Towards Evaluation of Multi-party Dialogue Systems

Authors:

Khyati Mahajan ✉ kmahaja2@uncc.edu
Sashank Santhanam
Samira Shaikh

Main Motivation

- Prolific research in NLG evaluation
  - Multiple taxonomies presented[1, 2, 3, 4]
  - Studies towards importance of automatic and human metrics[5, 6, 7, 8]
  - + Confusion surrounding inconsistent evaluation methods used[9]
- However, not much work towards evaluation specifically for Multi-party Conversation (MPC) evaluation
  - = Need for discussing MPC specific challenges and needs

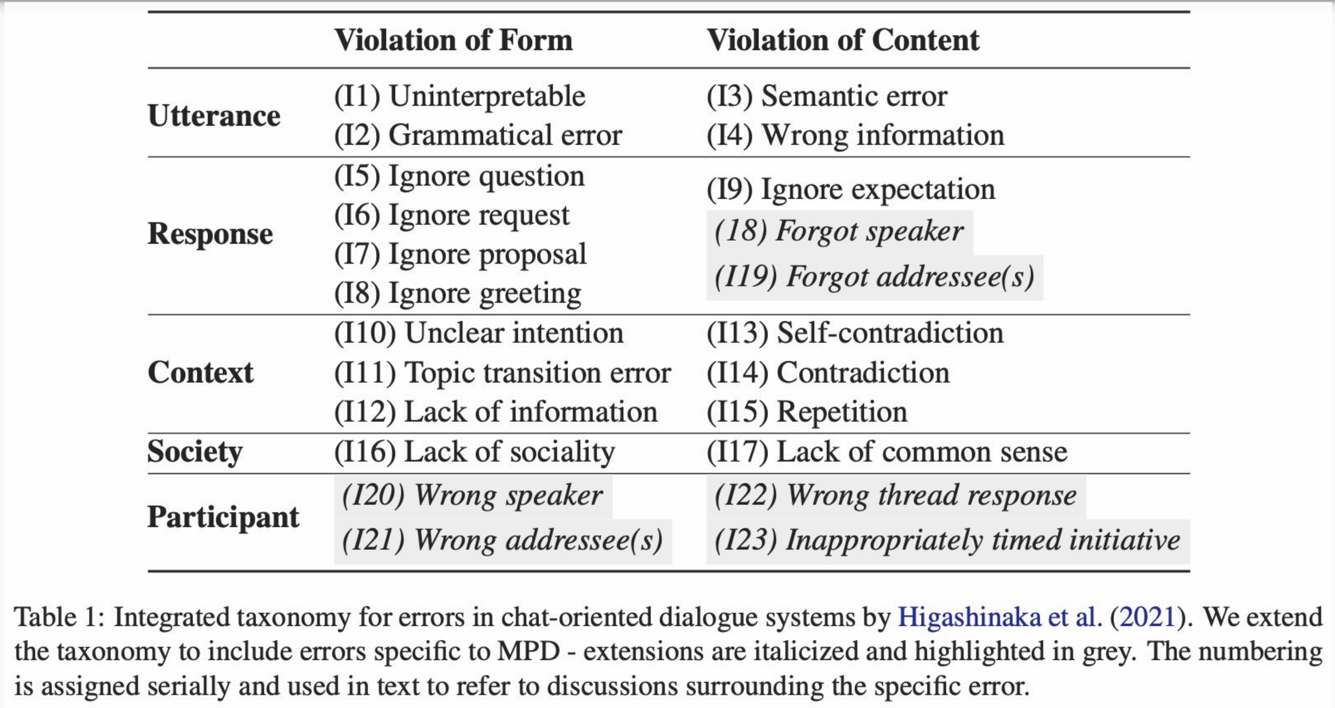
MPC Challenges

- The presence of multiple participants introduces new and interesting challenges from a dialogue modeling perspective
    - Participant roles - need to maintain speaker-specific and addressee-specific information jointly with dialogue modeling
    - Conversation structure - more graph-like than sequential
    - Threads within conversation - multiple topic threads could co-exist within sub-groups

Contributions

- 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
- Synthesize evaluation measures utilized in existing MPC research, and relate them to the expanded taxonomy introduced
  - Report important inconsistencies in current research

Expanded Taxonomy



	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 *(18) 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 extend 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.

Example Snippets

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.

Forgot Speaker S: U2 mentions factor C will be important to take into consideration for case X.

Forgot Addressee S: Thanks for bringing factors A, B and C up for case X, U1.

Wrong speaker S: U1 mentions factor C will be taken into consideration for case X.

Wrong addressee S: Interesting insight on factor C U1.

Example Snippets (Contd)

Wrong thread response

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!

Inappropriately timed initiative

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?

Survey of existing literature

- Surveyed evaluation metrics utilized in past MPC modeling research tackling the tasks:
    - Speaker Identification
    - Response Selection or Generation
    - Addressee Recognition
    - ~15 papers total

Survey findings

- Most common metrics reported
  - BLEU
  - ROUGE
  - Classification reports
  - Yet most are reported on different properties (ex *n* is different for *n*-gram comparisons)
- We report inconsistencies across all literature

Need for better error reporting

- Most metrics reported are not consistent across the main task they focus
  - Even when reporting on shared task (DSTC-8 Track 2 NOESIS challenge)
  - Cannot compare across SOTA claims
- Not all models are publicly released
  - Difficult to re-evaluate even with possible new benchmarks

Next steps

- Formalize errors towards MPC modeling benchmark
    - Introduce automatic evaluation metrics
        - Classification reports for Speaker Identification and Addressee Recognition
        - Track interactions between the group
        - Graph similarity for conversation structure and thread management

Next steps (Contd)

- Formalize errors towards MPC modeling benchmark
  - Introduce human evaluation metrics
    - Naturalness
    - Belonging
    - Engagement
    - Initiative
    - + Towards all participants

Thank you!

References

[1] Sai, A. B., Mohankumar, A. K., & Khapra, M. M. (2022). A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys (CSUR)*, *55*(2), 1-39.

[2] Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

[3] Deng, M., Tan, B., Liu, Z., Xing, E., & Hu, Z. (2021, November). Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 7580-7605).

[4] Higashinaka, R., Araki, M., Tsukahara, H., & Mizukami, M. (2021, July). Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 89-98).

[5] Novikova, J., Dušek, O., Curry, A. C., & Rieser, V. (2017, September). Why We Need New Evaluation Metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2241-2252).

[6] Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*(pp. 355-368).

[7] Howcroft, D. M., Belz, A., Clinciu, M. A., Gkatzia, D., Hasan, S. A., Mahamood, S., ... & Rieser, V. (2020, December). 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* (pp. 169-182).

[8] Belz, A., Mille, S., & Howcroft, D. M. (2020, December). 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*(pp. 183-194).

[9] van Miltenburg, E., Clinciu, M. A., Dusek, O., Gkatzia, D., Inglis, S., Leppänen, L., ... & Wen, L. (2021, January). Underreporting of errors in NLG output, and what to do about it. In *INLG*.