Personalized Dialogue Generation with Persona-Adaptive Attention

Qiushi Huang[1,2], Yu Zhang[2*], Tom Ko[3], Xubo Liu[1], Bo Wu[4], Wenwu Wang[1], H Tang[1*]

[1] University of Surrey
[2] Southern University of Science and Technology
[3] ByteDance AI Lab
[4] MIT-IBM Watson AI Lab
{qiushi.huang,xubo.liu,w.wang,h.tang}@surrey.ac.uk,{yu.zhang.ust,tomkocse}@gmail.com, bo.wu@ibm.com

Code:https://github.com/hqsiswiliam/persona-adaptive-attention

AAAI 2023

2023. 12. 21 • ChongQing

Leibniz-Institut für Sozialwissenschaften

Reported by JiaWei Cheng

Persona-based dialogue systems aim to generate consistent responses based on historical context and predefined persona.

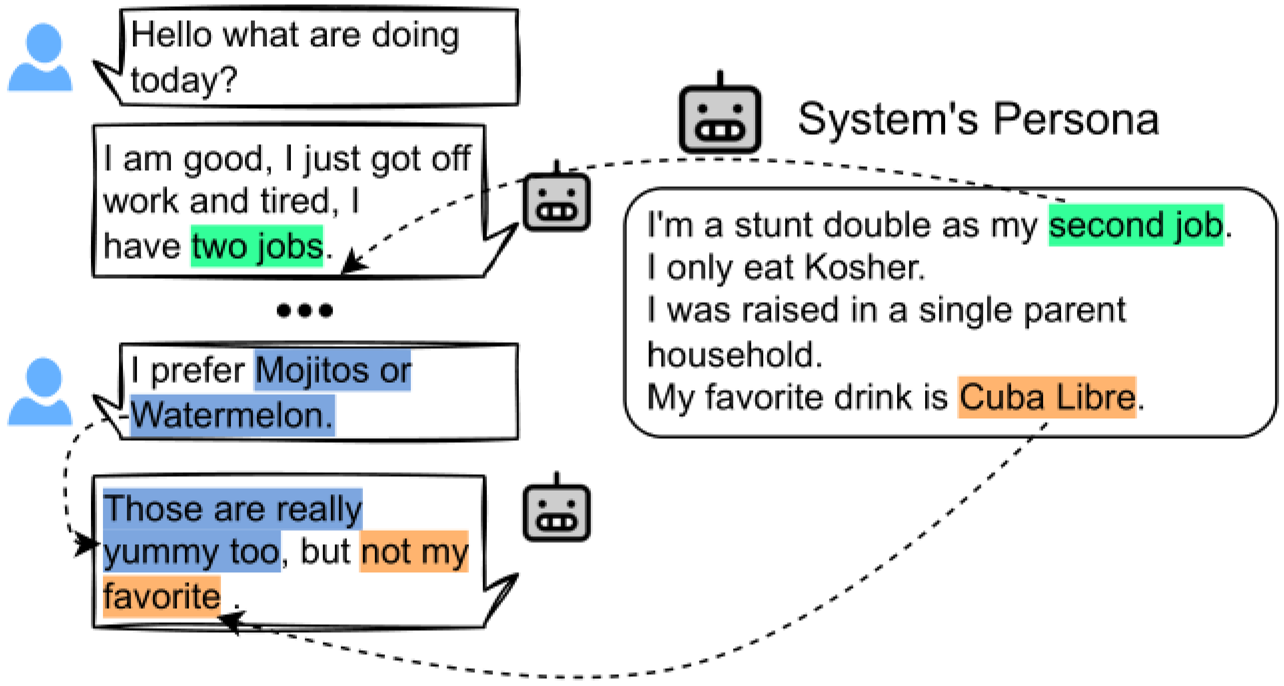


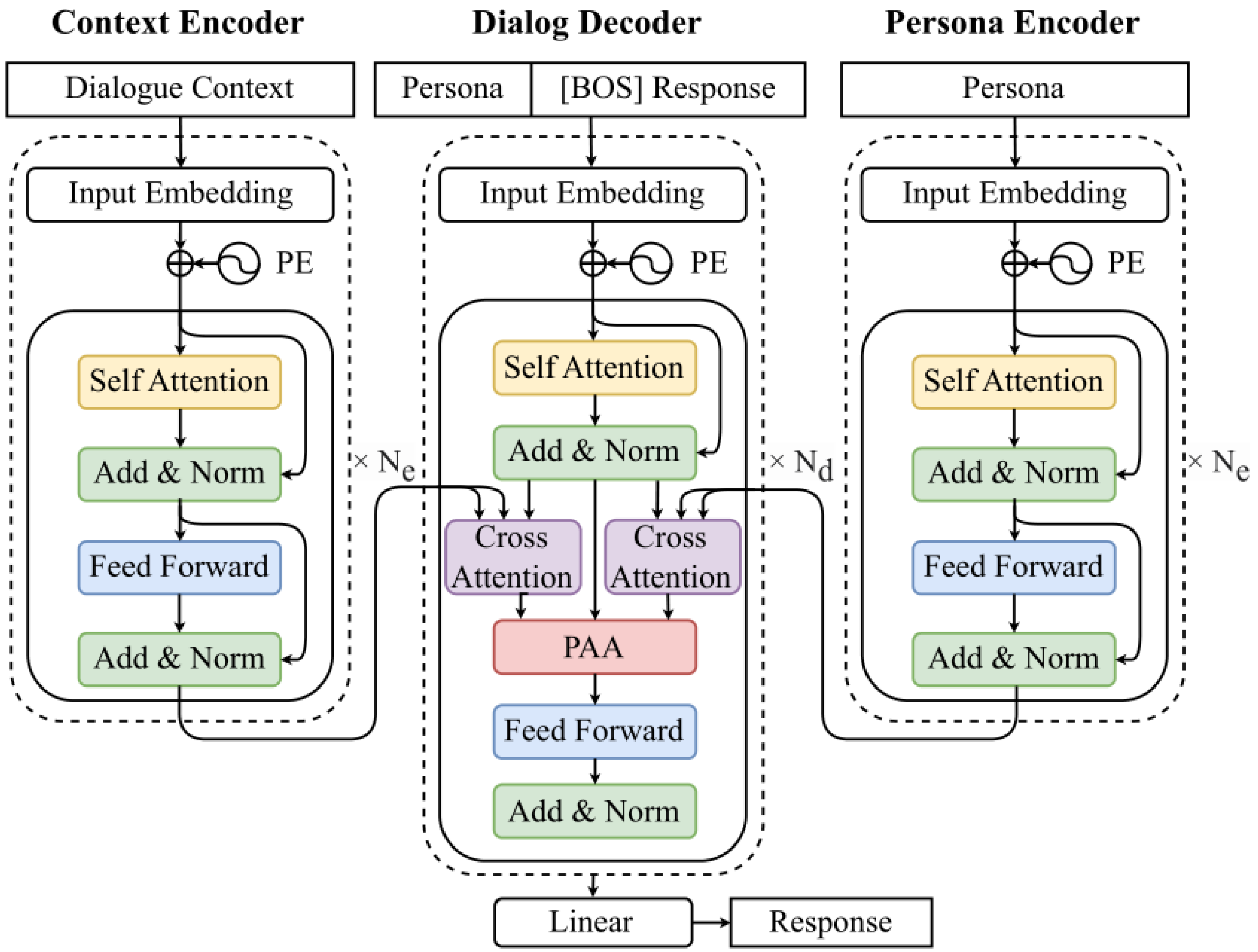
Figure 1: An example from the ConvAI2 dataset.

