Extraction of Political Activity of Assemblyman from Minutes of Municipal Assemblies Using the Political Category

Keiichi Takamaru
Utsunomiya Kyowa University
1-3-18, Odori, Utsunomiya
320-0811, Tochigi, Japan
takamaru@kyowa-u.ac.jp

Hideyuki Shibuki
Yokohama National University
79-7 Tokiwadai, Hodogaya-ku,
240-8501, Yokohama, Japan
shib@forest.eis.ynu.ac.jp

Yasutomo Kimura
Otaru University of Commerce
3-5-21, Midori, Otaru
047-8501, Hokkaido, Japan
kimura@res.otaru-uc.ac.jp

Dai Hasegawa
Hokkaido University
Kita-ku Kita 14 Nishi 9, Sapporo
060-0814, Hokkaido, Japan
hasegawadai@media.eng.hokudai.ac.jp

Hokuto Ototake
Hokkaido University
Kita-ku Kita 14 Nishi 9, Sapporo
060-0814, Hokkaido, Japan
hokuto@media.eng.hokudai.ac.jp

Kenji Araki
Hokkaido University
Kita-ku Kita 14 Nishi 9, Sapporo
060-0814, Hokkaido, Japan
araki@media.eng.hokudai.ac.jp

Abstract

This paper describes extraction of political activities of an assemblyman from minutes of municipal assemblies using the political category. The extraction system is oriented to the political information supporting service between local assemblymen and inhabitants through the World Wide Web. At first, we have constructed the local political categories based on the name of the committees in assemblies and their subjects. The annotation to the minutes has been carried out. The target are 7,084 paragraphs in the minutes of Otaru city assembly in 2007. We have carried out the experiment using SVMs with several new features for minutes. The experiment on the extraction of political activities from minutes using estimated political categories has been carried out. The correspondence rate between annotators' result and system-estimated result is 91.7% when second highest rank is permitted.

1 Introduction

Recently the decentralization is being promoted in Japan. Local politics is important for inhabitants' daily life as well as national politics, and municipal assemblymen are elected by inhabitants as well as the members of the Diet. However there are little information about local politics on mass media such as TV and newspaper. It is because the mass media have time and space limitation and they have to provide information for the nationwide audience. On the other hand, the World Wide Web stores huge amount of various information. Inhabitants can obtain desired political information by a search engine. However it is difficult to find the information which inhabitants are interested in, since they have a huge variety of political interests. Therefore the system which provides local political information in accordance with each inhabitant's interest is needed. We are aiming to construct the political information supporting service between local assemblymen and inhabitants through the World Wide Web. Activities of assemblymen are showed in minutes of assemblies. In our preceding works (Shibuki et al. (2007) and Kimura and Shibuki (2008)), we are proposing a matching system between assemblymen and inhabitants. The system extracts assemblyman's keywords from minutes released on the WWW. They express the activity of each assemblyman. Then, the system extracts political keywords from inhabitants' weblogs. They express political interests and requests of inhabitants. Then, our system connects the inhabitant's political keywords with the assemblyman's keywords. The inhabitants can obtain a list of assemblymen whose activities are close to their interests. However, the full match of both keywords has difficult, because text styles of minutes is different from those of weblogs. Then the extracted keywords are expanded by *Bunrui Goihyo* (The National Institute for Japanese Language (2004)) to increase matching rate in our previous work (Kimura and Shibuki (2008).) To realize furthermore increase of matching rate, we consider that a list of categories of a specific field should be used instead of the *Bunrui Goihyo*, a one of Japanese general thesaurus. Local politics highly reflects locality of each area. Therefore, it is not enough to use general political term dictionary as categories for local politics. We have to originally construct the political category which is specialized in local politics.

In this paper, we construct political categories for extracting the activities of assemblymen described in local assemblies' minutes. Then, we carry out a fundamental experiment for the assemblies' minutes applying natural language processing technique. The experiment is automatic category estimation of the paragraphs in the minutes. Then the experiment on the extraction of political activities from minutes using estimated political categories is carried out.

