Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2064–2068
Marseille, 11–16 May 2020

Extraction of the Argument Structure of Tokyo Metropolitan Assembly Minutes: Segmentation of Question-and-Answer Sets

Keiichi Takamaru[1], Yasutomo Kimura[2], Hideyuki Shibuki[3], Hokuto Ototake[4], Yuzu Uchida[5], Kotaro Sakamoto[6], Madoka Ishioroshi[3], Teruko Mitamura[7], Noriko Kando[3,8]
[1]Utsunomiya Kyowa University, Japan [2]Otaru University of Commerce, Japan
[3]National Institute of Informatics, Japan [4]Fukuoka University, Japan [5]Hokkai-Gakuen University, Japan
[6]Yokohama National University, Japan [7]Carnegie Mellon University, USA [8]SOKENDAI, Japan
takamaru@kyowa-u.ac.jp

Abstract

In this study, we construct a corpus of Japanese local assembly minutes. All speeches in an assembly were transcribed into a local assembly minutes based on the local autonomy law. Therefore, the local assembly minutes form an extremely large amount of text data. Our ultimate objectives were to summarize and present the arguments in the assemblies, and to use the minutes as primary information for arguments in local politics. To achieve this, we structured all statements in assembly minutes. We focused on the structure of the discussion, i.e., the extraction of question and answer pairs. We organized the shared task "QA Lab-PoliInfo" in NTCIR 14. We conducted a "segmentation task" to identify the scope of one question and answer in the minutes as a sub task of the shared task. For the segmentation task, 24 runs from five teams were submitted. Based on the obtained results, the best recall was 1.000, best precision was 0.940, and best F-measure was 0.895.

Keywords: Argument mining, Shared task, Local assembly, Tokyo Metropolitan assembly

1. Introduction

Numerous 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 the internet can be used to compare local autonomies and identify their individual characteristics. Local assembly minutes are crucial for determining such characteristics because they include various representatives' positions on the policies enforced by an autonomy. Some studies have been conducted by political scientists and econometricians to analyze local assembly minutes . (Kawaura et al., 2018). However, these studies have raised issues concerning the analysis methods of such minutes. One issue is concerned with 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. For this reason, in this study, we attempt to build the corpus of Japanese local assembly minutes. All speeches in an assembly are transcribed into a local assembly minutes based on the local autonomy law. Therefore, the local assembly minutes result in an extremely large amount of text data. For example, the number of characters of the Tokyo Metropolitan Assembly minutes on February 26, 2013 is over 300,000. Therefore, reading all the minutes becomes a challenge for the residents. Our ultimate goal is to achieve the following two objectives using NLP. One is to summarize and present the arguments in the assemblies. The other is to use the minutes as primary information for arguments in local politics.

In numerous Japanese local assemblies, an assembly person states opinions and questions, all at the same time. Then, a governor and officials answer all the stated questions. A statement of an assembly person often exceeds tens of thousands of characters. The answers are assigned to a governor and an official as necessary. Extracting the argument structure of a local assembly is an extremely crucial task. However, it is not easy to match a question with an answer in the conditions mentioned above.

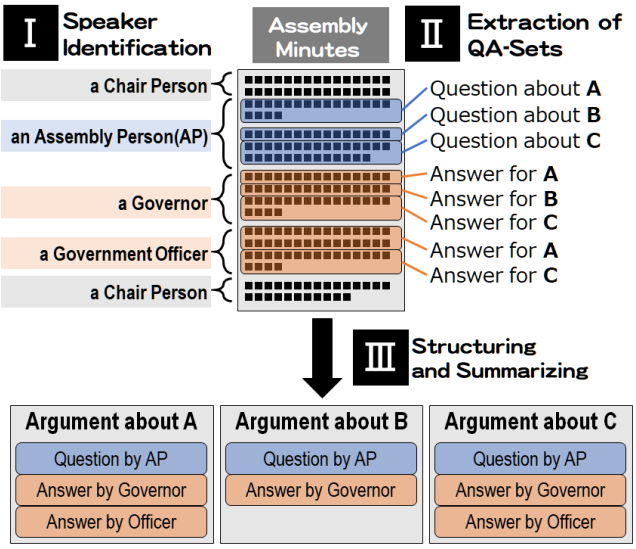

Figure 1: Argument Structure of Local Assembly Minutes

We aim to structure all the statements, starting from the opening statement to the closing statement, in assembly minutes. We focus on the structure of the discussion, i.e., on the extraction of question and answer pairs. The process flow of this study is illustrated in Fig. 1. First, speaker identification is executed in [I] as shown in the Fig. 1. The speaker of each sentence is specified. Next, the sets of ques-