Motivation

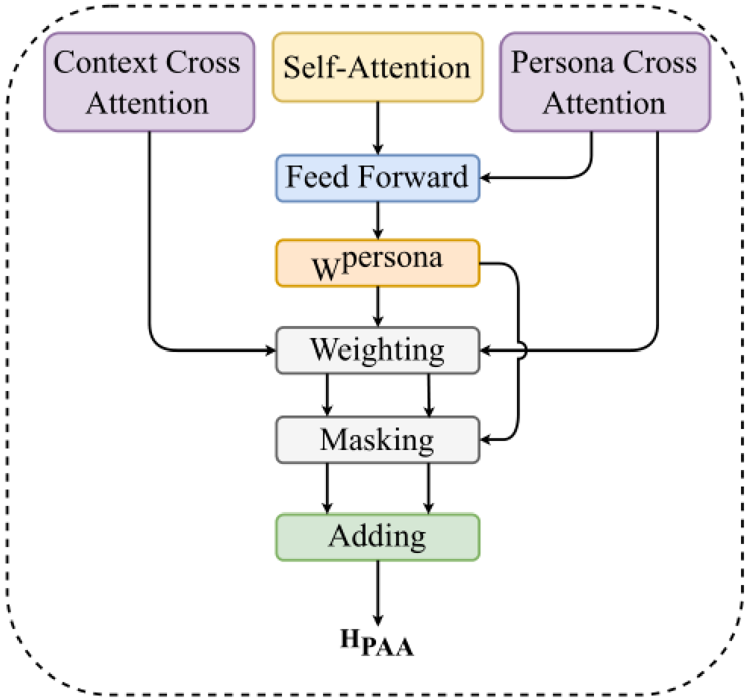
(1)： One challenge in persona-based dialogue generation is that the related datasets are usually small

(2)： Another challenge is to choose the weights between the persona and context.

