# Large Scale Evaluation of End-to-End Pipeline of Speaker to Dialogue Attribution in Japanese Novels

**Yuki Zenimoto Shinzan Komata Takehito Utsuro**
Degree Programs in Systems and Information Engineering,
Graduate School of Science and Technology, University of Tsukuba
{s2220753, s2320742}@u.tsukuba.ac.jp
utsuro@iit.tsukuba.ac.jp

## Abstract

The speaker to dialogue attribution task, which identifies the speaker of an utterance in a novel, is an essential task for the analysis of novels and their characters. Speaker to dialogue attribution task is composed of three tasks: utterance extraction, attributing character mentions to utterances, and clustering character mentions into character entities. However, there are no prior studies targeting Japanese novels that have conducted all of these tasks. Furthermore, the lack of shared evaluation data has made it difficult to compare methods. In this study, we propose a first end-to-end speaker to dialogue attribution pipeline applied to the publicly available speaker information annotation data from the "Balanced Corpus of Contemporary Written Japanese". We evaluate the performance of our approach and assess its limitations.

## 1 Introduction

The speaker to dialogue attribution task, which identifies the speaker of an utterance in a novel, is an essential task for the analysis of novels and their characters. Although Miyazaki et al. (2016); Ishii et al. (2021) manually collected a large amount of specified speakers' utterance for a systematic analysis of character traits and development of persona-based dialogue systems, automatically collecting a large amount of various speakers' utterance through speaker to dialogue attribution definitely enables detailed analysis of more diverse character traits and behaviors.

Speaker to dialogue attribution task is composed of three tasks as shown in Figure 1:[1] 1) utterance extraction, which extracts strings representing utterances, 2) attributing character mentions to utterances, which extracts strings representing speakers located around utterances as character mentions, and 3) clustering character mentions into character entities, which clusters character mentions that refer to an identical character entity.

Prior researches on speaker to dialogue attribution for English novels mainly used patterns like quote-mention-verb in narratives, which are the sentences other than utterances, to extract speakers (He et al., 2013; Muzny et al., 2017). Additionally, Muzny et al. (2017) proposed and released a deterministic sieve-based system. They also constructed a new publicly available dataset for speaker to dialogue attribution. Cuesta-Lazaro et al. (2022) proposed a complete pipeline to extract characters in a novel and link them to their utterances. This approach is the first application of deep learning to speaker to dialogue attribution, and it overcomes the previous rule-based approach.

However, there are no prior researches for Japanese novels that have conducted all three tasks. Furthermore, the lack of shared evaluation data for speaker to dialogue attribution has made it difficult to conduct research and compare methods. Zenimoto and Utsuro (2022) proposed a gender-specific language model, which classifies the gender of the speaker of a given utterance. They used the gender classification model for speaker to dialogue attribution and demonstrated that speech styles are effective in speaker to dialogue attribution. However, this method requires prior information on the speech styles of the speakers to be clustered and is not adaptable to arbitrary speakers. In addition, this method has only been tested against only two characters in one novel.

In this study, we present a first end-to-end[2] speaker to dialogue attribution pipeline for

[1] Terms specific to this domain such as *narrative, character, character mention*, and *character entity* are defined in the caption of Figure 1.

[2] In this study, by the term "end-to-end pipeline", we represent the notion of evaluating all the way through the three constituent tasks of speaker to dialogue attribution task, i.e.,

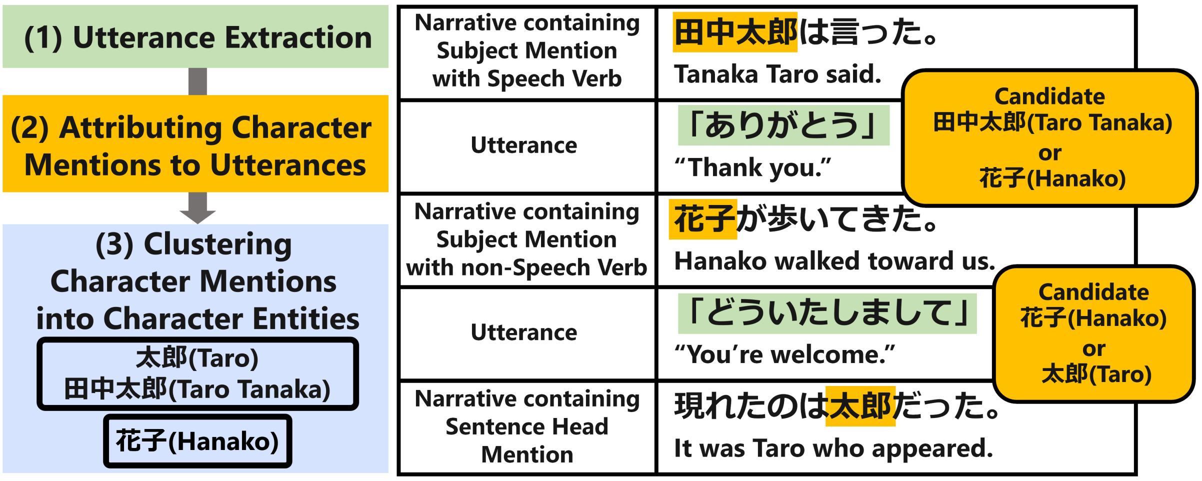

(1) Utterance Extraction
(2) Attributing Character Mentions to Utterances
(3) Clustering Character Mentions into Character Entities
太郎(Taro)
田中太郎(Taro Tanaka)
花子(Hanako)
Narrative containing Subject Mention with Speech Verb
田中太郎は言った。
Tanaka Taro said.
Utterance
「ありがとう」
"Thank you."
Candidate 田中太郎(Taro Tanaka) or 花子(Hanako)
Narrative containing Subject Mention with non-Speech Verb
花子が歩いてきた。
Hanako walked toward us.
Utterance
「どういたしまして」
"You're welcome."
Candidate 花子(Hanako) or 太郎(Taro)
Narrative containing Sentence Head Mention
現れたのは太郎だった。
It was Taro who appeared.

