based architecture that used for image, such as GCN, VGG, and CNN itself. While all the architecture are different, the general idea is the same for almost all research. Let's say multimodality used is text and image. DL used for text, let's say Bidirectional GRU will extract feature from text to vector (usually the weight of the DL), CNN also will extract feature from the image data (the weight learned or the convolution vector after convolution layer). These two sources of data that has been converted to vector, can be used to further trained with another architecture for specific task (classification or response generation) that included in the dialogue management.

Based on papers being analyzed, there are several challenges found. One, there is loss of information while combining multimodality. Second, there is a need to enhance model capability for handling case such as misclassification for sentiment or image captioning while handling image data. Third, stream of conversation. Current multimodal still focused on single chat from user. However, human sometimes send multiple chat bubble in one context in one stream before expecting reply, and there needs to be a mechanism to handle it. Lastly, multimodal chatbot performance is still being tested in specific domain or task. There's a need to find out the performance of chatbot while handling another task to see if chatbot being built is general enough to be translated to multiple domain tasks.

4. CONCLUSION

Based on analysis of the publications, the research questions explained multimodality in dialogue system can be used for certain domain and not limited to experiment only. The example of implementation from the case is in retail and fashion retail, in which customers often ask using combination of text prompt, and visual image of the object they need to find.

Deep learning architectures are mostly used to handle the multimodalities in the publications, such

as CNN and NLP-based variations like RNN and Transformer. Based on the discussions, the combination of modalities used played a role in the architecture being used. Combination of text and image is the most used modality, and therefore RNN based architecture was used to handle text modality and CNN based architecture was used to handle image modality.

Future studies should address the challenges faced by current multimodal dialogue systems. The primary challenge in multimodal learning is minimizing information loss that occurs when combining multiple modes of data. Additionally, creating a more natural conversational experience for humans through context-aware dialogue systems, possibly utilizing deep learning or software engineering approaches, is also essential. These challenges hopefully provide valuable insights for future researchers in developing multimodal dialogue systems.

Aside from that, future studies for similar review paper can address several aspects not discussed in this paper. First, the dataset can be put as the highlight. Second, another review can be focused on certain techniques, such as discussing state-of-the-art deep learning architecture for multimodal learning for chatbot.

5. AUTHOR CONTRIBUTION

Each author of this paper has contributed during the research. Muhammad Fikri Hasani as the first author contributes the ideation, literature curation and paper writing. The second author, Galih Dea Pratama, is mainly involved in literature curation and paper writing, evolving the findings from the first author. As for Erna Fransisca Angela Sihotang and Franz Adeta Junior as third and fourth author, they contribute through writing and proofreading of this paper.

REFERENCES:

[1] S. E. Brennan, "Conversation and Dialogue," in *Encyclopedia of Mind*, H. Pashler, Ed. SAGE Publications, 2010, pp. 1–9. doi: 10.4324/9780203718179-43.

[2] M. McTear, *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. New York: Morgan & Claypool Publishers, 2021.

[3] H. yeung Shum, X. dong He, and D. Li, "From Eliza to XiaoIce: challenges and opportunities with social chatbots," *Front. Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 10–26, 2018, doi: 10.1631/FITEE.1700826.

[4] Q. Huo, B. Ma, and E.-S. Chng, "Chinese Spoken Language Processing," in *5th International Symposium, ISCSLP 2006*, 2006, p. 724.

[5] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for Audio-Visual Speech Recognition," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2015-Augus, pp. 2130–2134, 2015, doi: 10.1109/ICASSP.2015.7178347.

[6] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[7] W. Zhang, Y. Zhang, L. Ma, J. Guan, and S. Gong, "Multimodal learning for facial expression recognition," *Pattern Recognit.*, vol. 48, no. 10, pp. 3191–3202, 2015, doi: 10.1016/j.patcog.2015.04.012.

[8] X. Sun, X. Long, D. He, S. Wen, and Z. Lian, "VSRNet: End-to-end video segment retrieval with text query," *Pattern Recognit.*, vol. 119, p. 108027, 2021, doi: 10.1016/j.patcog.2021.108027.

[9] I. Papaioannou and O. Lemon, "Combining chat and task-based multimodal dialogue for more engaging HRI: A scalable method using reinforcement learning," *ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 365–366, 2017, doi: 10.1145/3029798.3034820.

