Overview of the NTCIR-15 QA Lab-PoliInfo-2 Task

Yasutomo Kimura
Otaru University of Commerce, Japan
RIKEN, Japan
kimura@res.otaru-uc.ac.jp

Hideyuki Shibuki
National Institute of Informatics, Japan
shib@nii.ac.jp

Hokuto Ototake
Fukuoka University, Japan
ototake@fukuoka-u.ac.jp

Yuzu Uchida
Hokkai-Gakuen University, Japan
yuzu@eli.hokkai-s-u.ac.jp

Keiichi Takamaru
Utsunomiya Kyowa University, Japan
takamaru@kyowa-u.ac.jp

Madoka Ishioroshi
National Institute of Informatics, Japan
ishioroshi@nii.ac.jp

Teruko Mitamura
Carnegie Mellon University, U.S.A
teruko@andrew.cmu.edu

Masaharu Yoshioka
Hokkaido University, Japan
yoshioka@ist.hokudai.ac.jp

Tomoyoshi Akiba
Toyohashi University of Technology, Japan
akiba@cs.tut.ac.jp

Yasuhiro Ogawa
Nagoya University, Japan
yasuhiro@is.nagoya-u.ac.jp

Minoru Sasaki
Ibaraki University, Japan
minoru.sasaki.01@vc.ibaraki.ac.jp

Kenichi Yokote
HITACHI, Japan
kenichi.yokote.fb@hitachi.com

Tatsunori Mori
Yokohama National University, Japan
mori@forest.eis.ynu.ac.jp

Kenji Araki
Hokkaido University, Japan
araki@ist.hokudai.ac.jp

Satoshi Sekine
RIKEN, Japan
satoshi.sekine@riken.jp

Noriko Kando
National Institute of Informatics, Japan
SOKENDAI, Japan
kando@nii.ac.jp

ABSTRACT

The NTCIR-15 QA Lab-PoliInfo-2 aims at real-world complex Question Answering (QA) technologies using Japanese political information such as local assembly minutes and newsletters. QA Lab-PoliInfo-2 has four sub tasks, namely Stance classification, Dialog summarization, Entity linking and Topic detection. We describe the used data, formal run results, and comparison between human marks and automatic evaluation scores.

TEAM NAME

Task Organizers

SUBTASKS

Overview

1 INTRODUCTION

The QA Lab-PoliInfo-2 (Question Answering Lab for Political Information 2) task at NTCIR-15 aims at complex real-world question answering (QA) technologies, to show summaries of the opinions of assembly members and the reasons and conditions for such opinions, from Japanese regional assembly minutes.

We reaffirm the importance of fact-checking because of the negative impact of fake news in the recent years. The International Fact-Checking Network of the Poynter Institute established that April 2 would be considered as International Fact-Checking Day from 2017. In addition, fact-checking is difficult for general Web search engines to deal with because of the ‘filter bubble’ developed by Eli Pariser[7], which keeps users away from information that disagrees with their viewpoints. For fact-checking, we should confirm primary sources such as assembly minutes. The description of the Japanese assembly minutes is a transcript of a speech, which is very long; therefore, understanding the contents, including the opinions of the members at a glance is difficult. New information access technologies to support user understanding are expected, which would protect us from fake news.

We provide the Japanese Regional Assembly Minutes Corpus as the training and test data, and investigate appropriate evaluation metrics and methodologies for the structured data as a joint effort of the participants.

The QA using Japanese regional assembly minutes has the following challenges to consider:

1: comprehensible summary of a topic;
2: beliefs and attitudes of assembly members;
3: mental spaces for other assembly members;
4: contexts, including reasons;
5: several topics in a speech; and

	CLEF-2020 Profiling Fake News Spreaders on Twitter	CLEF-2020 CheckThat! Lab	NTCIR QA Lab-PoliInfo-2
Dataset	Twitter	Political debate	Assembly minutes, newsletter and Wikipedia
Task	Profiling Fake News Spreaders on Twitter	Task1: Check-Worthiness on tweets Task2 : Verified claim retrieval Task3 : Evidence retrieval Task4 : Claim verification Task5 : Check-Worthiness on debates	Task1: Stance Classification Task2: Dialog Summarization Task3: Entity Linking Task4: Topic Detection
Number of training data	300 training cases	Task1 88 tweets Task2 1,003 tweets Task5 50 documents	Stance Classification : 2,622 items on the agenda Dialog Summarization : 428+325 topics Entity Linking : 260,366 morphemes (2.7MB)
Language	English and Spanish	English and Arabic	Japanese

Figure 1: Comparison with related shared tasks

6: colloquial Japanese including dialect and slang.

In addition to the QA technologies, this task will contribute to the development of a semantic representation, context understanding, information credibility, automated summarization, and dialog systems.

In the NTCIR-15 QA Lab PoliInfo-2 (hereinafter called "PoliInfo2"), stance classification, dialog summarization, entity linking and topic detection sub tasks were held. The stance classification task is an expansion of the classification task in the NTCIR-14 QA Lab-PoliInfo[4]. Although the classification task aimed to infer individual political policies of assembly members from their speech, the stance classification task aims to infer political party stances of bills from speeches of assembly members in the party. The dialog summarization task is a combinational expansion of the segmentation and the summarization tasks in the NTCIR-14 QA Lab-PoliInfo. The segmentation task aimed to find a description related to a given short text from speeches as source document, and the summarization task aimed to summarize a description from a speech without changing the meaning. The dialog summarization aims to find and summarize descriptions related to a given topic word from question and answer speeches without changing the meaning. In the NTCIR-14 QA Lab-PoliInfo, We observed several inconsistent spellings with the same meaning. To deal with this, we held the entity linking task that is to extract and map descriptions to law names. The topic detection task is an additional task to study a role of political information in order to cope with the outbreak of COVID-19.

