Towards Evaluation of Multi-party Dialogue Systems

Khyati Mahajan
UNC Charlotte
kmahaja2@uncc.edu

Sashank Santhanam
UNC Charlotte
ssantha1@uncc.edu

Samira Shaikh
UNC Charlotte
sshaikh2@uncc.edu

Abstract

Recent research in the field of conversational AI has emphasized the need for standardization of the metrics used in evaluation. In this work, we focus on evaluation methods used for multi-party dialogue systems. We present an expanded taxonomy focusing on multi-party dialogue based on the need for evaluation dimensions that address challenges associated with the presence of multiple participants. We also survey the evaluation metrics utilized in current multi-party dialogue research, and present our findings with regards to inconsistencies within existing work. Furthermore, we discuss the subsequent need to have more consistent evaluation methodologies and benchmarks. We motivate how consistency will contribute towards a better understanding of progress in the field of multi-party dialogue systems.

1 Introduction

There has been much discussion lately in the field of Natural Language Generation (NLG) focusing the need for evaluation benchmarks and standards, as evidenced by the prolific literature focusing on the issues surrounding human evaluation (Howcroft et al., 2020; Belz et al., 2020; Clark et al., 2021; Hämäläinen and Alnajjar, 2021; van der Lee et al., 2021), as well as recently proposed benchmarks (Gehrmann et al., 2021; Khashabi et al., 2021; Liu et al., 2021; Mille et al., 2021). These are important and necessary debates - however, work has focused mainly on two-party dialogue systems. Multi-party dialogue (MPD) systems, which aim to model conversations between groups (>2 participants) have received less attention, especially in the area of evaluation. Additionally, while there is existing work towards modeling MPD, evaluation strategies are not consistent across existing literature, making it harder to place the progress of the field. In the context of multi-party dialogue (MPD), we discuss both automatic and human evaluation metrics used for evaluating the three main sub-tasks described in detail in Section 2.

Thus, in this paper, we foreground the challenges faced by the presence of multiple participants in a conversation, and how this property affects the evaluation of systems which aim to model group conversations. We present an expansion to the integrated taxonomy (Table 1) proposed by Higashinaka et al. (2021). We use (Higashinaka et al., 2021) as a baseline owing to their extensive study of data-driven and theory-driven error analysis, and the empirical validation of the proposed integrated taxonomy drawn from both these error analysis paradigms (Higashinaka et al., 2015, 2019). However, we find that the integrated taxonomy does not account specifically for the challenges faced by MPD modeling systems, and thus we propose an expansion specifically keeping these challenges in mind.

We then draw attention to specific shortcomings of evaluation metrics utilized in existing work, such as the lack of consistent reporting within similar evaluation metrics (such as $Recall_n@k$), and the lack of public availability of the proposed methodologies, making it harder to place the progress of the field even if an evaluation benchmark is proposed. Thus, there is a severe gap towards a consistent evaluation framework in Multi-Party Dialogue (MPD) which needs to be addressed. Our main contributions include:

1. We propose an expanded taxonomy focusing on the specific challenges introduced by multi-party dialogue, or group conversations (such as the need to maintain speaker-specific context and recognize the proper addressees), and provide examples for each newly introduced category.
2. We synthesize evaluation measures currently used in MPD research, and relate them to the expanded taxonomy introduced.

To study evaluation metrics in existing work, we surveyed over 338 research papers in the field

of MPD (Github link[1]). We obtained the initial pool based on a keyword search for variations of "multi-party dialogue", with 258 papers focused on work in English, and most of them published at *CL, LREC, and related conferences. The papers included in this article include only those which (a) focus on the English language, (b) include *multiple speakers in the majority of conversations*, and (c) which focus on text-based approaches (thus excluding research which uses multi-modal cues towards the aforementioned sub-tasks). This paper does NOT focus on multilingual corpora, or approaches which solely focus on concepts such as speech recognition or synthesis. We also limit discussion to research published within the past decade for a more relevant understanding of the current progress in MPD modeling, and aim to build upon limited prior work in MPD evaluation, which we discuss further in Section 3.3. With this filtering, we find a total of 15 papers whose aim is one or more of the sub-tasks of Speaker Identification, Addressee Recognition and Response Selection/Generation.

