Overview of the NTCIR-16 QA Lab-PoliInfo-3 Task

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

Hideyuki Shibuki
BESNA Institute Inc.
Japan
shib@besna.institute

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

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

Ken-ichi Yokote
HITACHI
Japan
kenichi.yokote.fb@hitachi.com

Kazuma Kadowaki
The Japan Research Institute, Limited
Japan
kadowaki.kazuma@jri.co.jp

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

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

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

Satoshi Sekine
RIKEN
Japan
satoshi.sekine@riken.jp

ABSTRACT

The goal of the NTCIR-16 QA Lab-PoliInfo-3 task is to develop real-world complex question answering (QA) techniques using Japanese political information such as local assembly minutes and newsletters. QA Lab-PoliInfo-3 consists of four subtasks: QA Alignment, Question Answering, Fact Verification, and Budget Argument Mining. In this paper, we present the data used and the results of the formal run.

TEAM NAME

Task Organizers

SUBTASKS

Overview

1 INTRODUCTION

The aim of the Question Answering Lab for Political Information 3 (QA Lab-PoliInfo-3) task of NTCIR-16 is to develop complex real-world question answering (QA) techniques. In this task, the participants extract and summarize the utterances of the National Diet of Japan and local assembly members, verify the authenticity of the utterances, and analyze the structure of the discussions.

Fact checking has become increasingly important due to the growing concern of fake news. In 2017, the International Fact-Checking Network of the Poynter Institute established April 2 as International Fact-Checking Day. Fact-checking is difficult for general Web search engines because of the "filter bubble" as coined by Pariser [17], which keeps users away from information that disagrees with their viewpoints.

We suggest using primary sources such as assembly minutes for fact checking. Japanese assembly minutes are speech transcripts, which are very long, and it can be difficult to understand the contents at a glance, such as the opinions of the members. New information access technologies to support user understanding are expected, which would protect us from fake news.

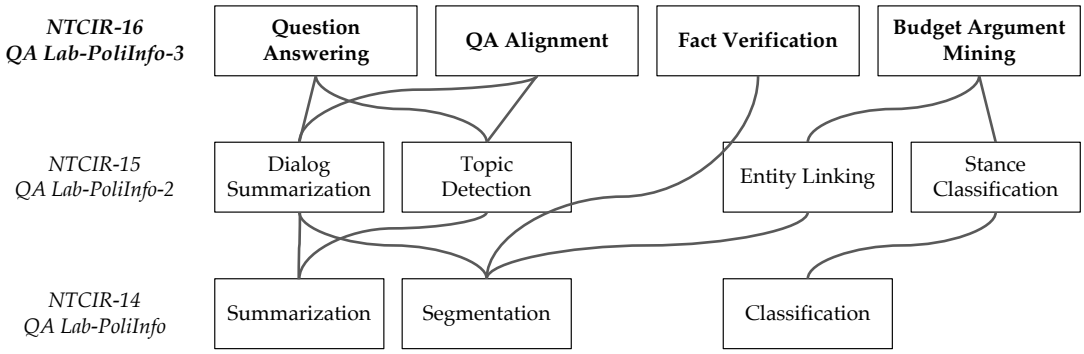

Figure 1: Relations between subtasks

We provide a Japanese 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.

QA using minutes from the Japanese assembly should be able to:

1: Provide an understandable summary of the topic;
2: Estimate the scope of each member's utterance;
3: Fact check each member's utterance;
4: Find the evidence for each member's utterance;
5: Link to different language resources; and
6: Deal with colloquial Japanese, including dialect and slang.

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

Figure 1 shows the relations between the subtasks. We have designed several subtasks on political information in NTCIR-14 and NTCIR-15. NTCIR-16 QA Lab-PoliInfo-3 includes the QA Alignment, Question Answering, Fact Verification, and Budget Argument Mining subtasks. The QA Alignment subtask is an important task that is required for the Question Answering subtask and NTCIR-15 QA Lab-PoliInfo-2's Dialog Summarization subtask. The Question Answering subtask is a combinational expansion of the Dialog Summarization and Topic Detection subtasks in the NTCIR-15 QA Lab-PoliInfo-2. In the QA Alignment subtask, the goal is to extract the appropriate range of related topics and the correspondence between questions and responses. The Fact Verification subtask is related to the Segmentation subtask in the NTCIR-14 QA Lab-PoliInfo. The purpose of Fact Verification is to verify the credibility of political claims using a predefined primary source. The Budget Argument Mining subtask is related to the Entity Linking and Stance Classification subtasks. The purpose of Budget Argument Mining is to identify argumentative components related to a budget item and then classify these argumentative components on the basis of their argumentative roles.

