Chat Translation Error Detection for Assisting Cross-lingual Communications

Yunmeng Li[1] **Jun Suzuki**[1,3] **Makoto Morishita**[2] **Kaori Abe**[1]
Ryoko Tokuhisa[1] **Ana Brassard**[3,1] **Kentaro Inui**[1,3]
[1]Tohoku University [2]NTT [3]RIKEN
li.yunmeng.r1@dc.tohoku.ac.jp

Abstract

In this paper, we describe the development of a communication support system that detects erroneous translations to facilitate cross-lingual communications due to the limitations of current machine chat translation methods. We trained an error detector as the baseline of the system and constructed a new Japanese–English bilingual chat corpus, **BPersona-chat**, which comprises multi-turn colloquial chats augmented with crowdsourced quality ratings. The error detector can serve as an encouraging foundation for more advanced erroneous translation detection systems.

1 Introduction

With the expansion of internationalization, there is an increasing demand for cross-lingual communication. However, while machine translation technologies have demonstrated sound performance in translating documents (Barrault et al., 2019, 2020; Nakazawa et al., 2019), current methods are not always suitable for translating chat (Läubli et al., 2018; Toral et al., 2018; Farajian et al., 2020; Liang et al., 2021). When a translation system generates erroneous translations, the user may be unable to identify such errors, which can lead to confusion or misunderstanding. Thus, in this study, we developed a cross-lingual chat assistance system that reduces potential miscommunications by detecting translation errors and notifying the users of their occurrences. As a critical component of such a system, we propose the erroneous chat translation detection task and conduct an empirical study to model error detection. An illustration of the baseline task is shown in Figure 1. When the translation system generates a translation that is suspected to be incorrect or not well-connected to the context, we prompt users on the source language side that the translation may be incorrect. The warning message is expected to encourage users to modify their text into a better translatable form. Simultaneously, users on the target language side receive the same warning message to indicate that unusual words or passages are likely translation errors.

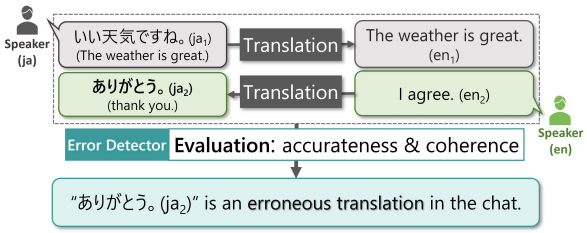

Figure 1: Illustration of the error detector predicting erroneous translations. The detector evaluates whether translation ja_2 is accurate and coherent in the chat.

To support this line of research, we created a new parallel chat corpus, **BPersona-chat**[1], which comprises multi-turn colloquial chats augmented with manually produced gold translations and machine-generated translations with crowdsourced quality labels (*correct* or *erroneous*). In an experiment, we trained an error detection model that classifies a given translation in a bilingual two-utterance chat as either correct or erroneous (Figure 1) and evaluated its performance on the BPersona-chat dataset. Our primary contributions are summarized as follows. (1) We propose the erroneous chat translation detection task. (2) We construct that BPersona-chat parallel chat corpus. (3) We trained the error detector, thereby providing a foundation to develop more sophisticated communication support systems.

2 Task Definition

As the baseline task, we define a *chat* as a two-utterance colloquial dialog between two humans using different languages. Here, we focus on predicting whether the second utterance, i.e., the response, was translated correctly. The preceding context, the translation of the context, the response, and the translated response are input to the error

[1]https://github.com/cl-tohoku/BPersona-chat

Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems, pages 88 - 95

detector. Then, the detector predicts the translated response using the other utterances as reference data. The detector then outputs whether the translated response is erroneous.

Figure 1 shows an example target task of evaluating the Japanese translation of an English utterance. Here, the Japanese speaker's initial utterance ja_1 is translated into en_1, and the English speaker's response en_2 is translated into ja_2. In this example, the detector is assessing the utterance "ありがとう。 (*Thanks.*)," which is not an accurate translation of the utterance "I agree." The detector is given the preceding context (ja_1, en_1, and en_2) as reference data to predict whether the translation is both accurate and coherent. If the detector is predicting the translation en_2 of response ja_2, the reference data include en_1, ja_1, and ja_2 in the opposite.

3 Related Work

Translation quality estimation task Our target task is a new setting compared to quality estimation tasks (Specia et al., 2020; Fonseca et al., 2019), which primarily focus on written text, e.g., Wikipedia articles and Amazon reviews. In contrast, the target task attempts to detect errors in chat translation systems; thus, we must understand the contexts of casual conversational settings.

