Character-Aware English-to-Japanese Translation of Fictional Dialogue Using Speaker Embeddings and Back-Translation

Ayuna Nagato **Takuya Matsuzaki**
Tokyo University of Science
1425527@ed.tus.ac.jp matuzaki@rs.tus.ac.jp

Abstract

In Japanese, the form of utterances often reflect speaker-specific character traits, such as gender and personality, through the choise of linguistic elements including personal pronouns and sentence-final particles. However, such elements are not always available in English and a character's traits are often not directly expressed in English utterances, which can lead to character-inconsistent translations of English novels into Japanese. To address this, we propose a character-aware translation framework that incorporates speaker embeddings. We first train a speaker embedding model by masking the expressions in Japanese utterances that manifest the speaker's traits and learning to predict them. The resulting embeddings are then injected into a machine translation model. Experimental results show that our proposed method outperforms conventional fine-tuning in preserving speaker-specific character traits in translations.

1 Introduction

Neural machine translation (NMT) has made remarkable progress in recent years. However, translating fictional narratives, such as novels, still poses substantial challenges. Prior work has pointed out difficulties such as preserving long-range coherence, maintaining consistent tone across chapters, and handling figurative or culturally specific expressions (Thai et al., 2022; Liu et al., 2023; Karpinska and Iyyer, 2023). One of the less-studied issues is the preservation of character-specific linguistic style, especially in the translation of dialogue.

This problem becomes particularly apparent when translating from English into Japanese. Japanese dialogue often encodes rich speaker characteristics – such as gender, personality, and social status – through a variety of linguistic devices, including first- and second-person pronouns, honorifics, and sentence-final particles. In contrast, English dialogue tends to be less explicit in expressing such traits. As a result, standard translation models often produce Japanese utterances that contradict the original speaker's identity, leading to unnatural or inconsistent character portrayals.

One major obstacle in addressing this issue is the scarcity of high-quality bilingual dialogue corpora in the literary domain. To overcome this, we employ a back-translation (Sennrich et al., 2016) strategy: we first translate a large collection of Japanese novels into English using a high-performing neural MT system. This enables us to construct a large-scale pseudo-parallel corpus of Japanese-English fictional dialogue, which serves as the foundation for training our models.

While fine-tuning translation models on such in-domain dialogue can partially alleviate the problem, it is insufficient to capture the nuanced stylistic variation required for faithful character portrayal. We hypothesize that explicitly modeling the speaker's identity can help resolve this issue.

In this work, we propose a character-aware translation model that integrates speaker embeddings into the translation process. These embeddings are learned from Japanese utterances by masking the expressions that manifest the speaker's traits and training the model to predict them, capturing latent speaker traits in a data-driven way. We inject these embeddings into a Transformer-based translation model and fine-tune it on bilingual literary dialogue data. Experimental results show that our approach produces translations that better preserve speaker-specific character traits compared to conventional fine-tuned baselines.

The contributions of this paper are as follows:

- We introduce a novel speaker embedding model tailored to Japanese dialogue.
- We incorporate these embeddings into a neural translation model.
- We utilize back-translated Japanese-English

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 180–190
November 8–9, 2025.

novel data to overcome the lack of existing bilingual corpora.

- We demonstrate both qualitative and quantitative improvements in preserving character consistency in Japanese translations.

2 Background: Character Expressiveness in Japanese Utterances

Japanese is a language rich in surface-level variation that reflects the speaker's social identity, personality, and emotional stance. In spoken language, particularly in literary dialogue, this variation is often encoded through the choice of **personal pronouns**, **honorific expressions**, and **sentence-final particles**. These elements do not simply convey information but actively construct the speaker's character. In this section, we outline each of these linguistic mechanisms and explain how they contribute to speaker characterization in Japanese.

2.1 Personal Pronouns

Unlike English, Japanese personal pronouns vary widely depending on the speaker's gender, formality, and social distance. Even within first- and second-person references, different pronouns evoke distinct speaker personas.

For example, for the first-person pronoun "I", speakers may choose from:

- 私 **(watashi)** : neutral or formal
- 僕 **(boku)** : typically used by polite males
- 俺 **(ore)** : rough or masculine tone
- あたし **(atashi)** : casual and feminine

and many more.

For second-person references:

- あなた **(anata)** : formal or neutral
- おまえ **(omae)** : rough, informal, sometimes aggressive
- あんた **(anta)** : casual, often used by women
- きみ **(kimi)** : gentle, sometimes condescending depending on context

2.2 Sentence-Final Particles

Sentence-final particles such as よ **(yo)**, ね **(ne)**, の **(no)**, ぞ **(zo)**, and わ **(wa)** play a key role in expressing pragmatic and emotional nuance.