2 RELATED WORK

Fake news detection and fact checking have emerged as research topics of importance. Research on fake news is related to political information, question answering, text alignment, fact checking, argument mining, and more. Here, we provide a brief description of each of these areas.

2.1 Political Information

Fake news detection and fact checking are often associated with political information such as public debates and meeting minutes. Fact checking tasks have been implemented in articles on the 2016 U.S. presidential debate [1]. Although minutes from Japan's National Diet can be collected using Web API (JSON or XML), Japanese local assembly minutes are difficult to access without crawling and scraping. Thus, a dataset that can be used for research is in development. The corpus contains minutes from the local assemblies of 47 prefectures in Japan from April 2011 to March 2015 [9]. These minutes can be used as primary information as they contain records of who said what, when, and where.

2.2 Question Answering

The Stanford Question Answering Dataset (SQuAD) 1.0 contains 100,000+ questions posed by crowdworkers on a set of Wikipedia articles [19]. SQuAD 2.0 combines the existing SQuAD with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones [18]. HotpotQA is a question answering dataset which contains 113k Wikipedia-based question-answer pairs, the purpose of which is to facilitate the development of QA systems capable of performing explainable, multi-hop reasoning over diverse natural language [26][15].

2.3 Text Alignment

Chousa et al. formalized the sentence alignment problem as the independent predictions of spans in the target document from sentences in the source document [5].

2.4 Fake News Detection

Fake news detection is a crucial and socially relevant task. Numerous studies have been conducted on the detection of fake news. There are also a number of survey papers related to fake news. Zhou and Zafarani reviewed and evaluated methods for detecting fake news from four aspects: incorrect statements, writing style,

propagation patterns, and the credibility of the sources [27]. Oshikawa et al. investigated the difference between fake news detection and other related tasks, and the importance of NLP solutions for fake news detection [16]. The Fake News Challenge[1] included a Stance Detection task for estimating the relative perspective (or stance) of two pieces of text relative to a topic, claim, or issue. The organizers of Profiling Fake News Spreaders examined how to detect fake news by profiling authors [20]. Sharma et al. compiled a list of available datasets around fake news detection and summarized their characteristic features [21].

2.5 Fact Checking

FEVER is a Fact Extraction and VERification Shared dataset which classifies whether human-written factoid claims could be supported or refuted using evidence retrieved from Wikipedia [23]. The FEVER 2.0 task was to both build systems to verify factoid claims using evidence retrieved from Wikipedia and to generate adversarial attacks against other participants' systems [24]. The CLEF-2018 Fact Checking Lab conducted Check-worthiness and Factuality tasks in both English and Arabic, using debates from the 2016 U.S. presidential campaign [1]. 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 [6][2].

2.6 Argument Mining

Research on argument mining has garnered considerable attention as a logic-based approach to natural language processing (NLP) to capture the structure of arguments [25][7]. Argument structure analysis is a typical task in argument mining that assigns labels (claim, premise) to discourse units of sentences and clauses [10]. Common processes in argument mining analysis include the identification of argumentative components, clause attributes, and relationships between clauses [11]. IBM Research AI presented "Project Debater", an autonomous debating system that can engage in a competitive debate with humans [22].

2.7 Financial Documents

There has been growing interest in applying NLP techniques to financial documents. FinNum-2 is a task for fine-grained numeral understanding in financial social media data [4]. Numeral attachment is a task for identifying the attached target of the numeral. FinCausal 2020 is a shared task that identifies causality in financial datasets [13]. Bentabet et al. organized a shared task at the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (FNP-FNS 2020) [3]. The aim of the shared task was to extract a table of contents (TOC) from investment documents by detecting the document titles and organizing them hierarchically into a TOC.

3 TASK DESCRIPTION

We designed the QA Alignment, Question Answering, Fact Verification, and Budget Argument Mining subtasks. We consider they include basic technologies of political information systems that ensure the credibility of information and perform fact-checking.

For evaluation, we introduced a leader board for each of the subtasks, which were published on the QA Lab-PoliInfo-3 website[2], so that participants could verify their results immediately during the dry and formal runs. Participants could post their system results five times a day.

Table 1: Data used in QA Alignment subtask

Dataset		Utterances	Year
Train		143,798	2011 – 2016
Test	Dry run	24,302	2019
	Formal run	27,187	2020

3.1 QA Alignment

3.1.1 *Purpose.*

The aim of the QA Alignment task is to associate each question with its answer in the minutes.

The minutes of a Japanese local assembly resemble a transcript. In a question and answer session, an assembly member asks several questions at a time, and a prefectural governor or a superintendent answers the questions. As a result, each question is not directly associated with its answer in the minutes.