Parallel dialog corpus There are bilingual dialog corpora, e.g., Business Scene Dialog (Rikters et al., 2019), which includes business negotiation scenes in both Japanese and English. However, our task requires data that include cross-lingual colloquial chats with both appropriate and erroneous translations. To the best of our knowledge, no such dataset exists; thus, we must prepare a new evaluation dataset to evaluate the proposed task.

4 Evaluation Dataset

To mitigate the construction time and cost, we took advantage of existing chat corpora as a starting point. We first filtered out inappropriate chats, then asked professional translators to perform utterance-by-utterance translations in consideration of the contexts to acquire correct translation candidates. In addition, we prepared utterance-by-utterance machine translations, without considering chat contexts to acquire incorrect translation candidates. Finally, we evaluated the translations to see if they were acceptable chat translations. The details of each process are described in the following.

Speaker	Utterance
person 1	I do not like carrots. I throw them away.
person 2	really. I can sing pitch perfect. *(incoherent: carrots → sing)*
person 1	I also cook, and I ride my bike to work. *(incoherent: sing → ride)*
person 2	great! I had won an award for spelling bee. *(incoherent: ride → spelling)*

Table 1: Example of incoherent chat from Persona-chat.

4.1 Base Datasets

We constructed Japanese–English bidirectional chat translation datasets. Specifically, we focused on Persona-chat (Zhang et al., 2018) and JPersona-chat (Sugiyama et al., 2021) as our base datasets. These datasets contain multiturn chat data in English and Japanese, respectively[2]. Each chat was performed between two crowd workers assuming artificial personas. The speakers discuss a given personality trait, including but not limited to self-introduction, hobby, and others.

4.2 Filtering Incoherent Data

A preliminary manual review of the Persona-chat dataset revealed occasionally incoherent chats, e.g., unnatural topic changes or misunderstandings (Table 1). We removed such examples from the dataset by asking crowd workers to flag passages they deemed incoherent. Here, we defined "incoherence" as questions being ignored, the presence of unnatural topic changes, one speaker not addressing what the other speaker said, responses appearing to be out of order or generally difficult to follow.

We scored each chat according to the workers' answers and selected the top 200 among $1,500$ chats[3]. The selected 200 chats were marked as accurate and coherent by at least seven of the 10 workers.

4.3 Bilingual Chats with Human Translations

To construct a parallel Japanese–English chat corpus, we combined the selected top 200 top chats ($2,940$ utterances in total) from the Persona-chat dataset and 250 chats ($2,740$ utterances in total) from the JPersona-chat dataset. We then translated them into their respective target languages[4].

[2]Persona-chat and JPersona-chat are not translations of each other.

[3]See Appendix C for additional details about the crowdsourcing process.

[4]We sought consent to translate JPersona-chat with the authors.

Speaker	Original utterance in Perosona-chat (en)	Translation by professional translators (ja)
person 1	Good evening, how has your day been?	こんばんは、今日はどうだった？
person 2	It was good I met up with some friends to larp	よかったよ、ライブRPGで友達と集まった。
person 1	I wish I had time for that, working 40 hours in a bank is killing me.	そんな時間があればなあ、銀行で４０時間勤務は死にそうだよ。
person 2	... ...	... ...

Table 2: Example of the top 200 coherent chats from the Persona-chat dataset as rated by crowdsourcing workers and translated to Japanese by professional translators.

Here, we commissioned professional translators proficient in Japanese and English to ensure high-quality translations. We asked the translators to consider both the accuracy of the translation and the coherence of the dialog. The translators were given information about the personas to help adjust the speaking styles. As a result, we obtained a parallel corpus of 450 dialogs (5,680 utterances) and their translations, which we refer to as the Bilingual Persona-chat (BPersona-chat) corpus. Table 2 shows a sample from the BPersona-chat corpus.

4.4 Bilingual Chats with Neural Machine Translation Translations

The task of the error detector is to distinguish between accurate and poor (potentially harmful) translations. The BPersona-chat corpus provides examples of the former. Given professionally-translated bilingual chats, we also prepared low-quality alternative translations generated using a machine translation model. Here, we trained a Transformer-based neural machine translation (NMT) model A on OpenSubtitles2018 (Lison et al., 2018), achieving a BLEU score (Papineni et al., 2002) of 4.9 on the BPersona-chat corpus[5]. Note that this BLEU score is relatively low because domain mismatch is possible between OpenSubtitles2018 and the BPersona-chat corpus. However, it was a preferable setting because we required poor translations to construct our dataset. In addition, we prepared better translations with a translation model B, which achieved a BLEU score of 26.4.

4.5 Human Evaluation of Translations

To confirm that the alternative translations generated by NMT model A were erroneous to the crowds, we asked crowd workers proficient in both English and Japanese to rate each translation in the chat as either good or bad. We qualified the workers to ensure they could reach the level of native Japanese, and the level of business and academic English.