[10] L. Liao, Y. Ma, X. He, R. Hong, and T. S. Chua, "Knowledge-aware multimodal dialogue systems," *MM 2018 - Proc. 2018 ACM Multimed. Conf.*, pp. 801–809, 2018, doi: 10.1145/3240508.3240605.

[11] A. Saha, M. M. Khapra, and K. Sankaranarayanan, "Towards building large scale multimodal domain-aware conversation systems," *32nd AAAI Conf. Artif. Intell. AAAI 2018*, pp. 696–704, 2018.

[12] C. Cui, M. Huang, W. Wang, X. S. Xu, X. Song, and L. Nie, "User attention-guided multimodal dialog systems," *SIGIR 2019 - Proc. 42nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 445–454, 2019, doi: 10.1145/3331184.3331226.

[13] L. Tavabi, K. Stefanov, S. N. Gilani, D. Traum, and M. Soleymani, "Multimodal learning for identifying opportunities for empathetic responses," *ICMI 2019 - Proc. 2019 Int. Conf. Multimodal Interact.*, pp. 95–104, 2019, doi:

10.1145/3340555.3353750.

[14] L. Nie, W. Wang, R. Hong, M. Wang, and Q. Tian, "Multimodal dialog system: Generating responses via adaptive decoders," *MM 2019 - Proc. 27th ACM Int. Conf. Multimed.*, pp. 1098–1106, 2019, doi: 10.1145/3343031.3350923.

[15] M. Firdaus, N. Thakur, and A. Ekbal, "MultiDM-GCN: Aspect-guided Response Generation in Multi-domain Multi-modal Dialogue System using Graph Convolutional Network," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2318–2328. doi: 10.18653/v1/2020.findings-emnlp.210.

[16] R. Bhushan *et al.*, "Odo: Design of multimodal chatbot for an experiential media system," *Multimodal Technol. Interact.*, vol. 4, no. 4, pp. 1–16, 2020, doi: 10.3390/mti4040068.

[17] S. Moon *et al.*, "Situated and Interactive Multimodal Conversations," pp. 1103–1121, 2021, doi: 10.18653/v1/2020.coling-main.96.

[18] Q. Sun *et al.*, "Multimodal Dialogue Response Generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, vol. 1, pp. 2854–2866. doi: 10.18653/v1/2022.acl-long.204.

[19] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, 2017, doi: 10.1109/MSP.2017.2738401.

[20] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *Vis. Comput.*, vol. 38, no. 8, pp. 2939–2970, 2022, doi: 10.1007/s00371-021-02166-7.

[21] S. Ali, S. Tanweer, S. Khalid, and N. Rao, "Mel Frequency Cepstral Coefficient: A Review," in *Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2020*, 2021, pp. 1–10. doi: 10.4108/eai.27-2-2020.2303173.

[22] A. Ahmed *et al.*, "A review of mobile chatbot apps for anxiety and depression and their self-care features," *Comput. Methods Programs Biomed. Updat.*, vol. 1, no. March, p. 100012, 2021, doi: 10.1016/j.cmpbup.2021.100012.

[23] T. P. Nagarhalli, V. Vaze, and N. K. Rana, "A Review of Current Trends in the Development of Chatbot Systems," *2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020*, pp. 706–710, 2020, doi: 10.1109/ICACCS48705.2020.9074420.

[24] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. 2015. doi: 10.4018/978-1-60960-165-2.ch004.

[25] G. Skantze and S. Al Moubayed, "IrisTK: a Statechart-based Toolkit for Multi-party Face-to-face Interaction," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, 2012, pp. 69–76. doi: 10.1145/2388676.2388698.

[26] R. J. William, "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning," *Mach. Learn.*, vol. 8, no. 3, pp. 229–256, 1992, doi: 10.1023/A:1022672621406.

[27] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," in *NIPS 2014 Deep Learning and Representation Learning Workshop*, 2014, pp. 1–9. [Online]. Available: http://arxiv.org/abs/1412.3555

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–14. doi: 10.48550/arXiv.1409.1556.

[29] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering Key Laboratory of Complex Systems Modeling and Simulation ," in *International Conference on Computer Vision*, 2017, pp. 1821–1830. doi: 10.1109/ICCV.2017.202.

