Local Government Name Disambiguation on Japanese Regional Assembly Minutes

Fumitoshi Ashihara*, Yasutomo Kimura, Kenji Araki*****

*,***Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Sapporo Hokkaido, 060-0814, Japan
Email: *ashihara@media.eng.hokudai.ac.jp, ***araki@ist.hokudai.ac.jp
**Department of Information and Management Science, Otaru University of Commerce
Midori 3-5-21,Otaru Hokkaido, 047-8501, Japan
Email: kimura@res.otaru-uc.ac.jp

Abstract

In this research, we aim to automatically investigate local government's involvement in important political tasks by retrieving utterances citing other local government actions from corpus of assembly minutes from all over Japan. To achieve this goal, we need to retrieve utterances including local government name first, but there are ambiguous names of local government units. In this paper, we demonstrate location name disambiguation on the corpus local of assemblies minutes. Our approach to the disambiguation is to collect ambiguous local government names from place name data and determine correct place name candidate in the utterances including collected ambiguous locations. Determining correct place name candidate utilizes superordinate place name, the co-occurrence of place names, and the assembly address. We achieved 0.946 point F-measure by combining the superordinate place name and the assembly address information.

1. Introduction

In this paper, we solve ambiguous name problem in local government names on the corpus local of assemblies minutes to extract utterances citing a political action for supporting decision making on local government. Our method utilizes three information; pattern matching to find superordinate place name, place names included in utterance, the assembly address. In experiment, we compare the effectiveness of three information by evaluating precision, recall, F-measure. The experiment results shows the assembly address information is most effective for local government name disambiguation. Furthermore, we achieved 0.946 point F-measure by the combination of pattern matching and the assembly address information.

2. Background

In recent years, Japanese local governments use websites to disclose administrative information to the public, and information about local councils is also available on the Web. Local councils have the functions of determining the basic principles of the local government and monitoring and evaluating the local government. In the regional assembly minutes, discussions on topics such as political issues and budget decisions are recorded during all meetings and sessions. However, despite the fact that local assembly minutes are openly available on the Web, these records are rarely read, because of technical contents and spoken language descriptions which are not elaborated on. Therefore, effective use of regional assembly minutes is needed for people concerned with local governments.

One of usages of regional assembly minutes is to compare own local government action with other local governments. However, comparing own local government with other local governments is a difficult task for general people because people do not know the local governments doing effective actions in political tasks.

> The next action to consider for the forest is *the industrial forest*, which is conducted by Wakayama prefecture first in the country in 2003.
> (In original Japanese: *Tsuginaru shinrin no torikumi nitsuite chumoku subeki ha Kigyo no mori, kore ha, Wakayamaken ga yahari heisei 15 nen ni zenkoku ni sakigakete jisshi itashimashita.)*

Figure 1: An example of citing other local government actions.

In this research, we solve this problem by extracting the local governments that attract the other local government and automatically estimating the local governments doing effective actions in political tasks. One approach to achieve this goal is retrieving utterances citing other local government actions from regional assemblies minutes from all over the country. Figure 1 shows an example of citing other local government actions in regional assemblies minutes.

By collecting utterances such as the example of Figure 1, we can investigate which local government is well referred by the others. For example, if we want to know which local government is doing effective action about aging problem, we search regional assembly minutes with queries about aging problem. Figure 2 shows an example of a search result with query "senior citizens OR care OR population aging OR low birth rate" from collected citing utterances. In this case, Inagi city is recognized as the most cited city and has the system that the city gives points to senior citizens who volunteer as care workers.