The workers rated 5,088 of NMT model A's 5,680 (89.58%) translations, 1,718 of NMT model B's 5,680 (30.25%) translations, and 597 of the 5,680 (10.51%) human translations as bad[6]. Then, each utterance-translation pair was marked as erroneous or correct based on human evaluations.

According to our task settings, an utterance cannot be used as the referenced preceding context if none of it is correct. Thus, we deleted the 159 utterances whose human translations, model A's translations, and model B's translations were all erroneous. As a result, we obtained 2,674 English utterances with 8,022 corresponding labeled Japanese translations, where 3,406 of the translations were labeled as erroneous, and the remaining 4,616 translations were labeled as correct. In addition, we obtained 2,397 Japanese utterances with 7,190 corresponding labeled English translations, where 3,096 translations were labeled as erroneous, and 4,094 were labeled as correct. These labeled data were used to evaluate the error detector in our subsequent experiments.

5 Baseline Error Detecting Classifier

As a baseline approach, we trained and evaluated a binary BERT-based (Devlin et al., 2019; Wolf et al., 2020) classifier as the error detector[7]. Here, the input was structured as "ja_1 [SEP] en_1 [SEP] en_2 [SEP] ja_2" to predict the Japanese translation ja_2 of the corresponding source utterance en_2. The input was structured as "en_1 [SEP] ja_1 [SEP] ja_2 [SEP] en_2" to predict the translation en_2 of the corresponding source utterance ja_2 in the opposite translating direction[8].

[5]Refer to Appendix A for additional details about training NMT model A.

[6]Refer to AppendixC for additional details about the crowdsourcing process.

[7]Refer to AppendixB for additional details about training this classification model.

[8] [SEP] was used to indicate different utterances, [CLS] was used to indicate the beginning of the data and [PAD] was

	ja→en	en→ja
Majority class	56.94	57.54
Minority class	43.06	42.46
Error detector	76.27	77.06

Table 3: Accuracy of the majority class classifier, minority class classifier, and error detector.

	ja → en			en → ja		
	F	**(Pre**	**Rec)**	**F**	**(Pre**	**Rec)**
Error detector	73.30	(71.10	75.65)	75.03	(69.75	81.18)

Table 4: F-score, precision, and recall of the error detector on BPersona-chat dataset.

Similar to the original experimental settings for BERT, we applied the SoftMax function to the classification result to obtain the final prediction.

We used the OpenSubtitles2018 dataset for training with approximately one million utterances. Here, we generated negative samples with the low-quality translation model A (Section 4.4), and we fine-tuned the multilingual BERT model provided by HuggingFace[9] to construct the error detector for both the English-to-Japanese and Japanese-to-English directions.

6 Experiments

In this section, we report on our trial of the chat translation error detection task (Section 2) using the model described in Section 5. The task was evaluated with the dataset described in Section 4.

6.1 Evaluation Metrics

Majority class and minority class classifiers To confirm that the error detector is not simply making lucky guesses, we calculated the accuracy of the majority class classifier, the minority class classifier, and the error detector. Note that the majority class of the data is the correct translation, and the minority class is the erroneous translation.

F-score, precision and recall We evaluated the performance of the error detector according to the F-score (**F**). We also show the precision (**Pre**) and recall (**Rec**) values for reference. The truth (T) is set as the erroneous translation, and the positive case (P) is detecting the erroneous translation.

Confusion matrix To evaluated the performance of the error detector on different types of translations, we provide confusion matrices according to whether the translation was translated by the human translator, NMT model A, or NMT model B.

6.2 Results

The results demonstrate that the error detector is capable for classifying erroneous translations in chats. According to the accuracy values given in Table 3, we conclude that the error detector gained higher performance compared to the majority and minority classifiers. The results suggest that the current method can solve the task without relying on lucky guesses. According to the F-score, precision, and recall values shown in Table 4, the error detector could identify erroneous translations in the BPersona-chat dataset.

However, although the detector could distinguish translations with terrible translation or coherence issues, it could not successfully identify errors that were not obvious. The confusion matrix of the results is shown in Table 5, where the row headers are the actual annotations, and the column headers are the labels predicted by the detector. As can be seen, the error detector did not perform well when attempting to predict the translations generated by the high-quality NMT model B. Here, the detector labeled more than half of the erroneous translations generated by NMT model B as correct. One possible reason for this is that the detector was trained on a dataset whose erroneous examples were generated by model A, which generated low-quality translations.