2 RELATED WORK

Fake news detection and Fact-checking have recently received significant research attention. Fake News Challenge[1] and CLEF-2018 Fact Checking Lab[2] are shared tasks dealing with political information. Fake News Challenge conducted the Stance Detection task estimating the relative perspective (or stance) of two pieces of text relative to a topic, claim or issue. CLEF-2018 Fact Checking Lab conducted the Check-worthiness and Factuality tasks in both English and Arabic, based on debates from the 2016 US Presidential Campaign[1].

Figure 1 shows a comparison with the related shared tasks such as Profiling Fake News Spreaders, CheckThat! and QA Lab PoliInfo-2. The organizers of Profiling Fake News Spreaders addressed the problem of fake news detection from the author profiling perspective[8]. CheckThat! addressed the development of technology capable of spotting check-worthy claims in English political debates in addition to providing evidence-supported verification of Arabic claims [3][2].

3 JAPANESE REGIONAL ASSEMBLY MINUTES CORPUS

Kimura et al.[5] constructed the Japanese Regional Assembly Minutes Corpus that collects minutes of plenary assemblies in 47 prefectures of Japan from April 2011 to March 2015. Figure 2 shows an example of the minutes of the Tokyo Metropolitan Assembly. Japanese minutes resemble a transcript. In the question-and-answer session, a member of assembly asks several questions at a time, and a prefectural governor or a superintendent answers the questions under his/her charge at a time. A speech is too long to understand the contents at a glance; therefore, information access technologies such as QA and automated summarization, will aid in understanding. For the QA Lab-PoliInfo task, we distributed a subset of the corpus, which is narrowed down to the Tokyo Metropolitan Assembly.

4 TASK DESCRIPTION

We designed the stance classification, the dialog summarization, the entity linking and the topic detection tasks. We put the tasks at the elemental technologies of information credibility or fact-checking for political information systems. Figure 3 shows a relation of the tasks.

Human evaluation has advantage in terms of detailing and deep understanding, while automatic evaluation has advantage in terms of labor and time savings. We used automatic evaluation so that participants could confirm their results immediately during the dry

[1] http://www.fakenewschallenge.org/
[2] http://alt.qcri.org/clef2018-factcheck/

Speaker	Question	Line	Utterance
Yamashita	1	266	まず、**東日本大震災**における被災地支援と東京の防災対策について伺います。
		267	三月十一日、マグニチュード九・〇、最高震度七の強く長い揺れが東日本一帯を襲うとともに、大津波、海砂を巻き込んだ黒く重い海水の塊が太平洋沿岸の防波堤を軒並み破壊し、海水や瓦れきが市街地に流れ込み、甚大な被害を引き起こしました。
		268	福島第一原子力発電所にも大津波が押し寄せ、冷却電源を失った原子炉建屋は爆発、格納容器が損傷して、放射性物質が広範に拡散しました。
		269	原発周辺の住民の皆さんは、自宅があるのに帰れない深刻な状況が続いています。
		270	私たちは、この未曾有の複合災害に対していち早く被災地支援と都内の震災対策を充実させること、そして補正予算の編成を知事に申し入れいたしました。
		271	
		272	また、各議員は、党の被災地支援活動やＮＰＯと連携した取り組みを行うなど、被災地支援に取り組んでまいりました。
		273	そこで伺います。
		274	都は、児童生徒への心のケアや、災害時要援護者の救護など、医療人材の継続的な派遣や、地元雇用を推進する自治体事業、キャッシュ・フォー・ワークといった取り組みへの支援をするなど、被災者の皆さんが希望を見出し、一歩踏み出すことのできるよう、生活再建をともにサポートしていくことが重要です。
		275	また、各県が創造的復興、もしくは再生を目指し、独自復興計画を策定、実現させていくことを都が後押しし、安全な地域社会の再建に寄与していく必
		276	要があります。
		277	このように被災地が取り組むべき課題は山積し、日々刻々地域ごとに状況が変化しております。
		278	被災地のニーズを的確に把握し、**被災地**、被災者**が真に必要とする支援に**今後とも**継続して取り組むべき**と考えますが、**知事の見解**を伺います。
		279	現在、都内には福島県などから自主避難してきた約五千名の避難者の皆さんが都営住宅などに仮住まいをしていらっしゃいます。
		280	故郷から遠く離れ、いつ帰れるのかという思いを持って生活している皆さんに、都は寄り添う形でその生活を支えていくべきと考えます。
		281	避難者は、見知らぬ東京での生活が不安であり、特に高齢者の方々については、引きこもりがちになるなど、孤立化も懸念されます。
		282	先日、特別区の都営住宅で、自治会の皆さんが避難者と懇談会を開き、福島での共通の話題で盛り上がりました。
		283	こうしたかかわり合いをふやす場でもあるミニ懇談会を開催し、避難者同士や地域との交流機会を創造することを求められております。
			また、福祉も含めた総合的な相談を区市町村や災害復興まちづくり支援機構、ＮＰＯなどと連携して開催するなど、広い協働の形で避難者の暮らしを支えることも重要と考えます。
			都は、コミュニティにも配慮した避難者に対する支援の取り組みを行っていくべきと考えますが、都の見解を伺います。
Yamashita	2	284	東日本大震災を教訓に、東京においても発災時における社会対応力の強化や防災リーダーなど、地域人材の育成などに一層取り組み、東京を災害に強い持続可能な都市としていかなければなりません。
		285	現在、各道府県や市町村で地域防災計画などを見直す動きが出ています。
		286	今回の震災による大津波は、近年研究が進みつつあった平安期の貞観地震に類似したものといわれています。
		287	高知県や茨城県では、既に江戸期の地震の実例を盛り込み、地域防災計画の策定や浸水想定を行っています。
		288	東京においても、江戸期に三連動地震による大津波、これに続く暴風雨や富士山噴火による複合災害が起きており、過去の災害分析からも改めて被害想定を研究すべきと考えます。
		289	実践的訓練やライフラインの耐震化、減災化のさらなる推進も必要です。
		290	福島原発事故を踏まえるのであれば、近い将来必ず起きるといわれている東海地震による静岡県浜岡原発事故リスクをも想定した放射能対策も行わなければなりません。
		291	地震、津波の被害想定の検討や防災対策の総点検、そして**東京の総合防災力**を**さらに高める取り組みが必要**だと考えますが、知事の見解を伺います。