In QA Alignment, the goal is to align member's question with its corresponding answer from a governor or superintendent. It acts as a pre-process for other tasks, such as Summarization and Topic Detection.

3.1.2 *Data.*

In this task, we use the minutes of the Tokyo Metropolitan Assembly[3]. The Tokyo Metropolitan Assembly has released a dataset titled "Main Meeting Net Report," where utterances in the minutes are rearranged so that each question corresponds to its answer[4]. We consider the correspondence in this net report as the gold standard.

We prepared the minutes of the Tokyo Metropolitan Assembly from April 2011 to December 2016 for training data and those of 2019 and 2020 for test data. For a full description of the structure, see Appendix A.1 section.

Input. 2019 assembly minutes (dry run) and 2020 assembly minutes (formal run)

Output. The `QAID` field of the input data will contain the same value for the corresponding question and answer.

Data size. See Table 1.

3.1.3 *Evaluation.*

We evaluate the correspondence following the evaluation in [14]. An example is provided below.

Table 2 shows the questions and answers consisting of 19 corresponding sentences, where the value of `QAID` represents the correspondence. Here, the correspondence of the correct data and that

[1] http://www.fakenewschallenge.org/

[2] https://poliinfo3.github.io/

[3] https://www.gikai.metro.tokyo.jp/record/proceedings/ (in Japanese)

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

Table 2: An example correspondence between correct data and system output

Sent.	QorA	QAID		Sent.	QorA	QAID	
		correct	output			correct	output
1	Q	1	1	11	A	0	0
2	Q	1	1	12	A	1	1
3	Q	1	2	13	A	1	1
4	Q	0	0	14	A	1	2
5	Q	2	3	15	A	0	0
6	Q	2	3	16	A	2	3
7	Q	2	3	17	A	3	3
8	Q	3	4	18	A	3	4
9	Q	3	4	19	A	4	5
10	Q	4	5				

Table 3: Conversion from Table 2 to *alignment*

correct		output		correct		output	
[QAID=1]	1-12	[QAID=1]	1-12	[QAID=2]	5-16	[QAID=3]	5-16
	1-13		1-13		6-16		6-16
	1-14				7-16		7-16
	2-12		2-12				5-17
	2-13		2-13				6-17
	2-14						7-17
	3-12			[QAID=3]	8-17		
	3-13				9-17		
	3-14	[QAID=2]	3-14		8-18	[QAID=4]	8-18
					9-18		9-18
				[QAID=4]	10-19	[QAID=5]	10-19

of the system output are different. In the correct data, there are four questions; for example, sentences 1, 2, and 3 form one question, and the corresponding answers are sentences 12, 13, and 14. However, in the system output, there are five questions, and the first question of the correct data is divided into two. To compare the system output with the correct data, consider *alignment* as follows. For example, sentences 1 and 2 and sentences 12 and 13 of the system output correspond to each other. In this case, there are four *alignments*: 1-12, 2-12, 1-13, and 2-13.

The first question in the correct data corresponds to three question sentences and three answer sentences for a total of nine *alignments*. Table 3 shows all *alignments* converted from the correspondence in Table 2. By converting to *alignment*, we can deal with different QAIDs between the correct data and the system output. In this case, in Table 2, the QAID of sentences 10 and 19 of the correct answer data is 4, whereas it is 5 in the system output, but it is detected as the same 10-19 *alignment* shown in Table 3.

We evaluated precision, recall, and F-measure on the basis of this alignment, and we used F-measure in the leader board. In this example, precision is 78.5% (11/14), recall is 64.7% (11/17), and F-measure is 71.0%.

3.1.4 *Baseline System.*

The baseline method uses rule-based separating and character n-gram matching. First, the questioner's utterance and the answerers' utterances are divided into paragraphs on the basis of the regular expressions in Table 4 proposed by Kanasaki et al. [8]. Here, 'Opening' indicates expressions that appear at the beginning of a paragraph, and if a sentence matches these regular expressions, the utterance is split immediately before it. Similarly, 'Closing' indicates expressions that appear at the end of a paragraph, and if a sentence matches, the utterance is split immediately after it. However, in this division method, a sentence such as "最初に、オープンデータについて伺います" ("First, I would like to ask about open data") becomes a one-sentence paragraph. If a paragraph consists of only one sentence and matches 'Header', it is merged with the next paragraph. When answerers answer multiple questions, they may produce an utterance such as "二点のご質問にお答えいたします" ("I will answer two questions"). We do not include this sentence in any paragraph.