- よ **(yo)** : adds emphasis or confidence
- ね **(ne)** : invites agreement or shared understanding
- の **(no)** : softens a statement, often used by female speakers
- ぞ **(zo)**, ぜ **(ze)** : express strong masculine emphasis
- わ **(wa)** : indicates a feminine or classical tone depending on usage

2.3 Honorifics and Politeness Levels

Japanese exhibits a highly stratified system of honorifics, including respectful, humble, and polite forms. These levels express not only social hierarchy but also character traits in literary dialogue.

Politeness can be expressed by an auxiliary verb or a light verb:

- です **(desu)**/ます **(masu)** : basic politeness
- ございます **(gozaimasu)** : highly respectful
- くださる **(kudasaru)** : humble expressions

as well as the choice of a verb:

- 食べる **(taberu)** ↔召し上がる **(mesiagaru)** : normal ↔polite form of "eat"
- 言う **(iu)** ↔おっしゃる **(ossyaru)** : normal ↔polite form of "say"

2.4 Variation in Utterances Reflecting Speaker Character

The above linguistic features often appear together, shaping the overall tone and personality of the speaker. Table 1 shows different utterances of "Who are you?" that reflect various speaker identities through the use of pronouns, honorifics, and sentence-final particles.

Although all the examples convey the same core meaning, the speaker's personality, social stance, and emotional intensity vary drastically. In English–Japanese translation of literary dialogue, these nuances have to be properly differentiated by the

Table 1: Examples of speaker-dependent utterance variation in Japanese

Utterance	Pronoun	Honorifics	Final Particle	Character Impression
あなたはどなた**ですか**？	あなた (anata)	です (desu)	か (ka)	Formal, respectful
おまえ、誰だ？	おまえ (omae)	none	none	Rough, masculine
あんた、誰**よ**？	あんた (anta)	none	よ (yo)	Strong-willed female
きみは誰な**の**？	きみ (kimi)	none	の (no)	Friendly, gentle

choices of linguistic features according to the character, even if the source English expression is the same. Conventional systems often produce translations that fail to align with the character's original persona. This motivates our approach to incorporate speaker embeddings into translation to better preserve character-specific traits.

3 Method

Figure 1 illustrates the architecture used for constructing speaker embeddings and utilizing them in English-Japanese translation. We detail the method in what follows.

3.1 Speaker Embedding Construction

To incorporate speaker-specific characteristics into the translation process, we construct a speaker embedding model trained on Japanese literary dialogue. The aim is to learn embeddings that capture the personality traits expressed in each character's speech. The training proceeds as follows:

Step 1: Creating a Japanese-English Parallel Corpus Due to the scarcity of parallel corpora of Japanese and English novels, we create a pseudo-parallel corpus by translating 13,772 Japanese novels from Aozora Bunko[1] into English. We translated these novels by using a Transformer-based large model trained on JParaCrawl v3 (Morishita et al., 2022). This pseudo-parallel corpus is also used in the training of the speaker-aware translation model described in Section 3.2.2.

Step 2: Speaker Identification Using Stanford CoreNLP (Manning et al., 2014), we extract speaker-utterance pairs and the sentences where the subject is one of the speakers from the English translations. Specifically, we extract (i) utterances and their speakers through quote attribution, and (ii) declarative sentences where one of the speakers in a novel is marked as the subject (nsubj). These English sentences are then aligned with their Japanese counterparts to create bilingual dialogue pairs.

Step 3: Speaker-Sensitive Masking We mask parts of each Japanese utterance that tend to encode speaker-specific traits using the following rules:

- **Pronouns (1st and 2nd person):** Tokens tagged as "pronoun" by MeCab morphological analyzer (Kudo et al., 2004) are replaced with [MASK].
- **Sentence-final particles:** Tokens tagged as "sentence-final particle" are masked.
- **Honorific expressions:** Polite or honorific expressions, including verbs and sentence-final copulas such as です (desu) and ます (masu) are masked. If none of these forms are present at the sentence end, an empty [MASK] token is inserted to maintain output consistency.

Step 4: Embedding Extraction To obtain the embedding of speaker X, all utterances by X and the sentences with X as the subject are extracted from an English novel, concatenated using [SEP] as separators, truncated to 512 tokens, and input to English BERT [2]. The resulting [CLS] token embedding is used as the speaker's embedding vector.