We first present an expanded taxonomy with error reporting drawn from the challenges presented in MPD (Traum, 2003) and (Branigan, 2006), adding categories specifically relevant and important towards MPD evaluation to the taxonomy presented by (Higashinaka et al., 2021). Next, we observe the evaluation metrics utilized in existing work in Section 4, whose error reporting strategies we relate to the proposed expanded taxonomy (Table 1) and note the lack of evaluation for important categories.

2 Overview: Challenges in MPD Evaluation

Evaluation for MPD has often focused on specific sub-tasks that are integral to the working of any conversational system participating in a group conversation. A lot of existing research focuses either on one or more of the sub-tasks: 1) *Speaker Identification* which concerns with how an MPD chatbot is able to track the speakers for each utterance as well as predict who the next speaker could be, 2) *Response Selection* which concerns with the selecting the correct next utterance from a set of choices or *Response Generation* which concerns with generating the next utterance from scratch given the context of the conversation, and 3) *Addressee Recognition* which concerns with being able to find the addressee(s) for the next utterance. All Speaker Identification, Response Selection and Addressee Recognition can be framed as classification tasks (evaluation would need to check whether the correct participant(s) were chosen from the group), whereas Response Generation requires evaluation metrics similar to response generation for two-party dialogue. Recently, systems trained towards jointly modeling one or more of the above tasks have been proposed, however as mentioned before the evaluation strategies lack consistency, and require further thought. While evaluating the classification could provide important indicators of the performance of the dialogue model itself, robust evaluation is needed to understand how well the system would perform in a real life setting. Some leading questions which venture into this challenge faced by MPD systems include:

1. Is the system able to maintain long-term context from all participants in the group? Is the selected/generated response relevant to the prompt and the context of the MPD participants while being grounded in the ongoing conversation? (Pointing to the need for managing speaker information)
2. Is the system able to respond to every participant's prompt, whether implicitly or explicitly mentioned? Conversely, is it able to learn to not respond (yet remember for context) to the relevant utterances? (Pointing to the need for managing addressee information)
3. Does the system contribute towards making the conversation successful? This success could be attributed to either making the conversation easier for the group by providing information when needed, measuring the interactivity introduced by the presence of MPD dialogue systems, and helping the group achieve the objective which led to the conversation. (Pointing to the need for evaluating appropriate timing and thread management abilities)

Keeping these challenges in mind, we present an expansion of error reporting categories which would be the first step towards accounting for the performance of a system which operated in the multi-party conversation. We briefly summarize the error reporting taxonomy for dialogue agents presented by (Higashinaka et al., 2021), and then discuss how the expansion accounts for errors specific to multi-party dialogue in Section 3.

[1] https://github.com/khyatimahajan/mpd-references

3 Expanded Taxonomy of Errors for Multi-Party Dialogue

Recently, Higashinaka et al. (2021) introduced an integrated taxonomy of errors in chat-oriented dialogue systems (Table 1). Their work focuses on responses given by a chatbot (conversing with one user) which could cause a breakdown in the conversation (Higashinaka et al., 2015). They empirically validate the resulting integrated taxonomy by asking the same annotators who annotated breakdowns to rate the breakdown for each error category (Higashinaka et al., 2019). While the resulting taxonomy is quite exhaustive, we find that it does not account for challenges specific to MPD, such as the need to know whether the user is able to attribute utterances to each participant correctly. Thus, we expand the taxonomy presented by Higashinaka et al. (2021), focusing specifically on how the presence of multiple participants affects the possible errors which occur in a group conversation.

