Developing a Japanese-English Literary Parallel Corpus from Aozora Bunko and Project Gutenberg: AoGu

〇Guanyu Ouyang[1]　Xiaotian Wang[1]　Takehito Utsuro[1]　Masaaki Nagata[2]

1. 筑波大学大学院 システム情報工学研究群
2. NTTコミュニケーション科学基礎研究所

- Background
- Related Works
- The Proposal
- Statistics and Baseline Experiment
- Translation case analysis
- Conclusion and future works

Category

- Background
- Related Works
- The Proposal
- Statistics and Baseline Experiment
- Translation case analysis
- Conclusion and future works

Background

- Compared to general machine translation tasks, literary translation presents unique challenges.

- It demands addressing complex discourse-level phenomena.
  -Pronoun resolution, inter-sentential consistency, and topic coherence. [1,2,3,4]

[1] E. Matusov. The challenges of using neural machine translation for literature. In Proc. the Qualities of Literary Machine Translation, pp. 10–19, 2019.
[2] K. Thai, M. Karpinska, K. Krishna, B. Ray, M. Inghil- leri, J. Wieting, and M. Iyyer. Exploring document-level literary machine translation with parallel paragraphs from world literature. In Proc. EMNLP, pp. 9882–9902, 2022.
[3] 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.
[4] 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.

Background

- Some research turned to context-aware and document-level translation approaches.
  - -Incorporate broader contextual information into the translation process.[2,5]

- And highlighted that literary translation as an ideal testbed for advancing context-aware MT.
  - -Inherent complexity and abundance of discourse-level phenomena in literary texts. [2,5]

[5] 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.

Background

- Lin et al. [6] noted that the **poor performance** of context-aware MT models
  -Often not relate to **the capability of model.**
  -But from the **sparsity of discourse-level phenomena in existing datasets.**

This underscores the critical need for datasets that include such complex linguistic features.

- However, resources for Japanese-English literary translation remain scarce.

- The only related existing dataset, the "English-Japanese Translation Alignment Data" [7]

[6] J. Lin, J. He, J. May, and X. Ma. Challenges in context- aware neural machine translation. In Proc. EMNLP, p. 15246–15263, 2023.
[7] Utiyama M. and Takahashi M. English-Japanese translation alignment data., 2003.

The purpose of this research

- Aims to resolve limitations in sentence-level aligned corpora.
  -Which miss crucial discourse-level information for document, context-aware machine translation.

- Aims to build an open literary domain parallel dataset.
  - Promoting the literary translation and context-aware MT of Japanese-English language pair.

- Background
- Related Works
- The Proposal
- Statistics and Baseline Experiment
- Translation case analysis
- Conclusion and future works

Existing datasets [Lin+,EMNLP2023],[Thai+,EMNLP2022],[Jin+,arxiv2024],[Jiang+,ACL2023]

Recent Literary Parallel Corpora

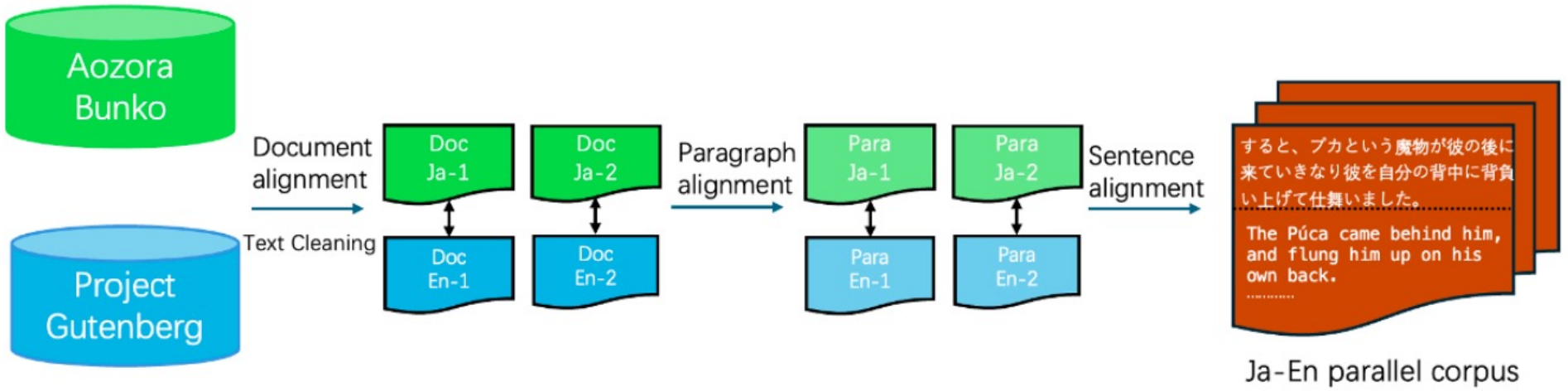
- **Jin et al. (2023)**: paragraph-aligned **Chinese-English** dataset with **10,545** parallel paragraphs from **six public-domain novels.**
- **Thai et al. (2022)**: **multilingual dataset from public-domain novels**. Japanese-English portion is **relatively small.**
- **Jin et al. (2024)**: **Chinese-English** dataset containing **5,373** paragraphs from several **open-source novels.**
- **Jiang et al. (2023)**: **Chinese-English** corpus with **15,095** discourse-level annotations across **80 documents (~150K words).**

