AoGu: A Japanese-English literary parallel corpus from Aozora Bunko and Project Gutenberg

Guanyu Ouyang[1] Xiaotian Wang[1] Takehito Utsuro[1] Masaaki Nagata[2]
[1]Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
[1]s2420829_@_u.tsukuba.ac.jp [1]s2320811_@_u.tsukuba.ac.jp
[1]utsuro_@_iit.tsukuba.ac.jp [2]masaaki.nagata_@_ntt.com

Abstract

This paper introduces a Japanese-English parallel corpus composed of literary works, constructed mainly using bilingual texts from Aozora Bunko and Project Gutenberg. Existing Japanese-English parallel datasets, such as JParaCrawl, JaParaPat, and ASPEC [1, 2, 3], offering coverage of common, patent, and academic domains, they lack resources specifically designed to address discourse-level phenomena and context-aware translation challenges which are existed in literary translation task. To bridge this gap, we build upon the "English-Japanese Translation Alignment Data" [1)] developed over a decade ago, updating and expanding it to better support research in discourse-level literary translation and document-level context modeling. Baseline experiments with transformer models on the constructed dataset demonstrate limited performance, highlighting the inherent challenges of literary translation and underscoring the need for more advanced methodologies and resources to enhance translation quality for literary texts.

1 Introduction

Neural Machine Translation (NMT) has advanced significantly in recent years, driven by innovations in neural architectures and the availability of large-scale parallel corpora. While these developments have greatly improved general translation tasks, literary translation presents unique challenges. It demands capturing nuanced semantic meanings and addressing complex discourse-level phenomena, such as pronoun resolution, inter-sentential consistency, and topic coherence [4, 5, 6, 7]. Traditional MT models often struggle with these aspects, resulting in translations that lack stylistic fidelity, contextual awareness, and narrative coherence. To address these issues, researchers have increasingly turned to context-aware and document-level translation approaches that incorporate broader contextual information into the translation process [8, 5].

Lin et al. [9] noted that the poor performance of context-aware MT models often stems not from their inability to handle long-distance dependencies but from the sparsity of discourse-level phenomena in existing datasets. This underscores the critical need for datasets that include such complex linguistic features, alongside advancements in translation models. Meanwhile, recent studies [8, 5] have highlighted literary translation as an ideal testbed for advancing context-aware MT, given the inherent complexity and abundance of discourse-level phenomena in literary texts.

However, resources for Japanese-English literary translation remain scarce. The only existing dataset, the "English-Japanese Translation Alignment Data" [10], was developed over a decade ago and lacks the scale and depth required for modern research. To address this gap, this study builds upon and significantly expands the existing dataset, providing a more comprehensive resource for Japanese-English literary translation. The updated dataset aims to better support research into context-aware and document-level translation methods for Japanese-English language pair.

2 Related Works

Jin et al. [9] developed a paragraph-aligned Chinese-English dataset containing 10,545 parallel paragraphs extracted from six public-domain novels. This dataset aims to

1) https://att-astrec.nict.go.jp/member/mutiyama/align/index.html

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

promote research into paragraph-level context-aware MT.

Thai et al. [5] introduced Par3, a multilingual dataset of 121,385 paragraphs from public-domain novels, . Despite its broad scope, the Japanese-English portion remains small, with only 1,857 paragraphs with averaging 4.4 sentences per paragraph(~8,170 sentences).

Jin et al. [11] constructed a large Chinese-English dataset with 5,373 paragraphs, consisting of 548.5K English and 700.9K Chinese sentences. They proposed the challenging chapter-to-chapter (Ch2Ch) translation setting, which showcases the importance of datasets reflecting complex discourse phenomena for literary texts.

Jiang et al. [12] extended the existing BWB[13] corpus with 15,095 discourse-level annotations across 80 documents (~150K words) to better explore the literary MT.

3 Dataset

3.1 Aozora Bunko

Founded in 1997, Aozora Bunko[2)] is a digital library providing access to a vast array of public domain works, with a current collection exceeding 17000 items. Moreover, literary works dominate the collection, accounting for approximately 72.4% of the total, with 15,696 titles categorized under this genre alone.

3.2 Project Gutenberg

Project Gutenberg[3)], established in 1971 by Michael S. Hart, is the first large-scale digital library dedicated to providing free access to public domain works. It offers over 60,000 texts across genres such as literature, philosophy, history, and science. A notable feature is its collection of professionally translated texts, which ensures high-quality translations for research and linguistic analysis.