We elaborate on each error from a MPD point of view, providing examples demonstrating the need for further research. We draw from perspectives presented by Traum (2003) and Branigan (2006), relating the challenges presented for realizing the differences between two-party and multi-party dialogue evaluation. Specifically, we expand on Response-level errors (I18 and I19) which are affected by the speaker and addressee(s), and add a new dimension with Participant-level errors (I20, I21, and I22), which showcase errors from a participant point of view. We include all these, italicized and highlighted, in Table 1, and include details for each error with examples in this section.

3.1 Response-level Errors

This subsection focuses on response level errors, which apply to the semantic meaning of the complex information contained in responses in MPD.

3.1.1 Violation of Content

We maintain the definition presented in Higashinaka et al. (2021), and thus violation of content errors indicate that even though the surface form of the utterance may be appropriate, it could lead to confusion during the conversation.

(I18) Forgot speaker: The utterances made by a specific user are often ignored. This relates specifically to the challenge of Speaker Identification (Traum, 2003), and is an extremely important property for maintaining context in MPD, since it could create confusion for the system downstream if the utterance is referred to again and the user feels ignored. In the example below, the System (S) forgets the utterance made by User 1 (U1) in the beginning of the conversation. Failure to remember the correct speaker for an utterance could lead to critical downstream errors.

(1) U1: We need to consider factors A and B for making a decision in case X.
U2: Factor C would also be interesting and important to consider along with A and B.
S: U2 mentions factor C will be important to take into consideration for case X.

(I19) Forgot addressee(s): The system forgets to mention the correct addressee(s), relating to the Addressee Recognition challenge (Traum, 2003), and specifically forgets one or more addressees it should have mentioned. If the system was prompted by multiple speakers on a similar topic, but the system responded only to some, this counts as an error since it could make forgotten participants feel alienated from the conversation. In the example below the System (S) forgets to address User 2 (U2), although it should have included both U1 and U2.

(2) U1: We need to consider factors A and B for making a decision in case X.
U2: Factor C would also be interesting and important to consider along with A and B.
S: Thanks for bringing factors A, B and C up for case X, U1.

3.2 Participant-level Errors

We introduce a new broad category of errors towards MPD evaluation called Participant-level errors. The categories of errors introduced in this section stem from the inherently entangled nature of responses in MPD - a response contains not only the content and context of the utterance but also information for who the speaker(s) and addressee(s) of the utterance are. We elaborate further in this subsection.

3.2.1 Violation of Form

The violation of form at the participant level indicates that the system involves the wrong partici-

	Violation of Form	Violation of Content
Utterance	(I1) Uninterpretable (I2) Grammatical error	(I3) Semantic error (I4) Wrong information
Response	(I5) Ignore question (I6) Ignore request (I7) Ignore proposal (I8) Ignore greeting	(I9) Ignore expectation *(I18) Forgot speaker* *(I19) Forgot addressee(s)*
Context	(I10) Unclear intention (I11) Topic transition error (I12) Lack of information	(I13) Self-contradiction (I14) Contradiction (I15) Repetition
Society	(I16) Lack of sociality	(I17) Lack of common sense
Participant	*(I20) Wrong speaker* *(I21) Wrong addressee(s)*	*(I22) Wrong thread response* *(I23) Inappropriately timed initiative*

Table 1: Integrated taxonomy for errors in chat-oriented dialogue systems by Higashinaka et al. (2021). We expand the taxonomy to include errors specific to MPD - extensions are italicized and highlighted in grey. The numbering is assigned serially and used in text to refer to discussions surrounding the specific error.

pant(s) in the current state, leading to confusion.

(I20) Wrong speaker: The system credits the wrong speaker for an utterance, also relating to Speaker Identification (Traum, 2003), but specifically mistaking one user for another. This can lead to very important mistakes, since different participants could express differing views which could be contrasting in nature. In the example, the System (S) credits the utterance to U1 instead of U2, which diminishes U2's contributions.

(3) U1: We need to consider factors A and B for making a decision in case X.
U2: Factor C would also be interesting and important to consider along with A and B.
S: U1 mentions factors C will be taken into consideration for case X.

