probability distribution over the generic vocabulary V_g at the decoding time step t is computed as,

$$P_{v_g} = \text{softmax}\big(\text{FC}_1(\mathbf{o}_t)\big), \tag{4}$$

where FC_1 is a fully connected layer.

To calculate the probability distribution over the history-aware vocabulary V_h, we adopt a max-pooling layer over the context-updated history memory $\mathbf{C}_s$, a fully connected layer, and a soft-max function as follows,

$$P_{v_h} = \text{softmax}\big(\text{FC}_2(\text{max-pooling}(\mathbf{C}_s))\big), \tag{5}$$

where FC_2 is a fully connected layer.

The final word probability distribution at time step t is computed by using a switching mechanism between P_{v_g} and P_{v_h} as follows,

$$P = \alpha_{v_g} * P_{v_g} + \alpha_{v_h} * P_{v_h}, \tag{6}$$

where α_{v_g} and α_{v_h} is the switching probability of generating from generic vocabulary or copying from history conversations. α_{v_g} and α_{v_h} is calculated as follows,

$$[\alpha_{v_g}, \alpha_{v_h}] = \text{softmax}\big(\text{FC}_3([o_j; \text{max-pooling}(\mathbf{C}_s)])\big), \tag{7}$$

where FC_3 is a fully connected layer, and [;] is a concatenation operation over the last dimension.

3.4 Model Training

We train the model to maximize the generation probability of the target response, given the current conversation context and history conversations in an end-to-end manner. The loss function of HAHT is defined as,

$$\mathcal{L} = -\sum_{t=1}^{n_y} \log\big(P(y_j|X, H, y_{<t})\big), \tag{8}$$

where X denotes the current conversation context, H denotes all history conversations, $y_{<t}$ denotes tokens before time step t, and n_y denotes the length of the ground truth response.

4 Experimental Settings

In this section, we introduce the experimental dataset, evaluation metrics, baseline methods, and model settings.

Session number	Train		Valid		Test	
	Conv.	Utter.	Conv.	Utter.	Conv.	Utter.
1*	8939	131,438	1000	7,801	1015	6,634
2	4000	46,420	500	5,897	501	5,939
3	4000	47,259	500	5,890	501	5,924
4	1001	11,870	500	5,904	501	5,940
5	-	-	500	5,964	501	5,945
Total	-	236,987	-	31,456	-	30,382

Table 1: The statistics of Facebook Multi-Session Chat (Facebook MSC) Dataset. Session number i indicates there are i-1 history conversation sessions that happen before the last conversation session. *: Session 1 does not contain history conversation sessions.

4.1 Experimental Dataset

The experiments are performed on a large dataset, *i.e.,* Facebook MULTI-SESSION CHAT (Facebook MSC) (Xu et al., 2022). It is a crowdsourced dataset consisting of multi-session conversations, where the interlocutors learn about each other's interests and discuss the things they have understood from past sessions. The number of history conversations in Facebook MSC varies from 1 to 4. Session number i indicates there are i-1 history conversations happening before the last conversation session. The statistics of the Facebook MSC dataset are summarized in Table 1. As session 1 does not have history conversations, we evaluate our model on session 2-5.

4.2 Evaluation Metrics

We conduct both automatic and human evaluations to demonstrate the effectiveness of the proposed model. For automatic evaluations, we leverage BLEU-2, BLEU-3 (Papineni et al., 2002), and ROUGE-L (Lin and Och, 2004) to measure word overlaps between the generated response text and ground truth text.

Moreover, we also randomly sample 50 MSCs from the test set to conduct human evaluations. We present all the history conversation sessions, current conversation context, and the generated responses to three well-educated annotators. The annotators will evaluate the quality of the generated responses from the following three aspects:

- **Readability**: measures whether the generated responses are natural and fluent.
- **Context Relevancy**: measures whether the generated responses are correlated with the current conversation context.
- **History Relevancy**: measures whether the generated responses are correlated with history con-