Figure 2: Example of the minutes of the Tokyo Metropolitan Assembly

Table 1: Data fields used in the assembly minutes of the stance classification task

Field name	Explanation
Date	Date
Prefecture	Prefecture name
ProceedingTitle	Title of proceeding
Proceeding	List with Speaker and Utterance as elements
URL	Tokyo Metropolitan website

and formal runs. After the formal run, human evaluation was used for detailed analysis.

For automatic evaluation, we introduced leader boards of the tasks, which were published on the QA Lab PoliInfo-2 website[3]. Participants could post their system results once a day.

4.1 Stance classification

4.1.1 Purpose. Stance classification task aims at estimating politician's position from politician's utterances. In PoliInfo2, system participating in the task estimates the stances of political parties from the utterances of the members of the Tokyo Metropolitan Assembly. Given the Tokyo Metropolitan Assembly, topics (agenda), member's list and political denomination list, and the systems classify their stance into two categories (agreement or disagreement) for each agenda.

[3]https://poliinfo2.net/

4.1.2 Data. We distributed the assembly minutes and an answer sheet. Table 1 and 2 show the data fields of the minutes and the answer sheet, respectively. The data sizes of them are shown in Table 3 and 4, respectively. Examples of the minutes and the answer sheet are shown as below.

ソースコード **1: Minutes for Stance Classification**

```
[
    {
        "Date": "2001/8/8",
        "Prefecture": "東京都",
        "ProceedingTitle": "平成十三年第一回臨時会会議録",
        "URL": "https://www.gikai.metro.tokyo.jp/record/extraordinary/2001-1.html",
            "Proceeding": [
            {
                "Speaker": " 議会局長（細渕清君）",
                "Utterance": " 議会局長の細渕でございます。"
            }
            ]
    }
]
```

ソースコード **2: Answer sheet for Stance Classification**

```
[
    {
        "ID":"PoliInfo2-StanceClassification-JA-Dry-Training-02543",
        "Prefecture":"東京都",
```

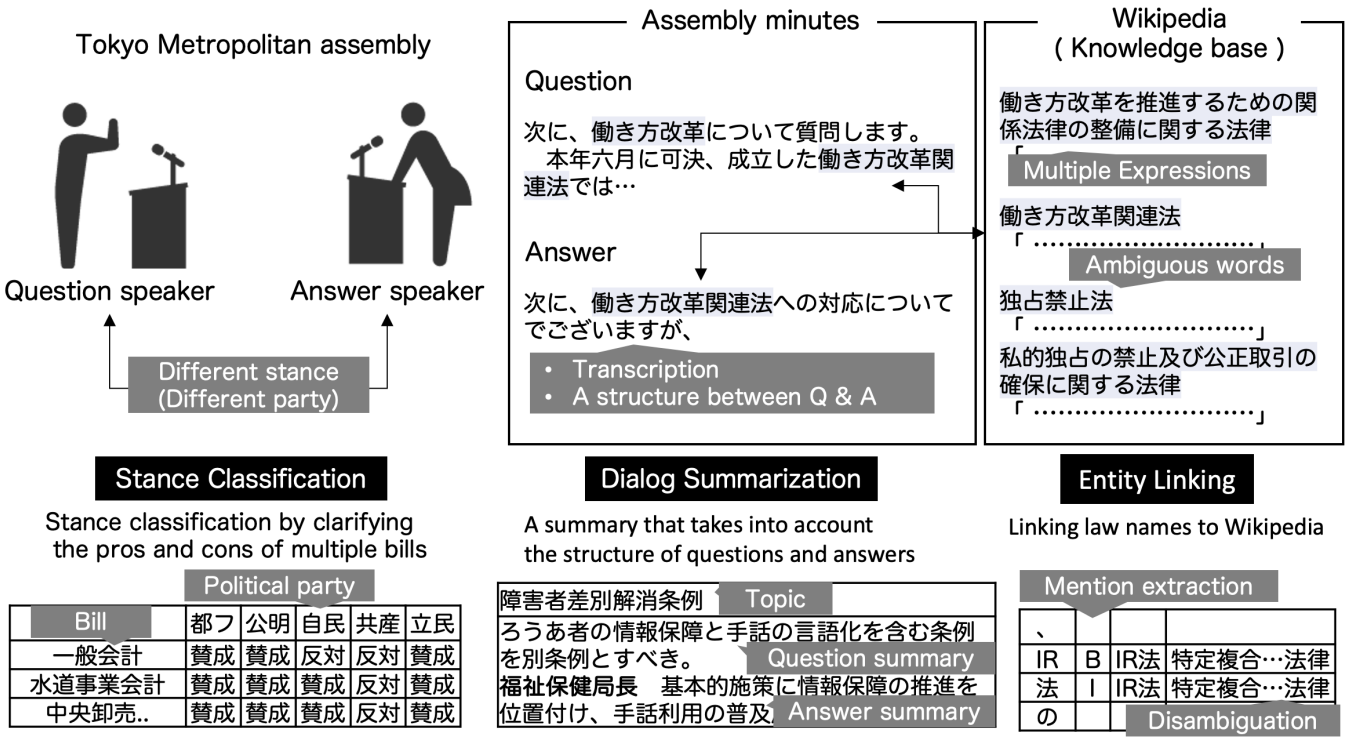

Figure 3: Relation of the three tasks

Table 2: Data fields used in the answer sheet of the stance classification task

Field name	Explanation
ID	Identification code
Prefecture	Prefecture name
Meeting	Meeting name
MeetingStartDate	Start date of the meeting (Date type)
MeetingEndDate	End date of the meeting (Date type)
Proponent	Proponent is either a governor or an assembly member .
BillClass	Class
BillSubClass	Sub class
Bill	Bill name
BillNumber	Bill identification
SpeakerList	Assembly member and political party
ProsConsPartyListBinary	**Answer section** agreement or disagreement (binary)
ProsConsPartyListTernary	**Answer section** agreement, disagreement or NS (ternary)