Figure 1: Diagram of our speaker attribution pipeline (*A narrative* is defined as a sentence other than utterances. *A character* is defined as a person in a novel. *A character mention* is defined as an expression to represent a character name in narratives. *A character entity* is defined as an entity to represent a unique character in a novel. In this study, we only focus on dominant character mentions: "Subject Mention" and "Sentence Head Mention", where the "Subject Mention" is further divided into two types according to the type of verb.)

Japanese novels. We propose an utterance extraction model using BERT, an attributing system based on surrounding narratives, and a character mention clustering method for Japanese names. Dataset is obtained from the "Balanced Corpus of Contemporary Written Japanese" (BCCWJ) (Maekawa et al., 2014) [3]. Our approach is optimized on the train data, and evaluated on the test data to assess the generalization performance. The results demonstrate interesting insights: 1) the subject mention of the speech verb is mostly the correct speaker, indicating that the speech verb serves as an important clue, 2) the mention of the correct speaker is preferred to be in the sentence subsequent to the utterance rather than preceding the utterance, 3) allowing to attribute multiple utterances to one character mention improves accuracy. The code used in this study is available at the following URL[4].

Our contributions are as follows:

1. We propose and release the end-to-end speaker to dialogue attribution pipeline for Japanese novels. We show that attributing speakers using surrounding narratives can appropriately classify about half of the utterances from the same speaker.

2. We conduct a large-scale analysis of the effectiveness of narratives in speaker to dialogue attribution for Japanese novels. This analysis shows that narratives play a crucial role and that their effectiveness depends on the position and structure of the narrative.

## 2 Related Work

Early works on speaker to dialogue attribution in English novels mainly utilize patterns like quote-mention-verb. Glass and Bangay (2007) addressed speaker identification in fiction texts. Their approach involves rule-based extraction of speaker mentions using major speech verbs and linking them to character entities in a character list. While their method achieves high recall in extracting explicit speakers, it does not address implicit speakers at all. Elson and McKeown (2010) took important first steps towards automatic quote attribution and address the implicit speakers using rule-based and statistical learning. Their method achieved

1) utterance extraction, 2) attributing character mentions to utterances, and 3) clustering character mentions into character entities. Here, given the text of a novel as the input, the task of 1) utterance extraction is performed and its output is provided as the input to the task of 2) attributing character mentions to utterances, where its output is then provided as the input to the task of 3) clustering character mentions into character entities. Due to the sake of simplification of evaluation, however, in the experimental evaluation of this paper, we chose to provide oracle utterance data as the input to the task of 2) attributing character mentions to utterances, and then conduct the evaluation of the subsequent tasks 2) and 3). Thus, our future work definitely includes the overall end-to-end evaluation of the whole constituent tasks 1), 2), and 3).

[3] https://clrd.ninjal.ac.jp/bccwj/

[4] https://github.com/Zeni-Y/speaker-classification

83.0% accuracy overall, but used oracle data at test time. O'Keefe et al. (2012) conducted experiments without using unrealistic oracle data that were used in Elson and McKeown (2010). They proposed a sequence labeling approach for quote attribution. However, their approach could not overcome their baseline in the literary domain.

He et al. (2013) proposed a supervised machine learning approach utilizing various features, such as the generative actor topic model (Celikyilmaz et al., 2010) and speaker appearance count. Muzny et al. (2017) proposed a speaker classification system based on multiple rules and constructed a new publicly available dataset for speaker classification. However, they assume the existence of a predefined list of characters, including the name, aliases, and gender of each character. Finally, Cuesta-Lazaro et al. (2022) proposed the first speaker classification system using deep learning and demonstrated more accurate speaker classification results compared to existing rule-based methods. In the utterance extraction task, they show that simple rules that detect opening and closing quotation marks can achieve 0.98 ± 0.01 F1-score against their entire dataset. For the identification of identical individuals from the extraction results, they constructed an out-of-domain co-reference resolution system and a simple rule-based clustering system. Additionally, They created a novel dataset and evaluated their proposed method on novels from diverse genres and time periods.

Previous studies on speaker classification for Japanese novels mainly utilize the speech style of utterances. In the studies conducted by Miyazaki et al. (2021) and Zenimoto and Utsuro (2022), they attempt to classify speakers based on the similarity of speech styles using utterances in light novels. These studies solve the classification task of determining which speaker a given utterance belongs to, using already extracted utterances. Due to the requirement of prior knowledge about the speech style of the target speakers, they cannot be applied to classify arbitrary speakers. Furthermore, there are multiple formats of utterances in Japanese novels, making the task of utterance extraction itself challenging. Therefore, in speaker classification targeting Japanese novels, it is necessary to construct an utterance extraction model and evaluate its performance. Additionally, the task of identifying identical individuals by clustering the extracted character mentions has not been conducted. Consequently, there are no prior researches targeting Japanese novels that have conducted all these three tasks. Moreover, each study employs its own undisclosed evaluation data, making it difficult to compare these methods.

| Data Type | #Novel | #Utterance | Avg. Word Length |
|---|---|---|---|
| train | 1,707 | 100,643 | 23.5 |
| validation | 569 | 33,914 | 24.1 |
| test | 569 | 33,360 | 23.7 |
| total | 2,845 | 167,917 | 23.7 |

Table 1: Statistics of the Dataset

| Bracket Symbol | #Utterance | #Non-Utterance |
|---|---|---|
| 「—」 | 92,456 | 3,807 |
| (—) | 1,071 | 1,700 |
| 『—』 | 337 | 1,635 |
| ⟨—⟩ | 312 | 1,044 |
| "—" | 79 | 1,078 |
| indirect utterance | 1,464 | — |

Table 2: Statistics on Bracket Type in the Training Data

## 3 Dataset

In this study, we use the existing BCCWJ speaker annotation data (Maekawa et al., 2014), which contains a total of 2,845 novels (published between 1986 and 2005) and 167,917 utterances with annotated speakers. This dataset is divided into a train/validation/test split (60%/20%/20%). Table 1 shows statistics of the dataset.