To extract utterances citing a political action, we should retrieve utterances which contain local government name and political actions. However, there is ambiguous name problem in local government names. For example, if the

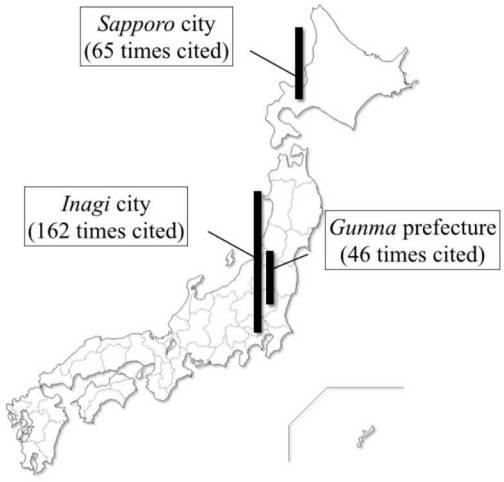

Figure 2: An example of search results with the queries about aging problem in citing utterances.

cited local government is "Shimizu town", there exist 2 local governments and 77 place name called "Shimizu town".

In this paper, we solve local government name disambiguation problem in Japanese regional assembly minutes.

In section 3, we introduce related research and our method for local government name disambiguation. Next, in section 4, we describe an experiment to investigate which method is effective to solve this problem and discuss for improving our methods. Finally, in section 5, we conclude.

3. Local Government Name Disambiguation

In this section, we firstly describe related research about place name disambiguation, and then show our methods, dataset, and experiment for local government name disambiguation.

3.1. Related Research

Place name disambiguation is an important step for geocoding. Typical disambiguation step uses the importance of the place, the main countries the text is about, and the co-occurrence place names (Li, 2003; Forslund, 2006). In recent years, the place name disambiguation in social media is also researched (Chandra, 2011; Han, 2012). These research utilize the words related to place or reply information.

In our method, the main local government name the text is about (superordinate place name information) and the co-occurrence place names are utilized too. In addition to these information, the assembly address, which is attached to all utterances in our regional assembly minutes corpus, is available on our corpus. In our experiment, we investigate which information is effective and whether the combination of methods is valid or not.

3.2. Dataset

At first, we construct ambiguous local government name dictionary using place names and administrative division names. In order to create the dictionary automatically, we search administrative division names from place name dictionary and extract the same name with the target administrative division names.

As a result, We constructed ambiguous local government name dictionary containing190 entries. Table 1 shows example entries of this dictionary. In addition, we divide the candidates into local government name and place name in Table 1.

To clarify the number of ambiguous name candidate, we show the frequency table of ambiguous name candidates in Figure 3. The X-axis shows the number of entries and the Y-axis shows the number of candidates.

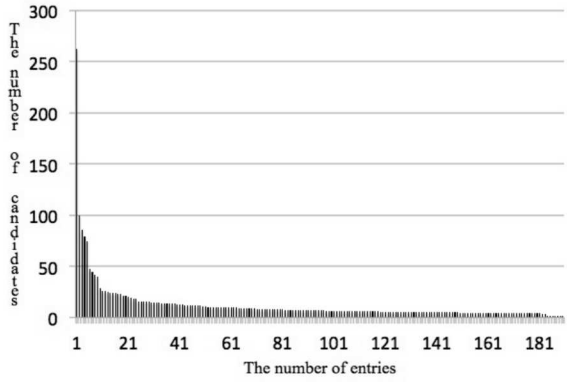

Figure 3: The frequency table of ambiguous name candidates.

Next, we extract utterances including ambiguous local government names from 434 regional assembly minutes from 2010 by using this dictionary. As a result, 10,150 utterances are extracted. Figure 4 shows one of extracted utterances.

> For example, the area around Kitahanada station in the *Kitaku* (ambiguous), where my office exists, is remarkable in population outflow ...
> (In original Japanese: *Tatoeba watashi no jimusho ga gozaimasu kitahanada eki shuhen, kitaku*nai desune, konoatari ha tokuni jinkou ryunyu ga kencho de gozaimashite...)

Figure 4: An example of extracted utterances.

3.3. Disambiguation Method

As described above, we use three pieces of information, superordinate place name, the co-occurrence place names, and the assembly address. These information is utilized with three methods given below.