(I21) Wrong addressee(s): The system mentions the wrong addressee when responding to a prompt, leading to miscommunication. Relating to the challenge of Addressee Recognition (Traum, 2003) as well, mentioning the wrong addressee could lead to a breakdown since this means the system was unable to maintain contextual information which it should have in the conversation. This could lead to a breakdown especially if the addressee who is mentioned does not wish to be mentioned/take part in the current conversation. In the example the System (S) mentions the wrong addressee U1 instead of U2.

(4) U1: We need to consider factors A and B for making a decision in case X.
U2: Factor C would also be interesting and important to consider along with A and B.
S: Interesting insight on factor C, U1.

3.2.2 Violation of Content

A violation of content means that the system makes an error which might seem appropriate in the conversation, but is incorrectly placed, therefore leading to confusion.

(I22) Wrong thread response: MPD can have communication ongoing in multiple threads within the same conversation (Thread/Conversation Management in Traum (2003)). If the system talks about the wrong topic when participating in a different thread, this could lead to confusion and interrupt the desired flow of conversation. In the example below there exist two threads of conversation: one whose topic is sports (U1, U2, U3) and the other whose topic is movies (U4, U5). There are sub-groups of users within the conversation who are participating in different threads, and the System (S) makes an error by mentioning a topic in the wrong thread and sub-group.

(5) U1: This football season has been going great!
U2: I agree, for most teams anyway. Which one is your favorite?
U3: I prefer soccer instead. Anyone here a soccer fan?
U4: I don't really pay much attention to sports. My main hobby is movies!
U5: Yeah, and Knives Out was a great one!
S: I agree U5! The Rams are doing so well this year!

(I23) Inappropriately timed initiative: MPD systems need to figure out when to take the floor in a conversation without causing an abrupt change in the conversation. Secondly, while they could be prompted to speak, it is also important to take the lead to get a conversation started since participants could be yielding the floor to other participants. This relates specifically to the challenge of Initiative Management (Traum, 2003), since the system needs to learn when to take initiative and introduce new topics without which the conversation might come to a halt. In the example the conversation flow is smoothly going on for fiction (U1, U2, and U3), but the System (S) interrupts with a contrasting topic.

(6) U1: I love documentaries and it has been great seeing so many come out in recent years.
U2: They do seem informative. I'm particularly interested in performative documentaries, they seem more personal.
U3: I also enjoy performative documentaries, like Supersize Me. Have you watched it U2?
S: Does anyone here like fiction?

3.3 Discussion

In recent research, we observe the prevalence of the aforementioned errors within MPD research. We notice how the need to account for multiple participants affects the response selection/generation pipeline for systems modeling MPD, and thus discuss error reporting in existing research in the section to highlight our observations. Since there is limited existing research in the field of MPD response selection/generation, we reserve experimental validation of the expanded taxonomy for future work. However, one research paper of particular interest to this discussion is Traum et al. (2004, 2006). They are the first to propose evaluations for interactions between virtual multi-party systems and users: 1) User Satisfaction via rated survey questions (accounting for Response-level errors I5-I9, I18, & I19, Society-level errors I16 & I17, and Participant-level errors I20-I23); 2) Intended Task Completion via predefined task success and inter-rater reliability (accounting for I4 and I12); 3) Recognition Rate via classification F-score (accounting for I19 and I21); and 4) Response Appropriateness via a custom defined scale (accounting for Context-level errors I10-I15 and I22-I23). This paper presents a great first step in evaluations for MPD systems which interact in the real world, and we hope to draw from such studies for future work (Section 5).

