Abstract

Recently, we have been creating a corpus of Japanese prefectural assembly minutes. The corpus contains assembly minutes of all 47 prefectures between April 2011 and March 2015. This four-year period represents one term of office for the assembly members in most prefectures. In prefectural assembly minutes, the name of the speakers can be recorded in several different ways such as Japanese Hiragana and Katakana characters, and Chinese characters. Our purpose of this study is to uniquely identify the speakers by hand, reconciling the different representations of their names. This paper describes how we annotated a Japanese political corpus with speaker identity information.

Keywords: Political documents, local assembly minutes, speaker identification

1. Introduction

Many local autonomies in Japan provide open access to a variety of political documents via their websites. These documents include basic urban development plans, local assembly minutes, and ordinances. Such information obtained through internet can be used to compare local autonomies and identify the individual characteristics of the autonomies. Local assembly minutes are crucial in determining such characteristics because they include various representatives’ positions on the policies enforced by that body.

Some studies to analyze local assembly minutes have been conducted by political scientists (Masuda, 2012). However, these studies have raised issues with the analysis methods of such minutes. One issue concerns the different ways in which these minutes are released to the public. There are 47 prefectures and several cities, towns, and villages in Japan; local assembly minutes are made available in a variety of ways. Gathering local assembly minutes and presenting the data collected in a unified format for analysis at a national level is therefore expensive. In this paper, we aim to create a corpus of Japanese local assembly minutes. Our overall objective is to develop a corpus that can be used for a broad range of interdisciplinary research.

Our goal is to create a prefectural assembly corpus with that is both accurate and complete. To do this, we identify which speaker has made each statement by hand. The issue we would like to solve.

Why is this topic interesting?

• Once we have identified each speaker, we will be able to answer the following three questions.

Q1 How many assembly members spoke in each Japanese prefectural assembly?

Table 1 shows the number of registered assembly members in the 47 Japanese prefectural assemblies for the period between April 2011 and March 2015. We classified the speakers into three categories: 25-44 years old (young), 45-64 years old (middle-aged), and 65-84 years old (old). The age composition of the assembly members during this period was as follows: 598 people were 25-44 years old, 1,790 people were 45-64 years old, 482 people were 65-84 years old, and 76 people were of unknown age. Table 2 shows the numbers of assembly members who were mentioned in the minutes as speakers, while Table 3 shows the numbers of assembly members who were not mentioned in the minutes as speakers. Table 4 divides the members who were mentioned in the minutes as speakers into the following categories: “governor”, “vice-governor”, “chairperson”, “member” and “other (governor’s agent).”

Table 1: Total number of members listed for the 47 Japanese prefectural assemblies for the period between April 2011 and March 2015.

<table>
<thead>
<tr>
<th>Age</th>
<th>25-44</th>
<th>45-64</th>
<th>65-84</th>
<th>Unknown</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>547</td>
<td>1,616</td>
<td>457</td>
<td>71</td>
<td>2,691</td>
</tr>
<tr>
<td>Female</td>
<td>51</td>
<td>174</td>
<td>25</td>
<td>5</td>
<td>255</td>
</tr>
<tr>
<td>Total</td>
<td>598</td>
<td>1,790</td>
<td>482</td>
<td>76</td>
<td>2,946</td>
</tr>
</tbody>
</table>

Q2 Is it possible to identify differences in speech content based on gender? Figure 1 shows the mean frequencies of particular keywords per person by gender, which are normalized using the keyword frequencies for men. This shows that women often used keywords such as “tax increase”, “poverty”, “consumption tax”
Table 2: Number of assembly members who were listed in the minutes as speakers.

<table>
<thead>
<tr>
<th>Age</th>
<th>25-44</th>
<th>45-64</th>
<th>65-84</th>
<th>Unknown</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>538</td>
<td>1,539</td>
<td>396</td>
<td>70</td>
<td>2,543</td>
</tr>
<tr>
<td>Female</td>
<td>50</td>
<td>172</td>
<td>23</td>
<td>5</td>
<td>250</td>
</tr>
<tr>
<td>Total</td>
<td>588</td>
<td>1,711</td>
<td>419</td>
<td>75</td>
<td>2,793</td>
</tr>
</tbody>
</table>

Table 3: Number of assembly members who were NOT listed in the minutes as speakers.