Next, we create a correspondence between each paragraph of the question and the answer. To do this, we consider the similarity between paragraphs and pair those with the highest similarity. Here, we convert the sentences in a paragraph into a set of character unigrams and bigrams and set the size of the intersection as the similarity. If the question with the highest similarity to the answer matches the answer with the highest similarity to the question, we remove them as a pair from the candidates. Repeat this until the number of pairs does not increase.

3.2 Question Answering

3.2.1 *Purpose.*

The purpose of the Question Answering task is to answer a question based on the contents of the minutes. Thus, the goal is to identify question utterances similar to the input question and return a summarization of its answer utterances. However, as described in Section 3.1, each question is not directly associated with its answer in the minutes, so participants are permitted to use the results of the QA Alignment tasks.

In this task, we used *Togikaidayori*[5], a newsletter from the Tokyo Metropolitan government that summarizes minutes. This task requires short answers, and participants need to summarize the answer utterances. Thus, this task can be considered a successor to PoliInfo and PoliInfo-2's Summarization subtasks.

3.2.2 *Data.*

Input. A question summary from *Togikaidayori* and its related information: date, questioner name, and answerer name. We also gave the participants the Tokyo Metropolitan Assembly Minutes, and we used 2019 data for the dry run and 2020 data for the formal run. For a complete description of the structure, see Appendix A.2 section.

Output. A summary of answer utterances in the original minutes corresponding to the input question.

Data size. See Table 5.

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

Table 4: Regular expressions used to find cue phrases [8]

Pattern	Regular expressions
Opening	^まず|^最初に|^初めに|^次に|^次いで|^最後に|^終わりに |^[一二三四五六七八九十]+点目 |^[^、]+についてで(す|あります|ございます)(が|けれど) |^終わり(ま|で)す。|^以上で|^ありがとうございま |他の質問に(ついて|つきまして)は
Closing	伺い[^、]*ます。|お尋ね[^、]*します|お答えください。 |(見解|所見|答弁)を求め[^、]*ます。 |(いかがで|どうで)(しょうか|すか)。 |.+質問を(終わります|終了します)。
Header	(^まず|^最初に|^初めに|^次に|^次いで|^最後に|^終わりに).*伺い[^、]*ます。

Table 5: Statistics of data for the Question Answering subtask

Dataset		Speeches		Sentences		Summaries	Date
		Question	Answer	Question	Answer		
Dry run	Train	611	1,758	60,979	28,283	2,765	June 2011 – December 2018
	Test	85	260	9,068	4,514	391	February 2019 – December 2019
Formal run	Train	1,465	4,842	150,194	72,128	7,627	September 2001 – December 2019
	Test	93	272	9,205	4,697	416	February 2020 – December 2020

3.2.3 *Evaluation.*

For this subtask, we conducted automatic evaluation and human evaluation.

Automatic evaluation. We consider the answer summary in *Togikaidayori* as the gold standard and calculated ROUGE scores [12]. On the leader board, we used the ROUGE-1 F-measure of content words.

Manual evaluation. Each participant evaluated the results, including the other participants' results as well as summaries from *Togikaidayori*, in the following four aspects and gave a grade of A, B, or C, with A being the highest and C being the lowest.

- **Correspondence** Whether the expression is an answer to a question or request, regardless of the authenticity of the content. The focus is on the format of the answer, such as "Yes / No" for "Do you ... ?" and "Because ..." for "Why ... ?". If the question is in the form of a request, determine whether the text is trying to answer the request appropriately.
- **Content** How much of the output includes the important content of the answer in the minutes.
- **Well-formed** The correctness of the expressions and grammar.
- **Overall** The appropriateness of the output as a comprehensive and summarized answer to the question, including the expression, length, content, and grammar.

3.2.4 *Baseline System.*

For the baseline system, we extracted the last 40 characters of the utterances made by the given answerer in response to the given questioner in the given meeting. We retrieved the characters from the end of the utterances because the answerers often expressed their conclusion at the end, and the number of characters is determined from the average length of the training dataset for the dry run (38.95).

3.3 Fact Verification

3.3.1 *Purpose.*

The Fact Verification subtask is to verify the credibility of political claims using predefined primary sources. Considering real-world settings, it would be better to allow participants to access any external information source. However, collecting the information is time consuming, and the main purpose of Fact Verification is not developing large scale search technologies. Therefore, to focus on methods for assessing information credibility and identifying misinformation, we simplify the problem settings. Our subtask is constructed in three steps. First, we define the documents that are considered primary sources and provide them in advance. Second, the participants try to extract sentences relevant to a given claim from the primary source. Finally, the participants classify the claim as true or false. When a participant classifies the claim as true, they are also required to present the extracted sentences for evidence as well as the classified label.