Step 5: Training the Speaker Embedding Model Masked Japanese utterances are input to Japanese BERT [3]. Each output token vector is combined with the speaker's English BERT embedding by vector addition and passed through the language modeling head to predict the masked tokens. Cross-entropy loss is computed only at masked positions. Figure 2 presents an overview of Step 4 and 5. All components—Japanese BERT, English BERT, and the language modeling head—are jointly fine-tuned.

For consistency with the downstream translation model, we also fine-tune the speaker embedding model using the aligned NICT corpus described in Section 3.2.2.

[1] https://www.aozora.gr.jp

[2] https://huggingface.co/google-bert/bert-base-uncased

[3] https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

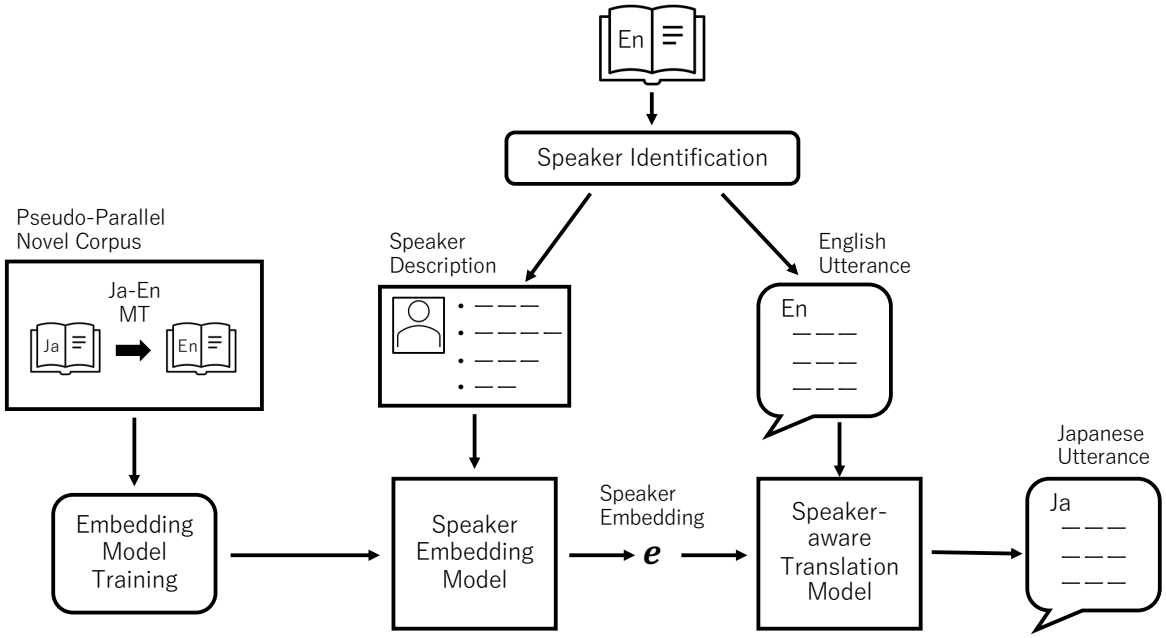

Figure 1: Overview of our proposed method.

For optimization, we used AdamW with hyperparameters of $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$, and learning rate of 10^{-5}. The batch size was set to 32. We chose the model achieved minimum loss on the validation set within five epochs.

3.2 Speaker-Aware Translation Model

3.2.1 Model Architecture

Our translation model adopts a Transformer encoder-decoder architecture. To incorporate character-specific style into the output, we inject speaker embeddings into the decoder. Specifically, the speaker embedding is added to the hidden states of the decoder's last layer before the final linear projection:

$$z_i' = z_i + e_s$$

where z_i is the decoder hidden state at time step i, and e_s is the speaker embedding vector. This additive integration enables the model to condition generation on speaker traits such as personality or social role.

3.2.2 Training Procedure

We employ a three-stage training process designed to balance general translation quality with speaker-sensitive stylistic control. These stages are: (1) pretraining of a general-domain English-to-Japanese model, (2) training on back-translated literary dialogue, and (3) fine-tuning on sentence-aligned, human-translated dialogue from novels.

Pretraining on JParaCrawl (English to Japanese) We initialize our model using a Transformer-based large model trained on JParaCrawl v3 (Morishita et al., 2022), in the English-to-Japanese direction. This model provides a strong general translation foundation but does not incorporate speaker-specific information. It serves as the backbone for subsequent adaptation.