Table 3: Tokyo Metropolitan assembly minutes for the stance classification task

Minutes	Number of files	File size
regular and extra meetings	2	97MB
committee meetings	30	462MB

Table 4: Data size of the answer sheet in the stance classification task

Answer sheet	Number of questions	File size
Training	2,622	8.1MB
Test	479	1.4MB

```
5        "Meeting":"平成３１年第１回定例会、第１回臨時会",
6        "MeetingStartDate":"2019/2/20",
7        "MeetingEndDate":"2019/3/28" ,
```

```
        "Proponent" : "知事提出議案",
        "BillClass" : "予算",
        "BillSubClass": "３１年度予算",
        "Bill" : "一般会計",
        "BillNumber" : "第一号議案",
        "SpeakerList" :{
            "増子ひろき" : "都ファースト",
            "吉原修" : "自民党",
        },
        "ProsConsPartyListBinary" :
        {
            "都ファースト" : "賛成",
            "公明党" : "賛成",
            "自民党" : "反対",
            "日本共産党" : "反対",
        },
        "ProsConsPartyListTernary" :
        {
            "都ファースト":null,
            "公明党":null,
            "自民党":null,
            "日本共産党":null,
        }
    }
]
```

Input	Tokyo Metropolitan assembly minutes (Proceedings, Committees)
	Answer sheet
Output	Binary : agreement or disagreement
	Ternary : agreement, disagreement or NS
Evaluation	The average of accuracy

4.1.3 Evaluation. For automatic evaluation, pros and cons of bills published by assembly secretariat were used as gold standard data.

$$Score = \frac{1}{NB}\sum_{i=1}^{NB}\frac{NCA(i)}{NPP(i)} \quad (1)$$

where NB is Number of Bill, $NCA(i)$ is Number of Correct Answers for $Bill_i$ and $NPP(i)$ is Number of Political Party,

4.2 Dialog summarization

4.2.1 Purpose. Dialog summarization task aims at summarizing the transcript of local assembly, taking the structure of dialogue into account. In PoliInfo2, systems participating in this task summarize the transcript based on the dialogue structure, which consists of "Members' questions" and "Governor's answer". Given the transcript and summary conditions (speaker name and number of summary characters etc), they generate the structured document.

4.2.2 Data. For the dialog summarization task, 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*[4], a public relations paper of the Tokyo Metropolitan Assembly are provided. Table 5 and 6 show the data fields of the minutes and the answer sheet, respectively. The data sizes of them are shown in Table 7 and 8, respectively. An example of the answer sheet is shown as below.

[4]https://www.gikai.metro.tokyo.jp/newsletter/ (in Japanese)

Table 5: Data fields used in the assembly minutes of the dialog summarization task

Field name	Explanation
ID	Identification code
Line	Line number
Prefecture	Prefecture name
Volume	Volume
Number	Day of the meeting
Year	Year
Month	Month
Day	Day
Title	Title
Speaker	Speaker
Utterance	Utterance

Table 6: Data fields used in the answer sheet in the dialog summarization task

Field name	Explanation
ID	Identification code
Date	Date
Prefecture	Prefecture name
Meeting	Meeting name
MainTopic	Main topic
QuestionSpeaker	Question speaker
SubTopic	Sub topic
QuestionSummary	Summary of question
QuestionLength	Limit length of summary
QuestionStartingLine	Starting line of question
QuestionEndingLine	Ending line of question
AnswerSpeaker	Answer speaker
AnswerSummary	Summary of answer
AnswerLength	Limit length of summary
AnswerStartingLine	Starting line of question
AnswerEndingLine	Ending line of question

Table 7: Tokyo Metropolitan assembly minutes for the dialog summarization task

Minutes	Number of files	File size
Proceedings	1	42MB

Table 8: Data size of answer sheet in the dialog summarization task

Answer sheet	Number of questions	File size
Training with segment	438	414KB
Training without segment	325	292KB
Test	254	161KB

ソースコード **3: Minutes for the dialog summarization**

```
{
"ID":"130001_230617_2",
"Line":2,"Prefecture":"東京都",
"Volume":"平成 23年_第 2 回",
"Number":"1",
"Year":23,
"Month":6,
"Day":17,
"Title":"平成 23年_第 2 回定例会(第 7 号)",
"Speaker":"和田宗春",
"Utterance":"ただいまから平成二十三年第二回東京都議会定例会を開会いたします。"
}
```

ソースコード **4: Answer sheet for the dialog summarization**

```
[
{
  "AnswerEndingLine": [ 532 ],
  "AnswerLength": [ 50 ],
  "AnswerSpeaker": [ "知事" ],
  "AnswerStartingLine": [ 528 ],
  "AnswerSummary": [
    "全国の先頭に立ち刻苦する被災地を支援するのは当然。今後も強力に後押しする。"
  ],
  "Date": "2011-06-23",
  "ID": "PoliInfo2-DialogSummarization-JA-Dry-Training-Segmented-00001",
  "MainTopic": "東京の総合防災力を更に高めよ<br>環境に配慮した都市づくりを",
  "Meeting": "平成 23年第 2回定例会",
  "Prefecture": "東京都",
  "QuestionEndingLine": 276,
  "QuestionLength": 50,
  "QuestionSpeaker": "山下太郎（民主党）",
  "QuestionStartingLine": 266,
  "QuestionSummary": "被災地が真に必要とする支援に継続して取り組むべき。知事の見解は。",
  "SubTopic": "東日本大震災"
}
]
```

4.2.3 Evaluation. For the Leader board for automatic evaluation, we used ROUGE-1 Recall[6] to calculate the score as . We used a summary of newsletter as the gold standard data.

