sponses, which can help to elicit long-term commitments and develop emotional attachments from users to sustain close relationships over time. To this end, we propose the History-Aware Hierarchical Transformer (HAHT) for multi-session open-domain dialogue systems, which can effectively leverage history conversations to conduct more engaging MSCs. HAHT maintains a long-term memory to store historical conversational contexts, which is updated when a new session is conducted. Based on the long-term memory and the context in the current session, relevant tokens in historical contexts are selected to adapt the current response.

Specifically, as the number of tokens in a conversation utterance and the number of turns in a conversation are usually not very long[1], we first encode the history conversation hierarchically into the history memory using Transformer (Vaswani et al., 2017). The history memory serves as a high-level representation of history conversations. Secondly, as history conversations usually can facilitate the understanding of the current conversation context, we design a history-aware context encoder. The context encoder encodes conversation context, considering both history conversations and the current conversation, by adopting the transformer attention over the history memory and current conversation context. Then, the context encoder also updates the history memory based on the current conversation context. Finally, we design a history-aware decoder to fuse learned history information into the response generation process. The history-aware decoder can switch between two strategies, *i.e.,* generating a word from the generic vocabulary or directly copying a word from history conversations.

Experimental results on the large-scale Facebook MSC dataset show that the proposed HAHT model outperforms previous multi-session open-domain dialogue systems in various evaluation metrics. Human evaluation results support that HAHT generates more readable, context-relevant, and history-relevant responses than baseline models. In addition, the ablation study confirms that both the hierarchical encoding of history conversations and the history-aware decoder contribute greatly to HAHT's performance on MSCs and help it leverage historical information more effectively.

2 Related Work

Open-domain dialogue systems aim to perform chit-chat without task and domain restrictions (Ritter et al., 2011) and establish long-term relationships with users (Clark et al., 2019; Roller et al., 2020). They are generally divided into two groups: generation-based systems and retrieval-based systems. Retrieval-based systems seek to find a suitable response from a large response candidate set (Zhou et al., 2016; Yuan et al., 2019; Zhong et al., 2020; Zhu et al., 2021; Qian et al., 2021), whereas, generation-based systems focus on generating responses from scratch based on the dialogue history (Serban et al., 2016; Shum et al., 2018; Adiwardana et al., 2020; Roller et al., 2020; Xu et al., 2022). In this paper, we focus on generation-based systems.

Early approaches to response generation include template-based generation methods (Higashinaka et al., 2014) and statistical machine translation (SMT) methods (Ritter et al., 2011). With the development of deep learning, sequence-to-sequence (Seq2seq) models have been applied to generation-based dialogue systems and achieved great performance (Li et al., 2016; Vinyals and Le, 2015; Serban et al., 2017). Recently, with the increasing availability of large-scale dialogue datasets (Li et al., 2017; Zhang et al., 2018; Dinan et al., 2019; Huang et al., 2020), Transformer-based language models pretrained with large-scale corpora, such as Meena (Adiwardana et al., 2020), BlenderBot (Roller et al., 2021), DialogueGPT (Zhang et al., 2020), and PLATO (Platonov et al., 2020), have made significant progress in the area of open-domain dialogues.

Despite the advancements in the field, current state-of-the-art generative pre-trained models are designed for and trained on large datasets of single-session conversations with a small number of turns. As a result, most existing models employ short token truncation lengths, such as 128 tokens for Meena (Adiwardana et al., 2020), and are unable to encode and utilize historical contexts in MSCs effectively. In addition, there is also a lack of public MSC datasets. Xu et al. released the first multi-session conversation dataset, *i.e., Facebook MULTI-SESSION CHAT (Facebook MSC)*, and explored different retrieval-augmented generative models on the dataset (Lewis et al., 2020; Shuster et al., 2021), which achieved better results than the standard Transformer (Vaswani et al., 2017). However, the experimental results demonstrate that

[1] On average, conversations have 13 turns and conversation utterances have 16 tokens in Facebook MSC dataset.