4 Inconsistency of Evaluation Metrics in Existing Research

Papers focusing on specific tasks within MPD have been observed to employ mostly automatic evaluation measures, with very few including human evaluations. Repeated observations within mainly two-party NLG evaluation have shown that automatic and human evaluations do not correlate well (Belz and Reiter, 2006; Reiter and Belz, 2009; Novikova et al., 2017; Santhanam and Shaikh, 2019; Santhanam et al., 2020), leading to arguments about automatic evaluations being unsuitable for assessing linguistic properties (Scott and Moore, 2007). Owing to these, van der Lee et al. (2021) survey the field and present arguments towards how the inclusion of human evaluations gives a more complete picture of the performance of systems whose main purpose is to participate in human conversations. With research in MPD severely lacking this reporting, it is difficult to place the success of systems which have been proposed to perform well in real-world scenarios. Moreover, owing to the complex nature of group conversations, this lack of reporting exacerbates the effect towards understanding the progress of MPD. Thus, this section illustrates research focusing on the core task of MPD modeling, drawing attention to the evaluation strategies followed by them. We provide a brief synthesis on currently formalized tasks, and relate the errors from the expanded taxonomy (Table 1).

4.1 Evaluation Metrics in Sub-tasks

We organize this section by including sub-task focused discussions to get a clearer idea of the evaluations reported for each sub-task, and how these relate to the expanded taxonomy of errors. We start with the joint formalized task introduced by Ouchi and Tsuboi (2016) - Addressee Recognition and Response Selection, Section 4.1.1 - which is the one of the most consistent research area with regards to error reporting. We then focus specifically on Response Selection in Section 4.1.2, then moving to Response Generation in Section 4.1.3, and lastly Speaker Identification in Section 4.1.4. Lastly, we wrap up by discussing the overall takeaways in Section 4.2.

4.1.1 Addressee Recognition and Response Selection

Ouchi and Tsuboi (2016) first formalized the task of Addressee and Response Selection (ARS) as a joint task, with the input consisting of the (responding agent, context, candidate responses) and the output consisting of the (addressee, response). They evaluate accuracy of their Dual Encoder based RNN model (called Dynamic RNN) over addressee selection (ADR) limited to the addressee of the last utterance, and response selection (RES), as well as a mix of both with addressee-response pair selection (ADR-RES). Zhang et al. (2018b) utilize the same framework for their evaluation, improving their model by including speaker embeddings, called SI-RNN. Le et al. (2019) focus on identifying addressees within the same task, but for all utterances, also reporting accuracy (with n-grams, n=5, 10, and 15) and $Precision@1$. They additionaally involve limited human evaluations, comparing the consistency between human and model outputs, along with significance tests. Gu et al. (2021) introduce MPC-BERT, introducing pre-trained models and fine-tuning for downstream tasks within MPD systems. They follow the same evaluation strategy established by Zhang et al. (2018b).

Thus most papers in this line of research focus on measuring errors towards I18, I19, I20, and I21, with some including human evaluations for a subjective understanding of the success of their models.

4.1.2 Response Selection

Wang et al. (2020) and Gu et al. (2020) focus on response selection as a classification task, with the former proposing Topic-BERT and the latter proposing SA-BERT, two very similar frameworks. The main difference between the approaches is that Topic-BERT build topic-sentence pairs as input, while SA-BERT instead build speaker embeddings as input - both utilize the basic embeddings for BERT pre-training (segment, position, and token embeddings). Both report recall as defined by the response selection task proposed in DSTC-8[2] (Kim et al., 2019) sub-tasks 1 and 2, using $Recall_n@k$ for reporting recall for matching n available candidates to k best-matched responses (the official leaderboard utilizes MRR and $Recall@10$ with $n = 100$). However, their is still no overlap in the evaluation results for response selection on DSTC-8 reported by both papers, with Wang et al. (2020) reporting $Recall@1, Recall@5, Recall@10$ and MRR (assuming all these are reported for $n = 100$ - only mentioned in Section 4.1 of the paper) which details the pre-training for Topic-BERT; and Gu et al. (2020) reporting only $Recall_2@1$, $Recall_{10}@1$, $Recall_{10}@2$, and $Recall_{10}@5$, although they do mention $Recall_{100}@1$ once in Section 1. Both papers do however mention $Recall_{10}@1$, $Recall_{10}@2$, and $Recall_{10}@5$ for the Ubuntu V1 corpus, which does allow partial comparison for results. Additionally, Wang et al. (2020) also report BLEU (Papineni et al., 2002) and $Precision@n$ (n=1, 2, 3, 4) scores for incorrectly selected responses, checking the relevance of the Topic-BERT retrieved results.