For human evaluation, we used the following quality questions by the participants. The quality questions were assessed by a three-grade evaluation (*A*, *B* and *C*) for content, well-formed, non-twisted, sentence goodness and dialog goodness, respectively. However, for the content evaluation, we prepared an extra grade *X* because a summary that does not include contents of gold standard data may be acceptable. The quality question score $QQ(v)$ from viewpoint v was calculated using the following expressions:

$$QQ(v) = \frac{\sum_{s \in S} g(s,v)}{|S|} \tag{2}$$

$$g(s,v) = \begin{cases} 2 & (gradeA) \\ 1 & (gradeB) \\ 0 & (gradeC) \\ a & (gradeX) \end{cases} \tag{3}$$

where S is a set of summaries the participants assessed, and a is a constant representing whether acceptable summaries that are different from the gold standard summary are regarded as correct or not. If such summaries are regraded as correct, a is 2; otherwise, a is 0.

Input	1. Tokyo Metropolitan assembly minutes 2. Answer sheet in Json format
Output	A summary that takes into account the structure of the dialogue between question and answer.
Evaluation	ROUGE-1 and human marks

4.3 Entity Linking

4.3.1 Purpose. Entity linking task aims at identifying political terms included in politicians' statements, and is to resolve mention recognition, disambiguation and linking the mention with the knowledge base. In PoliInfo2, Entity linking is the task of assigning a unique identity of "law name" which is one of the political terms. Given local assembly member's utterances, and systems extract a mention of "law name" and link the mention with the list of law names or Wikipedia.

4.3.2 Data. Table9 shows data fields used in the answer sheet of of the entity linking task. The answer sheet is TSV format, which is similar to AIDA CoNLL-YAGO Dataset.The data size is shown in Table 10. Figure 4 shows an example of the answer sheet.

私			
の			
方			
から			
は			
、			
ＩＲ	B	ＩＲ法	特定複合観光施設区域の整備の推進に関する法律
法	I	ＩＲ法	特定複合観光施設区域の整備の推進に関する法律
の			
導入			
に			
伴う			
変化			

Figure 4: Example of the entity linking file

4.3.3 Evaluation. Because the gold standard data was made by human workers and checked by participants, human evaluation did not conducted.

$$Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

$$Precision = \frac{NCM}{NOM} \tag{5}$$

$$Recall = \frac{NCM}{NGSM} \tag{6}$$

where NCM is the number of correct mentions, NOM is the number of outputted mentions, and NGSM is the number of gold standard mentions. The correct mention means that the expression is

Table 9: Data fields used in the answer sheet of the entity linking task (TSV format)

Field name	Explanation
Column1	Word segmented tokens by morphological analysis
Column2	A tag is either B (beginning of a mention) or I (continuation of a mention)
Column3	A full mention used to search for candidate entities
Column4	Wikipedia title
Column5	Wikipedia URL

Table 10: Data size the entity linking task

Answer sheet	Number of morphemes	File size
Training	260,366	2.7MB
Test	209,862	1.9MB

Input	1. Answer sheet in TSV format 2. Wikipedia titles (2019-12-01)
Output	Column 2 and 3 : An entity mention extraction Column 4 : Wikipedia title for the mention
Evaluation	End-to-end evaluation from column2 to column4

exactly matched the gold standard and that the mapped entity is correct.

4.4 Topic detection

4.4.1 Purpose. In order to cope with the outbreak of COVID-19, it is important to speedily provide the latest information of COVID-19 for citizens. Considering the possibility of information access technologies, we held another task using the local assembly minutes, namely, topic detection task.

Newsletters issued by local government can consistently provide arguments in local assembly to citizens, while it takes a long time to make them. Although local government also provides newsflashes for providing arguments promptly, there is room for improvement of comprehensibility at a glance. Therefore, the topic detection task aims to make a list of argument topics from newsflashes of local assembly minutes.

Input Newsflashes
Output Lists of dialog topic words/phrases per speaker

4.4.2 Data. An example of the answer sheet is shown as below.

ソースコード **5: Minutes for the topic detection**

```
[
    {
        "Date": "2020/2/19",
        "Prefecture": "東京都",
        "ProceedingTitle": "令和二年東京都議会会議録第一号",
        "URL": "https://www.gikai.metro.tokyo.jp/record/proceedings/2020-1/01.html",
            "Proceeding": [
            {
                "Speaker": "議長（石川良一君）",
                "Utterance": "ただいまから令和二年第一回東京都議会定例会を開会いたします。\n　これより本日の会議を開きます。\n"
            }
    }
]
```

4.4.3 Evaluation. We did not conduct quantitative evaluation. Task organizers and participants discussed appropriate topic words and the application.