<table>
<thead>
<tr>
<th>Age</th>
<th>25-44</th>
<th>45-64</th>
<th>65-84</th>
<th>Unknown</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>9</td>
<td>7</td>
<td>61</td>
<td>1</td>
<td>148</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>9</td>
<td>63</td>
<td>1</td>
<td>153</td>
</tr>
</tbody>
</table>

Q3 Is it possible to clarify the difference of the speech content depending on age? Figure 2 shows the mean frequencies of particular keywords per person by generation, which are normalized using the keyword frequencies for the middle aged (45-60 years old) group. Young members (25-44 years old) often used keywords such as “internet” and “children wait-listed.” Middle aged members (45-64 years old) made many remarks involving keywords such as “nursing care” and “medical expenses”. Old members (65-84 years old) tended to use keywords such as “regional revitalization” and “self-defense forces”.

Again, we note that it is important to identify the speakers in the local assembly minutes. Only after annotating the minutes with speaker information can we answer the above three questions accurately and completely.

Why could not previous research solve this issue?

1. No other corpus of local assembly minutes which is both accurate and complete exists.

2. Previous research has not analyzed the differences between the 47 prefectural governments.

Our contributions can be summarized as follows.

- We create a political corpus that includes everything that was said in the prefectural assemblies; we are also careful in ensuring that is both accurate and complete (Table 2).
- We identify all the expression ways in which each speaker’s name is expressed (Section 3).
- We identify characteristic statistical distributions using our corpus (Figure 1 and Figure 2).
- We create a training dataset for speaker identification.

2. Creating the corpus

In this section, we used to create our corpus of Japanese prefectural assembly minutes.

2.1. Prefectural assembly minutes

Japan is divided into 47 prefectures. The corpus of prefectural assembly minutes provides a language resource which clearly indicates who spoke in the assembly, as well as when, where and what they said. Figure 3 shows an example of assembly minutes from the Yamagata prefecture. There are a lot of 1,788 local governments throughout Japan, including prefectures, cities, towns, villages, and special wards in Tokyo. Approximately 86% (of these 1,542 municipalities) publish their local assembly minutes on the internet. Minutes are created for several types of public local assembly, including plenary assemblies (regular and extraordinary meetings) and committees (e.g., budget and audit committees). In this paper, we focus on minutes for regular prefectural meetings.

We have collected minutes from the regular meetings in all 47 prefectures that were held between April 2011 and March 2015. This four-year period represents one term of office for the assembly members in most prefectures. The total amount of collected text data is about 1.8 GB.

3. Speaker identification

We have identified all the speakers who were listed in the Japanese prefectural assembly minutes. In this section, we discuss the ways in which we classified and identified the speakers.

3.1. Speaker type

We classified the speakers as follows: “governor”, “vice-governor”, “chairperson”, “member” and “other (governor’s agent).” Table 4 shows the number of speakers of each type. Note that the number of “other” speakers is larger than the total number of “members”.

3.2. Identifying the speakers

As mentioned in the previous section, the speakers include not only just politicians but also other people, such as governor’s agent. The speakers’ names are also expressed in different ways in the local assembly minutes. Furthermore,
these name-expression patterns are different for each municipality, and they also include typographical errors.

In this section, we discuss the reasons why names are expressed in different ways in the minutes. We also annotate the speakers to identify them in the minutes based on the election information given on local government websites.

Identification can be difficult when the names are expressed differently in the minutes. The reasons for these discrepancies are as follows: the name has been written using Japanese syllabary characters; the name is a professional one; old-style characters have been used; or there is a typographical error either in the local assembly minutes or in the election information website. Japanese name can write both Chinese character and Japanese syllabary characters. Some Chinese characters used in Japanese names have old-style characters. We can change them alternately. Therefore, “Japanese syllabary characters” and “old-style characters” are unique problems on Japanese language. We have classified the discrepancies in terms of these five factors, and Table 6 shows the number of examples in each category. Note that several of these factors may contribute to a single discrepancy. Now, we discuss these five contributing factors, and we describe each one using an example.

3.2.1. Names written using Japanese syllabary characters

Commonly-used names written using Japanese syllabary characters may give rise to discrepancies when the last or first name is written using different characters, such Hiragana or Katakana characters. For example, one local assembly member is recorded as “Inamoto” using Japanese
Table 6: Number of discrepancy examples by category.