## 4 Utterance Extraction

Firstly, we introduce our utterance extraction method. In order to capture the characteristics of utterances in Japanese novels, we investigated the statistics of the bracket symbols at both ends of the utterances in the training data, as well as the statistics of non-utterance sentences enclosed by the same symbols as shown in Table 2. Table 2 shows that extracting strings enclosed by specific symbols as utterances would result in erroneously extracting many non-utterance expressions, such as just emphasis or quotations. Moreover, this approach would make it impossible to extract the indirect utterances, which are utterances not enclosed by specific symbols at both ends. Therefore, we attempt to extract utterances without erroneously extracting non-utterance expressions by constructing an utterance extraction model based on sequence labeling by BERT (De-

| Bracket Symbol | rule-based | | | BERT | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| 「—」 | 95.5 | 99.8 | 97.6 | 98.9 | 96.8 | 97.8 |
| (—) | 49.1 | 99.3 | 65.7 | 93.5 | 98.8 | 96.0 |
| 『—』 | 16.9 | 100 | 28.9 | 83.4 | 88.5 | 85.9 |
| 〈—〉 | 12.4 | 100 | 22.0 | 87.9 | 92.1 | 89.9 |
| “—” | 2.4 | 81.8 | 4.7 | 36.4 | 36.4 | 36.4 |
| indirect utterance | — | — | — | 26.5 | 20.9 | 23.3 |
| micro-average | 90.4 | 96.3 | 93.2 | 96.7 | 94.0 | 95.3 |

Table 3: Comparison of the Performance of Utterance Extraction

vlin et al., 2019). We compare the following two utterance extraction models:

**rule-based** All strings enclosed by the five types of brackets with the highest frequency of occurrence in the training data (「—」, (—), 『—』, 〈—〉, “—”) are extracted as utterances.

**BERT** We use the pre-trained BERT (Devlin et al., 2019) model for utterance extraction. Specifically, we used Tohoku University's Japanese version of BERT-base[5] which is trained on Japanese Wikipedia. Training is conducted on the training data, and the model with the minimum loss on the validation data is used for evaluation on the test data.

### 4.1 Evaluation Results

Table 3 shows the precision, recall, and F1-Score on the test data. We use micro-average for the evaluation scores of the all utterances. Table 3 shows that the BERT model has higher precision and F1-Score for all bracket symbols compared to the rule-based model, while the recall is lower than the rule-based model. It was found that the BERT model can extract utterances with fewer extraction errors. However, it is also evident that the BERT model has low extraction performance for utterances enclosed by “—” and indirect utterances. Additionally, as for the overall performance, there is not a significant difference in performance between the rule-based model and BERT model since most of the utterances are enclosed by 「—」.

## 5 Speaker Attribution

In this section, we describe the procedure of speaker attribution using surrounding narratives. We use the `ja_ginza_electra` model of GiNZA[6] for morphological and dependency analysis of the sentences. Then, we follow the procedures below of character mention extraction, determining verbs representing speech, attributing character mentions to utterances, and clustering character mentions into character entities.

### 5.1 Character Mention Extraction

Firstly, we extract character mentions from the narrative using the following two methods:

**Extraction of proper nouns representing person**

GiNZA's Named Entity Recognition system uses Extended Named Entity Hierarchy[7] by Sekine et al. (2002). We extract words detected with the tags “Person”, “Position-Vocation” (e.g. “警察 (police)”, “医者 (doctor)”), “Nationality” (e.g. “日本人 (Japanese person)”), and “Name-Other”[8] as person names.

**Extraction of common nouns and pronouns representing person**

We create a dictionary of words representing person and extract words in this dictionary as person names. This dictionary is created by automatically collecting a large number of words representing person from Japanese WordNet[9] (Bond et al., 2009), followed by manually adding 374 appropriate words and removing 181 inappropriate words, resulting in 7,215 words listed in this dictionary.

Subsequently, to cover honorifics (e.g. “田中 (Tanaka) さん (Mr.)”) and compound nouns (e.g. “花子 (Hanako) 先生 (teacher)”) as character mentions, the entire sequence of words, including extracted words and those with `compound` dependency relationship is considered as the final character mentions.

[5] https://github.com/cl-tohoku/bert-japanese

[6] https://megagonlabs.github.io/ginza/

[7] http://liat-aip.sakura.ne.jp/ene/ene8/definition_jp/html/enetree.html

[8] Unusual person names are often given this tag.

[9] We recursively collect words and sub-synsets in synsets containing three specified words (“human”, “person” and “people”).

### 5.2 Determining Verbs Representing Speech

When the verb that is the dependent of the character mention extracted in section 5.1 represents speech, the character mention is more likely the speaker of a certain utterance. Therefore, we examine whether the verb represents speech. In this study, we focus only on verbs and do not consider adjectives such as "大きい (loud)" in "花子の声は大きい (Hanako's voice is loud)".

We create a dictionary of verbs representing speech, and consider verbs in this dictionary as speech verb and all other words as non-speech verb. This dictionary is created by automatically collecting a large number of words representing speech from Japanese WordNet[10] (Bond et al., 2009), followed by manually adding two appropriate words and removing 112 inappropriate words, resulting in 3,084 verbs listed in this dictionary.

### 5.3 Attributing Character Mentions to Utterances

#### 5.3.1 Proposed Model

In our preliminary analysis before we design the framework of attributing character mentions to utterances, we found that dominant mention types are mostly restricted to subject mention and sentence head mention. Following this analysis, we define three types of mention based on the dependency relationships between the character mentions and the verb as shown in the examples in Figure 1.

**Subject Mention with Speech Verb** When the verb is a speech verb and its `nsubj` relation points to a character mention, we extract the character mention as the target mention.

**Subject Mention with Non-Speech Verb** When the verb is a non-speech verb and its `nsubj` relation points to a character mention, we extract the character mention as the target mention.