2 Related Works

There are several preceding researches. Yamamoto and Adachi (2005) have researched on automatic summarization methods of the Diet records using distinctive surface expression. Distinctive surface expressions may also exist in municipal assembly's minutes. However, surface expressions of minutes have the difference among municipalities. The category is needed as common expression of contents of the minutes.

Tomobe and Nagao (2005) and Motomura et al. (2005) have researched on discussion mining methods for semi-automatic creation of minutes and reusing the data as knowledge resources. These technologies make it efficient to use the information on the assemblies' minutes in the future. We have to use the raw texts of assemblies' minutes released on the WWW so far.

3 Construction of the Political Category

At first we construct the provisional category system based on the standing committees and their subjects in Obihiro city assembly[1], since the committees in Obihiro have the detailed lists of subjects. Obihiro city assembly has four kinds of standing committees, "general affairs and education committee", "public welfare committee", "industry and economics committee" and "construction committee." The name of committees and their subjects are treated as the categories. Then there are 59 categories in our provisional category list.

The provisional category list has highly dependency on one city's locality. For example, category group "industry and economics" has the category "*banei*[2]." There is not the category "seaport" though there is "airport" because of geographical factor. We modify the provisional category list by following criteria.

- Categories which have similar meanings are combined into one category.
- A category which have plural meanings is divided into plural categories.
- New categories are added to the list by comparing with minutes of other cities.
- Categories which are concerned with only one city are deleted.

As a result, there are 96 political categories in 5 groups.

4 Researches of Political Categories in Assemblies' Minutes

4.1 Annotation of Categories to Minutes

The purpose of our study is to obtain assemblymen's activities from minutes. The unit of annotation is set to a paragraph, since one paragraph tends to have one topic in the minutes. The annotator basically puts one category to a paragraph. He/She also annotates keywords which are clues to decide the category. The minutes which have

[1] http://www.city.obihiro.lg.jp/sigikai/inkaisyokan.jsp

[2] A kind of public horse race which is held only in Obihiro

Table 1: The Top 5 Political Categories Appeared in the Target Minutes

Rank	Name of Category (ID)	Number of paragraph (%)
1	financial affairs (1010)	615 (11.1%)
2	hospital servise (1101)	273 (4.9%)
3	education (1120)	273 (4.2%)
4	school (1121)	209 (3.8%)
5	medical treatment (1100)	204 (3.7%)

the categories and the keywords in each paragraph are stored by the XML form. These tagged data are also used for training data of the category estimation process in the chapter 5.

We have prepared the minutes from 59 municipalities in Hokkaido. We annotate to subset of them in this paper. We use the minutes of first through fourth regular meetings of Otaru assembly in 2007. The number of overall paragraphs is 7,084. Each paragraph is annotated by two annotators to cancel out the subjective decisions. The categories of each paragraph are the union of categories annotated by two annotators. The total number of annotators is 8. All of them are university students.

As a result of the annotation, there are 2,061 paragraphs which have no category. They are judged not to have any political topics. There are 3,225 paragraphs which have plural categories.

The assemblymen who work for the first regular meeting and second through fourth meetings are not same because of the election held in the year. There are 17 assemblymen who work in both terms. We use only the paragraphs which are uttered by these 17 assemblymen in 4.2 and 4.3.

4.2 Ranking of the Categories

Table 1 shows the rank of the number of the political categories which are contained in the target minutes. The category "financial affairs" ranks first. It forms 11.1%. It seems that "financial affairs" is a major topic in every municipality. The second ranked category "hospital service" is the unique topic in Otaru in 2007. Because there was a problem on a removal of the municipal hospital at that time. Comparing the ranking of categories among municipalities, the local political problems are appeared clearly. The most of assemblymen at the municipality are highly concerned about higher ranked political problems. The categories which express each assemblyman's interest have to take into consideration.