<table>
<thead>
<tr>
<th>Japanese syllabary characters</th>
<th>Professional names</th>
<th>Old-style characters</th>
<th>Error in minutes</th>
<th>Error on the website</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>447</td>
<td>3</td>
<td>186</td>
<td>1</td>
<td>39</td>
<td>663</td>
</tr>
</tbody>
</table>

Figure 3: Example of assembly minutes from the Yamagata prefecture, as posted online.

syllabary characters in the minutes; however, the member has been mentioned as “Inamoto” using Chinese characters on the election website.

3.2.2. Professional names
When members use professional names such as pen names or stage names, this can give rise to discrepancies. These professional names cannot be identified without creating a correspondence table between the real and professional names. For example, one local assembly member is recorded as “Pretty Nagashima” (his professional name) in the minutes but is mentioned as “Kaoru Kataoka” (his real name) on the election website.

3.2.3. Names written using old-style characters (Kanji variants)
Using old-style characters in one place and the current one in another place causes discrepancy and confusion. For example, one local assembly member is recorded in the minutes as “Sakae 稲” using the old-style characters but is mentioned in the election website as “Sakae 稲” using the current characters.

3.2.4. Typographical error in local assembly minutes
Typographical errors in the local assembly minutes can lead to discrepancies. For example, one local assembly member is recorded as “Ogiwara 葛原” (using a similar Chinese character) in the minutes but is mentioned as “Hagiwara 葛原” (a similar Chinese characters) on the election website.

3.2.5. Typographical errors on the election website
Typographical errors on the election website can also lead to discrepancies. For example, one local assembly member is recorded as “Shin-ichiro 真一朗” (using a similar Chinese character) in the minutes but is mentioned as “Shin-ichiro 真一郎” (using a Chinese character) on the election website.

4. Related work
The speaker identification work is an important issue in information science. Speaker identification has three major tasks as follows: (i) Speaker identification in the story such as novel and children’s story (Iosif et al., 2016; He et al., 2013), (ii) speakers clustering for speech recognition systems (Ahmed et al., 2017), and (iii) identification of speaker’s name in several different ways. Our speaker identification is to resolve (iii) identification of speaker’s name in several different ways. For example, identification of different expressions for a speaker’s name is an important task in NDL (National Diet Library). NDL in Japan provides Web NDL Authorities as the authority control system in Japan.1

Recently, some studies have explored document analysis, sentiment analysis, and political debates from a political viewpoint (Yano et al., 2009; Chambers et al., 2015; Cano-Basave et al., 2016). These studies used various document datasets as political corpora. In this section, we describe corpora that include political information.

Political document analysis studies have employed various document-collections methods (such as blogs (Yano et al., 2009)) on the web; probabilistic models have been proposed for generating both blog posts and comments on blog sites. Hassanali et al. (2010) proposed a technique for automatically tagging political blog posts using support vector machines and named-entity recognition. They used blog documents as a corpus. Chambers et al. (2015) modeled sentiment analysis in the social sciences using Twitter data (over two billion tweets) as corpus. Lerman et al. (2008) automatically predicted the impact of news on the public perception about the political candidates using daily newspaper articles as corpus. Cano-Basave et al. (2016) used semantic frames to model argumentation in speaker discourse. Their presidential political debates corpus comprises 20 debates that took place between May 2011 and February 2012. Iyyer et al. (2014) applied a recursive neural network framework to detect political positions. They performed experiments using the dataset of Congressional debates and an original political dataset as a corpus. As mentioned above, political corpora typically comprise blogs, Twitter data, newspaper articles, and original political-document datasets. Our political corpus, constructed from local assembly minutes, is therefore a novel and valuable source of political information.

5. Conclusion
In this paper, we have described the annotation process we used to identify the speakers in Japanese prefectural assembly minutes. We focused on the minutes from regular meetings in the 47 Japanese prefectures between April

1https://id.ndl.go.jp/auth/ndl/
2011 and March 2015. This four-year period represents one term of office for the assembly members in most prefectures. The speakers’ name were recorded in a variety of ways such as using professional names, Japanese Hira-gana and Katakana, and Chinese characters. Taking this into account, we have created a corpus of prefectural assembly minutes that is both accurate and complete. We have been publishing the website for the corpus as follows: http://local-politics.jp/47pref.

Acknowledgments
This work was supported by JSPS KAKENHI Grant Number JP16H02912 and JP17K02739.

6. Bibliographical References