3.3.2 *Data.*

Predefined primary source. We use Tokyo Metropolitan Assembly Minutes as the primary source, which is structured data of meeting transcripts. For a full description of the structure, see Appendix A.3 section.

Input. For truth claims, we use a set from *Togikaidayori*[6], each of which contains summaries of specific parts of the Tokyo Metropolitan Assembly Minutes given by a human. For misinformation, we

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

Table 6: Statistics of data for the Fact Verification subtask

Dataset		Truth	Misinformation
Dry run	Train	596	428
	Test	166	132
Formal run	Train	596	427
	Test	226	184

prepared two different instances: wrong metadata and wrong content. An instance of wrong metadata is a summary in the *Togikai-dayori*, in which we falsified the `Meeting` field so that the summary reflects what the assembly member said but at a different meeting. To create instances of wrong content, we used an open-access language model, GPT-J-6B[7]. We fed the model an utterance in the minutes, with a prompt such as "tl;dr:" or "Summary:" prefixed, to generate its summary. Due to the limited capability of the current model, it outputs either appropriate, factually incorrect, or ungrammatical summaries. We manually checked the model's outputs and used grammatical summaries that are factually inconsistent with the input as the misinformation (wrong content) claims for our dataset. Note that we did not manually edit any texts that the model generated.

Output. In this subtask, participants are required to output three values. `DocumentEntailment` label indicates whether the given claim is true or false. `StartingLine` and `EndingLine` are a range of sentences extracted for positive classification. If the classification is negative, participants assign a default value (-1).

Data size. See Table 6.

3.3.3 Evaluation.

We calculate macro-averaged Precision, Recall, and F1 scores. Suppose that the predicted outputs and corresponding gold standard are as follows:

$$\begin{aligned} StartGS &= \{startgs_1, ..., startgs_N\} \\ EndGS &= \{endgs_1, ..., endgs_N\} \\ LabelGS &= \{lgs_1, ..., lgs_N\} \\ StartPRED &= \{startpred_1, ..., startpred_N\} \\ EndPRED &= \{endpred_1, ..., endpred_N\} \\ LabelPRED &= \{lpred_1, ..., lpred_N\}. \end{aligned}$$

We calculated the scores using the following equation:

$$\begin{aligned} \text{Recall} &= \frac{1}{N}\sum_i \frac{lineoverlap(i)}{endgs_i - startgs_i + 1} \\ \text{Precision} &= \frac{1}{N}\sum_i \frac{lineoverlap(i)}{endpred_i - startpred_i + 1} \end{aligned}$$

$$\text{F1} = \frac{1}{N}\sum_i Hm\left(\frac{lineoverlap(i)}{endgs_i - startgs_i + 1}, \frac{lineoverlap(i)}{endpred_i - startpred_i + 1}\right),$$

where N denotes the number of test data, $Hm(a, b)$ denotes harmonic-mean of a and b, and $lineoverlap(i)$ denotes the number of predicted lines in the range between $startgs_i$ and $endgs_i$.

[7]https://huggingface.co/EleutherAI/gpt-j-6B

3.3.4 Baseline System.

The baseline method predicts false for all input claims. We submitted the result of this method as the TO (task organizer) team in the formal run.

3.4 Budget Argument Mining

3.4.1 Purpose.

The goal of Budget Argument Mining is to identify argumentative components related to a budget item and then classify these argumentative components on the basis of their argumentative roles when budget information and minutes are given.

One of the major responsibilities of the government is creating a budget that determines how their funds will be spent considering income and expenditures. National and local budget deliberations are held in the National Diet and local assemblies. The national budget is drafted by the cabinet and discussed in the National Diet before it is officialized. The budgets of local governments are proposed by the governors or mayors and are discussed and approved in the assembly. However, most citizens have difficulty understanding the background of the proposed budget, as well as the discussions that lead to the final budget.

Budget Argument Mining connects published budget documents with the discussions included in the meeting minutes. Specifically, when a budget item (amount, name of competent ministry/department, explanation, and others) is given, the politicians' statements related to the budget (statements referring to the amount of money) are identified in the meeting minutes, and an argumentative role such as Claim, Premise, or Other is assigned.

3.4.2 Data.

Input. The Budget Argument Mining subtask takes Budget Information and Minutes as input. Budget information includes the "date," "budget item," "previous year's budget amount," "current year's budget amount," etc. Minutes are from either the National Diet or local assembly. The minutes contain information such as speaker, utterance, monetary expression, related ID that links a budget item to the relevant argumentative component, and more. To construct the dataset for this subtask, we automatically identified argumentative components, which includes the `MONEY` expression, using the Japanese NLP library GiNZA[8].