Training with Back-Translated Literary Dialogue As described in Section 3.1, we utilize a pseudo-parallel corpus of Japanese novels and their translations to English. We use this back-translated dataset—composed of Japanese original dialogue and its English translation—to fine-tune our English-to-Japanese translation model with speaker embeddings. The Japanese side contains rich stylistic expressions, and the English side provides automatically extracted speaker labels via quote attribution and syntactic parsing.

This stage allows the model to learn how speaker-specific stylistic features in Japanese correspond to the more neutral English dialogue, and how to generate speaker-aware Japanese output based on speaker embeddings.

For optimization, we used Adam with hyperparameters of $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-8}$, and learning rate of 5×10^{-5}. The batch size was set to 6000 tokens. We trained up to 20k updates and averaged the last eight checkpoints for the final model.

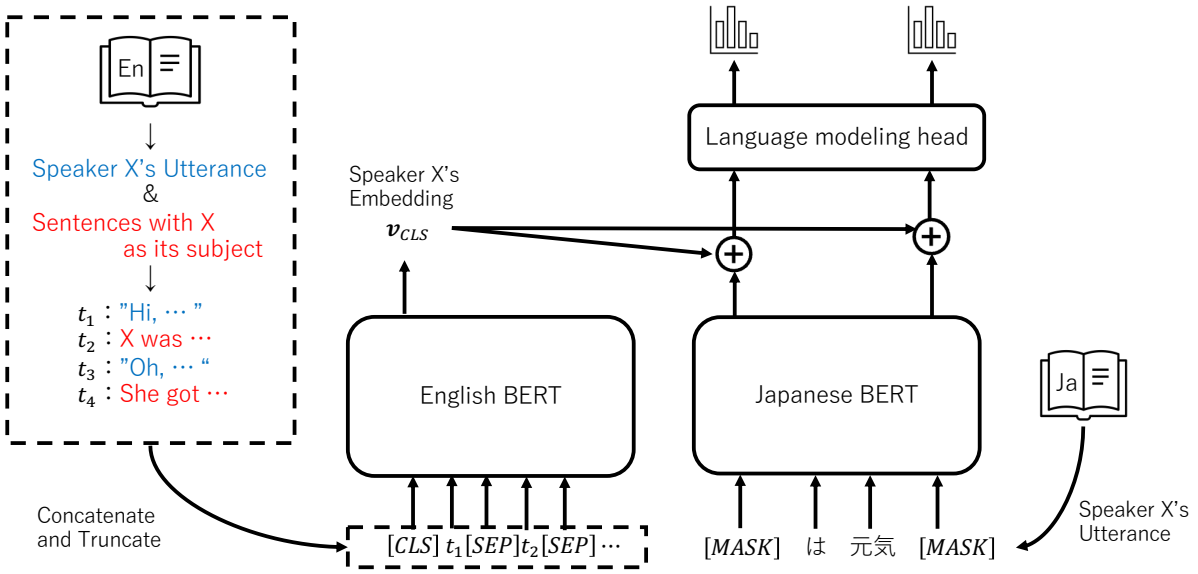

Figure 2: Architecture of the speaker embedding model. The masked Japanese input is processed by Japanese BERT, while the speaker embedding from English BERT is integrated during token prediction.

Fine-Tuning with Sentence-Aligned Modern Dialogue Although the back-translated Aozora data is rich in stylistic variation, it tends to reflect older literary styles. To adapt the model to contemporary Japanese, we fine-tune it using a small set of manually sentence-aligned English-Japanese novel data that was developed by Utiyama and Takahashi (2003) and is distributed by the National Institute of Information and Communication Technology (NICT). We henceforth call this data the NICT corpus. It mostly consists of modern fictional texts aligned at the sentence level, but without explicit speaker annotations.

We extract speaker information automatically using Stanford CoreNLP on the English side, as described in Section 3.1, and generate speaker embeddings using our trained embedding model. This final step improves fluency, modernity, and alignment with contemporary character dialogue styles.

For optimization, we used Adam with hyperparameters of $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-8}$, and a smaller learning rate of 10^{-5} to avoid overfitting on the limited NICT corpus. The batch size was set to 2000 tokens. We trained up to 2k updates and averaged the last three checkpoints for the final model.

Data Splits All training, validation, and test sets for the English-to-Japanese translation consist exclusively of utterances from dialogue segments, with narrative text excluded (except as the input to the speaker embedding model). For both the back-translated Aozora corpus and the aligned NICT corpus, we split the data into training, validation, and test sets in an 8:1:1 ratio. Importantly, the split is done on a per-work basis (i.e., by novel title) to avoid information leakage across sets. Furthermore, to ensure corpus independence, we exclude from the Aozora corpus any works that are also included in the NICT corpus.