Method (i) (Pattern matching) : Pattern matching by AB or A □ B (B in A) patterns to find superordinate place name A of ambiguous name B□

Method (ii) (Place names) : Selecting the candidate which has minimum mean distance to place names.

Table 1: Example entries of ambiguous local government name dictionary.

Place name	Local government name	Place name	Number of candidates
Sakae town	Sakae town in Inba county of Chiba prefecture	Sakae town in Kameyama city of Mie prefecture, Sakae town in Nabari city of Mie prefecture,...	263
Nishiki town	Nishiki town in Kuma county of Kumamoto prefecture	Nishiki town in Shimokawa town in Kamikawa county of Hokkaido, Nishiki town in Biei town in Kamikawa county of Hokkaido, ...	100
Showa town	Showa town in Nakakoma county of Yamanashi prefecture	Showa town in Fukuchiyama city of Kyoto prefecture, Showa town in Ono city of Hyogo prefecture,...	86
.....			
Minato-ku	Minato-ku of Tokyo, Minato-ku in Nagoya city of Aichi prefecture, Minato-ku in Osaka city of Osaka prefecture	None	3
Asahi-ku	Asahi-ku in Yokohama city of Kanagawa prefecture, Asahi-ku in Osaka city of Osaka prefecture	None	2
Date city	Date city of Hokkaido, Date city of Hukushima prefecture	None	2

Method (iii) (Assembly address) : Selecting the candidate which has minimum distance to the assembly address.

Let $\mathbf{a} = \{a_1, a_2, ..., a_i, ..., a_n\}$ denote ambiguous place name candidates. The method (i) determines the correct answer by detecting a superordinate place name from the former word. To identify the superordinate place name, we use 1,894 entries place name dictionary automatically collected from address data. If a_i has superordinate place name A, correct answer is chosen as a_i.

Let $\mathbf{l} = \{l_1, l_2, ..., l_j, ..., l_m\}$ denote the place names extracted from the utterance. The method (ii) is represented in following Form (1).

$$a_i = \arg\min_i \frac{\sum_{j=1}^{m} distance(a_i, l_j)}{m} \quad (1)$$

The GIS information for calculating distance is extracted from place name strings by using Distributed Address Matching System[1]. The function $distance(x, y)$ calculates the distance between x and y by using Haversine formula. (Sinnott, 1984).

Figure 5 shows the calculation example of method (ii) when the utterance is “There already exists left hothouses in Hukaya city (place name) and Yori town (ambiguous) for two years in the similar situation.”.

[1]http://newspat.csis.u-tokyo.ac.jp/geocode/modules/dams/index.php

Figure 5: The Calculation Example of Methods (ii).

Firstly, we get GIS information by using Distributed Address Matching System. Next, we calcuate the distance between the place Hukaya city (Saitama prefecture) and Yori town candidates which are ambiguous place name. As a result, Yori town in Oosato county of Saitama prefecture, which has minimum distance to Hukaya city (11.87 kilometer), is selected as the correct answer.

Table 2: Experimental Results

	Method (i) Pattern maching	Method (ii) Place names	Method (iii) Assembly address
Precision	1.000 (6 / 6)†	0.788 (26 / 33)	0.851 (74 / 87)
Recall	0.069 (6 / 87)‡	0.379 (33 / 87)	1.000 (87 / 87)
F-measure	0.129	0.512	0.920

† The numerator shows the number of correct answer and the denominator is the number of correct answers our system outputs.
‡ The numerator shows the number of correct answer and the denominator is the number of data.

The method (iii) utilizes the assembly address s and is represented as below form (2).

$$a_i = \arg\min_i distance(a_i, s) \quad (2)$$

Figure 6 shows the calculation example of method (iii) when the utterance is "I specifically introduce Wake town (ambiguous) which takes Prime Minister's Award." and the assembly address of this utterances is Bizen city of Okayama prefecture.

