A Supplementary Material

A.1 Proof of Eq. 3

$$
\begin{aligned}
&\lambda_1 I(H,c)+\lambda_2 I(c,F)\\
\geq&\lambda_1 I(\tilde{H},c)+\lambda_2 I(c,\tilde{F})\\
=&\lambda_1\mathbb{E}_{p_\phi(\tilde{H}c)}\log\frac{p_\phi(\tilde{H}|c)}{p(\tilde{H})}+\lambda_1\mathbb{E}_{p_\phi(c\tilde{F})}\log\frac{p_\phi(\tilde{F}|c)}{p(\tilde{F})}\\
=&\lambda_1\mathbb{E}_{p_\phi(\tilde{H}c)}\log p_\phi(\tilde{H}|c)+\lambda_1\mathbb{H}(\tilde{H})+\lambda_2\mathbb{E}_{p_\phi(c\tilde{F})}\log p_\phi(\tilde{F}|c)+\lambda_2\mathbb{H}(\tilde{F})\\
\geq&\lambda_1\mathbb{E}_{p_\phi(\tilde{H}c)}\log p_\phi(\tilde{H}|c)+\lambda_2\mathbb{E}_{p_\phi(c\tilde{F})}\log p_\phi(\tilde{F}|c)\\
=&\lambda_1\mathbb{E}_{p_\phi(\tilde{H}c)}\log p_\gamma(\tilde{H}|c)+\lambda_1 KL(p_\phi(\tilde{H}|c)||p_\gamma(\tilde{H}|c))+\lambda_2\mathbb{E}_{p_\phi(c\tilde{F})}\log p_\gamma(\tilde{F}|c)+\lambda_2 KL(p_\phi(\tilde{H}|c)||p_\gamma(\tilde{H}|c))\\
\geq&\lambda_1\mathbb{E}_{p_\phi(\tilde{H}c)}\log p_\gamma(\tilde{H}|c)+\lambda_2\mathbb{E}_{p_\phi(c\tilde{F})}\log p_\gamma(\tilde{F}|c)\\
=&\mathbb{E}_{p_\phi(\tilde{H}u_i\tilde{F},c)}[\lambda_1\log p_\gamma(\tilde{H}|c)+\lambda_2\log p_\gamma(\tilde{F}|c)]
\end{aligned}
$$

A.2 Proof of Eq. 4

$$
\begin{aligned}
I(u_i,c|H)&=\mathbb{E}_{p(H)}\mathbb{E}_{p_\phi(u_ic|H)}\log\frac{p_\phi(u_i|Hc)}{p(u_i|H)}\\
&=\mathbb{E}_{p(H)}\mathbb{E}_{p_\phi(u_ic|H)}\log p_\phi(u_i|Hc)+\mathbb{H}(u_i|H)\\
&\geq\mathbb{E}_{p(Hu_iF)}\mathbb{E}_{p_\phi(c|Hu_iF)}\log p_\phi(u_i|Hc)\\
&=\mathbb{E}_{p(Hu_iF)}\mathbb{E}_{p_\phi(c|Hu_iF)}\log p_\gamma(u_i|Hc)+\mathbb{E}_{p_\phi(HcF)}KL(p_\phi(u_i|Hc)||p_\gamma(u_i|Hc))\\
&\geq\mathbb{E}_{p(Hu_iF)}\mathbb{E}_{p_\phi(c|Hu_iF)}\log p_\gamma(u_i|Hc)
\end{aligned}
$$