4.3 The Categories for the Assemblymen

In the next step, we confirm if the political interests of each assemblyman can be obtained from categories for paragraphs in the minutes. The cross-tabulation between political categories of paragraphs and assemblymen is shown in Table 2. Table 2 shows the 12 assemblymen who had higher utterance frequencies of these 17 assemblymen. The boxed values are the highest ones in each category. The total utterance frequency of each assemblyman has large difference. However it mainly depends on the parties which they belong to. The total utterance frequency is not an appropriate clue to find assemblyman's activity. We consider that the difference of utterance frequencies in every category expresses assemblyman's feature. For example, the assemblyman A has talked about "education", "school" and "movement of local residents" a lot. The priority of his political activities seems to be those problems. As well, the assemblyman G has the highest frequency at "officer", and the assemblyman L has the highest frequency at "sightseeing." We cannot know their actual political interests. However the minutes include their official utterances. The analysis of minutes is one of unbiased way to know the assemblymen's activity. We consider his/her utterance frequency of the category as intensity of assemblyman's activity.

5 Experiments on Category Estimation

As discussed in the chapter 4, the category for each paragraph in minutes is needed to obtain the assemblymen's activity. However, we cannot put category tags for overall minutes in Japanese municipalities. So natural language processing technique has to be applied to obtaining the categories.

Then we carry out a fundamental experiment on assemblies' minutes. The experiment is estimation of political categories for the paragraphs.

We try to classify the political paragraphs and their categories using SVMs, since SVMs are generally used for various classification tasks in NLP (V.Vapnik (1995), Kudo and Matsumoto (2001) and Joachims (1998).) We use "TinySVM

Table 2: The cross-tabulation between Political Categories of Paragraphs and Uttered Assemblymen in Minutes of Otaru in 2007

Rank	Category	The number of utterance of assemblemen (A descending order of the total number of utterance)											
		A	B	C	D	E	F	G	H	I	J	K	L
1	financial affairs	94	$\boxed{122}$	44	73	40	43	30	17	19	20	32	10
2	hospital service	26	$\boxed{63}$	25	59	19	21	3	13	2	8	7	1
3	education	$\boxed{39}$	2	22	19	18	22	4	0	23	9	4	1
4	school	$\boxed{53}$	14	20	19	12	19	5	0	19	5	4	0
5	medical treatment	24	$\boxed{43}$	2	38	21	17	2	13	7	9	5	0
6	general administration	23	20	$\boxed{24}$	20	19	$\boxed{24}$	22	9	7	2	0	3
7	facility	25	$\boxed{27}$	22	6	24	26	10	7	3	6	7	17
8	budget	22	$\boxed{33}$	20	12	5	9	12	10	12	2	0	0
9	officer	10	18	17	14	8	18	$\boxed{20}$	2	10	6	0	0
10	movement of local residents	$\boxed{23}$	15	13	11	20	2	12	5	12	9	2	2
11	sightseeing	14	1	10	8	1	15	19	0	3	1	6	$\boxed{31}$
12	construction	21	$\boxed{24}$	16	14	8	16	3	6	0	2	1	4
	-omitted-												
	TOTAL	829	803	605	541	527	464	326	213	182	175	159	143

(ver.0.09)[3]” in our experiments. The ”linear kernel” provided in TinySVM is used as the kernel function.

5.1 Preparation of the Data Set

Experimental data are the minutes of first through fourth regular meetings of Otaru city assembly in 2007 which are used in chapter 4. They have 7,084 paragraphs.