Jia et al. (2020) also tackle response selection, with more focus on dialogue dependency to organize the conversation into contextually aware threads, proposing the Thread-Encoder model (built with Transformer based BERT-base, same as Wang et al. (2020) and Gu et al. (2020). They utilize similar data (Ubuntu V2 and DSTC-8), and report evaluations for response selection, reporting $hits@k$ (similar to $Recall@k$ as per the paper and ParlAI[3] metrics, k = 1, 2, 5), and MRR for Ubuntu V2 and $hits@k$ (similar to $Recall@k$, k = 1, 5, 10, 50) and MRR for DSTC-8 (with n=100).

Since most papers working on response selection essentially work on a classification task, naturally the reporting is limited to classification metrics. However, even research conducted around the same time, over the same task, reports different metrics with only partial overlaps which could be used to partially compare performance. However, we do

[2] https://github.com/dstc8-track2/NOESIS-II/

[3] https://parl.ai/docs/tutorial_metrics.html

not consider this evaluation to count towards any of the expanded taxonomy since none of the classification metrics specifically look for performance consciously in any of the dimensions included in the taxonomy - they just measure whether the system was able to choose the next utterance given the previous utterances and a possible list of the right next utterance. Breaking down the evaluation into components presented in the taxonomy, i.e. measuring success keeping in mind the speaker, addressee, and content & context of the selected utterance would help understand the performance in a more robust manner - like Wang et al. (2020) report BLEU for the incorrect responses.

4.1.3 Response Generation

Zhang et al. (2018a), Liu et al. (2019) and Hu et al. (2019) tackle response generation, taking in previous utterances as input and the next utterance as output (Liu et al. (2019) also specifically include the responding speaker and target addressee in the inputs and outputs). Zhang et al. (2018a) report the BLEU-n (n based on n-grams, n = 1, 2, 3, 4) and METEOR (Banerjee and Lavie, 2005) scores (mentioning that the evaluation could be supplemented); Liu et al. (2019) report BLEU, ROUGE (Lin, 2004), noun mentions, and length of generated response, along with limited human evaluations for fluency, consistency, and informativeness; and Hu et al. (2019) report BLEU-n (n = 1, 2, 3, 4), METEOR, ROUGE-L (L for longest common subsequence), along with human evaluations for fluency, grammaticality, and rationality. Qiu et al. (2020) focus on the dialogue thread structures which are utilized in Hu et al. (2019), utilizing structured attention with Variational RNN, reporting the same automatic metrics BLEU-n (n = 1, 2, 3, 4), METEOR, ROUGE-L (L for longest common subsequence). They also find that they are able to perform speaker identification and addressee recognition without specifically training towards these tasks.

Yang et al. (2019) tackle response generation along with speaker identification, proposing LSTMs to build an encoder, a contextual RNN, a speaker encoder, and a decoder, called Multi-role Interposition Dialogue System (MIDS). They report accuracy for speaker identification; and perplexity and loss for response generation.

Even with the majority of papers reporting the basic automated evaluation metrics most common for generation (BLEU, METEOR, ROUGE (van der Lee et al., 2021)), these are not always reported. Moreover, Liu et al. (2016) also show that the aforementioned metrics show either weak or no correlation with human judgements. Human evaluations are also limited, although they do cover some of the most reported metrics (fluency, consistency, informativeness, grammaticality, rationality (van der Lee et al., 2021)). Most research thus cover major aspects of the expanded taxonomy, namely Utterance-level I1-I4, Context-level I10-I15, and Society-level I17. Some papers also report speaker identification and addressee recognition, accounting for I18, I19, and Participant-level I20-I23 with thread management.