Output. Participants of this subtask are required to output an argument class and a relatedID for each argumentative component. Argument classification is to classify argumentative components into the following seven argument classes:

(1) Premise : Past and decisions
(2) Premise : Current and future / estimates
(3) Premise : Other
(4) Claim : Opinions, suggestions, and questions
(5) Claim : Other
(6) Not a monetary expression
(7) Other.

RelatedIDs are given to link a budget item to the relevant argumentative component.

Data size. See Table 7.

[8]https://github.com/megagonlabs/ginza

Table 7: Number of argumentClasses and relatedIDs

Dataset	Year	National/local government	argumentClass								relatedID	
			Premise			Claim		Other	Not Money	Total	Non-empty	Count
			Past	Future	Other	Opinion	Other					
Train	2019	Otaru City	23	84	15	18	0	0	4	144	31	38
		Ibaraki Prefecture	26	80	39	0	0	0	2	147	5	13
		Fukuoka City	63	138	40	29	12	3	9	294	211	286
	2020	National Diet	45	92	14	10	0	0	4	165	11	13
		Otaru City	4	55	9	16	0	0	1	85	16	17
		Ibaraki Prefecture	21	70	31	3	0	0	4	129	12	12
		Fukuoka City	78	103	64	22	11	3	3	284	64	83
	SubTotal		260	622	212	98	23	6	27	1,248	350	462
Test	2019	Otaru City	43	74	49	9	2	1	12	190	18	20
		Ibaraki Prefecture	5	3	25	1	0	0	0	34	1	1
		Fukuoka City	8	13	13	1	2	0	3	40	3	3
	2020	National Diet	11	25	4	21	0	0	4	65	1	1
		Otaru City	31	53	25	9	0	1	4	123	21	26
		Ibaraki Prefecture	2	8	21	0	0	0	3	34	2	2
		Fukuoka City	1	20	8	1	0	0	4	34	1	5
	SubTotal		101	196	145	42	4	2	30	520	47	58
Total			361	818	357	140	27	8	57	1,768	403	520

3.4.3 Dataset.

We used budget information and minutes from the National Diet, Otaru City, Ibaraki Prefecture, and Fukuoka City. Table 7 presents the number of argumentClasses and relatedIDs. The training data contained a total of 1,248 money expressions (moneyExpressions), among which 1,083 for the local governments and 165 for the National Diet. The test data contained a total of 520 money expressions (moneyExpressions), among which 455 for the local governments and 65 for the National Diet.

3.4.4 Evaluation.

We designed the score of Budget Argument Mining to consider both argument class labeling (AC) and relatedID linking (RID). The score is calculated by the following equation:

$$\text{Score} = \frac{1}{|S_{RID}|} \sum_{x,y \in S_{RID}} \{ACC(x,y) \times RIDC(x,y)\}.$$

x and y are the labels given to the same monetary expression of the system output and the gold standard data, respectively. S_{RID} is a set of monetary expressions in the gold standard data whose RIDs are not null, as shown in the following equation:

$$S_{RID} = \{y | y.RIDs \neq null\}.$$

ACC indicates whether the AC of a monetary expression is correct or not, as shown in the following equation:

$$ACC(x,y) = \begin{cases} 0 & (x.AC \neq y.AC) \\ 1 & (x.AC = y.AC). \end{cases}$$

RIDC indicates whether an RID output by the system is included in the RIDs of the gold standard data or not:

$$RIDC(x,y) = \begin{cases} 0 & (x.RID \notin y.RIDs) \\ 1 & (x.RID \in y.RIDs). \end{cases}$$

3.4.5 Baseline System.

Our baseline system randomly sets argumentClass and relatedID for each of the money expressions. For argumentClass, we randomly chose one of the seven classes by uniform selection. For relatedID, we chose one budget item from the same government for the same year as when the meeting was held. However, because many of budget items have no relatedIDs in the training/gold datasets, we only assign one at a rate of 1/3.

4 SCHEDULE

The NTCIR-16 QA Lab-PoliInfo-3 task has been run following this timeline:

March 24, 2021: QA Lab-PoliInfo-3 first round table meeting
March 29, 2021: NTCIR-16 kickoff meeting
June 15, 2021: QA Lab-PoliInfo-3 second round table meeting
June 15, 2021: Dataset release

Dry Run

August 10–November 12, 2021: Dry run
November 1–12, 2021: Evaluation by participants (Question Answering)
November 15, 2021: Evaluation result release

Formal Run

November 22, 2021: Update of dataset for formal run
November 22–30, 2021: Formal run
November 30, 2021: Task registration due for formal run (not required for dry run participants)

NTCIR-16 CONFERENCE

December 6–17, 2021: Evaluation by participants
December 18–19, 2021: Evaluation by organizers
December 20, 2021: Evaluation Result Release
February 1, 2021: Task overview paper release (draft)