tions and answers are identified in [II] in Fig. 1. Finally, these sets are organized, summarized, and presented to the residents in [III] in Fig. 1.

Step [I] was presented by (Kimura et al., 2018) at LREC 2018. An environment in which the registered corpora could be maintained was created. In this paper, we aim to realize step [II]. To this end, we organized a shared task "QA Lab-PoliInfo" in NTCIR 14. As a sub task of the shared task, we conducted a "segmentation task" to identify the scope of a question and provide an answer to it in the minutes. This paper describes the concept of the structure of the discussion in the minutes and the results of the segmentation task conducted at the shared task NTCIR14 "QA Lab-PoliInfo."

In summary, our main contributions can be listed as follows:

- Segmentation of assembly minutes based on argument mining is essential to the application of a minute, i.e. making a summary and providing important political information.
- We organized Shared Task "QA-Lab Poli-Info" in NTCIR 14 to extract question-and-answer sets in local assembly minutes.
- We released the Segmentation task dataset [1]. The details of the annotation procedure are described in the Dataset section (Sec.4).

2. Structure of Assembly Minute

A local assembly is held four times a year. It usually lasts for 10 - 20 days. Plenary sessions and several committees are held per a local assembly. In the committees, detailed discussion is carried out on individual regulation plan, measures and so on.

In a plenary session, a chair person elected from assembly person proceeds with an assembly. The plenary session contains "deliberation of the bill", "representative questions" (questions as representatives of local political party to the local government), "general questions" (opinions and questions from individual assembly members) and so on.

In the deliberation of the bill, a draft ordinance, a draft budget, petitions from citizens are proposed, discussed and voted. In the representative questions and general questions, several assembly persons ask questions in one day. an assembly person states opinions and questions, all at the same time. Then, a governor and officials answer all the stated questions(Fig 1).

Therefore, following steps are needed in order to determine the argument structure in the Minutes. First, The minutes are divided into "deliberation", "questions and answers", and so on. Then the deliberation sessions are divided into "proposition" and "discussion" and "voting".

The "question and answer" sessions are divided based on questioner. Then questions/answers are divided into individual question/answer. Finally, the set of a question and an answer are identified.

3. Related work

The process of selecting sentences in a document is called segmentation or sentence boundary detection(Zeldes et al., 2019; Azzi et al., 2019). In this section, we discuss the studies related to the process of segmentation or sentence boundary detection.

Sentence boundary detection is a widely used technique. A variety of systems use lists of hand-crafted regular expressions and abbreviations(Aberdeen et al., 1995) Gillick discussed the challenges associated with it, in addition to the relevant features, (Gillick, 2009).

Azzi et. al. organized FinSBD-2019, which focused on sentence boundary detection in PDF noisy text in the financial domain(Azzi et al., 2019). Their shared task aimed at collecting systems for extracting well segmented sentences from financial prospectuses by detecting and marking their beginning and ending boundaries. Sentences are basic units of a written language, and detecting the beginning and end of sentences, or sentence boundary detection (SBD) is a foundational first step in several natural language processing (NLP) applications, such as POS tagging; syntactic, semantic, and discourse parsing; information extraction; or machine translation.

Judge et. al. proposed a Shared Task on structure extraction from financial documents(Juge et al., 2019). Systems participating in this shared task were given a sample collection of financial prospectuses with different levels of structures and lengths. The participant's systems extract structural information and build a table of content.

4. Dataset

In this study, we aim to extract question-and-answer sets (QA-Sets) for structuring and summarizing.