4.1.4 Speaker Identification

Ma et al. (2017) and de Bayser et al. (2019) focus on speaker identification, with the former using RNN and CNN to identify speakers with a sitcom dataset, and the latter using MLE, SVM, CNN, and LSTM architectures to model sitcom, finch and multibotwoz datasets. While both utilize a variety of features (such as surrounding utterance concatenation, agent and content information) with the models to improve predictions, Ma et al. (2017) report accuracy and F1 (+ F1 towards each participant and a confusion matrix to better analyze wrong predictions), and de Bayser et al. (2019) report accuracy. They extend their work in de Bayser et al. (2020) by integrating MLE, CNN, and FSA-based architectures, for multibotwoz, reporting accuracy.

Classification for speaker identification does help response selection and generation, counting towards errors I18 and I20 from the expanded taxonomy. However, it would be helpful to include more classification metrics (like Ma et al. (2017) who report the confusion matrix) to allow for more robust evaluations.

4.2 Discussion

It is imperative to observe the various kinds of evaluation metrics which have been used to evaluate different tasks within MPD. Most metrics reported are not consistent across the main task they focus on, sometimes even when they report performance on a shared task such as DSTC-8 (Kim et al., 2019). It is important to note that these inconsistencies lead to confusion when it comes to looking for the current state-of-the-art systems, as well as for making important performance comparisons such as significance testing. Additionally, we find that there is a 50-50% (8:7) division of the code in the

papers being publicly available (if we include broken links, the unavailability goes up, but we count these as attempts to provide reproducible methods). This means that even with re-evaluation given a benchmark, there is a possibility that comparison across existing research will not be able to provide a full picture of the progress in each sub-task.

All these issues draw attention to the need for more shared tasks and robust benchmarks which report errors in a manner fitting the proposed taxonomy. We postulate that this would allow better comparisons across tasks, and overall performance towards building systems able to participate in MPD - although we reserve the evaluation of our proposed extensions to the taxonomy itself for future work. We aim to follow methods similar to the ones described by (Higashinaka et al., 2019) to maintain the standards they set up for validation of error analysis.

5 Conclusion and Future Work

We have presented an expansion - which focuses specifically on errors important in multi-party dialogue - to the integrated taxonomy of errors proposed by Higashinaka et al. (2021). We include examples for each newly introduced error in Section 3, and relate the errors to the challenges detailed by Traum (2003). We then present inconsistencies in the evaluation strategies reported in existing research (Section 4), organized by the sub-tasks they focus on. We observe the difficulty in comparisons across the proposed methods owing to inconsistencies in error analysis. We also relate the reported errors to the expanded taxonomy, drawing parallels for an overall comparison.

We observe how the challenges introduced by the presence of multiple participants affect the need for more robust evaluations (Section 3.3) which are capable of reporting how well the approach performs, and find that (Traum et al., 2004, 2006) provide a great discussion surrounding these errors, albeit more focused on interactions between virtual systems and users. We also find that even with defined tasks, inconsistencies could arise in reporting errors (Section 4.2), leading to confusion when placing the progress of research in MPD.

We note that while our presented taxonomy is relevant to the errors reported in current literature, there is a need to evaluate their effectiveness empirically, which is the main limitation for this paper and proposed future work. Another big limitation of this work which is also a part of proposed future work is the formalization of the proposed expanded errors specific to MPD from this paper (Table 1), and the validation of the formalization towards a proposed benchmark. The first shared task DSTC-8 (Kim et al., 2019) focused on the response selection sub-task, however there is the need for future shared tasks which account for all three sub-tasks (speaker identification, response selection/generation and addressee recognition), and related sub-tasks (such as disentanglement, thread management, and coreference resolution).

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Anja Belz, Simon Mille, and David M Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *11th conference of the european chapter of the association for computational linguistics*, pages 313–320.

Holly P. Branigan. 2006. Perspectives on multi-party dialogue. *Research on Language and Computation*, 4:153–177.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human'is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296.

