Architecture	2-to-2 Transformer (Vaswani et al., 2017; Tiedemann and Scherrer, 2017)
Enc-Dec layers	6
Attention heads	8
Word-embedding dimension	512
Feed-forward dimension	2,048
Share all embeddings	True
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$) (Kingma and Ba, 2015)
Learning rate schedule	Inverse square root decay
Warmup steps	4,000
Max learning rate	0.001
Initial Learning Rate	1e-07
Dropout	0.3 (Srivastava et al., 2014)
Label smoothing	$\epsilon_{ls} = 0.1$ (Szegedy et al., 2016)
Mini-batch size	8,000 tokens (Ott et al., 2018)
Number of epochs	20
Averaging	Save checkpoint for every 5000 iterations and take an average of last five checkpoints
Beam size	6 with length normalization (Wu et al., 2016)
Implementation	`fairseq` (Ott et al., 2019)

Table 5: List of hyper-parameters for training the NMT model

Architecture	BERT (base) (Devlin et al., 2019)
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$, weight decay=0.01) (Kingma and Ba, 2015)
Learning rate schedule	Inverse square root decay
Max learning rate	0.001
Mini-batch size	16 samples
Number of epochs	1
Implementation	`transformers` (Wolf et al., 2020)

Table 6: List of hyper-parameters for training the classification model

A Settings of Machine Translation Model

This section describes the details of the training neural machine translation model. Firstly, we tokenized the corpus into subwords with BPE (Sennrich et al., 2016). We set the vocabulary size to 32,000. Then we trained the 2-to-2 Transformer-based NMT model (Tiedemann and Scherrer, 2017), which outputs two consecutive given two input sentences to consider larger contexts. Table 5 shows the list of hyper-parameters.

B Settings of Classification Model

This section describes the details of the training classification model. Table 6 shows the list of hyper-parameters.

C Details of Crowd-sourcing Tasks

C.1 Filtering Persona-chat

We asked crowd workers on Amazon Mechanical Turk (`https://requester.mturk.com/`) to filter out incoherent data in Persona-chat. Here, we defined a chat as "incoherent" if:

- questions being ignored;
- the presence of unnatural topic changes;
- one is not addressing what the other said;
- responses seeming out of order;
- or being hard to follow in general.

Workers were instructed to disregard minor issues such as typos and focus on the general flow.

In the full round, we selected $1,500$ chats from Persona-chat. Each crowd worker was tasked to rate 5 chats at a time, and each chat was rated by 10 different workers. Eligible workers were selected with a preliminary qualification round.

C.2 Rating Translations

We asked crowd workers on Crowdworks (`https://crowdworks.jp/`) to label the human translation and the NMT translation in BPersona-chat as low-quality or high-quality. In the task, we defined a translation as bad if:

- the translation is incorrect;
- parts of the source chat are lost;
- there are serious grammatical or spelling errors that interfere with understanding;
- the person's speaking style changes from the past utterance;
- the translation is meaningless or incomprehensible;
- or the translation is terrible in general.

Workers worked on files in which one file included one complete chat; therefore, they could check the context and rate each utterance of the conversation.

To the limited number of workers, in the full round, crowd workers were tasked to rate around 50 to 300 chats in two weeks. Eligible workers were selected with a preliminary qualification round.