BPersona-chat: A Coherence-Filtered Japanese–English Dialogue Corpus

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

Abstract

Researchers have focused on translating casual texts such as chats in different forms from formal texts in recent years. We strive to improve the accuracy of chat translation while obtaining smooth and natural translations; however, the parallel corpora used for chat translation models are still very limited. In this research, we translated the existing monolingual dialogue corpora, Persona-chat and JPersona-chat, to construct a Japanese–English dialogue corpus named **BPersona-chat**. To ensure the quality of the corpus, we filter out incoherent dialogues from the Persona-chat dataset via crowdsourcing. Finally, we applied BPersona-chat to classifiers that can judge whether a pair of chats is accurate and natural for evaluations.

1 Introduction

With the development of natural language processing technology, machine translation models have gained sound performances in translating official documents such as news, academic papers, and legal files for languages with abundant resources. In recent years, researchers have turned their attention to translating colloquial dialogues with the existing methods for document translation. However, it has been pointed out that the sentence-level system and the document-level system are not entirely qualified for translating chats due to the unique characteristics of chats such as multi-speakers or information omitted [1, 2]. Considering the differences between documents and dialogues, when translating chats, we also have to pay attention to the coherence of dialogues in addition to the correctness of words and grammar. Hence, we need to evaluate chat translation based not only on the traditional BLEU [3] score but also on the coherence and consistency in the flow of chats. Consequently, we built classifiers to evaluate the translation of chats to improve the performance of chat translation in our previous research [4].

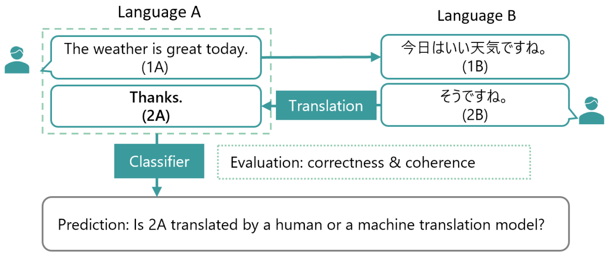

Figure 1 An example of evaluating a bilingual chat. The classifier is predicting the type of 2A referring to 1A, 1B and 2B.

In the previous research, we trained Japanese-English classification models to evaluate whether the translated response is accurate and coherent with respect to the chat flow. Figure 1 shows the system for translating and evaluating the translated response between two speakers using English and Japanese. In this system, 1B is the translation of 1A provided by a human, and 2A is the translation of 2B provided by a human or generated by a machine translation model. The classifier can predict the type of 2A with the reference data 1A, 1B, and 2B, which are from the parallel corpus.

To test the performance of our classifiers, we first applied the in-domain test data from **OpenSubtitles2018** [5, 6]. Nevertheless, the classifiers did not show strong agreement when predicting the human-translated data. Considering the quality and characteristics of OpenSubtitles2018, we decided to apply out-of-domain test data to check the performance of the classifiers.

Unfortunately, parallel corpora capable of chat translation are very limited. In particular, parallel corpora containing chats, such as OpenSubtitles, also contains numerous other types of data that are not suitable for evaluating chat translation. There are topical question-and-answer dialogue parallel corpora from past research, including data

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

with specific scenes and topics. Nevertheless, the topic of the dialogue is too strong to fit in our precondition of casual conversations. Therefore, to achieve our purpose, we decided to build a parallel dialogue corpus for evaluation in this research.

In order to solve this problem, we translated the existing monolingual corpus into bilingual to build a Japanese–English parallel dialogue corpus, named **BPersona-chat**. Since Persona-chat has noises, we selected understandable and coherent dialogues from the monolingual dataset through crowdsourcing for chat translation to ensure the quality.

To test the performance, we applied the **BPersona-chat** data to the classifiers. As a result, most human-translated data can be correctly recognized as coherent translations by the classifiers. The total accuracy is at most 95.57%, which is 12.65% higher than the highest accuracy from our previous results. The accuracy of the human-translated dialogue is at most 97.17%. These results show the performance and generality of our built classifiers. At the same time, they also show the correctness and coherence of the parallel data we built.

2 Related Work

To build an ideal parallel dialogue corpus containing high-quality chats, we surveyed existing dialogue corpora. We listed the existing corpora with a focus on their topics, domains, and languages.

BConTrasT and BMELD In the chat translation task of WMT2020[1)] [1], the organizers provided participants with an English-German parallel corpus, **BConTrasT**, containing the dialogue data only. The corpus is based on the **Taskmaster-1 corpus** [7], originally monolingual English language. It includes task-based dialogues in six domains, for example, ordering the pizza or making reservations. The organizers selected a subset of this dataset and translated it into German at the AI-powered Human-refined translation company, Unbabel[2)].