4 Experiments

This section presents a comprehensive evaluation of the proposed speaker-aware translation model. We first conduct qualitative and quantitative manual analyses to assess how incorporating speaker embeddings affects character-sensitive translation aspects such as pronoun choice and sentence-final particles. Next, a case study on a single character from a literary work examines consistency in character voice over multiple utterances. Additionally, we perform automatic evaluation using BLEU, ChrF, and COMET-22 scores on dialogue-heavy test data to quantify improvements over baseline models. Finally, we analyze the learned speaker embeddings via Principal Component Analysis (PCA) to interpret the linguistic and stylistic features captured in the embedding space.

Together, these evaluations demonstrate the effectiveness of integrating speaker embeddings in enhancing the fidelity and expressiveness of literary dialogue translation.

4.1 Quantitative Analysis: Manual Evaluation of Speaker-Sensitive Translation

To qualitatively and quantitatively evaluate the impact of incorporating speaker embeddings, we conducted a manual analysis of sampled outputs. From the 1,436 utterances in the NICT test set, we randomly selected 150 utterances and compared translations produced by two systems:

- **Baseline**: A standard Transformer model pretrained on JParaCrawl v3 and fine-tuned on back-translated Aozora data followed by fine-tuning on the NICT corpus.
- **Proposed**: The same model architecture and training procedure, but augmented with speaker embeddings during training and inference.

We evaluated each output along five criteria:

- **Pronouns**: Whether first- and second-person pronouns exactly matched the reference translation (surface differences such as 私 (watashi) vs わたし (watashi) were treated as equivalent).
- **Sentence-final particles**: Whether sentence-final particles exactly matched the reference translation.
- **Honorific usage**: Whether honorific or polite expressions were used when the reference translation employed them.
- **Intra-utterance consistency**: Whether a single utterance maintained consistent character traits (e.g., not mixing 僕 (boku) and 私 (watashi)).
- **Translation errors**: Whether the output contained critical errors such as omissions, repetitions, or untranslated segments.

The results are summarized in Table 2. For pronouns, sentence-final particles, and honorifics, we report accuracy (percentage of exact matches with the reference). For consistency and translation errors, we report the number of problematic utterances out of the 150 sampled.

Manual evaluation results, as shown in Table 2, demonstrate several notable improvements brought by our proposed speaker-aware translation model.

First, the accuracy of pronoun translation significantly increased from 25.6% in the baseline to 61.6% in our model. This suggests that incorporating speaker embeddings greatly improves the model's ability to select appropriate first- and second-person pronouns, which are crucial in reflecting character-specific traits. Similarly, accuracy on sentence-final particles improved from 40.5% to 52.5%, despite the strict criterion of requiring exact matches. In fact, our manual inspection revealed that some minor mismatches—such as わ (wa) vs. よ (yo) or だ (da) vs. ぞ (zo) —still produce utterances with similar character impressions, implying that the model's improvement may be underrepresented by exact matching metrics alone.

Interestingly, performance on honorific forms remained unchanged (28.1% accuracy in both models). One possible explanation is that honorific expression tends to depend more on the social relationship between the speaker and the hearer, rather than on the speaker's character identity alone. This highlights a potential limitation of our speaker-only embedding approach in capturing such pragmatic nuances.

The number of consistency errors —-such as inconsistent use of personal pronouns within the same utterance—- decreased from seven to zero. This indicates that the proposed model contributes to maintaining character consistency at the utterance level. Furthermore, the number of critical translation errors, including untranslated segments and repeated phrases, was reduced from four to zero.

Taken together, these findings support the effectiveness of integrating speaker embeddings in improving character-sensitive aspects of translation, especially for pronoun and sentence-final particle choices, while also enhancing consistency of character traits.

4.2 Case Study: Fatty Coon in *The Tale of Fatty Coon*

To further analyze the effect of our method on maintaining the consistency of character's traits, we conducted a case study on Fatty Coon, the protagonist of *The Tale of Fatty Coon* by Arthur Scott Bailey. This character is portrayed as an energetic young boy raccoon, often using casual language such as ぼく (boku) and sentence-final particles like よ (yo) in the Japanese translation.