Overview

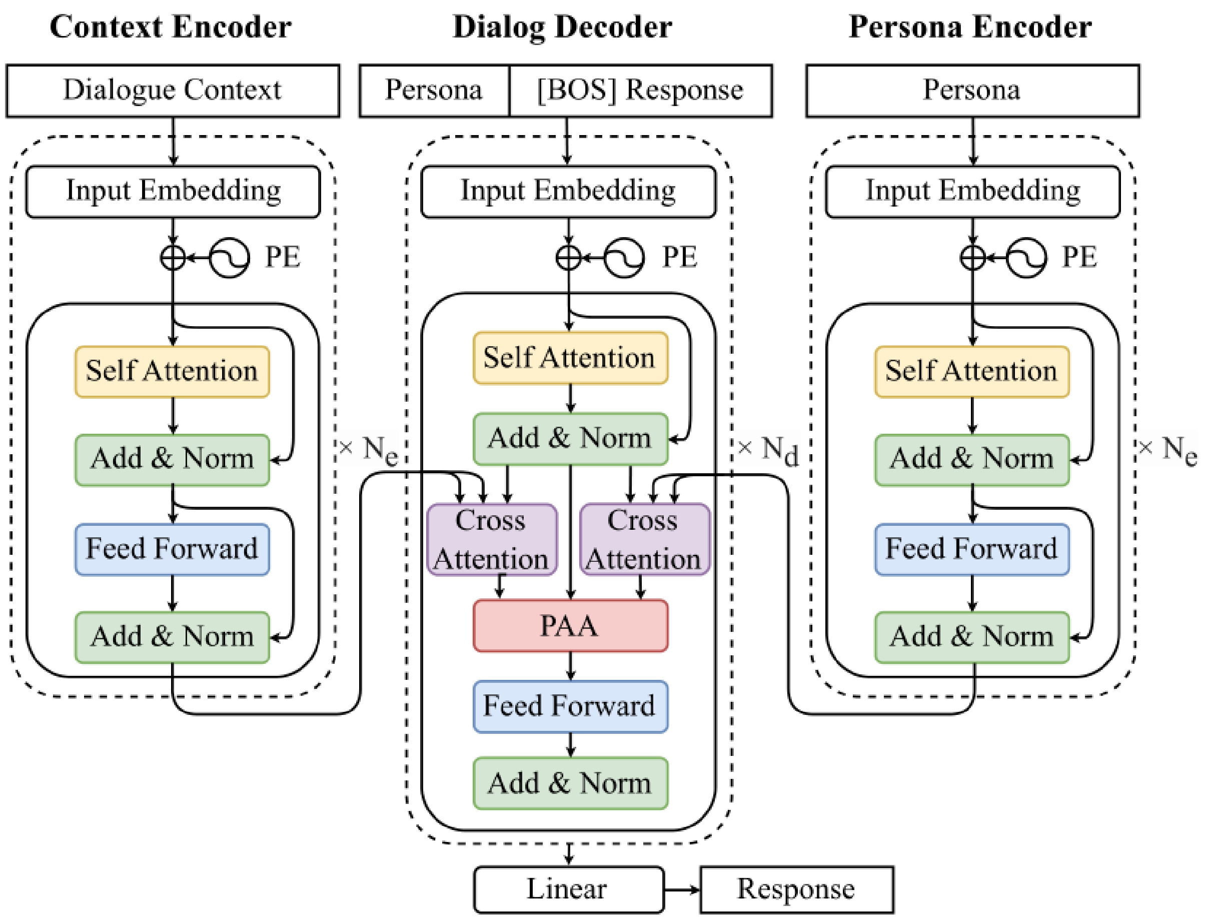


(a) The overview of our framework, PAA indicates the Persona-Adaptive Attention



(b) The architecture of Persona-Adaptive Attention, H_{PAA} is the module's output