📌

- Most literary parallel corpus are in Chinese-English language pair, the available source for Japanese-English remain limited.

- Background
- Related Works
- The Proposal
- Statistics and Baseline Experiment
- Translation case analysis
- Conclusion and future works

3-stage corpus construction pipeline



0. Preprocessing and cleaning

📝 **Japanese Works**

Cleaned information includes **header descriptions, symbol explanations, and phonetic annotations**.

🔷 **Examples:**

- **行右小書き注釈 (Small inline annotations)**
  - （1）［＃「(1)」は行右小書き］ -> 空
  - （*2）［＃「(*2)」は行右小書き］ -> 空
- **二倍踊り字 (Repetition marks)**
  - ／＼ -> くの字点く(U+3031)
  - ／″＼ -> （帯浊点）ぐ(U+3032)
- **外字マーク (External character marks)**
  - ※[「てへん＋垂」、168-14] → □
  - ※［＃「鼾のへん＋嗅のつくり」、第4水準2-94-73］《かぎ》 -> かぎ
- **注音記号 (Phonetic symbols)**
  - 確《しっ》かり -> 確かり
  - 十｜片《ペニー》 -> 十片
  - 確《かく》実《じつ》 -> 確実

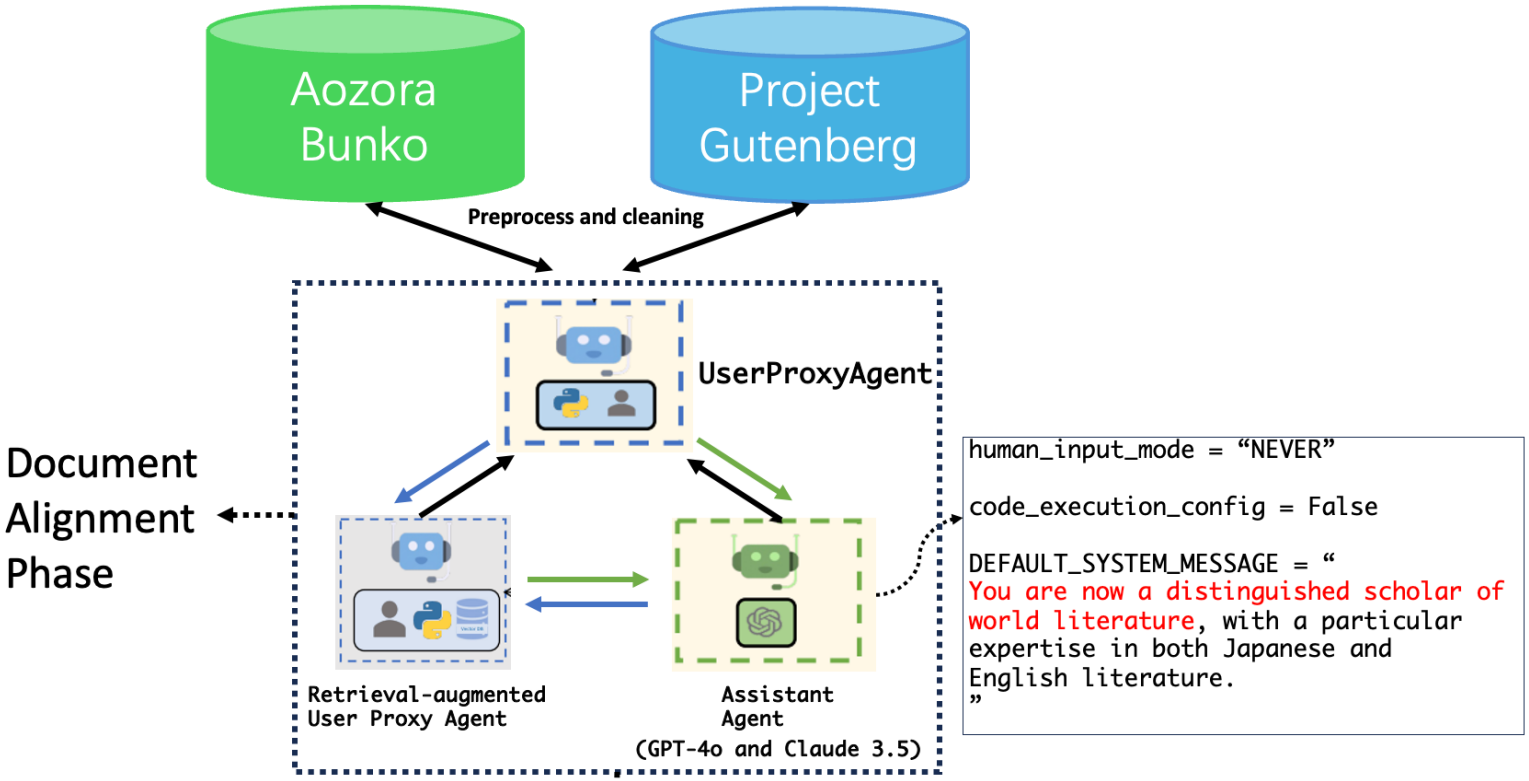
📖 **English Works**

All illustration tags and annotation information.

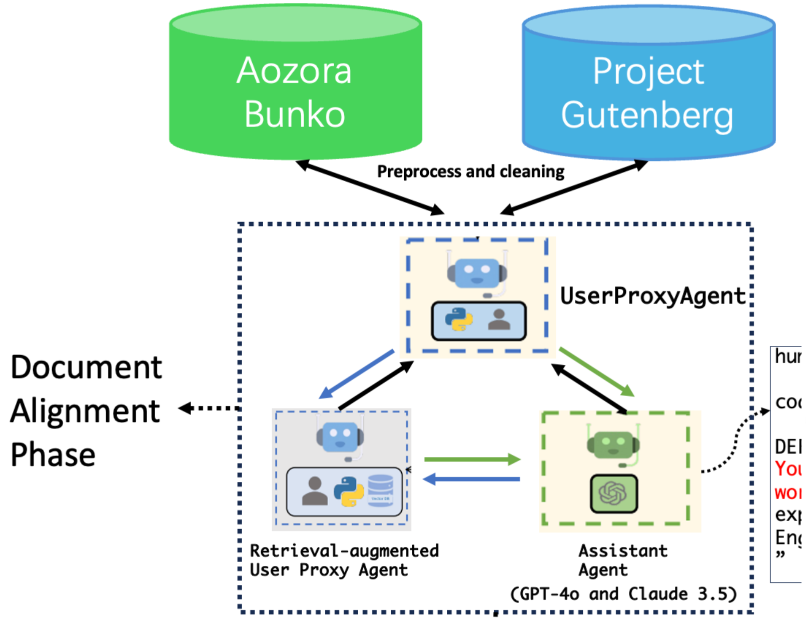
⚙️ **Other Preprocessing**

- **Standardized punctuation**
- **Unified text format**
- **Removed redundant symbols**