4.5 Schedule

The NTCIR-15 QA Lab-PoliInfo task has been run according to the following timeline:

September 30, 2019: QA Lab-PoliInfo Kickoff Meeting
October 18, 2019: First round table meeting in NII
December 15, 2019: Second round table meeting in NII
February 15, 2020: Dataset release

Dry Run
April 23, 2020: First online round table meeting
May 7, 2020: Dry Run
May 27, 2020: Second online round table meeting using zoom
June 27, 2020: Third online round table meeting using zoom
June 30, 2018: Submission Deadline for Dry Run

Formal Run
July 1- 12, 2020: Update of dataset
July 31, 2020: Task Registration Due for Formal Run (This is not required for Dry Run participants)
July 13 - 31, 2020: Formal Run (Stance classification, Dialog summarization and Entity linking)

NTCIR-15 CONFERENCE
August 1 - 7, 2020: Evaluation by participants
August 8 - 14, 2020: Evaluation by organizers
August 15, 2020: Evaluation Result Release
August 26, 2020: Fourth online round table meeting using zoom
September 1, 2020: Task overview paper release (draft)
September 20, 2020: Submission due of participant papers
November 1, 2020 Camera-ready participant paper due
December 8-11, 2020: NTCIR-15 Conference & EVIA 2020

5 PARTICIPATION

Eighteen teams were registered, but only 15 teams participated actively, namely, submitted any results. Table 11 shows the active participating teams.

6 SUBMISSIONS

Table 12 shows the number of submissions. The number in brackets means the number of late submissions. In the dry run, there were 19 submissions from 5 teams for the stance classification, 2 submissions from 2 teams for the dialog summarization and a submission from a team for the entity linking. In the formal run, there

Table 11: Active participating teams

akbl*	Toyohashi University of Technology
knlab	Shizuoka University
wer99	Tokyo Institute of technology
Ibrk*	Ibaraki University
LIAT*	RIKEN Center for Advanced Intelligence Project (AIP)
HUHKA*	Hokkaido University
JRIRD	The Japan Research Institute, Limited
selt	Waseda University
nukl*	Nagoya University
rnyk**	individuals
Forst*	Yokohama National University
SKRA	Hokkaido University
TKLB	Osaka Electric-Communication University
wfrnt*	HITACHI
TO*	task organizers

*Task organizer(s) are in the team
**only the dry run

Table 12: Number of submitted runs

Team ID	Dry run			Formal run			
	Stance classification	Dialog summarization	Entity linking	Stance classification	Dialog summarization	Entity linking	Topic detection
akbl	7	-	-	5	-(1)	-	2
knlab	1	-	-	6	-	-	-
wer99	4	-	-	9	-	-	-
Ibrk	2	-	-	4	-	-	1
LIAT	-	-	-	-	1(1)	-	-
HUHKA	-	-	1	-	-	8(4)	-
JRIRD	-	1	-	-	3	-	-
selt	-	-	-	-	-	4	-
nukl	-	-	-	-	4	1(1)	1
rnyk	5	-	-	-	-	-	-
Forst	-	-	-	2(3)	5(5)	4	-
SKRA	-	-	-	-	1	-	-
TKLB	-	-	-	-	-	-	1
wfrnt	-	-	-	-	2	-	-
TO	-	1	-	-	3	-	-
Sum	19	2	1	26(3)	19(7)	17(5)	6

were 29 submissions (including 3 late submissions) from 5 teams for the stance classification, 26 submissions (including 7 late submissions) from 8 teams for the dialog summarization and 22 submissions (including 5 late submissions) from 4 teams for the entity linking. Fpr the topic detection, there were 6 submissions from 5 teams. In total, there were 90 submissions from 15 teams.

7 RESULT

7.1 Dry run

We conducted only automatic evaluation in the dry run. Table 13, 14 and 15 show results of the stance classification, the dialog summarization and the entity linking in the dry run, respectively. Because the test data of the stance classification was corrected on May 22 and June 5, we separated the results according to the period.

Table 13: Accuracy in the stance classification task at the dry run

ID	team	Accuracy
from May 7 to May 21		
116	rnyk	.9457
72	akbl	.9437
90	wer99	.9375
114	rnyk	.9325
115	wer99	.9284
from May 22 to June 4		
120	rnyk	.9493
125	akbl	.9467
124	wer99	.9416
119	rnyk	.0001
from June 5 to July 4		
144	Ibrk	.9569
143	knlab	.9523
129	rnyk	.9499
139	akbl	.9494
126	akbl	.9472
132	akbl	.9466
131	akbl	.9422
130	wer99	.9382
136	akbl	.8927
140	Ibrk	.8839

Table 14: ROUGE-1-R scores in the dialog summarization task at the dry run

ID	team	ROUGE
141	JRIRD	.2865
137	TO	.2436

Table 15: F-measures in the entity linking task at the dry run

ID	team	ROUGE
108	HUHKA	.4049

7.2 Formal run

Automatic evaluation and human evaluation were conducted in the formal run. Table 16, 17 and 18 show automatic evaluation results of the stance classification, the dialog summarization and the entity linking in the formal run, respectively.

Table 19 shows a human evaluation result of the stance classification.

Table 20, 21 and 22 show human evaluation results of the dialog summarization. Table 23 shows the Cohen's kappa scores for the human evaluation of the dialog summarzation.

Although the deadline was July 31, we accepted submissions until August 31. They were treated as late submissions. Table 24, 25 and 26 show results of the late submissions of the stance classification, the dialog summarization and the entity linking, respectively.

Table 16: Accuracy in the stance classification task at the formal run

ID	team	Accuracy
175	wer99	.9976
177	wer99	.9976
202	wer99	.9976
191	wer99	.9970
196	wer99	.9952
186	wer99	.9923
182	wer99	.9910
205	Ibrk	.9650
180	Ibrk	.9644
149	Ibrk	.9600
167	Ibrk	.9598
203	knlab	.9531
214	knlab	.9531
199	knlab	.9529
158	knlab	.9520
160	knlab	.9520
156	akbl	.9498
204	akbl	.9498
218	akbl	.9496
198	akbl	.9492
153	wer99	.9481
154	wer99	.9461
193	knlab	.9452
169	akbl	.9399
171	Forst	.9388
164	Forst	.9382

8 OUTLINE OF THE SYSTEMS

We briefly describe the characteristic aspects of the participating groups systems and their contribution below.