3.3 Dataset construction

The main process of dataset construction, as shown in Figure 1, consists of four key steps: document alignment, text preprocessing, paragraph alignment, and sentence alignment.

3.3.1 Document alignment

A random inspection of works from Aozora Bunko and Project Gutenberg (English works) revealed notable differences in their textual characteristics. Most works in Aozora Bunko are partial chapters of novels, individual pieces from collections, or excerpts chosen based on the translator's preferences, rather than complete works. In contrast, most works in Project Gutenberg are complete novels or fully compiled series. This highlights that potential parallel document pairs often differ significantly in content, with an single Aozora Bunko work typically aligning to only a small portion of a single Project Gutenberg work. Based on this observation, rather than relying on traditional semantic text similarity methods for mining parallel document pairs, we leveraged the capabilities of pre-trained large-scale language models, specifically GPT-4o[4)] and Claude-3.5-Sonnet[5)], to assist in document alignment.

We adopt a 2-stage approach:

1. : For each work in Aozora Bunko, we extract the first 3–5 lines of the text, which typically include the title of the work, the original author's name, and the translator's name. We define a pre-trained model as a retrieve-agent, Using a predefined prompt, we aim for the retrieve-agent to provide the English title of the chapter, the potential associated work title, and the original author's name in English. The details of the prompt are shown in the Table 5 in Appendix B. Then we implemented an automated script to perform a global character-level match across all metadata of English works in Project Gutenberg using the retrieval information provided by the retrieve-agent. For cases where the retrieve-agent returns "No match" or there are no matching results in Project Gutenberg, we defined a RAG-agent, we first eliminates Japanese works for which they have matched English works. For the remaining Japanese works, we also request retrieval information from the retrieve-agent. If no matches are found, the RAG-agent extracts the first three and last three lines of the text body of Japanese work and sends an updated query to the retrieve-agent. The RAG-agent works to a maximum of three iterations for each Japanese work. The implementation involves RAG-agent module are based on the multi-agent open-

2) https://www.aozora.gr.jp/
3) https://www.gutenberg.org/
4) https://openai.com/index/gpt-4o-system-card/
5) https://www.anthropic.com/news/claude-3-5-sonnet

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

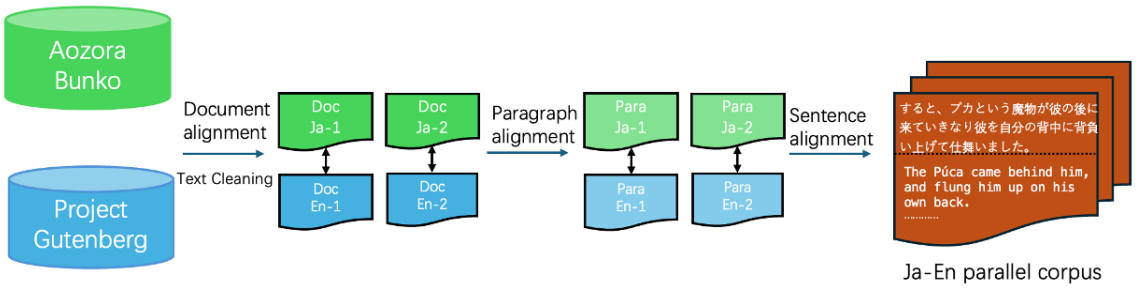

Figure 1 Pipeline of constructing the corpus

source framework AutoGen[14].

2. : We manually reviewed each parallel document obtained from Stage 1, labeled specific chapters in the English works that correspond to the Japanese works, and removed all non-parallel pairs as well as non-English documents from Project Gutenberg. As a result, we obtained a total of 632 parallel document pairs.

3.3.2 Text cleaning

For the Japanese works, we removed the header descriptions and symbol explanations, eliminated phonetic annotations (such as kana readings and kanji readings), deleted input annotations and special character marks, and removed copyright information at the bottom. Additionally, we replaced the iteration mark"／＼" with the vertical kana repeat mark (U+3031) and replaced "／″ ＼" with the vertical kana repeat with voiced sound mark(U+3032).

For the English works, we removed all illustration tags and all annotation information.

3.3.3 Paragraph alignment

Using the labeling information from Stage 2 of document alignment, we extracted paragraphs from the English documents. The final parallel paragraphs consist of the original documents of the Japanese works and the corresponding chapters from the English documents.