1. Document alignment



1. Document alignment



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.*

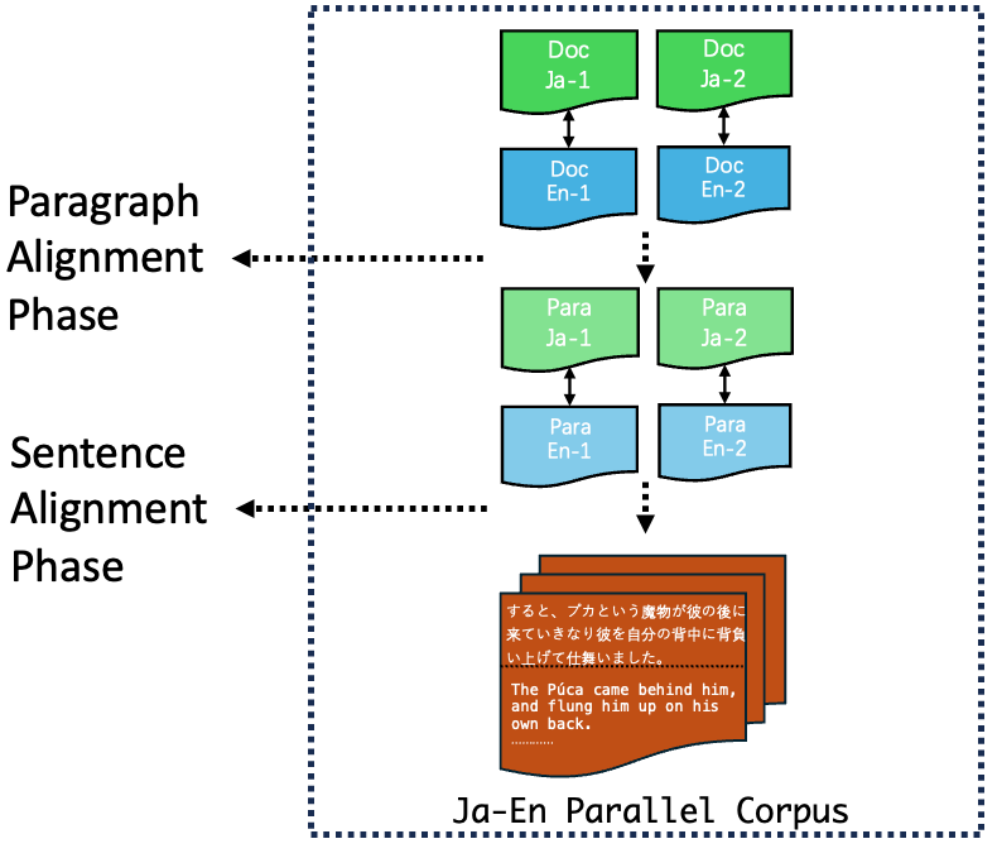
Iterate through all documents:

- **Step 1: Send Initial Request**
- **Step 2: Character-level Matching from the metadata of Project Gutenberg with character indexing**
- **Step 3: Deep Search (Up to 3 Attempts)**

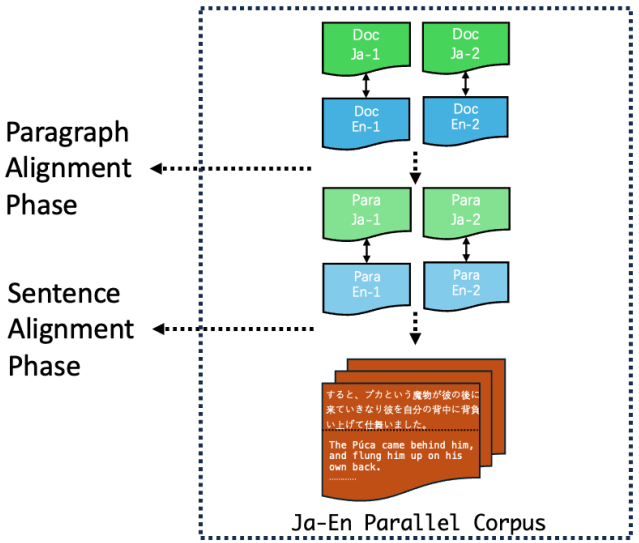
Return the aligned document set R.

  - If a match is found in the response, add it to R.

2. Paragraph Alignment & 3. Sentence Alignment



2. Paragraph Alignment & 3. Sentence Alignment



📌 Paragraph Alignment :

Extract paragraphs from the English documents using labeled information.
The final parallel paragraphs consist of :

- original documents of the Japanese works
- corresponding chapters from the English documents.

📌 Sentence Alignment :

- Finetuned sentence segmentator: *sat-12l-sm model* [8]
- Segmentator finetune dataset: Subset of Utiyama's data
- Sentence alignment tool: Vecalign [9]
- Embedding model: LaBSE and LASER2

⚙ Settings for sentence segmentation and alignment:

- Threshold: 0.01
- Overlapsize: 12
- Max size allowable aligned sentence: 12

[8] 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.
[9] B. Thompson and P. Koehn. Vecalign: Improved sentence alignment in linear time and space. In Proc. EMNLP and 9th IJCNLP, pp. 1342–1348, 2019.

Category

- Background
- Related Works
- The Proposal
- Statistics and Baseline Experiment
- Translation case analysis
- Conclusion and future works

Dataset statistics

Table 1. The statistics of the dataset

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

- The statistics are counted from 513 documents out of 634 documents.
- To compute the number of subwords, the tokenizer from the LaBSE was utilized.

Baseline Experiment

Table 2. Results on AoGu

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

- **Model:** 6-layer transformer with 8 multi-heads and 512 embedding dimension.

