Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Journal Track

CADS: A Systematic Literature Review on the Challenges of Abstractive Dialogue Summarization (Abstract Reprint)

Frederic Kirstein , **Jan Philip Wahle** , **Bela Gipp** and **Terry Ruas**
Georg-August University Göttingen, Papendiek 14, 37073 Göttingen, Germany
kirstein@gipplab.org, wahle@uni-goettingen.de, gipp@uni-goettingen.de, ruas@uni-goettingen.de

Abstract Reprint. This is an abstract reprint of a journal article by [Li and Zhou, 2025].

Abstract

Abstractive dialogue summarization is the task of distilling conversations into informative and concise summaries. Although focused reviews have been conducted on this topic, there is a lack of comprehensive work that details the core challenges of dialogue summarization, unifies the differing understanding of the task, and aligns proposed techniques, datasets, and evaluation metrics with the challenges. This article summarizes the research on Transformer-based abstractive summarization for English dialogues by systematically reviewing 1262 unique research papers published between 2019 and 2024, relying on the Semantic Scholar and DBLP databases. We cover the main challenges present in dialog summarization (i.e., language, structure, comprehension, speaker, salience, and factuality) and link them to corresponding techniques such as graph-based approaches, additional training tasks, and planning strategies, which typically overly rely on BART-based encoder-decoder models. Recent advances in training methods have led to substantial improvements in language-related challenges. However, challenges such as comprehension, factuality, and salience remain difficult and present significant research opportunities. We further investigate how these approaches are typically analyzed, covering the datasets for the subdomains of dialogue (e.g., meeting, customer service, and medical), the established automatic metrics (e.g., ROUGE), and common human evaluation approaches for assigning scores and evaluating annotator agreement. We observe that only a few datasets (i.e., SAMSum, AMI, DialogSum) are widely used. Despite its limitations, the ROUGE metric is the most commonly used, while human evaluation, considered the gold standard, is frequently reported without sufficient detail on the inter-annotator agreement and annotation guidelines. Additionally, we discuss the possible implications of the recently explored large language models and conclude that our described challenge taxonomy remains relevant despite a potential shift in relevance and difficulty.

References

[Li and Zhou, 2025] Jinlin Li and Xiao Zhou. Curegraph: Contrastive multi-modal graph representation learning for urban living circle health profiling and prediction. *Artificial Intelligence*, 340:104278, 2025.