[30] H. Pham, T. Manzini, P. P. Liang, and B. Poczós, "Seq2Seq2Sentiment: Multimodal Sequence to Sequence Models for Sentiment Analysis," in *Proceedings of the First Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, 2019, pp. 53–63. doi: 10.18653/v1/w18-3308.

[31] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, 2014, pp. 3123–3128.

[32] L. Lu, N. Neale, N. D. Line, and M. Bonn, "Improving Data Quality Using Amazon Mechanical Turk Through Platform Setup," *Cornell Hosp. Q.*, vol. 63, no. 2, pp. 231–246, 2022, doi: 10.1177/19389655211025475.

[33] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, 2019, doi: 10.18653/v1/N19-1423.

[34] F. Eyben *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016, doi: 10.1109/TAFFC.2015.2457417.

[35] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE - The Munich versatile and fast open-source audio feature extractor," *MM'10 - Proc. ACM Multimed. 2010 Int. Conf.*, pp. 1459–1462, 2010, doi: 10.1145/1873951.1874246.

[36] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, 2018, pp. 59–66. doi: 10.1109/FG.2018.00019.

[37] 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 of the Association for Computational Linguistics (ACL)*, 2005, vol. 7, no. 1, pp. 311–318. doi: 10.3917/chev.030.0107.

[38] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *HLT '02: Proceedings of the 2nd International Conference on Human Language Technology Research*, 2002, pp. 138–145. doi: 10.3115/1289189.1289273.

[39] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734. doi: 10.1128/jcm.28.4.828-829.1990.

[40] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daume III, "Deep Unordered Composition Rivals Syntactic Methods for Text Classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015, pp. 1681–1691. doi: 10.3115/v1/P15-1162.

[41] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.

[42] S. Cass, "Nvidia makes it easy to embed AI: The Jetson nano packs a lot of machine-learning power into DIY projects - [Hands on]," *IEEE Spectr.*, vol. 57, no. 7, pp. 14–16, 2020, doi: 10.1109/MSPEC.2020.9126102.

[43] P. Budzianowski *et al.*, "MultiWoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2018, pp. 5016–5026. doi: 10.18653/v1/d18-1547.

[44] C. Wang and H. Jiang, "The lower the simpler: Simplifying hierarchical recurrent models," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, vol. 1, pp. 4005–4009. doi: 10.18653/v1/n19-1402.

[45] A. Carta, A. Sperduti, and D. Bacciu, "Encoding-based memory for recurrent neural networks," *Neurocomputing*, vol. 456, pp. 407–420, 2021, doi: 10.1016/j.neucom.2021.04.051.

[46] N. Craswell, "Mean Reciprocal Rank," in *Encyclopedia of Database Systems*, Springer Science+Business Media, LLC, 2009, p. 1703. doi: 10.1007/978-1-4899-7993-3_228-2.

[47] C. Raffel *et al.*, "T5: Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, 2020, doi: 10.48550/arXiv.1910.10683.

[48] K. H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked Cross Attention for Image-Text Matching," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11208 LNCS, pp. 212–228, 2018, doi: 10.1007/978-3-030-01225-0_13.

[49] X. Zang, L. Liu, M. Wang, Y. Song, H. Zhang, and J. Chen, "PhotoChat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling," in *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Natural Language Processing, Proceedings of the Conference, 2021, pp. 6142–6152. doi: 10.18653/v1/2021.acl-long.479.

[50] S. Kulkarni *et al.*, "PPL Bench: Evaluation Framework For Probabilistic Programming Languages," in *Facebook AI*, 2020, pp. 1–9. [Online]. Available: http://arxiv.org/abs/2010.08886

[51] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, 2004, vol. 34, no. 12, pp. 74–81. doi: 10.1253/jcj.34.1213.

[52] D. J. Hand, P. Christen, and N. Kirielle, "F*: an interpretable transformation of the F-measure," *Mach. Learn.*, vol. 110, no. 3, pp. 451–456, 2021, doi: 10.1007/s10994-021-05964-1.

[53] A. Ramesh *et al.*, "Zero-Shot Text-to-Image Generation," in *Proceedins of the 37th International Conference on Machine Learning*, 2020, vol. 139, pp. 8821–8831. doi: 10.48550/arXiv.2102.12092.

[54] Y. Zhang *et al.*, "DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 270–278. doi: 10.18653/v1/2020.acl-demos.30.