Maíra Gatti de Bayser, P. Cavalin, C. Pinhanez, and Bianca Zadrozny. 2019. Learning multi-party turn-taking models from dialogue logs. *ArXiv*, abs/1907.02090.

Maıra Gatti de Bayser, Melina Alberio Guerra, Paulo Cavalin, and Claudio Pinhanez. 2020. A hybrid solution to learn turn-taking in multi-party service-based chat groups. *Interactions*, 10(4):2.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D

Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2041–2044.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. Mpc-bert: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3682–3692.

Mika Hämäläinen and Khalid Alnajjar. 2021. The great misalignment problem in human evaluation of nlp methods. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 69–74.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2019. Improving taxonomy of errors in chat-oriented dialogue systems. In *9th International Workshop on Spoken Dialogue System Technology*, pages 331–343. Springer.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. Integrated taxonomy of errors in chat-oriented dialogue systems. In *SIGDIAL*.

Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 87–95.

David M Howcroft, Anja Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. Gsn: A graph-structured network for multi-party dialogues. In *IJCAI*.

Qi Jia, Yizhu Liu, Siyu Ren, Kenny Zhu, and Haifeng Tang. 2020. Multi-turn response selection using dialogue dependency relations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1911–1920.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. 2021. Genie: A leaderboard for human-in-the-loop evaluation of text generation. *arXiv preprint arXiv:2101.06561*.

Seokhwan Kim, Michel Galley, Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, Mahmoud Adada, et al. 2019. The eighth dialog system technology challenge. *arXiv preprint arXiv:1911.06394*.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and R. Yan. 2019. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. In *EMNLP/IJCNLP*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Cao Liu, Kang Liu, Shizhu He, Zaiqing Nie, and Jun Zhao. 2019. Incorporating interlocutor-aware context into response generation on multi-party chatbots. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 718–727.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. 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 Language Processing*, pages 2122–2132.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021. Explainaboard: An explainable leaderboard for nlp. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289.

Kaixin Ma, Catherine Xiao, and Jinho D. Choi. 2017. Text-based speaker identification on multiparty dialogues using multi-document convolutional neural networks. In *ACL*.

Simon Mille, Kaustubh Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic construction of evaluation suites for natural language generation datasets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *EMNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Liang Qiu, Yizhou Zhao, Weiyan Shi, Yuan Liang, Feng Shi, Tao Yuan, Zhou Yu, and Song-chun Zhu. 2020. Structured attention for unsupervised dialogue structure induction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1889–1899.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Sashank Santhanam, Alireza Karduni, and Samira Shaikh. 2020. Studying the effects of cognitive biases in evaluation of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Donia Scott and Johanna Moore. 2007. An nlg evaluation competition? eight reasons to be cautious. In *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, pages 22–23.

D. Traum. 2003. Issues in multiparty dialogues. In *Workshop on Agent Communication Languages*.

David R Traum, Susan Robinson, and Jens Stephan. 2004. Evaluation of multi-party virtual reality dialogue interaction. In *LREC*.

David R Traum, Susan Robinson, and Jens Stephan. 2006. Evaluation of multi-party reality dialogue interaction. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE ....

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Weishi Wang, Steven CH Hoi, and Shafiq Joty. 2020. Response selection for multi-party conversations with dynamic topic tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6581–6591.

Qichuan Yang, Z. He, Zhiqiang Zhan, Jianyu Zhao, Y. Zhang, and C. Hu. 2019. Mids: End-to-end personalized response generation in untrimmed multi-role dialogue*. *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Haisong Zhang, Zhangming Chan, Yan Song, Dongyan Zhao, and R. Yan. 2018a. When less is more: Using less context information to generate better utterances in group conversations. In *NLPCC*.

Rui Zhang, H. Lee, L. Polymenakos, and Dragomir R. Radev. 2018b. Addressee and response selection in multi-party conversations with speaker interaction rnns. In *AAAI*.