3.3.4 Sentence alignment

In the presence of irregular line breaks within the text, including intra-sentence line breaks, we merged all lines within each paragraph for both English and Japanese works. Subsequently, we applied the sat-12l-sm model[15] from wtpsplit [16] to perform sentence segmentation on the merged paragraphs, setting a threshold of 0.01 to achieve finer-grained sentence segmentation. Because we aim to use Vecalign [17] to achieve a more reasonable granularity of parallel sentences.

For all segmented sentences, Vecalign was utilized to perform sentence alignment across all parallel paragraphs. The parameters were configured with an overlap size of 12 and a maximum allowable number of merged sub-sentences set to 12. The embedding models employed included the LaBSE model [18] and the LASER2 model [19].

3.4 Dataset statistics

We completed sentence alignment for 513 out of the 632 parallel documents. For sentence embedding, we employed both the LaBSE and LASER2 models. Table 1 presents detailed statistics of the sentence-level datasets initially constructed using these two embedding models. To compute the number of subwords, the tokenizer from the LaBSE model was utilized.

Table 1 Statistics of AoGu and Utiyama's dataset. #subword refers to the total number of subwords, #sent refers to the total number of sentence pairs.,#doc refers to the total number of document pairs

Embedding Model	#subword (Japanese)	#subword (English)	#sent	#doc	#subword/sent (Japanese)	#subword/sent (English)	#sent/doc
LaBSE	9.73M	7.37M	292,298	513	33.3	25.2	569.8
LASER2	9.72M	7.16M	311,265	513	31.2	23.0	606.8
Utiyama's dataset	2.44M	1.72M	109,431	160	22.3	15.8	683.9

In 2003, Masao Utiyama et al.developed a Japanese-English parallel corpus[6)] , aligned at the sentence level, utilizing resources from Aozora Bunko, Project Gutenberg, and Project Sugita Genpaku,et al. This corpus is primarily composed of literary works and poetry, encompassing a total of 160 documents in both Japanese and English. AoGu was built upon this foundation and further updated and expanded. To compare the specific differences, The rows of Utiyama's dataset in Table 1 presents the statistical information of the dataset developed by Masao Utiyama et al.

6) https://att-astrec.nict.go.jp/member/mutiyama/align

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

4 Baseline Experiment

We sampled the two datasets obtained using LaBSE and LASER2 with the LaBSE model, setting up two sampling groups with thresholds of 0.4 and 0.6. Four 6-layer transformer baseline models were trained on the sentence-level dataset using Fairseq [20]. The specific parameter settings are as follows: the Adam optimizer was used, with a label smoothing value of 0.1, a dropout rate of 0.3, an initial learning rate of 4e-4, 3000 warm-up update steps, a maximum of 6144 tokens per batch, an update frequency of 4, and a total of 50 epochs.For evaluation, the BLEU [21] and COMET [22] metrics were adopted, with a beam search size of 4. The COMET model used is wmt22-comet-da[23]. The specific results are shown in Table 2. All experiments are conducted on three A6000 GPUs.

Table 2 The baseline of the sentence-level dataset for 4 different configuration

Method	Dataset Size			Metrics	
	Train	Valid	Test	COMET	BLEU
Vecalign (LaBSE) + LaBSE sampling (>0.4)	260,802	13,041	13,041	0.683	8.08
Vecalign (LaBSE) + LaBSE sampling (>0.6)	201,083	10,055	10,055	0.688	8.18
Vecalign (LASER2) + LaBSE sampling (>0.4)	272,812	13,640	13,640	0.680	11.83
Vecalign (LASER2) + LaBSE sampling (>0.6)	224,702	11,235	11,235	0.685	11.64

From the Table 2, it can be observed that the BLEU scores for the four baseline settings are relatively low, while the COMET scores are comparatively higher. The result demonstrates that the baseline model still has significant room for improvement in its understanding of literary texts at the sentence level.

We also conducted testing on the out-domain ASPEC dataset, and the results are shown in Table 3. The results indicate that the model trained on literary sentence-level data has significantly limited generalization ability, highlighting the substantial differences in characteristics between literary and non-literary texts.

5 Discussion

The "document pairs" in this paper are defined as (Japanese source document - English source document, where the Japanese document corresponds to only part of the English document). "Paragraph pairs" refer to (Japanese source document - corresponding English sub-

Table 3 The baseline settings tested on out-domain ASPEC test set

Method	Dataset Size	Metrics	
	Test	COMET	BLEU
Vecalign (LaBSE) + LaBSE sampling (>0.4)	1,808	0.534	2.4
Vecalign (LaBSE) + LaBSE sampling (>0.6)	1,808	0.518	2.8
Vecalign (LASER2) + LaBSE sampling (>0.4)	1,808	0.539	2.24
Vecalign (LASER2) + LaBSE sampling (>0.6)	1,808	0.529	2.21

paragraph). Currently, only sentence-level alignment has been completed, as paragraph-level alignment, influenced by subjective factors, has not yet been performed due to significant differences in paragraph division between Japanese and English texts.