Out of 74 utterances attributed to Fatty Coon (based on CoreNLP speaker tagging), 9 were misattributed. We excluded these and manually analyzed

Table 2: Manual evaluation results on 150 randomly sampled utterances from the NICT test set. Accuracy is reported for categorical items (pronouns, sentence-final particles, honorifics), and raw counts are reported for consistency errors and critical translation errors.

Model	Pronouns	Final Particles	Honorifics	Consistency Errors	Translation Errors
Baseline	25.6% (32/125)	40.5% (98/242)	28.1% (9/32)	7	4
Proposed	**61.6% (77/125)**	**52.5% (127/242)**	28.1% (9/32)	**0**	**0**

the remaining 65 utterances.

We found that the baseline model frequently produced outputs that were inconsistent with the character's personality. Specifically, 29 out of the 65 utterances (44.6%) included language that was too formal or feminine, such as the use of 私 (watashi) or sentence-final particles like わ (wa), or even polite verb forms. In contrast, our proposed method produced consistent outputs aligned with the character's casual and boyish tone in nearly all cases, with only 9 utterances showing mismatches (e.g., use of honorifics).

Table 3 shows representative examples. In each case, the proposed method produces translations that are closer to the reference translation and consistent with the intended persona of Fatty Coon.

4.3 Automatic Evaluation on NICT Dialogue Segments

To complement our manual evaluation, we conducted automatic evaluation using BLEU (Papineni et al., 2002), ChrF (Popović, 2015), and COMET-22 (Rei et al., 2022) on the NICT test data, comprising 1,436 Japanese utterances. We report BLEU scores computed using `sacreBLEU` with `--tokenize=intl` option. COMET is a neural-based metric trained to predict human direct assessment scores, and has been shown to correlate more strongly with human evaluation than surface-overlap metrics such as BLEU (Rei et al., 2020).

Table 4 provides the results. The absolute BLEU and ChrF scores appear low, especially compared to typical scores reported in general domain English-to-Japanese translation. However, our task differs significantly in both content and style: the data is literary dialogue, which contains diverse speaker-specific expressions, idiosyncratic phrasing, and multiple valid translations. In such settings, surface-form overlap metrics like BLEU often underestimate translation quality (Toral and Way, 2018; Mathur et al., 2020; Thai et al., 2022).

Despite the limitations of these metrics in capturing speaker-specific style or consistency, the proposed method outperforms both the base model and the fine-tuned baseline by a noticeable margin on both BLEU and ChrF. This suggests that introducing speaker-aware information helps produce translations that better align with the reference utterances even in automatic evaluation metrics. In the case of BLEU score, the larger gain compared to the fine-tuned baseline also supports our hypothesis that speaker traits contribute to reducing ambiguities in character-driven translation.

In addition, the COMET-22 scores also show that our proposed method achieves the highest performance among the compared systems (0.802 vs. 0.778 for the baseline). This further supports that incorporating speaker information not only improves surface-level similarity but also yields translations that are semantically closer to human references, in line with human judgments of adequacy.

To assess the reliability of these improvements, we conducted paired bootstrap resampling tests. BLEU differences between the baseline and our proposed model were not statistically significant ($p = 0.14$). However, both ChrF ($p < 0.01$) and COMET ($p < 0.05$) confirmed that the improvements of the proposed model over the baseline are statistically significant. These results suggest that while BLEU may underestimate gains in this task, stronger metrics such as ChrF and COMET provide more robust evidence that speaker-aware information improves translation quality.

Nonetheless, given the nature of the task, we emphasize that manual evaluation and qualitative analysis (as described in the previous subsections) provide a more reliable assessment of character expressiveness and speaker consistency in translation.

4.4 Speaker Embedding Visualization via PCA

To investigate the structure of the learned speaker embeddings, we apply Principal Component Analysis (PCA) to the speaker vectors extracted from the test portion of the back-translated Aozora corpus. This test set was held out during model training.

We first applied our trained speaker embedding model to the machine-translated English versions

Table 3: Example translations of utterances by Fatty Coon. The proposed method produces outputs more consistent with the character's persona.