A.3 Derivation of Eq. 8

$$
\begin{aligned}
&\mathbb{E}_{p(\tilde{u_iF}|\tilde{H})}[\mathbb{E}_{p_\phi(c|\tilde{H}\tilde{u_iF})}\log p_\phi(\tilde{u_iF}|c)-KL(p_\phi(c|\tilde{H}\tilde{u_iF})||p_\theta(c|\tilde{H}))]\\
=&\mathbb{E}_{p(\tilde{u_iF}|\tilde{H})}[\mathbb{E}_{p_\phi(c|\tilde{H}\tilde{u_iF})}\log\frac{p_\phi(\tilde{u_iF}|c)p_\theta(c|\tilde{H})}{p_\phi(c|\tilde{H}\tilde{u_iF})}]\\
=&\mathbb{E}_{p(\tilde{u_iF}|\tilde{H})}[[\mathbb{E}_{p_\phi(c|\tilde{H}\tilde{u_iF})}\log\frac{p_\phi(\tilde{u_iF}|c)p_\theta(c|\tilde{H})p(\tilde{u_iF}|\tilde{H})}{p_\phi(\tilde{u_iF}|\tilde{H}c)p_\phi(c|\tilde{H})}]\\
=&\mathbb{E}_{q_\phi(c|\tilde{H})}KL(p_\phi(\tilde{u_iF}|\tilde{H}c)||p_\phi(\tilde{u_iF}|\tilde{H}c))-KL(p_\phi(c|\tilde{H})||p_\theta(c|\tilde{H}))-\mathbb{H}(\tilde{u_iF}|\tilde{H})
\end{aligned}
$$

A.4 Information Retrieval Technique for Multiple References

We collected multiple reference responses for each dialogue context in the test set by information retrieval techniques. References are retrieved based on their similarity with the provided context. Responses to the retrieved utterances are used as references. The process of retrieving similar context is as follows: First, we select 1000 candidate utterances using the tf-idf score. These candidates are then mapped to a vector space by summing their contained word vectors. After that, they are reranked based on the average of cosine similarity, Jaccard distance and Euclidean distance with the ground-truth context. The top 10 retrieved responses are passed to human annotators to judge the appropriateness.

A.5 Phrases that count as forming dull responses

1) i know

2) no __eou__(yes __eou__)

3) no problem

4) lol

5) thanks __eou__

6) don’t know

7) don’t think

8) what ?

9) of course

10) wtf

Utterances matching one of these phrases are treated as dull responses.

A.6 Effect of hyperparameter λ_1/λ_2

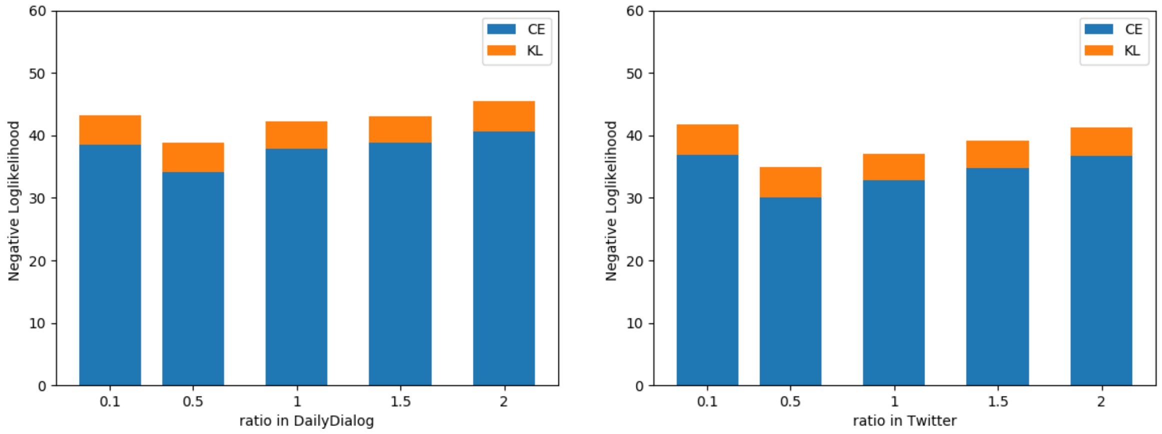

Figure 3: Effect of hyperparameter ratio λ_1/λ_2 on two datasets.

Figure 3 visualizes the effects of hyperparameters λ_1 and λ_2. The negative-log-likelihood is decomposed into two parts: decoding cross entropy (CE) as in Eq. 4 and KL divergence as in Eq. 7. The sum is a lower bound of the true log-likelihood. The optimal ratio is around 0.5 for both datasets, which means only half weights should be given to the history compared with the future context. Two reasons can explain this phenomena. Firstly, future vector is harder to infer than history as it is not explicitly exposed as an input in Eq. 3. Secondly, minimizing the KL divergence in Eq. 7 pushes the code space to discard information from the future context so that it could vanish to zero. Therefore, more weights should be given to the future context to maintain a balance.