The features used in this experiment are “the number of characters”, “nouns”, “verbs”, “adjectives”, “noun bigrams”, “keywords by hand”, “names of talkers and a subject at the head”. They are decided thourgh our preliminary experiments. A length of a paragraph seems to be in proportion to the number of topics in the paragraph. Therefore “the number of characters” is introduced to the features. It is 11 binary values which indicate whether the number of characters are within 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, and more than 200. The content words which correspond to action and modification is need for the features because the purpose of the system is extract activities of assemblymen. Therefore “verbs” and “adjectives” are introduced to features in addition to “nouns.” “MeCab” (Kudo et al. (2004)) is used to analyze morphology and part-of-speech. “Noun bigrams” is introduced to take into account with compound nouns.

“Keywords by hand”, “keywords by TermExtract”, “subjects”, “names of talkers” and “a subject at the head” are introduced to the features which are specialized to minutes. “Keywords by hand” is the keywords which are annotated in chapter 4. “Keywords by TermExtract” is the keywords which are extracted by “TermExtract”(Nakagawa et al. (2003)). “Subjects” is a noun preceding a particle “*wa*.” “Names of talkers” are written at the head of the paragraphs, when the talker is changed. It is extracted automatically using a rule-based method. “a subject at the head” is a binary value which indicates whether the paragraph starts with expression “[noun](*wa*)” which means “[noun] is.”

5.2 Estimation of Political Category

5.2.1 Experiment

A label for the feature vector is a binary value, “-1” or “1”. “1” means the paragraph belongs to a specific category. “-1” means the paragraph does not belong to a specific category. Therefore we prepare 96 different data sets for each category.

The SVM constructs a separating hyperplane using these training data. A real number is outputted, when a test data is inputted. The exper-

[3] http://www.chasen.org/ taku/software/TinySVM/

iment is carried out by 10-fold cross-validation. The threshold for the decision is set to -0.4 in this experiment according to our preliminary experiments.

Recall, precision and F-measure are 0.612, 0.581 and 0.596 each other.

5.3 Extraction of Activity of Assemblymen from Minutes

In this section, we examine the extraction of assemblymen's activities from the minutes. Utterance frequency of an assemblyman expresses intensity of his/her political activities. Therefore a cross-tabulation between the estimated political categories of paragraphs and the assemblymen is shown in Table 3. Table 2, which is the result of annotated categories, is referenced as the correct answer. When the tendency of the utterance frequencies of the both tables is corresponded each other, we conclude that assemblyman's activity can be extracted from estimated categories.

A list of assemblymen in the Table 3 is the same as Table 2. The categories which is shown in Table 2 are also indicated in Table 3. The boxed values are highest ones in every category. The assemblyman who has highest utterance frequency in the categories "education" and "school" is out of the list, since the table shows 12 of 17 assemblymen. When the assemblyman with highest utterance frequency in Table 3 corresponds with that in Table 2, the value in Table 3 is boxed by a double line. There are five such categories, "promotion of general administration", "officer", "movements of local residents", "construction" and "sightseeing." In the six categories, "financial affairs", "hospital service", "facility", "budget", "education" and "school", the assemblymen with the highest utterance frequency in the result of annotated categories correspond with the assemblyman with the second highest utterance frequency in the result of estimated categories.

As a result, in the 11 categories, assemblymen with the highest frequency utterance in the annotated result have the first or second highest frequency utterance in the estimated result. The correspondence rate is 91.7% when second highest rank in the estimated categories is permitted. We can extract the political activity of the assemblyman from estimated categories.

6 Conclusions

We have tried to extract political activities of an assemblyman from minutes of municipal assemblies using the political category. The extraction system is oriented to the political information supporting service between local assemblymen and inhabitants through the World Wide Web. At first, we have constructed the local political categories based on the name of the committees in assemblies and their subjects. The 96 political categories in 5 groups have been constructed. Then the annotation to the minutes has been carried out. The targets are 7,084 paragraphs in the minutes of Otaru city assembly in 2007. We have investigated the assemblymen's utterance frequency in each political category using the cross-tabulation. We have found that the utterance frequency of an assemblyman in every category can express the intensity of his/her political activity.

