Mapping the Landscape of Dialogue Research: A Meta-Analysis of ACL Anthology 2024

Amandine Decker[1,2], **Maxime Amblard**[1] **and Ellen Breitholtz**[2]
[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{amandine.decker, maxime.amblard}@loria.fr
[2]University of Gothenburg, CLASP
ellen.breitholtz@ling.gu.se

1 Introduction

Face-to-face conversation is at the core of human communication. It is the first form of interaction we learn as children and remains fundamental to how we connect with one another (Clark, 1996). However, despite its central role, it is difficult to study systematically. Capturing face-to-face conversation in ecologically valid settings presents substantial challenges: conversations are multimodal by nature—encompassing not only language, but also gesture, gaze, prosody, *etc.* (Mondada, 2019; Kendrick et al., 2023). These extra-linguistic modalities are complicated to record without disturbing the participants, making the conversation less natural. Guaranteeing privacy is also hard in natural settings, due to voice and/or face recording but also as daily life topics of conversation involve sharing personal information which can be used for deanonymisation purposes (Amblard et al., 2014).

Meanwhile, the use of new modalities of communication such as messaging and video conferencing tools is increasing (Joskowicz, 2023; OMIP-IDAN, 2024), and more and more conversational AI systems are developed. These new paradigms raise questions about how conversation is conceptualised, represented and evaluated within the NLP/CL community. In this meta-review, we investigate how conversation is currently studied in the field by analysing papers from the 2024 ACL Anthology. We focus on major venues[1] indexed in the anthology (excluding workshops) using metadata accessed via the ACL API[2]. Our goal is to identify what tasks are being studied, what types of corpora are being used, and how these tasks are evaluated. This enables us to identify the current gaps in the field and define future research directions.

2 Method

In our study we included papers published in 2024 across venues indexed in the ACL Anthology (an open source archive of research in computational linguistics and natural language processing). We used the ACL Anthology API to retrieve metadata for all papers published in 2024, excluding venues labelled as workshops in the metadata. Our goal was to identify papers dealing with dialogue.

We selected papers whose titles contained the stems of "conversation", "dialogue", and "discourse". We then excluded the ones that were not actually focusing on dialogue (*eg.* focus on monological discourse, or papers on *convers*ion). In the end we retrieved 445 papers, which represents about 6.3% of the papers published in the selected venues. Once retrieved, one person manually reviewed the papers to determine the conversational task addressed, the corpora used and the evaluation methods employed.

3 Tasks

Across the 2024 ACL Anthology, the majority of papers dealing with conversation focus on dialogue systems (51%). Emotion recognition in conversation (14%) and conversation modelling and analysis (13%) are still quite important. Moreover, some papers (5%) focus on casting traditional tasks into conversational ones, such as conversational recommender system which consists in getting a recommendation through a conversation with a dialogue system. The rest of the papers fall into various categories: summarisation (2%), presentation of corpora (8%) and tools to study conversations (2%), and domain specific papers (*eg.* medical field, 5%).

Within the dialogue systems category, task-oriented dialogue (TOD) system used for *e.g.* booking restaurants, are especially dominant. These systems typically aim to assist users in completing specific goals, and their performance is usually

[1] ACL, ClinicalNLP, CODI, EACL, EMNLP, INLG, JEP/-TALN/RECITAL, LREC/COLING, NAACL, SemEval, SIG-DIAL, TACL, and the Findings section of ACL, EACL, EMNLP, and NAACL.

[2] https://aclanthology.org/info/development/

measured through metrics like joint goal accuracy, which tracks whether the system correctly identifies all necessary slots in a dialogue (*ie.* place, time, *etc.*). However, the conversational dimension of such systems is harder to evaluate (see Section 5).

4 Corpora

The most widely used corpora in the papers we reviewed were variants of the MultiWOZ (Budzianowski et al., 2018; Eric et al., 2020; Zang et al., 2020) dataset, which is unsurprising given the high proportion of papers on TOD systems. A broad range of other corpora are also employed across different studies. Among the most frequently cited are DailyDialog (Li et al., 2017), SGD (Rastogi et al., 2020; Lee et al., 2022) – another TOD corpus –, several corpora based on the TV-show *Friends* (Zahiri and Choi, 2018; Poria et al., 2019), conversational threads from Reddit and QReCC (Anantha et al., 2021).