Method

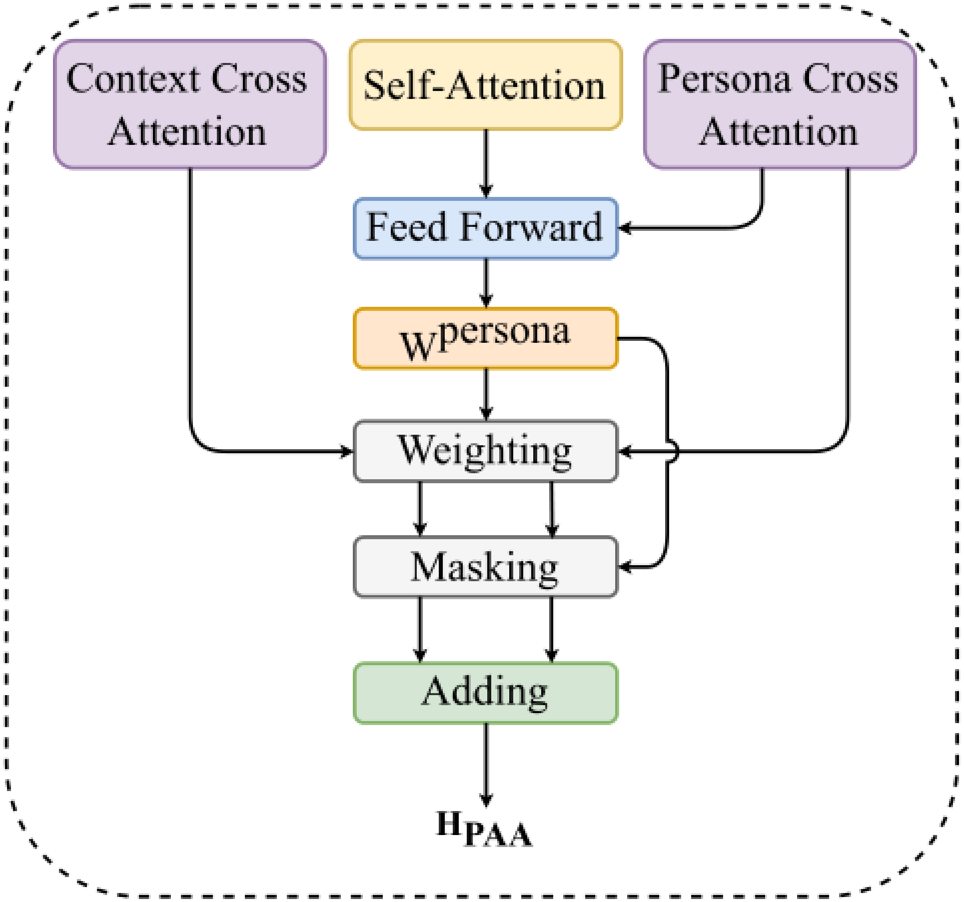


$$\begin{aligned} h_P &= \text{Encoder}_P(I_P), \\ h_U &= \text{Encoder}_U(I_U), \end{aligned} \quad (1)$$

$$\begin{aligned} h_R &= \text{Self-Attention}(I_R) + M_R, \\ \hat{h}_R &= \text{AddNorm}(h_R), \end{aligned} \quad (2)$$

$$\begin{aligned} o_P &= \text{Softmax}(\frac{Q_r K_p^\top}{\sqrt{d}})V_p, \\ o_U &= \text{Softmax}(\frac{Q_r K_u^\top}{\sqrt{d}})V_u, \end{aligned} \quad (3)$$

Method



$$m_p = FC([h_R; o_P]),$$
$$w_{persona} = \text{Sigmoid}(m_p). \quad (4)$$

$$\tilde{o}_P = w_{persona} o_P,$$
$$\tilde{o}_U = (1 - w_{persona}) o_U. \quad (5)$$

$$m_{persona} = \mathbb{M}(w_{persona} > \tau),$$
$$m_{context} = \mathbb{M}(1 - w_{persona} > \tau). \quad (6)$$

$$\tau = |I_U| / (|I_U| + |I_P|),$$

$$\hat{o}_P = m_{persona} \odot \tilde{o}_P,$$
$$\hat{o}_U = m_{context} \odot \tilde{o}_U, \quad (7)$$
$$H_{PAA} = \hat{o}_P + \hat{o}_U,$$

$$\mathcal{L}_{NLL} = -\log(p_\theta(I_R | I_P, I_U))$$
$$= -\sum_{i=1}^{|I_R|} \log(p_\theta(t_i^y | I_P, I_U, t_{<i}^y)), \quad (8)$$

Experiments

Method	PARAMS	PPL ↓	F1 ↑	BLEU-1 ↑	BLEU-2 ↑	Dist-1 ↑	Dist-2 ↑
Encoder-GPT2	182M	20.06	11.95	16.78	1.69	0.11	0.23
GPT2-SMALL	124M	18.10	11.83	20.36	3.97	**1.31**	**6.30**
GPT2-MEDIUM	355M	17.65	11.45	18.06	3.58	1.13	6.07
GPT2-LARGE	774M	16.98	10.93	5.99	0.79	0.42	2.62
Attn-Routing	254M	17.94	12.77	18.74	2.80	0.70	2.39
PAA (Ours)	254M	**14.03**	**17.36**	**20.50**	**4.17**	**1.31**	5.21

Table 1: Automatic evaluation results on ConvAI2 dataset over our implemented approach. Boldface indicates the best result in terms of the corresponding metrics. Attn-Routing means the Attention-Routing mechanism, the implementation details are described in Appendix.

Experiments

Method	PPL [5] ↓	Hits@1 ↑	F1 ↑
KVPM	-	54.8	14.25
DIM	-	78.8	-
LIC	-	17.3	17.79
TransferTransfo	17.51	82.1	19.09
P^2 Bot	15.12	81.9	**19.77**
BoB	**7.80**	-	-
GPT2-D3	15.69	-	-
PAA (Ours)	14.03	**93.9**	17.36

Table 2: Automatic evaluation results on ConvAI2 over published work.