4.1. Annotation procedure

We design a segmentation task to extract the QA-Sets in the assembly minutes. We create the QA-Sets using both Tokyo Metropolitan assembly minutes[2] and *Togikai-dayori* (Newsletter in Japanese)[3]. A newsletter can be regarded as a quoted sentence, and a correct summary of questions and answers created by the assembly office staff. Our target sources are the minutes of the Tokyo Metropolitan Assembly from April 2011 to March 2015 and a summary of a speech of a member of assembly described in *Togikai-dayori*, a public relations paper of the Tokyo Metropolitan Assembly are provided. In the annotation, two annotators find "Starting Line" and "Ending Line" While comparing with both Tokyo Metropolitan assembly minutes and *Togikai-dayori*.

4.2. Annotator Confidence

We requested two annotators to find segments lying between the "Starting Line" and "Ending Line". The number of target sentences in the Tokyo Metropolitan assembly minutes for four years from April 2011 to March 2015 is 115,750.An agreement between the "Starting Line" and "Ending Line" for the two annotators in the segmentation

[1] https://github.com/kmr-y/NTCIR14-QALab-PoliInfo-FormalRunDataset

[2] https://www.gikai.metro.tokyo.jp/record/proceedings/

[3] https://www.gikai.metro.tokyo.jp/newsletter/

task was observed. Cohen ' s Kappa statistic measures the agreement between the two annotators(Cohen, 1960)[4]. Therefore, the Cohen's Kappa for the two annotators (Weights: unweighted) is as follows:

- Subjects (boundaries) = 4,596
- Raters (annotators) = 2
- Kappa = 0.896

In the QA Lab-PoliInfo, the segmentation dataset did not include ambiguous segments because we had two annotators determine all the boundaries with discussion. Table 1 presents the training data and test data. 2

Table 1: Number of Segmentation set

Training data	Test data
298	83

4.3. Dataset release

The segmentation dataset was provided in JSON format. Figure 2 shows an example of the Json format used for segmentation task. The dataset comprised two different parts: the train set and test set, in Japanese. We released the Formal Run dataset on Mar 12, 2019 (https://github.com/kmr-y/NTCIR14-QALab-PoliInfo-FormalRunDataset).

5. Task description

The systems participating in the task are provided with the minutes along with a pair of summaries of a question and answer of a member of the assembly. Based on this, the systems identify the corresponding original speech from the minutes provided to them and answer the positions of the first and last sentences of the identified speech.

> **Input**: The minutes and a pair of summaries of a question and the answer of a member of assembly
>
> **Output**: The first and the last sentences of the original speech corresponding to each summary
>
> **Evaluation**: Recall, precision, and F-measure of the concordance rate of the first and last sentences

As a measure of evaluation, we use the recall R_{seg}, precision P_{seg}, and F-measure F_{seg} of concordance of the first and last sentences to the gold standard data. The aforementioned measures are calculated using the following expressions:

$$R_{seg} = \frac{N_{cp}}{N_{gsp}}, \quad (1)$$

$$P_{seg} = \frac{N_{cp}}{N_{sp}}, \quad (2)$$

$$F_{seg} = \frac{2R_{seg}P_{seg}}{R_{seg} + P_{seg}}, \quad (3)$$

[4]https://cran.r-project.org/web/packages/irr/index.html

Table 2: Data fields used in the Segmentation task

Field name	Explanation
ID	Identification code
Prefecture	Prefecture name
date	According to the Japanese calendar
Meeting	According to *Togikai dayori*
MainTopic	According to *Togikai dayori*
SubTopic	According to *Togikai dayori*
Speaker	Name of member of assembly
Summary	Description in *Togikai dayori*
QuestionSpeaker	Name of member of assembly
QuestionSummary	Description in *Togikai dayori*
AnswerSpeaker	Name of member of assembly
AnswerSummary	Description in *Togikai dayori*
QuestionStartingLine	**Answer section**
QuestionEndingLine	**Answer section**
AnswerStartingLine	**Answer section**
AnswerEndingLine	**Answer section**

Table 3: Participating teams

Team ID	Organization
nami	Hitachi, Ltd.
akbl	Toyohashi University of Technology
RICT	Ricoh Company, Ltd.
KSU	Kyoto Sangyo University
TO*	Task Organizers

*Task organizer(s) are in the team

where N_{cp} is the number of the first and last sentences of which the position is in agreement with the gold-standard position, N_{gsp} is the number of gold-standard positions, and N_{sp} is the number of sentence positions the participants submitted. Figure 3 shows the terms that denote the aforementioned evaluation measures: N_{gsp}, N_{cp} and N_{sp}.