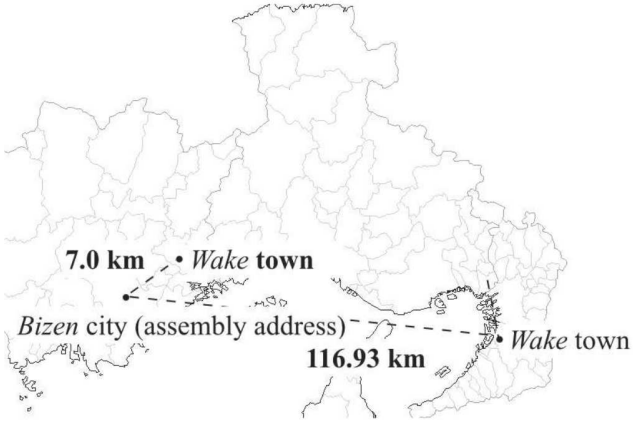

Figure 6: The Calculation Example of Methods (iii)

To calculate the method (iii), we get GIS information in the same way as in method (ii). Next, we calcuate the distance between the assembly address Bizen city of Okayama prefecture and Wake town (ambiguous) candidates. As a result, Wake town in Wake county of Okayama prefecture, which has minimum distance to Bizen city (7.0 kilometer), is selected as the correct answer.

4. Experiment

4.1. Experiment Settings

We evaluate precision, recall, and F-measure for all three methods. For evaluation, we extracted 100 utterances at random from 10,150 entries, and then we exclude 2 utterances, which we could not determine the correct answers and 11 utterances since the target local government name of these utterances is included in other place names.

4.2. Experiment Results

Table 2 shows the results of the experiment. The results for (iii) shows higher F-measure compared to (i) and (ii).

Next, we investigate effect of combining the methods. Table 3 shows the effect of (i) and (ii) when added to (iii). These results indicate that, (i) is considered to improve F-measure by being combined with (iii).

Table 3: The improved number of results utilizing method (i) and (ii) combined with (iii).

	Improved	Getting worsen
Method (i)	4	0
Method (ii)	2	2

4.3. Discussion

In this subsection, we discuss the results of each method and the effectiveness of combination of the methods.

The results of Table 2 shows the following three points.

- The method (i) is able to disambiguate precisely, but the number of utterances matched by this pattern is very low.
- The recall of method (ii) is not so high. The origin of the low recall of method (ii) is considered to be the lack of place name dictionary or the lack of place names in the text.
- The F-measure of method (iii) is the best.

The result of method (iii) indicates that the place of correct answer tends to be near to the assembly address. While most approach for place name disambiguation utilize content information such as place name in text, we show that the metadata of regional assembly minutes corpus is effective for local government name disambiguation task.

Next, we discuss the result of Table 3 which shows the improved number of results utilizing method (i) and (ii) combined with (iii). The result of the method (i) shows it has the ability to complement the method (iii) and it is able to increase F-measure to 0.946 point. Figure 7 shows an example of improved data.

With regard to making free of school lunch fee, ..., the school lunch fee is free in Kazuki town(ambiguous) of Yamaguchi prefecture...
(In original Japanese: *Gakko kyushokuhi no mushoka de arimasu keredomo, yamaguchiken no kazukicho nadodeha, gakko kyushoku wo mushoka ni shi te orimashite...*)

Figure 7: An example of improved data by combining the method (i) with the method (iii).

In this case, the information of assembly address is not available because Kai city of Yamanashi prefecture (assembly address) is not near by Kazuki town of Yamaguchi prefecture.

On the other hand, the method (ii) do not improve the result because of lower precision. However, there is room for improvement on the method (ii) by increasing place name dictionary.

Finally, we discuss the two examples that our system erroneously estimates. Figure 8 shows an example of failed data.