To compare the error detector with the traditional BLEU calculation, we calculated the sentence-BLEU score of each utterance in the BPersona-chat dataset using the method provided by NLTK (Bird et al., 2009). The results demonstrate that the detector can help distinguish an erroneous translation even when the translation has a high BLEU score. Table 6 shows an example of a translation en_2 with a high sentence-BLEU score but incorrectly translated the Japanese word "米" into "America" rather than "rice". We found that the detector helped distinguish this case as erroneous, as was expected.

6.3 Quality of the Evaluation Dataset

The reason a considerably high score was obtained on the NMT model A's translations is not entirely straightforward. Note that we trained the classification model on OpenSubtitles2018, which has a different distribution from BPersona-chat. This

used as the padding token.

[9] https://huggingface.co/

ja→en								
	Human			**NMT model A (low-quality)**			**NMT model B (high-quality)**	
	Correct	**Erroneous**		**Correct**	**Erroneous**		**Correct**	**Erroneous**
Correct	1879	207	**Correct**	11	155	**Correct**	1252	590
Erroneous	290	21	**Erroneous**	90	2140	**Erroneous**	374	181
en→ja								
	Human			**NMT model A (low-quality)**			**NMT model B (high-quality)**	
	Correct	**Erroneous**		**Correct**	**Erroneous**		**Correct**	**Erroneous**
Correct	2406	176	**Correct**	6	265	**Correct**	1005	758
Erroneous	83	9	**Erroneous**	53	2350	**Erroneous**	505	406

Table 5: Confusion matrix of the error detector on BPersona-chat data (row headers are the actual annotations, and column headers are the prediction made by the detector).

en_1 (context)	What did you have for dinner?
ja_1	晩ご飯に何を食べましたか？
ja_2 (source)	晩ご飯に米を食べました。
en_2 **(translation)**	**I had America as my dinner.**
(reference)	*(I had rice as my dinner.)*
sentence-BLEU	72.7 (compared to the reference)
classifier's prediction	erroneous

Table 6: Example where the error detector successfully predicted the erroneous translation en_2 even though it had a high sentence-BLEU score.

means that the training was performed using out-of-domain data. One potential reason for the high performance may be attributed to the nature of the automatically generated translations. As with the experimental results described in Section 6.2, it was difficult for the detector to distinguish the good translations generated using the high-quality NMT model B. To improve performance, it is important to clarify the exact issue with the erroneous translation.

7 Discussions and Future Work

In this paper, we have proposed the chat translation error detection task to assist cross-lingual communication. For this purpose, we constructed a parallel Japanese–English chat corpus as the backbone for evaluation, including high-quality and low-quality translations augmented with crowdsourced quality ratings. We trained the error detector to identify erroneous translations, and the detector could help detect the erroneous translations in chat.

While this is the first trial to realize a cross-lingual chat assistance system, we hope to promote research to complete the chat translation assistance system in the future, and we aim to advance the detector's ability to indicate the translation's critical error possibility. This will allow speakers to focus on translations with high error rates. In addition, we hope to identify specific errors in the translations for users. To achieve this goal, we would like to refine the BPersona-chat dataset with multiple labels corresponding to different translation errors. The binary classification model would also be improved into multi-label, which would enable the error detector to analyze concrete problems. Thus, we would be able to identify the exact error in the current speech for revisions. We will also consider providing translation suggestions as reference information to help users modify.

When both parties cannot understand each other's language, the advanced error detecting system is expected to alert them of possible errors and guide them to modify their texts, thereby reducing translation problems in multilingual chats. Finding a balance between coherence and accuracy is always difficult in chat translation. However, we believe that advancing and refining the error detector and the corresponding dataset will help us identify and solve specific problems in chat translation systems.

Acknowledgements

This work was supported by JST (the establishment of university fellowships towards the creation of science technology innovation) Grant Number JPMJFS2102, JST CREST Grant Number JPMJCR20D2 and JST Moonshot R&D Grant Number JPMJMS2011 (fundamental research). The crowdsourcing was supported by Amazon Mechanical Turk (https://www.mturk.com/) and Crowdworks (https://crowdworks.jp/).

References

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the WMT 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796.

Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Modeling bilingual conversational characteristics for neural chat translation.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1742–1748.

Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 1–9.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2019. Designing the business conversation corpus. In *Proceedings of the 6th Workshop on Asian Translation*, pages 54–61, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu

Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based japanese chit-chat systems.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 2818–2826.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems 31 (NIPS 2017)*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too?

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 7: List of hyper-parameters for training the NMT model A

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 8: 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 A (Tiedemann and Scherrer, 2017), which outputs two consecutive given two input sentences to consider larger contexts. Table 7 shows the list of hyper-parameters.

B Settings of Classification Model

This section describes the details of the training classification model. Table 8 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.