**Sentence Head Mention** When the character mention is detected as the sentence head, we extract the character mention as the target mention.

For those three mention types above, we conduct the procedure below and assign utterances to mentions:[11]

(P1) For each utterance, up to one narrative[12] that is preceding or subsequent to the utterance is searched for any of the three mention types. When there exist multiple mentions of those three mention types, we follow the preference below:[13] the mention in the sentence containing the utterance, the mention in the sentence subsequent to the utterance, and the mention in the sentence preceding the utterance. With the procedure above, only one mention or no mention is detected for each utterance.

(P2) For each mention detected in (P1), even if it is detected for two consecutive utterances[14], we assign both of those two consecutive utterances to the detected mention[15].

Next, to each of the remaining utterances that are not assigned to any mention after the procedures (P1) and (P2) for the three mention types above, we apply the following two procedures of utterance attribution in this order and assign each remaining utterance to a mention.

**Speaker Alternation** Suppose that the mention of the $n$-th utterance is unknown in a series of three consecutive utterances such as $n$-th, $n$+1-th, and $n$+2-th utterances or $n-2$-th, $n-1$-th, and $n$-th utterances[16], we assume

[10] We recursively collect words and sub-synsets in synsets containing eight specified words (e.g. "talk", "express").

[11] Evaluation on the validation data showed that the performance decreased with certain preferences in mention types such as preferring in descending order of "Subject Mention with Speech Verb," "Subject Mention with Non-Speech Verb," and "Sentence Head Mention.

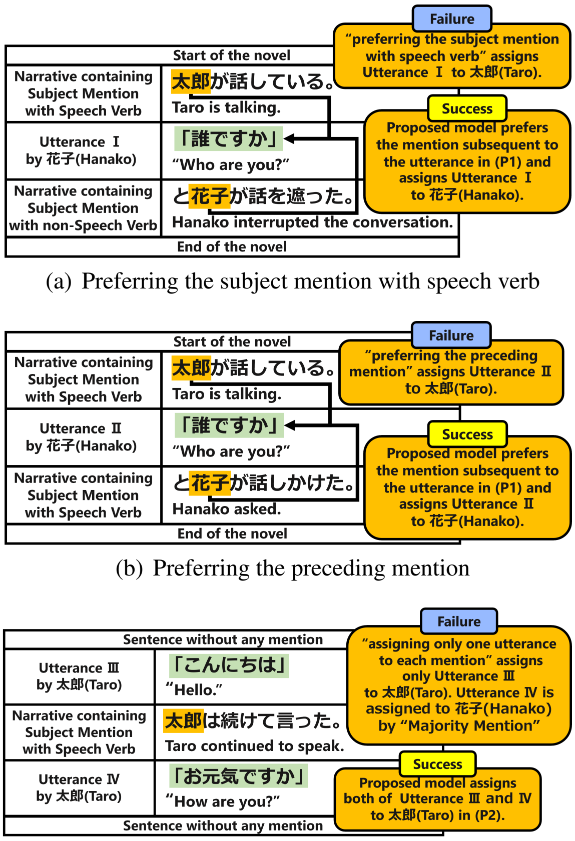
[12] Evaluation on the validation data showed that the performance decreased when the search range was set to two or more narratives.

[13] Evaluation on the validation data showed that the performance decreased with other variants of the preferences in positions of the mention relative to the utterance, such as preferring the mention in the sentence preceding the utterance rather than the one in the sentence subsequent to the utterance.

[14] In (P1), each mention is detected for up to two utterances. This simply follows by satisfying in (P1) the constraint below: i.e., from each utterance, mentions are not searched beyond any other utterance.

[15] Note that we do not consider the constraint of assigning only one utterance to each mention. When we consider this constraint except for "Speaker Alternation" and "Majority Mention" procedures, evaluation on the validation data showed that the performance decreased.

[16] Here, we do not allow this procedure of "Speaker Alternation" to be applied to the cases where one or more narratives are inserted between two consecutive utterances. When we allow this procedure of "Speaker Alternation" to be applied to the cases where one or more narratives are inserted between two consecutive utterances, evaluation on the validation data showed that the performance decreased.


Failure
"preferring the subject mention with speech verb" assigns Utterance Ⅰ to 太郎(Taro).
Start of the novel
Narrative containing Subject Mention with Speech Verb
太郎が話している。
Taro is talking.
Success
Proposed model prefers the mention subsequent to the utterance in (P1) and assigns Utterance Ⅰ to 花子(Hanako).
Utterance Ⅰ by 花子(Hanako)
「誰ですか」
"Who are you?"
Narrative containing Subject Mention with non-Speech Verb
と花子が話を遮った。
Hanako interrupted the conversation.
End of the novel
Start of the novel
Failure
"preferring the preceding mention" assigns Utterance Ⅱ to 太郎(Taro).
Narrative containing Subject Mention with Speech Verb
太郎が話している。
Taro is talking.
Utterance Ⅱ by 花子(Hanako)
「誰ですか」
"Who are you?"
Success
Proposed model prefers the mention subsequent to the utterance in (P1) and assigns Utterance Ⅱ to 花子(Hanako).
Narrative containing Subject Mention with Speech Verb
と花子が話しかけた。
Hanako asked.
End of the novel
Failure
Sentence without any mention
"assigning only one utterance to each mention" assigns only Utterance Ⅲ to 太郎(Taro). Utterance Ⅳ is assigned to 花子(Hanako) by "Majority Mention"
Utterance Ⅲ by 太郎(Taro)
「こんにちは」
"Hello."
Narrative containing Subject Mention with Speech Verb
太郎は続けて言った。
Taro continued to speak.
Success
Proposed model assigns both of Utterance Ⅲ and Ⅳ to 太郎(Taro) in (P2).
Utterance Ⅳ by 太郎(Taro)
「お元気ですか」
"How are you?"
Sentence without any mention

(a) Preferring the subject mention with speech verb

(b) Preferring the preceding mention

(c) Assigning only one utterance to each mention (When "花子 (Hanako) is the most frequent mention")

Figure 2: Examples of failures by variants of the proposed model and successes by the proposed model

that speakers are altered at each of those three consecutive utterances and assign the $n$-th utterance to the mention of the ($n$±2)-th utterances.