Source	Reference	Baseline	Proposed
"I'd like to eat all the corn in the world."	世界中のとうもろこしを全部食べちゃいたい**よ (yo)**。	世界中のとうもろこしをみんな食べてみたい**わ (wa)**。	世界中のとうもろこしがみんな食べたい**よ (yo)**。
"Look, Mother!"	見て、お母さん！	**ご ら ん な さ い (Honorifics)**、お母さん！	見てよ、おかあさん！
"Maybe you don't think I heard him screech–"	**ぼく (boku)** があいつの絶叫を聞いたっていうのも―――	たぶん、**わたし (watashi)** が金切り声を聞いたんじゃないと思うだろう――	**僕 (boku)** が金切声を聞いたと思わないだろう――

Table 4: Automatic evaluation scores on NICT test data

Model	BLEU	ChrF	COMET-22
JParaCrawl (pre-FT)	3.8	14.8	0.776
Baseline	3.4	16.7	0.778
Proposed	**5.7**	**18.2**	**0.802**

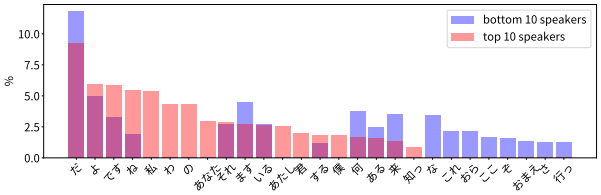

Figure 3: Proportion of [MASK] tokens in utterances of speakers with top and bottom PC Scores

of the test set novels. Each character's English utterances were concatenated and processed using BERT to produce a fixed-size speaker embedding, as described in Section 3.1. As a result, we obtained embeddings of 6,449 speakers.

We then performed PCA on these embeddings and analyzed the first principal component. Table 5 lists the top and bottom 10 characters based on their scores on the first principal component, along with their associated works and authors. We found that characters with high component scores tend to be female, while those with low scores tend to be male.

This interpretation is supported by examining the masked tokens in their utterances (Figure 3). Tokens like わ (wa), の (no), あなた (anata), and あたし (atashi) appear frequently in the utterances of characters with high PCA scores—expressions that are stereotypically feminine in Japanese. In contrast, utterances from characters with low PCA scores more often contain ぞ (zo), おまえ (omae), and おら (ora), which are typically masculine expressions.

Moreover, we observe that characters with similar component scores often come from the same literary work or author. This pattern suggests that the speaker embeddings encode not only individual character traits but also stylistic patterns associated with particular authors.

These results indicate that our speaker embedding space captures interpretable dimensions related to gendered language use and authorial style in literary dialogue.

5 Discussion

The proposed speaker-aware translation model showed improvements in preserving speaker-specific linguistic traits. However, several limitations and potential directions for future work have emerged.

Accuracy of Speaker Attribution. In this study, we relied on speaker attribution results from Stanford CoreNLP. As observed in the case study of "Fatty Coon", speaker misattribution occurred in 9 out of 74 utterances. These errors can directly affect the quality of the speaker embeddings, potentially leading to unnatural or inconsistent translations.

Effect of Character Description Length on Embedding Quality. To construct speaker embeddings, we used English character descriptions—specifically, sentences where the character was the subject and the character's utterances—from the beginning of each work. These sentences were concatenated until reaching the maximum input length of the English BERT model. As a result, the amount of contextual information varied depending on how much text was available for each character. We have not yet conducted a systematic analysis of how the amount or content of these character-descriptive sentences affects the quality of the embeddings or the final translation output. Investigating the impact of description length and exploring methods to prioritize more informative sentences (e.g., those rich in personality cues) may help enhance consistency,

Table 5: Top and bottom characters on the first PCA component

Bottom (Low PC score)			Top (High PC score)		
Speaker	Gender	Work	Speaker	Gender	Work
Kasuke	Male	"Irefuda" by Kikuchi Kan	Kuroe	Female	"Charako-san" by Hisao Juran
Koshu	Male	"Dai-bosatsu toge" by Nakazato Kaizan	her	Female	"Charako-san" by Hisao Juran
Iori	Male	"Zenigata Heiji" by Nomura Kodo	Yoshie	Female	"Charako-san" by Hisao Juran
his	Male	"The Escape of Terasaka Kichiyemon" by Naoki Sanjugo	Noriko	Female	"Sugiko" by Miyamoto Yuriko
Yasuke	Male	"The Woman Who Stepped on a Shadow" by Okamoto Kido	Haruko	Female	"Nozarashi" by Toyoshima Yoshio
Isuke	Male	"Quick-Eared Sanji" by Hayashi Fubo	Suzue	Female	"A History of a Couple" by Kishida Kunio
Iori	Male	"Miyamoto Musashi" by Yoshikawa Eiji	her	Female	"This Morning' s Snow" by Miyamoto Yuriko
Yoriharu	Male	"The Armor of Asahi" by Kunieda Shiro	Charako	Female	"Charako-san" by Hisao Juran
his	Male	"On Leisure" by Itami Mansaku	his	Male	"Sugigaki" by Miyamoto Yuriko
Yamada	Male	"My Private Taiheiki" by Yoshikawa Eiji	Madam	Female	"The Shadowless Criminal" by Sakaguchi Ango

especially for characters with limited text.