- **Hyperparameter:** adam-betas '(0.9, 0.98)', label smoothing 0.1, dropout rate 0.3, initial learning rate 4e-4, 3000 warm-up update steps, maximum of 6144 tokens per batch, update frequency 4, total of 50 epochs. For evaluation, beam search size is set with 4.

Baseline Experiment

Table 3. Results on out-of-domain ASPECT 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

- **Model:** 6-layer transformer with 8 multi-heads and 512 embedding dimension.

- **Hyperparameter:** adam-betas '(0.9, 0.98)', label smoothing 0.1, dropout rate 0.3, initial learning rate 4e-4, 3000 warm-up update steps, maximum of 6144 tokens per batch, update frequency 4, total of 50 epochs. For evaluation. beam search size is set with 4.

- Background
- Related Works
- The Proposal
- Statistics and Baseline Experiment
- Translation case analysis
- Conclusion and future works

Translation case analysis

Vecalign (LASER2) + LaBSE sampling with similarity > 0.4 setting

#	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."
	COMET = 0.681	だ。		
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."

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).

Translation case analysis

Vecalign (LASER2) + LaBSE sampling with similarity > 0.4 setting

#	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
	COMET = 0.790			

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.

Translation case analysis

Vecalign (LASER2) + LaBSE sampling with similarity > 0.4 setting

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 ."

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."

Translation case analysis

Vecalign (LASER2) + LaBSE sampling with similarity > 0.4 setting

#	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
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."

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.

- Background
- Related Works
- The Proposal
- Statistics and Baseline Experiment
- Translation case analysis
- Conclusion and future works

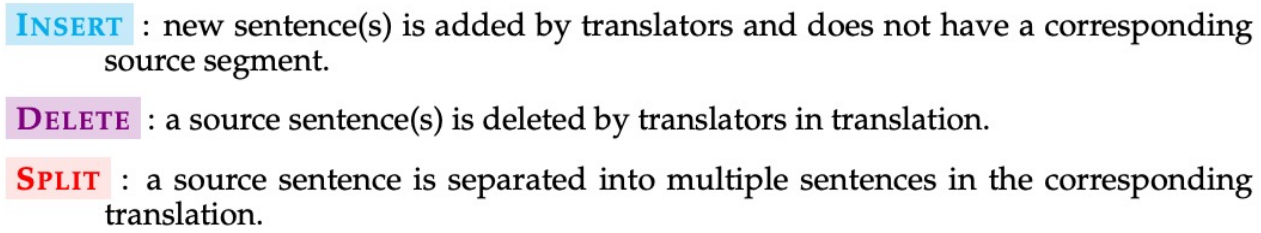
Conclusion

- This research develop a Japanese-English literary parallel dataset.
  -Mainly using the bilingual text from Aozora Bunko and Project Gutenberg.
- The baseline settings trained under sentence-level parallel data show limited performance in BLEU and COMET.
- Literary text shows its characteristic diversity and impose higher demands on translation models in terms of
  -Context awareness, complex semantic relationship modeling, and contextual coherence.

Future works

Current some problem:

Figure 1. some typical sentence misalignment types [9]



- In literary translation, to ensure the readability and fluency of the translated text, human translation experts often translate the source text according to different linguistic conventions and regional styles.

- As a result, translations **may not strictly adhere to sentence alignment.** And some misalignment types are listed in the Figure 1 above.

[9] L. Jin, Li A., and X. Ma. Towards chapter-to-chapter context-aware literary translation via large language models, arXiv2407.08978, 2024.

Future works

Current some problem:

Figure 1. some typical sentence misalignment types [9]

INSERT : new sentence(s) is added by translators and does not have a corresponding source segment.

DELETE : a source sentence(s) is deleted by translators in translation.

SPLIT : a source sentence is separated into multiple sentences in the corresponding translation.

- Free translation(意訳) and faithful translation(忠実翻訳) frequently alternate in the AoGu dataset, presenting a challenge for sentence-level alignment:

  ! Free translation(意訳) ： significant differences in expression between the source and target will pose a challenge to the modeling capability of the embedding model, which is trained on strictly parallel data.

  ! INSERT and DELETE types: will pose a challenge to the sentence alignment paradigm.

  ! etc,al.

[9] L. Jin, Li A., and X. Ma. Towards chapter-to-chapter context-aware literary translation via large language models, arXiv2407.08978, 2024.

Future works

- Refine the dataset by explore adequate alignment method to add the “context”.
  - which will be useful for handling the complex semantic relationship.
- Explore paragraph and document-level training and infra settings to enhance the performance of literary translation.
- Explore LLM-based training and infra settings on literary translation task.

Thanks for your listening!