**Majority Mention** Each of the remaining utterances that are not assigned to any mention after the procedure of "Speaker Alternation" is assigned to the most frequent[17] mention (i.e., the one with the highest count of attribution with utterances) throughout the entire utterances in the novel.

### 5.3.2 Variants of the Proposed Model

In our experiments, we compare the proposed model with the following three variants. For each of those three variants, Figure 2 shows an example where the result by the variant is failure, while the result by the proposed model is success.

**proposed model** This model attributes utterances to character mentions following the procedure described in section 5.

**preferring the subject mention with speech verb** This model follows the proposed model but with certain preferences in mention types in descending order of "Subject Mention with Speech Verb", "Subject Mention with Non-Speech Verb", and "Sentence Head Mention".

**preferring the preceding mention** This model follows the proposed model but with a variant of the preference in (P1), which is "preferring the mention in the sentence preceding the utterance rather than the one in the sentence subsequent to the utterance".

**assigning only one utterance to each mention** This model follows the proposed model but with the constraint of assigning only one utterance to each mention in (P2).

## 6 Clustering Character Mentions into Character Entities

Finally, we cluster character mentions that refer to an identical character entity. We decompose the character mention into four elements: "proper noun", "prefix"[18], "suffix"[19] and "others"[20] by using GiNZA's Named Entity Recognition and the person name dictionary described in section 5.1. Then, we cluster mentions that share an identical "proper noun" element or an identical "other" element into an identical character entity.

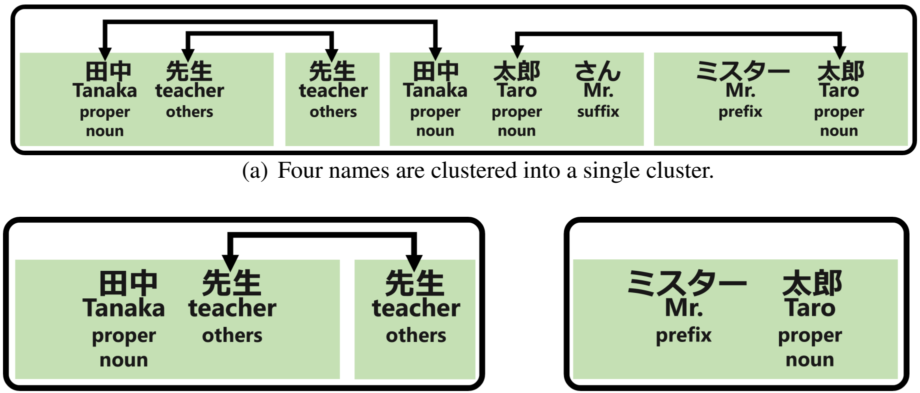
Figure 3 shows examples of character entity clustering. In Figure 3, "田中 (Tanaka) 先生 (teacher)", "先生 (teacher)", "田中 (Tanaka) 太郎 (Taro) さん (Mr.)" and "ミスター (Mr.) 太郎 (Taro)" are extracted and decomposed into the four elements. As shown in Figure 3(a), "田中 (Tanaka) 先生 (teacher)" and "先生 (teacher)" are considered to represent an identical person because they share an identical "others" element "先生 (teacher)". Similarly, the pair of "田中 (Tanaka) 先生 (teacher)" and "田中 (Tanaka) 太郎 (Taro) さん (Mr.)", and the pair of "田中 (Tanaka) 太

[17] Here, we increment the frequency of each mention if two mentions located at the distinct positions in the novel have an identical word sequence.

[18] Japanese language usually does not have prefix to person names. Exceptions include foreign words such as "ミスター (Mr.)"

[19] Morphemes belonging to Japanese suffix grammatical category (e.g., "さん", "君" (Mr., Ms. )).

[20] Other nouns representing role words typically used as suffixes (e.g., "先生 (teacher)", "将軍 (general)") or as one word without any proper noun.



(a) Four names are clustered into a single cluster.

(b) Three names are clustered into two clusters.

Figure 3: Examples of Character Entity Clustering

郎 (Taro) さん (Mr.)" and "ミスター (Mr.) 太郎 (Taro)" are each considered to represent an identical person. Finally, all of these four names are clustered into an identical character entity. If there is no "田中 (Tanaka) 太郎 (Taro) さん (Mr.)", "田中 (Tanaka) 先生 (teacher)", "先生 (teacher)" and " ミスター (Mr.) 太郎 (Taro)" are clustered into two clusters, as shown in Figure 3(b).

## 7 Evaluation Procedure

The performance of the speaker to dialogue attribution is evaluated on the test data, and we use two types of evaluation metrics: clustering evaluation and name matching accuracy. In this evaluation, the utterances to be clustered are provided as oracle data to evaluate only the performance of attributing character mentions to utterances and clustering character mentions into character entities.

### 7.1 Clustering Evaluation

In order to evaluate the clustering results, we use B3 measures Precision, Recall and F1-Score (Enrique et al., 2009), as in the study by Cuesta-Lazaro et al. (2022). The detailed procedure of evaluating character entity clustering is presented in section A.

### 7.2 Name Matching Accuracy

In order to evaluate the proportion of the extracted character mentions that refer to the correct character name, we define the name matching accuracy. The detailed procedure of evaluating the name matching accuracy is presented below. In the BCCWJ speaker annotation data, each utterance $u$ is annotated with a unique reference speaker name $r(u)$. In the results of character entity clustering of section 6, on the other hand, we denote the predicted speaker of the utterance $u$ as $p(u)$, which is predicted as a cluster of character mentions. Then, the name matching accuracy is measured as the proportion of utterances where the predicted speaker $p(u)$ is determined to be the same character entity as the reference speaker name $r(u)$ by the method described in section 6. We denote this evaluation metric as name matching accuracy (A).