Furthermore, the use of Vecalign introduces a penalty parameter that may cause contextually continuous sentences to be split in the alignment results. The baseline model is trained in the scenario of single-sentence translation without contextual information. Table 4 in Appendix A presents four examples and detailed case analysis under the Vecalign (LASER2) + LaBSE sampling with similarity > 0.4 setting. These cases reveal that the baseline model trained at the sentence level demonstrates limited capabilities in pronoun resolution, modeling complex semantic relationships, and capturing the stylistic and contextual nuances of literary texts. These limitations underscore the need for more advanced approaches, such as paragraph-level or context-aware training, to enhance the model's performance in literary translation tasks.

Future work will focus on exploring literary translation tasks in context-aware settings, and alignment will be conducted at the paragraph level, accompanied by a more refined approach to sentence-level alignment.

6 Conclusion

This paper introduces a parallel Japanese-English literary corpus, detailing its development process and statistical information. The baseline experimental results demonstrate that literary machine translation tasks impose higher demands on translation models in terms of context awareness, complex semantic relationship modeling, and contextual coherence.

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

References

[1] M. Morishita, K. Chousa, J. Suzuki, and M. Nagata. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In **Proc. 13th LREC**, pp. 6704–6710, 2022.

[2] M. Nagata, M. Morishita, K. Chousa, and N. Yasuda. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In **Proc. LREC-COLING**, pp. 9452–9462, May 2024.

[3] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara. ASPEC: Asian scientific paper excerpt corpus. In **Proc. 10th LREC**, pp. 2204–2208, 2016.

[4] E. Matusov. The challenges of using neural machine translation for literature. In **Proc. the Qualities of Literary Machine Translation**, pp. 10–19, 2019.

[5] K. Thai, M. Karpinska, K. Krishna, B. Ray, M. Inghilleri, J. Wieting, and M. Iyyer. Exploring document-level literary machine translation with parallel paragraphs from world literature. In **Proc. EMNLP**, pp. 9882–9902, 2022.

[6] M. Fonteyne, A. Tezcan, and L. Macken. Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In **Proc. 12th LREC**, 2020.

[7] Y. Liu, Y. Yao, R. Zhan, Y. Lin, and D. Wong. NovelTrans: System for WMT24 discourse-level literary translation. In **Proc. 9th WMT**, pp. 980–986, 2024.

[8] K. Marzena and I. Mohit. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In **Proc. 8th WMT**, pp. 419–451, 2023.

[9] J. Lin, J. He, J. May, and X. Ma. Challenges in context-aware neural machine translation. In **Proc. EMNLP**, p. 15246–15263, 2023.

[10] Utiyama M. and Takahashi M. English-japanese translation alignment data., 2003.

[11] L. Jin, Li A., and X. Ma. Towards chapter-to-chapter context-aware literary translation via large language models, 2024.

[12] Y. Jiang, T. Liu, S. Ma, D. Zhang, M. Sachan, and R. Cotterell. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In **Proc. 61st ACL**, pp. 7853–7872, 2023.

[13] Y. Jiang, T. Liu, S. Ma, D. Zhang, J. Yang, H. Huang, R. Sennrich, R. Cotterell, M. Sachan, and M. Zhou. BlonDe: An automatic evaluation metric for document-level machine translation. In **Proc. NAACL**, pp. 1550–1565, 2022.

[14] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. Awadallah, R. White, D. Burger, and C. Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In **Proc. 1st COLM**, 2024.

[15] M. Frohmann, I. Sterner, I. Vulić, B. Minixhofer, and M. Schedl. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In **Proc. EMNLP**, pp. 11908–11941, 2024.

[16] B. Minixhofer, J. Pfeiffer, and I. Vulić. Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation. In **Proc. 61st ACL**, pp. 7215–7235, 2023.

[17] B. Thompson and P. Koehn. Vecalign: Improved sentence alignment in linear time and space. In **Proc. EMNLP and 9th IJCNLP**, pp. 1342–1348, 2019.

[18] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic BERT sentence embedding. In **Proc. 60th ACL**, pp. 878–891, 2022.