Despite the variety of sources used to study conversation, a recurring issue is the lack of information w.r.t. the corpora used. Many papers omit essentials such as the language of the data, whether the dialogue is scripted or naturally occurring, and whether it is human-human or human-machine. For instance, DailyDialog is often described as a dataset for everyday chit-chat, but the dialogues were handcrafted by annotators rather than collected from real interactions. While using handcrafted data does not necessarily entail poor quality, it means that models are being trained and evaluated on data which reflects the *representation* we have of what a natural conversation is, rather than what it actually is, and people tend to underestimate the amount of e.g. pauses, repairs and other disfluencies, which are distinctive markers of conversation. A similar difference lies between a movie script and the actual transcriptions of the actors' speech. This observation points to a broader need: we must better define what conversational features are desirable in a model, and these criteria are likely to be highly task-dependent. For example, while hesitations and pauses may increase human-likeness, artificially inserting them into TOD system might not be necessary to satisfy users.

5 Evaluation

Our findings (unsurprisingly) show that conversational AI is the main focus of current research. However, many models are evaluated using benchmarks that may already be included in the training data of the models themselves, raising concerns about the validity of results.

Many studies use reference-based automatic metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). These metrics compare the system's responses to predefined gold-standard utterances based on lexical overlap. While they are convenient and widely used, they are limited in their ability to capture the interactive and context-sensitive nature of conversation. For instance, these metrics may penalize valid but diverse or creative responses that differ from the reference, even though such variation is a core feature of natural dialogue.

The growing complexity of conversational tasks calls for improved evaluation methods. LLMs have emerged as an alternative since they enable us to easily assess multiple conversational dimensions. Focusing on papers on dialogue systems, about 13% rely on such evaluation. Yet, about one third do not validate the LLM judgments with human input, even though LLMs might have divergent "expectations" of conversational quality in relation to humans. Regarding human evaluation, the most assessed dimensions are coherence (20% of the papers on dialogue systems including human evaluation), fluency (16%) and relevance (14%).

6 Conclusion

Conversation remains a central yet elusive object of study in computational linguistics. As our modes of communication evolve the way we study and evaluate conversation must also adapt. Our analysis of dialogue-related papers in the 2024 ACL Anthology reveals that much of the field's focus is currently on TOD and conversational AI, often evaluated with automatic metrics that inadequately reflect the richness and complexity of conversation. This highlights the need to create a task-sensitive framework for defining and evaluating conversational quality. What we want from conversational agents – be it accuracy, fluidity or efficiency – depends on the context. Without making these expectations explicit, both evaluation and model development risk drifting toward superficial or misleading objectives. Establishing clear dimensions to evaluate seems necessary to conduct meaningful evaluations.

References

Maxime Amblard, Karën Fort, Michel Musiol, and Manuel Rebuschi. 2014. L'impossibilité de l'anonymat dans le cadre de l'analyse du discours. In *Journée ATALA éthique et TAL*, Paris, France.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2024. Large language models suffer from their own output: An analysis of the self-consuming training loop. *Preprint*, arXiv:2311.16822.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. 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*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Herbert H. Clark. 1996. *Using language*. Cambridge university press.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Jose Joskowicz. 2023. Video conferencing technologies: Past, present and future.

Kobin H Kendrick, Judith Holler, and Stephen C Levinson. 2023. Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical transactions of the royal society B*, 378(1875):20210473.

Harrison Lee, Raghav Gupta, Abhinav Rastogi, Yuan Cao, Bin Zhang, and Yonghui Wu. 2022. Sgd-x: A benchmark for robust generalization in schema-guided dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10938–10946.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lorenza Mondada. 2019. Contemporary issues in conversation analysis: Embodiment and materiality, multimodality and multisensoriality in social interaction. *Journal of Pragmatics*, 145:47–62. Quo Vadis, Pragmatics?

SANUSI Bernice Oluwalanumi OMIPIDAN, Ismail Adewale. 2024. Rise of social media in the digital age: Whatsapp a threat to effective communication. *IMSU Journal of Communication Studies*, 8(1).

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *AAAI Workshops*, volume 18, pages 44–52.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.