Similar to **BConTrasT**, the **BMELD** dataset [2] is based on the English dialogue dataset in the **MELD** [8]. The authors crawled the corresponding Chinese translations from **MELD** and then manually post-edited them according to the dialogue history of the native Chinese speakers.

1) https://www.statmt.org/wmt20/chat-task.html
2) https://unbabel.com/

speaker	utterance
person 1	i am going for a horse ride tomorrow. do you like horses?
person 2	i never have juice, just water.
person 1	is that hard for you? i love sugar
person 2	yes i do i work on the baby floor an i want no kids lol

Table 1 An example of an incoherent chat from Persona-chat [10].

Business Scene Dialogue Corpus The **Business Scene Dialogue (BSD)** [9] corpus is a Japanese-English business conversation corpus that includes half of the monolingual scenarios initially written in Japanese and the other half written initially in English.

Persona-chat and JPersona-chat The **Persona-chat** dataset [10] contains multi-turn dialogues conditioned on personas. Each dialogue was performed between two crowdsourcing workers assuming artificial personas. The persona given to each worker is described by three to five profile sentences, such as "I like to ski," "I am an artist," "I eat sardines for breakfast daily."

Similarly, the **JPersona-chat** dataset [11], which includes multi-turn conditioned on given personas. is collected in Japanese.

In existing parallel dialogue corpora, dialogue data in **BConTrasT** and **BSD** occurred in a specific topic scene, such as meal ordering or business negotiation. We found that some dialogues were similar in Q&A format or formal texts that did not meet our standard casual conversations. However, the need for casual conversation data in **BMELD** is mainly in Chinese, therefore unsuitable for the models we trained in previous research [4]. For our motivation, we believe that **Persona-chat** and **JPersona-chat** are the most appropriate to build new Japanese–English parallel dialogue corpora. Note that dialogues in both corpora do not have a set topic context despite having a set personality premise. Most of these speakers discussed a given personality trait, including but not limited to self-introduction, hobby, and others.

3 Methods and Experiments

3.1 Crowdsourcing

We found that Persona-chat contains low-quality conversational data when we manually checked them. These

This work is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

data have incoherent parts of dialogues, unnatural change of topics, misunderstandings in the foreword, leading to an inability to continue chatting. Table 1 shows an example of incoherent chat from Persona-chat. The noise will significantly impact our results since we want to construct a dialogue database that features a natural and smooth chat with translations. Hence, we prioritized rating Persona-chat data with crowdsourcing.

We expected to eliminate incoherent or unnatural conversations when rating the Persona-chat data for subsequent translation work finally. However, it is hard to define "incoherence" clearly due to the complexity of the dialogue. In this research, we opted to focus on the overall dialogue from macro vision instead of treating a tiny error as incoherent. We assumed that if there are incongruent connections that influence dialogue comprehension, dialogue is incoherent. To make the crowdsourcing task easier to understand, we informed the workers with the following rules:

We defined "not meshing well (incoherent)" as

- questions are ignored,
- there are unnatural topic changes,
- one is not addressing what the other said,
- responses seem out of order,
- or is hard to follow in general.

Minor issues (grammar or spelling errors) are acceptable when they do not affect chat flow.

Based on these criteria, we invited crowdsourcing workers to label incoherent chats. We chose Amazon Mechanical Turk as our platform for crowdsourcing. As the Persona-chat we wanted to filter is in English, we set the basic qualification types to confirm that they were native English speakers or had adequate English proficiency, living in an English environment for prolonged periods. Since dialogues are ambiguous and the benchmark rules of this experiment are subjective, we first performed a qualification round before conducting a full round of experiments. We excluded some workers whose criteria were outliers by comparing workers' scores. We also ensured that workers entering the full round had positive and effective feedback using the control question. In the full round, we selected 1,500 dialogue datasets from Persona-chat. For each crowdsourcing task, we gave five chats to ten workers. If a worker marks a chat as not-meshing, it is recorded as one point of negative comments; otherwise, it is recorded as one point of positive comments. Finally, we selected high-quality dialogue data from the top 200 conversations with the highest positive ratings. These 200 conversations were marked as good by at least seven of the ten workers.

3.2 Translating

We obtained the top-ranked 200 chats considered natural and smooth by crowdsourcing workers from the $1,500$ dialogues of Persona-chat.

The top 200 chats are coherent with easy-to-follow flows compared to those rated less. Table 2 shows one of the top 200 chats that were rated higher by crowdsourcing workers and translated by professional translators afterward. For constructing the Japanese–English bilingual corpus, we translated 200 chats from Persona-chat and 250 chats from JPersona-chat. We commissioned professional translators who are proficient in both Japanese and English to ensure the quality of the translation. To ensure that translators could take into account the correctness of translation and the coherence of dialogue, we put the following precautions for translators.