An important point to note is that, for 22,124 (66.3%) utterances out of the overall 33,360 test data for which the reference speaker names do not appear within up to one preceding or subsequent narrative[21], our approach can not extract their correct mentions. Therefore, to analyze the effectiveness of our approach using surrounding narratives, we exclude those 22,124 (66.3%) utterances and evaluate on remaining 11,236 (33.7%) utterances, for which the reference speaker name appears within up to one preceding or subsequent narrative. We denote this evaluation metric as the name matching accuracy (B). An example of evaluating name matching accuracy is presented in section B.

Our final goal of the evaluation is the name matching accuracy. However, the name matching accuracy cannot properly evaluate cases where the extracted mentions are pronouns or nicknames because these mentions can not be properly clustered by the proposed clustering procedure even if these

[21] Those cases are roughly categorized into: 1) the proposed approach is not applicable because the utterances are consecutive and are not surrounded by narratives, 2) although utterances are surrounded by narratives, the reference speaker is not present in the surrounding narratives, 3) although the reference speaker is present in the surrounding narratives, its mention is referred to by pronouns.

| Model | #Utterance | #Correct Character Entity Cluster | #Predicted Character Entity Cluster | Precision (%) | Recall (%) | F1-Score (%) | Name Matching Accuracy (%) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | (A) | (B) |
| “proposed model with the highest name matching accuracy of (B)” | 58.6<br>±44.1 | 5.9<br>±4.2 | 6.5<br>±3.9 | 55.9<br>±14.7 | 63.9<br>±15.7 | 58.2<br>±12.3 | 44.5<br>±21.4 | **93.8** |
| “preferring the subject mention with speech verb” | 58.6<br>±44.1 | 5.9<br>±4.2 | 6.4<br>±3.8 | 55.3<br>±14.5 | 63.8<br>±15.7 | 57.8<br>±12.2 | 44.6<br>±21.2 | 93.0 |
| “preferring the preceding mention” | 58.6<br>±44.1 | 5.9<br>±4.2 | 6.7<br>±4.0 | 55.5<br>±14.8 | 62.8<br>±16.0 | 57.2<br>±12.4 | 43.6<br>±20.9 | 89.3 |
| “assigning only one utterance to each mention” | 58.6<br>±44.1 | 5.9<br>±4.2 | 6.5<br>±3.9 | 54.9<br>±15.0 | 66.1<br>±15.2 | 58.5<br>±12.2 | 43.1<br>±22.1 | 79.1 |

(a) Results of macro-average/standard deviation for all the novels in the test data

| Sample Novel ID | #Utterance | #Correct Character Entity Cluster | #Predicted Character Entity Cluster | Precision (%) | Recall (%) | F1-Score (%) | Name Matching Accuracy (%) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | (A) | (B) |
| LBj9_00220 | 58 | 3 | 6 | 79.3 | 46.0 | 58.2 | 58.6 | 85.0 |
| LBl9_00012 | 41 | 4 | 3 | 50.4 | 95.4 | 66.0 | 53.7 | 100.0 |
| LBp9_00033 | 171 | 4 | 10 | 52.8 | 48.4 | 50.5 | 6.4 | 90.9 |

(b) Results of three individual sample novels in the test data

| Mention Type / Utterance Attribution Procedure | #Applied (coverage rate (%)) | Name Matching Accuracy (%) $\left(\frac{\#\text{correct}}{\#\text{applied}}\right)$ |
|---|---|---|
| Subject Mention with Speech Verb | 8,343 (25.0) | 60.7 $(\frac{5{,}063}{8{,}343})$ |
| Subject Mention with Non-Speech Verb | 9,878 (29.6) | 50.8 $(\frac{5{,}018}{9{,}878})$ |
| Sentence Head Mention | 996 (3.0) | 45.6 $(\frac{454}{996})$ |
| Speaker Alternation | 5,063 (15.2) | 39.7 $(\frac{2{,}012}{5{,}063})$ |
| Majority Mention | 9,080 (27.2) | 26.6 $(\frac{2{,}418}{9{,}080})$ |
| Total | 33,360 (100.0) | 40.8 $(\frac{13{,}627}{33{,}360})$ |

(c) Results of each of the three mention types and the two utterance attribution procedures (for all the novels in the test data)

Table 4: Evaluation Results of the Speaker to Dialogue Attribution

mentions are correctly identified by the proposed method. Therefore, we also conduct the clustering evaluation to evaluate these cases.

## 8 Evaluation Results

### 8.1 Overview

Table 4 shows the speaker to dialogue attribution results of the macro-average and standard deviation for the all novels in the test data, as well as the results of each of the three individual sample novels in the test data. First, we analyze the results of the clustering evaluation in Table 4(a). Our proposed model achieves 58.2 F1-Score, indicating that more than half of the utterances are appropriately classified from an identical speaker. Next, we analyze the results of name matching accuracy (A) and (B). Our proposed model achieves 93.8% name matching accuracy (B), while the name matching accuracy (A) is just around 44%. The name matching accuracy (B) is measured mostly by applying the procedures (P1) and (P2) based on the three mention types using narratives. The name matching accuracy (A), on the other hand, is measured mostly by further applying the two procedures of “Speaker Alternation” and “Majority Mention” to the remaining two-thirds of the test data. The name matching accuracy (A) is just as half of that of the name matching accuracy (B), simply because the name matching accuracy of the two procedures of “Speaker Alternation” and “Majority Mention” is relatively lower[22] as shown in Table 4(c).

The name matching accuracy (B) listed in Table 4(a) is comparable to those reported in Muzny et al. (2017); Cuesta-Lazaro et al. (2022). There

[22] This is also the reason why there exists little difference in F1-Score of the clustering performance as well as the name matching accuracy (A) across multiple variants of the proposed model.

was little difference between the proposed model and the variant of "preferring the subject mention with speech verb", indicating that, according to the preference employed by the proposed model, most of the "Subject Mentions with Speech Verb" are located at the preferred position from the utterance compared with other mentions. In addition, the damaged result of the variant of "preferring the preceding mention" indicates that the correct character mentions tend to be in the sentence subsequent to the utterance rather in the sentence preceding the utterance. Finally, the damaged result with the variant of "assigning only one utterance to each mention" indicates that assigning more than one utterances to each mention helps improving the performance of the proposed model, showing that the proposed rules for those three mention types are still not reliable enough in their current implementation.