The akbl team tackled the Stance Classification, the Dialog Summarization, and the Topic Detection tasks. For the Stance Classification task, they used a rule-based analyzer on the opinion statements at first, then, for those left undetermined, they applied a BERT-based stance classifier on the debate statements. For the Dialog Summarization task, they firstly searched for the relevant segment, then extracted the final sentence to form the output summary. For the Topic Detection task, they employed a clustering algorithm on the BERT embeddings of initial topic candidates extracted by using regular expressions, then their final topics were selected based on the centroid of each cluster.

The knlab team tackled the Stance classification task. For the Stance Classification task, they designed features obtained from a sentiment dictionary and BERT, then trained LightGBM to classify the stances.

The wer99 team tackled the stance classification task. They designed a set of rules to recognize an explicit mention to a stance for a bill. When a party does not mention a stance explicitly, they use clues in the bill name to predict a stance.

Table 17: ROUGE-1-R scores in the dialog summarization task at the formal run

ID	team	ROUGE
189	JRIRD	.3208
185	JRIRD	.2980
195	JRIRD	.2980
216	nukl	.2581
148	TO	.2436
215	Forst	.2410
187	nukl	.2387
161	nukl	.2274
172	nukl	.2198
200	Forst	.2145
194	Forst	.2093
157	TO	.1331
208	wfrnt	.1171
151	TO	.1164
181	wfrnt	.1058
176	Forst	.0782
184	Forst	.0729
211	SKRA	.0696
206	LIAT	.0555

Table 18: F-measures in the entity linking task at the formal run

ID	team	F-measure
212	HUHKA	.6035
201	HUHKA	.4887
155	HUHKA	.4747
174	HUHKA	.4468
197	HUHKA	.4468
150	HUHKA	.4049
192	HUHKA	.3980
217	Forst	.3910
183	Forst	.3656
147	Forst	.3389
166	HUHKA	.3247
146	Forst	.3089
173	selt	.2980
178	selt	.2978
179	selt	.2978
213	selt	.2930
190	nukl	.2375

The ibrk team tackled the Stance Classification task. They develop rule-based system for the Stance Classification task by detecting the word "agree" or "disagree" with each bill in speaker's utterances. If its word is not obtained in the utterances about a bill, they categorize his/her opinion into "agree" or "disagree" according to some heuristics.

The LIAT team tackled the Summarization task. For the Summarization task, they took an approach of sentence extraction. They

Table 19: Accuracy of stance classification task at the formal run (human evaluation results)

ID	team	Accuracy
149	Ibrk	−−
153	wer99	−−
154	wer99	−−
156	akbl	0.668
158	knlab	0.805
160	knlab	0.834
164	Forst	0.144
167	Ibrk	−−
169	akbl	0.838
171	Forst	0.852
175	wer99	0.978
177	wer99	0.978
180	Ibrk	−−
182	wer99	0.978
186	wer99	0.978
191	wer99	0.978
193	knlab	0.834
196	wer99	0.978
198	akbl	0.675
199	knlab	0.805
202	wer99	0.982
203	knlab	0.805
204	akbl	0.892
205	Ibrk	−−
214	knlab	0.805
218	akbl	0.888

decomposed the task into border detection, topic matching, and extractive summarization and used an attention mechanism to solve each subtask.

The HUHKA team tackled the Entity Linking task. For Entity Linking task, they extracted mentions of "law name" with BERT, and filter the extracted mentions. For the extracted mentions, they performed disambiguation using exact match, Wikipedia2Vec, mention-entity prior, and e-Gov.

The JRIRD team tackled the Dialog Summarization subtask. For the Dialog Summarization subtask, they developed a BERT-based module that extracts candidate sentences, and a UniLM-based module that generates a summary from the extracted sentences.

The selt team tackled the Entity Linking tasks. For the Entity Linking task, they detected the mentions using fine-tuned BERT and disambiguate the entities of them with wikipedia2vec. To improve their system performance, they used some rules for mention and entity decision.

The nukl team tackled the Dialog Summarization, Entity Linking, and Topic Detection tasks. For the Dialog Summarization task, they applied Progressive Ensemble Random Forest (PERF) developed at the NTCIR-14 QA Lab-PoliInfo to sentence extraction and sentence reduction. For the Entity Linking task, they applied simple matching. For the Topic Detection tasks, they used a rule-based approach.

Table 20: Quality question scores of the dialog summarization in the formal run (Content, Well-formed and Sentence goodness)

ID	team	num of summaries	Content (X=2)		Content (X=0)		Well-formed		Sentence goodness	
148	TO	533	398.5	0.748	357.5	0.671	843.0	1.582	389.0	0.730
157	TO	533	258.0	0.484	210.0	0.394	803.0	1.507	228.5	0.429
176	Forst	533	188.5	0.354	146.5	0.275	748.0	1.403	138.0	0.259
181	wfrnt	533	157.0	0.295	138.0	0.259	688.5	1.292	146.0	0.274
185	JRIRD	533	540.5	1.014	479.5	0.900	975.5	1.830	555.5	1.042
187	nukl	533	422.5	0.793	375.5	0.705	867.5	1.628	423.5	0.795
189	JRIRD	533	576.5	1.082	519.5	0.975	990.5	1.858	601.5	1.129
206	LIAT	533	176.0	0.330	136.0	0.255	867.0	1.627	122.5	0.230
208	wfrnt	533	175.0	0.328	158.0	0.296	730.5	1.371	160.5	0.301
211	SKRA	533	184.5	0.346	135.5	0.254	888.5	1.667	151.5	0.284
215	Forst	533	414.5	0.778	355.5	0.667	906.5	1.701	415.5	0.780
216	nukl	533	442.0	0.829	398.0	0.747	896.0	1.681	445.5	0.836

Table 21: Quality question scores of the dialog summarization in the formal run (Non-twisted)