First, we asked translators to translate the chats based on the personas (profile sentences) to ensure the tone and role preference was similar to the original utterances. Secondly, considering the characteristics of the Japanese language, we allowed translators to modify the translated dialogue in English to keep it remain fluent and natural. For example, they could append subjects and phrases, or change the tone of sentences, as shown in Table 3. Finally, we requested translators to avoid translationese. Translators could choose appropriate English words instead of direct transliteration when encountering specific Japanese words. For example, "サラサラした髪" can be translated as "smooth hair". Same for translation from English to Japanese. As a result, we obtain a parallel corpus with 450 dialogues, named **BPersona-chat**. In total, there are $5,708$ utterances.

4 Results and Analysis

In the previous research, when applying classifiers on the test data extracting from **OpenSubtitles2018**, the classifiers could not correctly predict 2A that were taken from the corpus, which were supposed to be translated by human translators [4]. We consider the behavior is possibly related to the low quality of OpenSubtitles2018. Accord-

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

speaker	utterance (en)	utterance (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	yikes, you have to make time for friends and fun.	うわっ、友達と趣味の時間作らなきゃ。
person 1	i know but i am so focused on doing a good job that i forget to.	そうなんけど、いい仕事をすることに必死で忘れるんだ。
person 2	...	...

Table 2 An example of the top 200 coherent chat from Persona-chat, rated by crowdsourcing workers.

person	origin (ja)	translation (en)
person 1	将来は占い師になりたいと思っています。	I want to be a fortune-teller in the future.
person 2	占い師さんになりたいのですね。頑張ればきっと叶いますよ！	**I see**, you want to be a fortune-teller**?** If you do your best, it will surely come true!

Table 3 An example of adding sentences and changing tones when translating the original Japanese dialogue to English.

classifier	OpenSubtitles2018	BPersona-chat
2B-2A	0.8223	0.9464
1A-2A	0.7431	0.9525
1A-2B-2A	0.8238	0.9534
1A-1B-2B-2A	0.8292	0.9557

Table 4 Accuracy of classifiers on predicting whether 2A is model-translated or human-translated with two datasets.

ing to our previous research, data from OpenSubtitles2018 might contain utterances in one pair that is not a chat but a speech; there might be just a single speaker instead of two speakers or multiple speakers. In addition, the utterances may not initially be in Japanese or English. This is because the OpenSubtitles2018 is a corpus of multi-lingual movie subtitles. Data in OpenSubtitles2018 does not have to come from English movies or Japanese movies. Furthermore, the subtitles are collected using the OpenSubtitles website[3)], which means the subtitles do not have to be translated by professional translators. Considering the above reasons, the low quality of OpenSubtitles2018 may influence the test results.

Regarding the quality and contents of OpenSubtitles2018, we applied BPersona-chat to the classifiers to confirm the performance. As the classifiers can only predict a pair of utterances instead of the full dialogue, we split each dialogue into 5,229 pairs of two utterances.

The prediction results on **BPersona-chat** are shown in Table 4. Compared to the highest accuracy 82.92% in the previous research, the highest accuracy with **BPersona-chat** is 95.57%, which is 12.65% higher.

3) https://www.opensubtitles.org/en/search/subs

With respect to the accuracy for the human-translated label, the highest accuracy with BPersona-chat (97.17%) significantly outperformed that with OpenSubtitles2018 (72.77%). Also, the highest accuracy for model-translated label with BPersona-chat (95.70%) has slightly higher accuracy than that with OpenSubtitles2018 (93.38%).

Overall, BPersona-chat can be used for evaluating Japanese–English chat translation systems as out-of-domain data. The classifiers we have created before can gain good results with BPersona-chat on predicting the human-translated chats.

5 Conclusion and Future Work

In conclusion, we built a Japanese–English parallel dialogue corpus, **BPersona-chat**. The BPersona-chat is translated by professional translators based on Persona-chat and JPersona-chat. Compared to task-oriented dialogue datasets, such as BConTrast and Business Scene Dialogue, the BPersona-chat is a chit-chat dialogue corpus containing colloquial chats in Japanese and English. To ensure the chats from Persona-chat are high-quality casual chats, we evaluated 1,500 chats from it and picked the top 200 chats via crowdsourcing. Finally, we applied the data to the classifiers we had built before. Compared to our previous results, we gained a significant improvement on predicting the human-translated data with BPersona-chat.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP19H04425 and JP20J21694.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

References

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

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

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

[4] Yunmeng Li, Ryo Fujii, Makoto Morishita, Jun Suzuki, and Inui Kentaro. Towards detecting errors: Classifying model-generated output in chat translation. In **Proceedings of NLP2021**, pp. A3–4, 2021.