[19] K. Heffernan, O. Çelebi, and H. Schwenk. Bitext mining using distilled sentence representations for low-resource languages. In **Findings of EMNLP**, pp. 2101–2112, 2022.

[20] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A fast, extensible toolkit for sequence modeling. In **Proc. NAACL**, pp. 48–53, 2019.

[21] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proc. 40th ACL**, pp. 311–318, 2002.

[22] R. Rei, C. Stewart, A. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation. In **Proc. EMNLP**, pp. 2685–2702, 2020.

[23] R. Rei, J. C. de Souza, D. Alves, C. Zerva, A. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In **Proc. 7th WMT**, pp. 578–585, 2022.

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

A Error Analysis Of Vecalign (LASER2) + LaBSE sampling with similarity $>$ 0.4 setting

The Table 4 shows four specific translation output compared between the Reference and Hypothesis.

For case 1， The source sentence reflects the speaker's perspective (Gryde speaking), whereas the reference adopts the listener's perspective (people listening). The model maintained the source's perspective. Additionally, "夢中になって" can be ambiguous, describing either the speaker's state (chosen by the model) or the listener's state (chosen by the reference).

In case 2, the source text uses "手紙" (letter) as the pronoun, and the reference preserves "letter" in the same role. However, the model replaces it with "he," altering the original perspective. This demonstrates the model's insufficient understanding of contextual coherence.

For case 3, the model failed to handle pronouns correctly, and compared to the model's direct translation"put his foot to my house twice," the reference translation leans more toward a free translation: "you would never have put another foot." .Additionally, the reference tends to use the free translation rather than direct translation: "そいつぁ間違えっこなしだ。" -> "you may lay to that ."

For case 4, the reference translation's sentence structures are more diverse, reflecting the characteristics of literary texts, whereas the model's translation tends to adhere closely to the sentence structure of the source text.

Table 4 Cases for Vecalign (LASER2) + LaBSE sampling with similarity >0.4 settings

#	Metrics	Source	Hypothesis	Reference
1	BLEU = 41.80 COMET = 0.674	二十分間グライドは夢中になって喋った。	"II For twenty minutes Gryde was talking wildly."	"For twenty minutes Gryde was followed with rapt attention."
2	BLEU = 7.24 COMET = 0.674	ここまでは手紙はすこぶる落着いて書いてあったが、ここでペンが急に走り書きになって、筆者の感情が抑え切れなくなっていた。「	Up to this he had written a very quiet note , but here he scribbled a note , and the writer 's feelings relaxed .	So far the letter had run composedly enough, but here with a sudden splutter of the pen, the writer's emotion had broken loose
3	BLEU = 4.72 COMET = 0.681	「もしあんなような奴とつきあってたんなら、二度と己の家へ足を入れさすんじゃなかったぞ。そいつぁ間違えっこなしだ。	"If he had met such a fellow, he wouldn't have put his foot to my house twice, he would have been mistaken."	" If you had been mixed up with the like of that , you would never have put another foot in my house , you may lay to that .
4	BLEU = 9.85 COMET = 0.790	彼の考えそのものが間違いなのか、それとも彼は今、謎の核心へと導かれているのだろうか。」私はひとり考えた。	Was his thoughts doubtless mistaken, or he now led to the point of the mystery?" I thought.	"Either his whole theory is incorrect," I thought to myself, "or else he will be led now to the heart of the mystery."

B Prompt Setting

Table 5 The prompt for retrieve-agent

You are now a distinguished scholar of world literature, with a particular expertise in both Japanese and English literature.

Task:

I will provide you with the name of an author in Japanese and the title of their work in Japanese. Your task is to:

1. Identify the English name of the author.
2. Provide the corresponding English title for the work.
3. If the provided title represents a chapter or section of a larger work, also provide the title of the larger work to which it belongs.
4. If there is no match for one work, please just return "No match".
5. If you are not confident with the result, please list all possible result in each "Author", "Chapter Title" and "Parent Work Title" section.
6. You are also supported by a RAG-agent, in the case I sent the extra content of works, please using this information to further identify.

Guidelines:

Carefully analyze each input to determine whether the given title is a standalone work or part of a larger collection.

Provide accurate and internationally recognized English titles wherever possible.

Always follow the format demonstrated in the example below.

Example:

Q:

アーヴィング ワシントン

ウェストミンスター寺院

A:

Author: Irving, Washington

Chapter Title: Westminster Abbey

Parent Work Title: *The Sketch Book of Geoffrey Crayon, Gent.*

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