ID	team	num of evaluable	Non-twisted all		Non-twisted evaluable	
148	TO	259	539.0	1.011	429.5	1.658
157	TO	172	377.0	0.707	255.5	1.485
176	Forst	90	279.0	0.523	113.5	1.261
181	wfrnt	102	224.0	0.420	159.0	1.559
185	JRIRD	360	650.0	1.220	569.0	1.581
187	nukl	270	557.5	1.046	456.0	1.689
189	JRIRD	373	701.5	1.316	638.5	1.712
206	LIAT	104	247.0	0.463	147.0	1.413
208	wfrnt	109	248.5	0.466	172.5	1.583
211	SKRA	118	283.5	0.532	164.5	1.394
215	Forst	274	556.5	1.044	435.5	1.589
216	nukl	292	598.5	1.123	496.5	1.700

Table 22: Quality question scores of the dialog summarization in the formal run (Dialog goodness)

ID	team	num of topics	Dialog goodness	
148	TO	254	124.0	0.488
157	TO	254	69.5	0.274
176	Forst	254	33.5	0.132
181	wfrnt	254	22.0	0.087
185	JRIRD	254	215.5	0.848
187	nukl	254	138.5	0.545
189	JRIRD	254	238.0	0.937
206	LIAT	254	28.0	0.110
208	wfrnt	254	27.0	0.106
211	SKRA	254	43.0	0.169
215	Forst	254	153.5	0.604
216	nukl	254	156.5	0.616

The Forst team tackled the Stance Classification, the Dialog Summarization, and the Entity Linking tasks. For the Stance Classification task, they used a rule-based approach taking into account the date of assembly, speaker name and bill name. For the Dialog Summarization task, they extracted sentences using word embedding similarity between a sentence and a passage including it. For the Entity Linking task, they extracted mentions using BiLSTM-CRF model and disambiguated the entities using RNN model.

SKRA team tackled Dialog Summarization task. They extracted key sentences using an unsupervised extraction method based on EmbedRank++. The team TKLB tackled the Topic Detection task. For the task, they proposed to find differences of opinions and positions among the participants based on the co-occurrence graph. To reflect the broader contexts that all of the given discussions provide, they applied the Latent Dirichlet Allocation (LDA) to weight each word.

The wfrnt team tackled the Dialog-Summarization task. For the Dialog-Summarization task, they investigated whether heuristics of conclusion extraction in Japanese is useful to develop a baseline system for summarization. They quantitatively verified the validity of examination of language use such as "English begins with conclusion, Japanese begins with background."

9 CONCLUSIONS

We described the overview of the NTCIR-15 QA Lab-PoliInfo-2 task. The goal is realizing complex real-world question answering (QA) technologies, to show summaries of the opinions of assembly members and the reasons and conditions for such opinions, from Japanese regional assembly minutes. We conducted in a dry run and a formal run, which are including the stance classification, dialog summarization, entity linking and topic detection sub tasks. There were 105 submissions from 15 teams in total. We described the task description, the collection, the participation and the results.

REFERENCES

[1] Pepa Atanasova, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and

Table 23: Cohen's kappa scores for human evaluation of dialog summarization task in the formal run

Team	Content	Well-Formed	Non-Twisted	Sentence goodness	Dialog goodness
JRIRD	0.602	0.454	0.387	0.447	0.310
nukl	0.378	0.303	0.337	0.392	0.314
Forst	0.287	0.109	0.354	0.400	0.358
wfrnt	0.317	0.156	0.392	0.400	0.354
SKRA	0.328	0.311	0.369	0.349	0.325
LIAT	0.449	0.417	0.360	0.414	0.365
TO	0.586	0.417	0.325	0.477	0.369

Table 24: Accuracy of the late submissions of the stance classification task in formal run

ID	team	Accuracy
234	Forst	.9408
232	Forst	.9391
226	Forst	.8642

Table 25: ROUGE-1-R scores of the late submissions of the dialog summarization task in the formal run

ID	team	ROUGE
235	Forst	.1384
231	Forst	.1219
242	Forst	.1155
239	Forst	.1133
240	Forst	.1040
224	LIAT	.0946
237	akbl	.0621

Table 26: F-measures of the late submission of the entity linking task in the formal run

ID	team	F-measures
238	HUHKA	.5863
233	HUHKA	.5518
236	HUHKA	.5000
229	HUHKA	.3980
225	nukl	.3813

Verification of Political Claims. Task 1: Check-Worthiness. *CoRR* abs/1808.05542 (2018). arXiv:1808.05542 http://arxiv.org/abs/1808.05542

[2] Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of CheckThat! 2020 — Automatic Identification and Verification of Claims in Social Media. In *Proceedings of the 11th International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF '2020)*. Thessaloniki, Greece.

[3] Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019. Overview of the CLEF-2019 CheckThat!: Automatic Identification and Verification of Claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (LNCS)*. Lugano, Switzerland.

[4] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura, Noriko Kando, Tatsunori Mori, Harumichi Yuasa, Satoshi Sekine, and Kentaro Inui. 2019. Overview of the NTCIR-14 QA Lab-PoliInfo Task. *Proceedings of The 14th NTCIR Conference* (6 2019).

[5] Yasutomo Kimura, Keiichi Takamaru, Takuma Tanaka, Akio Kobayashi, Hiroki Sakaji, Yuzu Uchida, Hokuto Ototake, and Shigeru Masuyama. 2016. Creating Japanese Political Corpus from Local Assembly Minutes of 47 prefectures. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*. The COLING 2016 Organizing Committee, Osaka, Japan, 78–85. https://www.aclweb.org/anthology/W16-5410

[6] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013

[7] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Books.

[8] Francisco Rangel, Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2020. Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In *CLEF 2020 Labs and Workshops, Notebook Papers*, Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol (Eds.). CEUR-WS.org.