We have carried out the experiments to estimated political categories for the paragraphs using SVMs.The F-measure of the result was 0.596.

At last, the experiment on the extraction of political activities of an assemblyman from minutes of municipal assemblies using the estimated political category has been carried out. The assemblyman who has highest intensity to the category is correctly estimated in 5 categories of 12 categories. The correspondence rate between annotated result (the correct answer) and estimated result is 91.7% when second highest rank in the estimated categories is permitted. We have found that the estimated categories have the ability to express the activity of assemblyman.

In the next step, we will apply this result to develop the political information supporting service between local assemblymen and inhabitants through the World Wide Web.

Acknowledgments

This research was partly supported by the Strategic Information and Communications R&D Promotion Program under grant No.082301004 from the Ministry of Internal Affairs and Communications, Japan.

References

T. Joachims. 1998. Text categorization with support vector machines: Learning with manyrelevant features. *Proceedings of the European Conference on Machine Learning*.

Table 3: The cross-tabulation between Estimated Political Categories of Paragraphs and Uttered Assemblymen in Minutes of Otaru in 2007

Rank	Category	The number of utterance of assemblymen (A descending order of the total number of utterance)											
		A	B	C	D	E	F	G	H	I	J	K	L
1	financial affairs	$\boxed{108}$	91	32	31	29	39	31	14	19	14	23	9
2	hospital service	32	32	16	$\boxed{34}$	14	15	3	11	2	6	9	0
3	general administration	17	19	11	12	18	$\boxed{\boxed{21}}$	20	5	10	0	7	3
4	medical treatment	$\boxed{30}$	13	2	15	15	21	5	12	4	2	13	1
5	facility	21	28	19	1	19	$\boxed{35}$	14	11	2	7	10	9
6	budget	$\boxed{23}$	22	16	9	5	5	8	6	11	5	3	4
7	education	26	5	7	18	15	23	7	1	24	8	11	0
8	officer	13	11	10	6	10	7	$\boxed{\boxed{18}}$	5	15	3	2	0
9	school	26	9	7	19	13	17	4	1	20	5	5	1
10	movements of local residents	$\boxed{\boxed{27}}$	16	18	13	11	12	9	6	14	7	3	7
-omitted-													
14	construction	15	$\boxed{\boxed{23}}$	6	10	9	9	1	4	2	5	0	5
-omitted-													
16	sightseeing	14	0	6	8	0	11	9	0	3	1	5	$\boxed{\boxed{28}}$
-omitted-													
	TOTAL	671	524	349	298	376	367	230	219	191	122	221	145

Y. Kimura and H. Shibuki. 2008. A method of matching member activity with implicit political opinion extracted from blog. *Natural Language Understanding and Models of Communication*.

T. Kudo and Y. Matsumoto. 2001. Chunking with support vector machines. *Proceedings of NAACL*.

T. Kudo, K. Yamamoto, and Y. Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

K. Motomura, H. Tomobe, and K. Nagao. 2005. A meeting activation mechanism in a discussion mining system. *The 67th National Convention of IPSJ*.

H. Nakagawa, T. Mori, and H. Yumoto. 2003. Term extraction based on occurrence and concatenation frequency. *Journal of Natural Language Processing*, 10(1):27–45.

H. Shibuki, Y. Kimura, and N. Yamazaki. 2007. A proposal of a system for opinion word extraction from minutes. *Forum on Information Technology 2007*, 2.

The National Institute for Japanese Language, editor. 2004. *Bunrui Goihyo(Word List by Semantic Principles, Revised and Enlarged Edition)*. Dainippon-tosho.

H. Tomobe and K. Nagao. 2005. Discussion mining: Knowledge discovery from sets of minutes. *The 67th National Convention of IPSJ*.

V.Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag.

K. Yamamoto and Y. Adachi. 2005. Informative spoken language summarization of the diet minutes. *Journal of Natural Language Processing*, 12(1):51–78.