The fact that we can get the connection with Minato-ku where everyone want to get the connection means... (In original Japanese: *Minasan ga paipu wo mochitai minatoku ni ponto tsute ga dekita toiu noha...*)

Figure 8: The first failed example.

Our system fails to answer this example because there is no place name in the utterance and the assembly address is not near by correct answer (Minato-ku of Tokyo). In this case, there was the superordinate place name (Tokyo) in the former utterance. Considering this point, we should check the former utterance whether there is effective contextual information or not.

Next, we consider the other problem from the second example. Figure 9 shows another failed data.

I make an application in Ogaki city, Seki city, Yoro city, Ikeda town (ambiguous), and Sakahogi town because I should bring the application form. (In original Japanese: *Maitoshi, shinsei motte ikana naran toiu kotode, genzai, oogakishi, sekishi, sorekara yorocho, ikedacho, sakamogicho de yatte imasu.*

Figure 9: The second failed example.

The correct answer of this example is the local government candidate because there is the word "application". In this case, the place information is not available because the place names and the candidates are in same prefecture. Figure 10 shows the map of the assembly address and the candidates of this example.

From Figure 10, we can identify the correct answer is far from the assembly address. To improve precision, we should consider the data whose place information is not effective.

5. Conclusion

In this paper, we introduce local government name disambiguation on the regional assemblies minutes corpus.

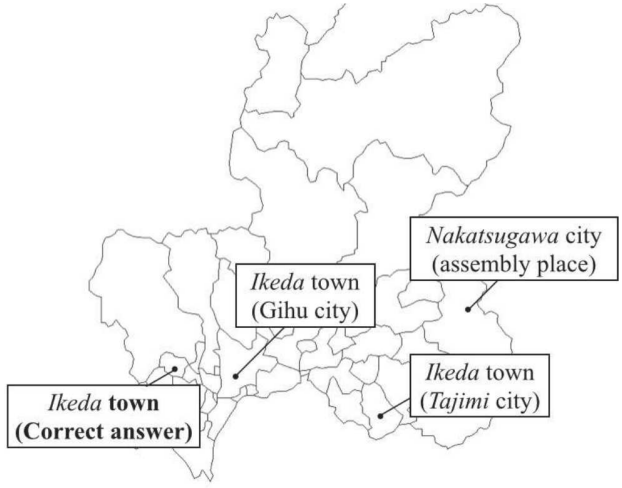

Figure 10: The assembly address and the candidates of the second failed example.

Our approach to the disambiguation is to collect ambiguous local government names from place name data and determine correct place name candidate in the utterances including collected ambiguous place names.

Determining correct place name candidate utilizes pattern matching to find superordinate place name, the place names in the utterance, and the assembly address. As a result, utilizing the assembly address achieves the best F-measure. Furthermore, we achieved 0.946 point F-measure by combining superordinate place name and the assembly address information.

In future work, we plan to improve the method (ii) by increasing place name dictionary and utilizing the effectiveness of place name information out of the utterance. Furthermore, we need to investigate effective methods for failed data whose place information is not available.

6. References

Li H. Srihari R. K. Niu and Li W, 2003. Infoxtract location normalization: A hybrid approach to geographic references in information extraction. *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*:pp. 39–44.

Chandra S. Khan L. Muhaya, 2011. Estimating twit- ter user location using social interactions–a content based approach. *Privacy, security, risk and trust (pas- sat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom)*:pp. 838–843.

Forslund B. Pouliquen M. Kimler R. Steinberger C. Ignat T. Oellinger K. Blackler F. Fuart W. Zaghouani A. Widiger A. and C. Best, 2006. Geocoding multilingual texts: Recognition, disambiguation and visualisation. *In Proceedings of LREC-2006, Genova, Italy*.

Han B. Cook and Baldwin T., 2012. Geolocation prediction in social media data by finding location indicative words. *Proceedings of COLING 2012*:pp. 1045–1062.

Sinnott, R. W., 1984. Virtues of the haversine. *Sky and Telescope*.