Experiments

Method	Flue. ↑	Info. ↑	Rele. ↑	Per.C. ↑
E-GPT2	4.37	2.54	1.97	0.31
GPT2-M	4.15	3.70	3.10	0.43
PAA	**4.80**	**4.54**	**3.69**	**0.70**

Table 3: Human evaluation results on sampled decoding response. The fluency, informativeness, relevance, and persona consistency are abbreviated as "Flue.", "Info.", "Rele.", and "Per.C.". E-GPT2 represents the Encoder-GPT2, and GPT2-M means GPT2-MEDIUM.

Experiments

Method	PPL ↓	F1 ↑
DirectSUM	23.15	11.37
PARAM	17.76	12.75
Dual	18.57	15.87
Skipped	14.73	17.30
Context	14.65	17.22
PAA	**14.03**	**17.36**

Table 4: The automatic evaluation results on PAA variants.

Experiments

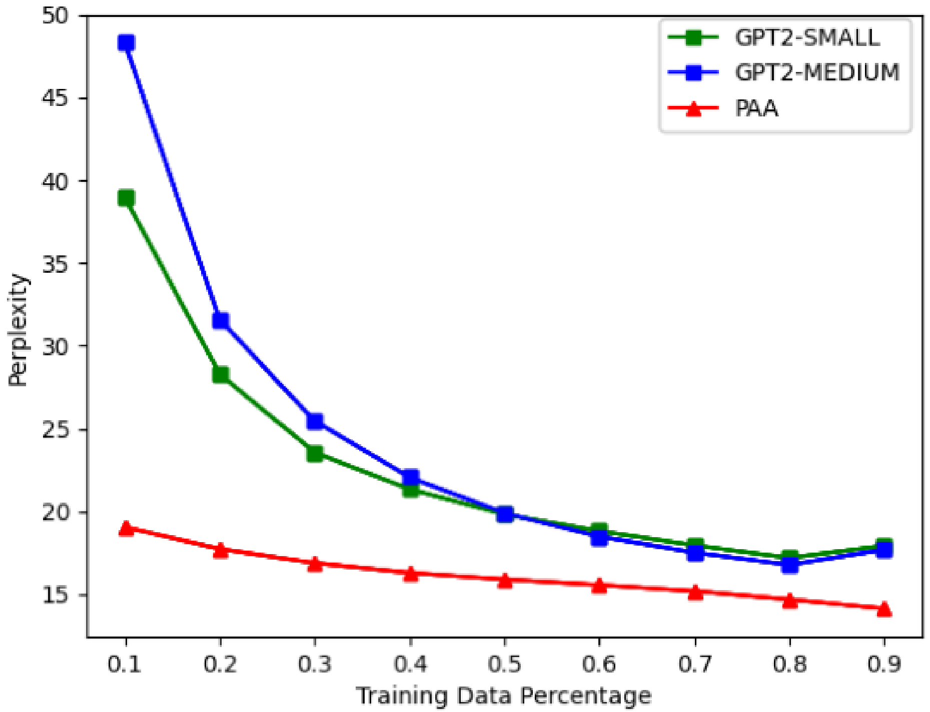


Figure 3: Comparison with GPT2 under low-resource scenario, we sampled 10% to 90% of training data to train GPT2-SMALL, GPT2-MEDIUM and PAA.

Experiments

Method	PARAMS	PPL ↓	F1 ↑
Reddit 2.7B	2.7B	18.90	12.60
BlenderBot 1	2.7B	**10.20**	18.30
R2C2 BlenderBot	2.7B	10.50	**20.50**
OPT-175B	175B	10.80	18.50
PAA (Ours)	254M	14.03	17.36

Table 5: Automatic evaluation results on the ConvAI2 dataset over large pre-trained language models.

Thanks!