Next, we analyze the results of speaker to dialogue attribution for each sample novel. From Table 4(b), the novel with the sample ID LBj9_00220 has high precision but low recall. This is because the number of the clusters of the predicted speakers are more than those of reference speakers, causing utterances from an identical speaker to be clustered into different persons. On the other hand, the novel with the sample ID LBl9_00012 has high recall but low precision. This is because almost all the predicted speakers were clustered into one person. Furthermore, the novel with the sample ID LBp9_00033 has a high F1-Score, but the name matching accuracy is significantly lower. This is because most of the extracted names were pronouns like "僕 (I)" and "姉さん (sister)", causing failures in the clustering of an identical person. To solve these issues, we can consider several approaches: identifying all the characters throughout a given novel even if they are not the speakers of utterances and exploiting them in the clustering of pronouns, utilizing speech styles and vocatives (expressions that indicate the party being addressed) in the utterances, and leveraging coreference resolution to determine the clusters including pronouns and proper nouns.

### 8.2 Analysis of the Effect of Mention Types and Utterance Attribution Procedures

This section analyzes the effect of each mention type and utterance attribution procedure. For each mention type and utterance attribution procedure, Table 4(c) shows coverage rates as well as name matching accuracies calculated from the numbers of utterances to which each mention type and utterance attribution procedure is applied and those of their correct application.

Coverage rates of the three mention types range in descending order of "Subject Mention with Non-Speech Verb", "Subject Mention with Speech Verb", and "Sentence Head Mention". Especially, "Sentence Head Mention" is rarely applied, while those of "Speaker Alternation" and "Majority Mention" is 42.4% in total, indicating that about half of utterances depend on the results of utterance attribution of other utterances.

Name matching accuracies of the three mention types range in descending order of "Subject Mention with Speech Verb", "Subject Mention with Non-Speech Verb", and "Sentence Head Mention". "Subject Mention with Speech Verb" is with the highest name matching accuracy, while "Majority Mention" is with the lowest name matching accuracy, indicating that it is not very effective.

## 9 Conclusion

This paper proposed a first end-to-end speaker to dialogue attribution pipeline for Japanese novels. In the utterance extraction task, we showed that the rule-based model, which extracts strings enclosed by specific symbols is insufficient. Then, we proposed a sequence labeling model by BERT, which can extract utterances in various formats with high accuracy. In speaker to dialogue attribution, we showed that extracting speakers using surrounding narratives can appropriately classify about half of the utterances from an identical speaker. In addition, our model could assign the 93.8% utterances, where the reference speaker name appears within up to one narrative preceding or subsequent to the utterance, to correct speaker. Moreover, we analyzed the coverage rate and name matching accuracy for each mention type and utterance attribution procedure in detail, and clarified the effectiveness of each pattern. As a future work, since speaker attribution based only on surrounding narratives is quite limited, it is expected to incorporate other features such as speech styles as well as embedding based characteristics.

## Acknowledgments

In this study, we used the data from the National Institute for Japanese Language and Linguistics project “Multifaceted Study of Spoken Language Based on a Large-Scale Daily Conversation Corpus” (Project Leader: Hanae Koiso) and the Japan Society for the Promotion of Science - Grant-in-Aid for Scientific Research “Expansion of Corpus by Annotating Speaker’s Information to Conversational Sentences” (15H03212).

## References

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the Japanese WordNet. In Proceedings of the 7th Workshop on Asian Language Resources (ALR7), pages 1–8, Suntec, Singapore. Association for Computational Linguistics.

Asli Celikyilmaz, Dilek Hakkani-Tur, Hua He, Grzegorz Kondrak, and Denilson Barbosa. 2010. The actor-topic model for extracting social networks in literary narrative. In Proceedings of the NIPS 2010 Workshop. Machine Learning for Social Computing.

Carolina Cuesta-Lazaro, Animesh Prasad, and Trevor Wood. 2022. What does the sea say to the shore? a BERT based DST style approach for speaker to dialogue attribution in novels. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5820–5829, Dublin, Ireland. Association for Computational Linguistics.

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

David K Elson and Kathleen R McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI’10, pages 1013–1019. AAAI Press.

Amigó Enrique, Gonzalo Julio, Artiles Javier, and Verdejo and Felisa. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. In Inf. Retr., pages 12(4):461–486.

Kevin R. Glass and Shaun Bangay. 2007. A naïve, salience-based method for speaker identification in fiction books. In PRASA 2007: Proceedings of the 18th Annual Symposium of the Pattern Recognition Association of South Africa, pages 1–6.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1312–1320, Sofia, Bulgaria. Association for Computational Linguistics.

Ryo Ishii, Ryuichiro Higashinaka, Koh Mitsuda, Taichi Katayama, Masahiro Mizukami, Junji Tomita, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Yushi Aono. 2021. Methods for efficiently constructing text-dialogue-agent system using existing anime characters. Journal of Information Processing, 29:30–44.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. Language Resources and Evaluation, 48(2):345–371.

Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, and Yoshihiro Matsuo. 2016. Towards an entertaining natural language generation system: Linguistic peculiarities of Japanese fictional characters. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 319–328, Los Angeles. Association for Computational Linguistics.

Chiaki Miyazaki, Saya Kanno, Makoto Yoda, Junya Ono, and Hiromi Wakaki. 2021. Fundamental exploration of evaluation metrics for persona characteristics of text utterances. In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 178–189, Singapore and Online. Association for Computational Linguistics.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 460–470, Valencia, Spain. Association for Computational Linguistics.

Timothy O’Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 790–799, Jeju Island, Korea. Association for Computational Linguistics.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC’02), Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

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 129–138. Association for Computational Lingustics.