Table 8: Active participating teams

Team	Organization
10807010	Tokyo Institute of Technology
AKBL*	Toyohashi University of Technology
ditlab	Denso IT Laboratory
Forst*	Yokohama National University
fuys*	Fukuoka University
Ibrk*†	Ibaraki University
JRIRD*	The Japan Research Institute, Limited
nukl*	Nagoya University
OUC*	Otaru University of Commerce
rVRAIN	Universitat Politècnica de València
SMLAB	National Agriculture and Food Research Organization
	The University of Tokyo
takelab	Osaka Electro-Communication University
TO*	task organizers

*Task organizer(s) are in team

†No submissions for formal run (only late submissions)

March 1, 2022: Submission due for participant papers
May 1, 2022: Camera-ready participant paper due
June 14–17, 2022: NTCIR-16 Conference

5 PARTICIPATION

Thirteen teams registered for the task, but only 11 teams participated actively, i.e., submitted results for the formal run. Table 8 shows the active participating teams.

Table 9: Number of submissions in dry run

Team	QAA	QA	FV	BAM	Total
AKBL	3	4	16	-	23
10807010	-	-	10	-	10
rVRAIN	-	-	-	9	9
ditlab	8	-	-	-	8
OUC	-	-	-	6	6
Forst	-	-	6	-	6
fuys	-	-	-	5	5
takelab	-	-	-	3	3
nukl	-	2	-	-	2
Subtotal	11	6	32	23	72
TO	1	1	1	3	6
Total	12	7	33	26	78

6 SUBMISSIONS

Tables 9 and 10 show the number of submissions for the dry run and the formal run, respectively. The number in brackets is the number of late submissions. In the dry run, there were 11 submissions from two teams for QA Alignment, six submissions from two teams for Question Answering, 32 submissions from three teams for Fact Verification, and 23 submissions from four teams for Budget Argument Mining. In the formal run, there were 54 submissions (and a late submission) from three teams for QA Alignment, 32 submissions (and a late submission) from three teams for Question Answering, 14 submissions (and four late submissions) from three teams for Fact Verification, and 43 submissions (and 16 late submissions) from six teams for Budget Argument Mining. In total, there were 143 submissions (and 22 late submissions) from 11 teams.

Table 10: Number of submissions in formal run

Team	QAA	QA	FV	BAM	Total
ditlab	32 (+1)	22	-	-	54 (+1)
OUC	-	-	-	21 (+4)	21 (+4)
AKBL	11	2	8	-	21
Forst	11	-	2 (+4)	-	13 (+4)
nukl	-	8 (+1)	-	-	8 (+1)
rVRAIN	-	-	-	7	7
SMLAB	-	-	-	6	6
takelab	-	-	-	5	5
10807010	-	-	4	-	4
JRIRD	-	-	-	3	3
fuys	-	-	-	1 (+1)	1 (+1)
Ibrk	-	-	-	- (+11)	- (+11)
Subtotal	54 (+1)	32 (+1)	14 (+4)	43 (+16)	143 (+22)
TO	1	1	1	1	4
Total	55 (+1)	33 (+1)	15 (+4)	44 (+16)	147 (+22)

7 RESULTS

Tables 11, 12, 14, and 15 show the automatic evaluation results of QA Alignment, Question Answering, Fact Verification, and Budget Argument Mining in the formal run, respectively.

Table 13 shows the human evaluation results of Question Answering.

8 OVERVIEW OF PARTICIPANTS' SYSTEMS

We briefly describe the characteristic aspects of the participating teams' systems and their contributions below.

The AKBL team participated in the QA Alignment, Question Answering, and Fact Verification subtasks. For the QA Alignment subtask, their method first divides the given question and answer texts into semantically consistent segments. Then they apply the Hungarian algorithm with the BM25 similarity metric to align the segments. For the Question Answering subtask, their system first selects a short segment relevant to a given question summary from the answer text and then converts it into the answer summary by using the abstractive summarizer based on the pre-trained BART. For the Fact Verification subtask, their most optimal system first retrieves a passage relevant to a given claim from the assembly minutes and then checks if the passage supports the claim by using a BERT-based textual entailment classifier.