6. Evaluation

NTCIR14 is an evaluation workshop in which multiple participants evaluate their methods using common data. In this section, we describe the segmentation task's results in the NTCIR14. The purpose is to clarify what methods are effective by comparing the participant's methods.

Table 3 shows Participating teams. Table 4 presents the results of the segmentation task in NTCIR14 QA Lab-PoliInfo. For this task, 24 runs from five teams were submitted. Table 4 lists the results of this task. The best recall was 1.000 of nami-11, best precision was 0.940 of nami-01, and best F-measure was 0.895 of RICT-01.

The following describes two highly accurate teams; "nami" and "RICT". The nami's team proposed a method that has filter-by-confidence step after assuming all text segments to be an argument, instead of argument detection step(Yokote and Iwayama, 2019). The nami's method avoided negative affect of noisy text classification process.

The RICT team regarded segments as retrieval objects, and utilized cue-phase-based semi-supervised learning method to detect segment boundary(Jiawei Yong and Shinomiya, 2019). The RICT method consists of the segmentation and the segment search. They prepared another dataset, which was annotated by themselves.

```
{
 "ID":"Segmentation-2018-JA-00020",
 "Prefecture":"Tokyo",
 "Date":"2012-2-28",
 "Meeting":"First regular meeting in 2012",
 "MainTopic":"Support for Foreign nurse candidates"
 "SubTopic":"Deafblind support",
 "QuestionSpeaker":"Yoshio Nakajimas",
 "QuestionSummary":"1. Collaboration with municipalities 2. National Center established in the country",
 "AnswerSpeaker":"Health and Welfare Director",
 "AnswerSummary":"1. Encourage the use of support centers 2. Encourace imporovement of support measures",
 "QuestionStartingLine":23037,
 "QuestionEndingLine":23048,
 "AnswerStartingLine":23234,
 "AnswerEndingLine":23239
}
```


Figure 2: An example of Json format for the Segmentation Task

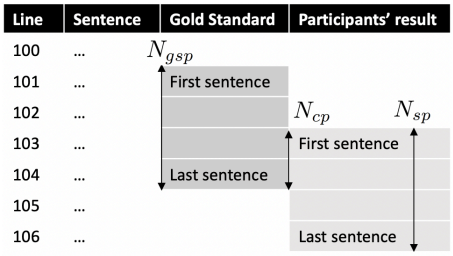

Figure 3: Terms for the evaluation measure : N_{gsp}, N_{cp} and N_{sp}

Table 4: Result of segmentation task

	R		P		F
nami-01	0.814	(1,433/1,761)	0.940	(1,433/1525)	0.872
nami-02	0.864	(1,521/1,761)	0.851	(1,521/1,788)	0.857
nami-03	0.984	(1,733/1,761)	0.499	(1,733/3,475)	0.662
nami-04	0.639	(1,125/1,761)	0.805	(1,125/1,398)	0.712
nami-05	0.553	(973/1,761)	0.931	(973/1,045)	0.694
nami-06	0.655	(1,153/1,761)	0.657	(1,153/1,754)	0.656
nami-07	0.797	(1,404/1,761)	0.933	(1,404/1,505)	0.860
nami-08	0.831	(1,464/1,761)	0.932	(1,464/1,570)	0.879
nami-09	0.875	(1,541/1,761)	0.843	(1,541/1,827)	0.859
nami-10	0.993	(1,749/1,761)	0.464	(1,749/3,769)	0.632
nami-11	1.000	(1,761/1,761)	0.112	(1,761/15,765)	0.201
akbl-01	0.768	(1,352/1,761)	0.538	(1,352/2,515)	0.633
akbl-02	0.847	(1,492/1,761)	0.455	(1,492/3,282)	0.592
akbl-03	0.656	(1,155/1,761)	0.519	(1,155/2,227)	0.580
RICT-01	0.882	(1,554/1,761)	0.909	(1,554/1,709)	0.895
RICT-02	0.856	(1,507/1,761)	0.889	(1,507/1,695)	0.872
RICT-03	0.853	(1,503/1,761)	0.780	(1,503/1,926)	0.815
RICT-04	0.780	(1,374/1,761)	0.746	(1,374/1,842)	0.763
RICT-05	0.936	(1,648/1,761)	0.712	(1,648/2,314)	0.809
KSU-01	0.779	(1,372/1,761)	0.243	(1,372/5,643)	0.370
KSU-02	0.759	(1,337/1,761)	0.268	(1,337/4,998)	0.396
KSU-03	0.820	(1,444/1,761)	0.661	(1,444/2,185)	0.732
KSU-04	0.797	(1,403/1,761)	0.922	(1,403/1,521)	0.855
TO-01	0.354	(623/1,761)	0.898	(623/694)	0.508