# A The Procedure of Evaluating Character Entity Clustering

Firstly, we describe the detailed definition of $B^3$ measures of precision and recall. In the BCCWJ speaker annotation data, each utterance $u$ is annotated with a unique reference speaker name $r(u)$. In this paper, the set $\mathbb{R}$ of the reference clusters of the utterances is represented as $\mathbb{R} = \{R_1, \ldots, R_k\}$, where each reference cluster $R_i$ ($i = 1, \ldots, k$) satisfies the following two constraints:

- $\forall u, \forall u' \in R_i,\ \ r(u)$ and $r(u')$ are identical.
- $i \neq j,\ \ \forall u \in R_i, \forall u' \in R_j,$ $r(u)$ and $r(u')$ are not identical.

The set $\mathbb{P}$ of the predicted clusters of the utterances, on the other hand, is represented as $\mathbb{P} = \{P_1, \ldots, P_l\}$, where each predicted cluster of the utterances is obtained from the results of character entity clustering of section 6. Now, in the evaluation of character entity clustering of section 6, we measure the degree of agreement between the set $\mathbb{P}$ of the predicted clusters of the utterances and the set $\mathbb{R}$ of the reference clusters of the utterances. As the clustering evaluation metrics, we use $B^3$ measures of precision, recall, and F1-Score (Enrique et al., 2009), as in the study by Cuesta-Lazaro et al. (2022).

Specifically, first we denote the result of character entity clustering of section 6 as the set $\mathbb{N} = \{N_1, \ldots, N_l\}$ of the clusters of person names. Next, let $N_i (\in \mathbb{N})$ denote a cluster of person names and $n (\in N_i)$ denote a person name in $N_i$. Also, for each person name $n (\in N_i)$, we denote $P(n)$ as the set of utterances, each of which has the person name $n$ as its predicted speaker. Here, we denote $u (\in P(n))$ as an utterance in $P(n)$ and $p(u)$ as its predicted speaker[23]. Note here that the following relation holds for the person name $n$ and the set $P(n)$ of utterances:

$$P(n) \quad = \quad \{u \mid p(u) = n\}$$

Then, the predicted cluster $P_i (\in \mathbb{P})$ of the utterances corresponding to $N_i (\in \mathbb{N})$ is represented as the union $\bigcup\limits_{n \in N_i} P(n)$ of $P(n)$ over the whole cluster $N_i$ of person names:

$$P_i \quad = \quad \bigcup_{n \in N_i} P(n)$$

Following those definition of notations, the definition of $B^3$ measures of precision and recall is given below. First, given a predicted cluster $P$ and an utterance $u (\in P)$ in $P$, "precision$\big(P, u(\in P)\big)$" is defined as the ratio of whether the reference speaker $r(u)$ of $u$ in $P$ is identical to the reference speaker $r(u')$ of other utterance $u'$ in $P$.

$$\text{precision}\big(P, u(\in P)\big) = \frac{\big|\{u' \in P \mid r(u) \text{ and } r(u') \text{ are identical}\}\big|}{|P|}$$

Then, "precision($\mathbb{P}$)" over the whole set $\mathbb{P}$ of the predicted clusters of the utterances is measured as the micro average of "precision$\big(P, u(\in P)\big)$" over all the utterances within $\mathbb{P}$ as below, and used as the precision of the clustering evaluation metrics.

$$\text{precision}(\mathbb{P}) \quad = \quad \frac{\sum\limits_{P \in \mathbb{P}} \sum\limits_{u \in P} \text{precision}(P, u)}{\sum\limits_{P \in \mathbb{P}} |P|}$$

Next, given a reference cluster $R$ and an utterance $u (\in R)$ in $R$, "recall$\big(R, u(\in R)\big)$" is defined as the ratio of whether the predicted speaker $p(u)$ of $u$ in $R$ is identical to the predicted speaker $p(u')$ of other utterance $u'$ in $R$.

$$\text{recall}\big(R, u(\in R)\big) = \frac{\big|\{u' \in R \mid p(u) \text{ and } p(u') \text{ are identical}\}\big|}{|R|}$$

Then, "recall($\mathbb{R}$)" over the whole set $\mathbb{R}$ of the reference clusters of the utterances is measured as the micro average of "recall$\big(R, u(\in R)\big)$" over all the utterances within $\mathbb{R}$ as below, and used as the recall of the clustering evaluation metrics.

$$\text{recall}(\mathbb{R}) \quad = \quad \frac{\sum\limits_{R \in \mathbb{R}} \sum\limits_{u \in R} \text{recall}(R, u)}{\sum\limits_{R \in \mathbb{R}} |R|}$$

# B An Example of Evaluating Name Matching Accuracy

This section presents an example of evaluating name matching accuracy. Suppose the case where

[23] The predicted speaker $p(u)$ of each utterance $u$ can be considered as one of the predicted clusters $N_1, \ldots, N_l (\in \mathbb{N})$ of person names.

all the person names in a novel are clustered into the two clusters $\{$“田中 (Tanaka) 先生 (teacher)”, “先生 (teacher)”$\}$ and $\{$“ミスター (Mr.) 太郎 (Taro)”$\}$ of Figure 3(b). Also suppose, for all the utterances $u$ in this novel, the reference speaker name $r(u)$ as “田中 (Tanaka)”. Then, according to the method of section 6, the reference speaker name “田中 (Tanaka)” is considered to be the same person as the first predicted cluster of person names $\{$“田中 (Tanaka) 先生 (teacher)”, “先生 (teacher)”$\}$ because they share the “proper noun” element “田中 (Tanaka)”. However, the reference speaker name “田中 (Tanaka)” is not considered to be the same person as the second predicted cluster of person names $\{$“ミスター (Mr.) 太郎 (Taro)”$\}$. Now, suppose the case where the numbers of the utterances for the first and the second predicted clusters of person names are 3 and 2, respectively, the name matching accuracy is measured as $3/(3+2) = 60\%$.