[5] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 923–929, 2016.

[6] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 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)**, pp. 1742–1748, 2018.

[7] Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. Taskmaster-1: Toward a realistic and diverse dialog dataset, 2019.

[8] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations, 2019.

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

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

[11] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chit-chat systems, 2021.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

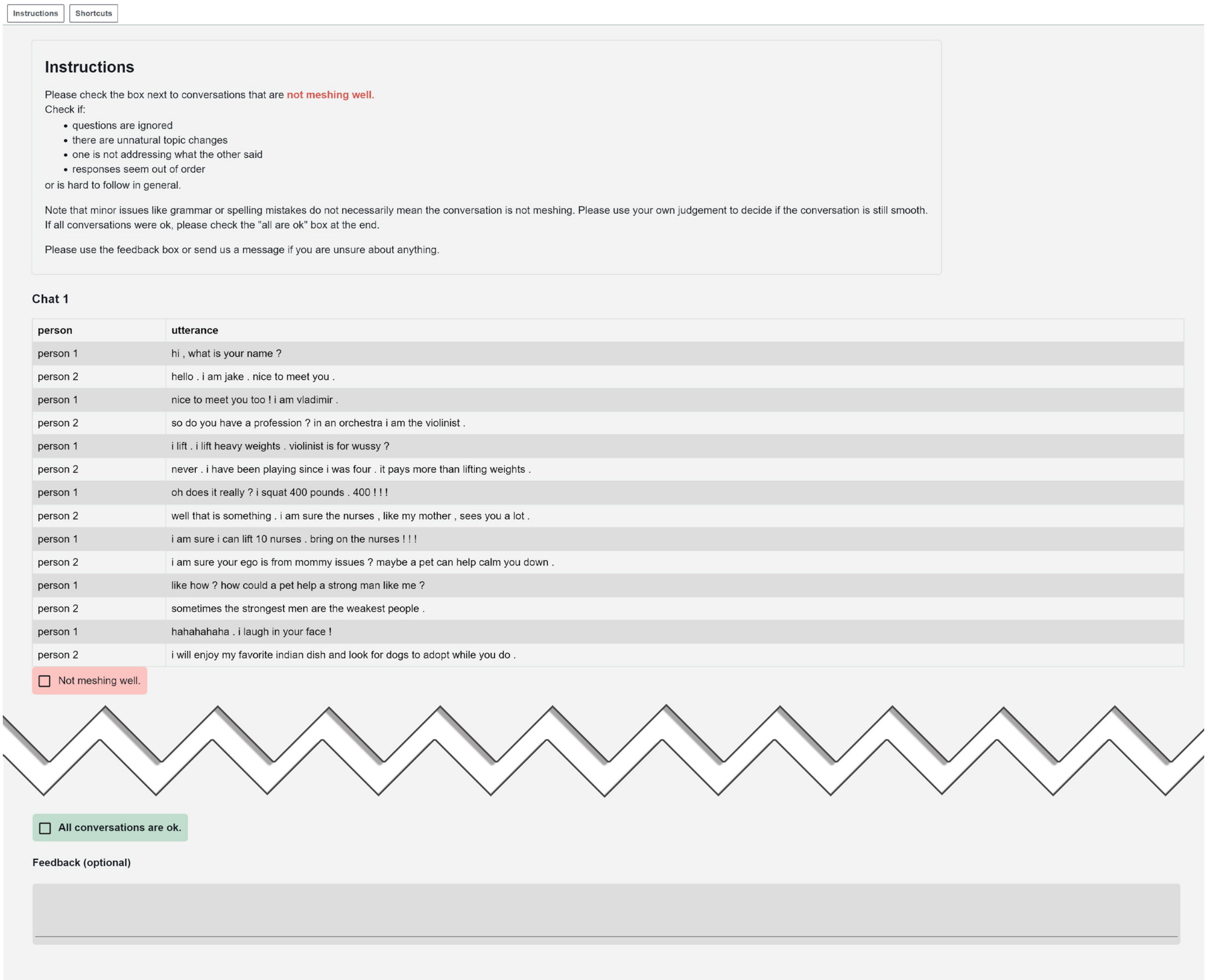

Figure 2 A preview of Amazon Mechanical Turk Working Screen. There are five chats in total.

A Appendix

A.1 Detail of Crowdsourcing

Figure 2 shows the working screen of Amazon Mechanical Turk workers. The instruction of the task is shown at the top of the page. In total, there are five chats per assignment. Each chat has a checkbox question at the bottom. If the worker thinks the chat is not meshing well, she or he can tick on the checkbox. At the bottom of the page, there is a check question to check the validity of answers. In the end, workers can write down their advice through the feedback box.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).