Limitations of Back-Translation Quality. One challenge we observed is that the back-translation process used to construct the training data—specifically, translating Japanese to English and then back to Japanese—sometimes produces unnatural Japanese sentences. In our pipeline, the major source of noise stems from the Japanese-to-English translation step. When this translation is inaccurate, it leads to poor-quality pseudo-parallel data, which in turn affects the quality of the final English-to-Japanese translation model. While using manually curated parallel data would be ideal, such data is extremely limited, especially for literary dialogue. Future work may improve the overall quality by employing stronger Japanese-to-English translation models, or by integrating automatic quality filtering mechanisms to reduce the impact of noisy samples.

6 Related Work

6.1 Speaker Attribution in Narrative Texts

Speaker attribution—the task of identifying who is speaking in a given utterance—is a crucial preprocessing step for speaker-aware translation. In narrative texts such as novels, explicit speaker tags are often absent, requiring automatic identification based on linguistic cues. For English texts, tools such as Stanford CoreNLP provide heuristic-based speaker tagging, but they often fail in the presence of figurative or indirect speech.

Speaker attribution in Japanese poses additional challenges due to frequent subject omission, flexible word order, and the use of sentence-final particles that vary by speaker. Ishikawa et al. (2024) addressed this by leveraging grammatical and contextual features to estimate speaker identity in Japanese novels. Zenimoto and Utsuro (2022) proposed a method for identifying the speakers of quoted utterances in Japanese novels using a gender classification model.

6.2 Machine Translation for Literary Texts

Discourse-level literary translation remains one of the most demanding tasks in natural language processing. Unlike general-domain texts, literary works require models to handle complex semantic phenomena such as figurative language, long-range dependencies, character voice, and culturally grounded expressions (Pang et al., 2025). These aspects place high demands on translation systems, which must not only be accurate but also preserve subtle stylistic and narrative consistency.

While recent progress in large language models (LLMs) has enabled strong performance on many NLP tasks, training or fine-tuning models specifically for literary translation remains costly and resource-intensive. In response to these challenges, the WMT2023 Shared Task on Discourse-Level Literary Translation was launched, highlighting the need for models that go beyond sentence-by-sentence translation (Wang et al., 2023). Results from the shared task demonstrated that even state-of-the-art systems struggle to maintain coherence, tone, and character consistency across longer texts. These findings suggest that further advances are needed in integrating discourse-level information and stylistic modeling, particularly for literature.

7 Conclusion

This paper proposed a speaker-aware machine translation framework aimed at preserving character-specific expressions in Japanese literary dialogue. By constructing speaker embeddings from English descriptions of each character and incorporating them into the translation model, our method promotes more consistent and personality-aligned outputs.

We evaluated our approach using the NICT English-Japanese translation alignment dataset. Manual analysis showed that our method improves the consistency of personal pronouns and sentence-

final particles, which are strongly associated with character identity. However, the use of honorifics did not improve as clearly, likely because honorific usage depends more on social context—such as the relationship between the speaker and hearer—than on character traits alone.

In future work, we plan to explore the integration of situational context (e.g., dialogue participants and relationships), adopt higher-quality translation models, and refine our speaker recognition pipeline to further enhance the character consistency and fluency of literary dialogue translation.

References

Kazuki Ishikawa, Kohei Ogawaa, and Satoshi Sato. 2024. Speaker identification using speech-style encoder. *Journal of Natural Language Processing*, 31(3):894–934.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Xuebo Liu, Yutong Wang, Derek F. Wong, Runzhe Zhan, Liangxuan Yu, and Min Zhang. 2023. Revisiting commonsense reasoning in machine translation: Training, evaluation and challenge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15536–15550, Toronto, Canada. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710, Marseille, France. European Language Resources Association.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Antonio Toral and Andy Way. 2018. *What Level of Quality Can Neural Machine Translation Attain on Literary Text?*, pages 263–287. Springer International Publishing, Cham.

Masao Utiyama and Mayumi Takahashi. 2003. English-japanese translation alignment data. https://www2.nict.go.jp/astrec-att/member/mutiyama/align/index.html.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.

Yuki Zenimoto and Takehito Utsuro. 2022. Speaker identification of quotes in Japanese novels based on gender classification model by BERT. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 126–136, Manila, Philippines. Association for Computational Linguistics.