7. Conclusion

We structured all the statements from opening to closing in assembly minutes in this paper. We presented the segmentation dataset to extract the argument structure of Tokyo Metropolitan Assembly Minutes. The segmentation dataset was created with the help of two annotators. Cohen ' s Kappa for the two annotators was 0.896. We organized a Shared Task "QA Lab-PoliInfo" in NTCIR 14 to extract QA sets in local assembly minutes. For the segmentation task, 24 runs from five teams were submitted. We released the segmentation task dataset [5]. The details of the annotation procedure are described in the Dataset section (Sec.4). As future work, we plan to summarize the sets of questions and answers to make available simply by residents. We intend to conduct the dialog summarization task in "QA Lab-PoliInfo2" at NTCIR 15[6]; this task summarizes the transcript of a local assembly by considering the dialogue structure. In PoliInfo2, the systems participating in this task summarize the transcript based on the dialogue structure, which comprises "Members' questions" and "Governor's answer." Given the transcript and summary conditions (speaker name, number of summary characters, and so on), the structured document can be successfully generated.

Acknowledgment

This research was partially supported by Secom Science and Technology Foundation and JSPS KAKENHI Grant Numbers JP17K02739, JP16H02912, JP16H01756.

Bibliographical References

Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P., and Vilain, M. (1995). MITRE: Description of the alembic system used for MUC-6. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Azzi, A. A., Bouamor, H., and Ferradans, S. (2019). The FinSBD-2019 shared task: Sentence boundary detection in PDF noisy text in the financial domain. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 74–80, Macao, China, 12 August.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

[5]https://github.com/kmr-y/NTCIR14-QALab-PoliInfo-FormalRunDataset

[6]https://poliinfo2.net/en/

Gillick, D. (2009). Sentence boundary detection and the problem with the U.S. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 241–244, Boulder, Colorado, June. Association for Computational Linguistics.

Jiawei Yong, Shintaro Kawamura, K. K. S. N. and Shinomiya, K. (2019). Rict at the ntcir-14 qalab-poliinfo task. In *Proceedings of The 14th NTCIR Conference*, Tokyo, Japan, 6.

Juge, R., Bentabet, I., and Ferradans, S. (2019). The FinTOC-2019 shared task: Financial document structure extraction. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 51–57, Turku, Finland, 30 September. Linköping University Electronic Press.

Kawaura, A., Kimura, Y., Takamaru, K., and Uchida, Y. (2018). Elected officials in the local assembly: Analysis of prefectural plenary session transcripts. *Doshisha University Center for the Study of the Creative Economy Discussion Paper Series*, (2018-02).

Kimura, Y., Uchida, Y., and Takamaru, K. (2018). Speaker identification for japanese prefectural assembly minutes. In Kiyoaki Shirai, editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Yokote, K.-I. and Iwayama, M. (2019). Nami question answering system at qa lab-poliinfo. In *Proceedings of The 14th NTCIR Conference*, Tokyo, Japan, 6.

Zeldes, A., Das, D., Maziero, E. G., Antonio, J., and Iruskieta, M. (2019). The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104, Minneapolis, MN, June. Association for Computational Linguistics.