The ditlab team participated in the QA Alignment and Question Answering subtasks. First, they developed a QA Alignment system that associates each question to its answer by using heuristic rules to make paragraphs composed of the related sentences. The heuristic rules were optimized for minutes. They prepared four features for matching. Next, they built a QA system that uses a similarity

Table 11: Scores of QA Alignment subtask in formal run

ID	Team	F-Measure	Precision	Recall
314	ditlab	**0.8329**	**0.8703**	0.8037
286	ditlab	0.8324	0.8691	0.8038
281	ditlab	0.8298	0.8664	0.8005
245	ditlab	0.8289	0.8651	0.8000
259	ditlab	0.8289	0.8651	0.8000
282	ditlab	0.8275	0.8502	0.8096
285	ditlab	0.8264	0.8650	0.7956
237	ditlab	0.8224	0.8559	0.7959
227	ditlab	0.8224	0.8556	0.7961
267	ditlab	0.8221	0.8562	0.7955
243	ditlab	0.8219	0.8555	0.7954
242	ditlab	0.8190	0.8542	0.7913
226	ditlab	0.8161	0.8476	0.7912
260	ditlab	0.8156	0.8204	0.8134
221	ditlab	0.8107	0.8376	0.7903
235	AKBL	0.8098	0.8000	0.8311
241	ditlab	0.8074	0.8053	0.8134
236	AKBL	0.8050	0.8307	0.7858
215	AKBL	0.8002	0.7780	0.8354
182	AKBL	0.7931	0.7615	**0.8435**
180	AKBL	0.7917	0.7990	0.7903
213	ditlab	0.7870	0.7970	0.7816
206	ditlab	0.7857	0.7974	0.7791
204	ditlab	0.7855	0.7938	0.7820
179	AKBL	0.7855	0.7927	0.7873
181	AKBL	0.7826	0.8001	0.7728
178	AKBL	0.7823	0.7598	0.8175
284	Forst	0.7753	0.7670	0.7883
294	AKBL	0.7750	0.8159	0.7450
198	Forst	0.7746	0.7854	0.7716
197	Forst	0.7746	0.7854	0.7716
210	ditlab	0.7718	0.7614	0.7875
283	Forst	0.7703	0.7615	0.7837
261	Forst	0.7703	0.7615	0.7837
262	Forst	0.7699	0.7594	0.7852
290	Forst	0.7662	0.7803	0.7608
218	ditlab	0.7649	0.7316	0.8085
199	ditlab	0.7638	0.7661	0.7659
216	AKBL	0.7591	0.7587	0.7694
205	Forst	0.7584	0.7774	0.7484
188	ditlab	0.7574	0.7567	0.7626
297	AKBL	0.7548	0.8263	0.7016
220	Forst	0.7522	0.7680	0.7446
186	ditlab	0.7343	0.7312	0.7427
184	ditlab	0.7336	0.7258	0.7473
185	ditlab	0.7322	0.7244	0.7454
196	Forst	0.7317	0.7366	0.7340
194	ditlab	0.6897	0.6646	0.7255
168	ditlab	0.6859	0.6721	0.7097
169	ditlab	0.6772	0.6651	0.6985
170	ditlab	0.6648	0.6191	0.7411
171	ditlab	0.6370	0.5956	0.7035
167	TO	0.6166	0.5991	0.6437
173	ditlab	0.5392	0.4887	0.6250
195	Forst	0.0000	0.0000	0.0000

measure to find the original question that is similar to the question summary. The QA system then identified the answers associated with the original question using the results of the QA Alignment described above. A Text-to-Text Transfer Transformer (T5) was used to summarize the associated answer.

The Forst team tackled the QA Alignment and the Fact Verification subtasks. For the QA Alignment, they used a rule-based approach using clue expressions to segment statements and using similarity to map the individual questions and answers. For the Fact Verification, they used a rule-based approach using word similarity between one sentence in the minutes and the summary sentence.

The fuys team assigned argumentClass and relatedID in several ways. To improve accuracy, they also added a dimension to determine whether a person is a member of Congress. RelatedID was assigned using key word extraction with TFIDF.

The Ibrk team participated in the Budget Argument Mining subtask and constructed a Bi-directional LSTM-CNNs-CRF model to predict argument labels and a BERT-based embedding model for assigning a related ID to each money expression.

The JRIRD team participated in the Budget Argument Mining subtask and constructed two independent BERT-based classification models for the two objectives. Their model for argument classification (AC) classifies each occurrence of a monetary expression into one of the seven classes. Their model for the related ID detection (RID) judges the likelihood that the pair of one candidate budget item and a sentence is valid.

The nukl team applied T5 to generate answer summaries using two inputs, the answerer's entire utterance and the answer text corresponding to the input question. The final output was determined from the length of the answerer's utterance.

The OUC team worked on the Budget Argument Mining subtask. For argument classification, they performed 7-level classification on ArgumentClass using a fine-tuned BERT. To link relatedIDs, they used TF-IDF vectorization of documents and cosine similarity calculation.

The rVRAIN team tackled the Budget Argument Mining subtask, consisting of a combination of classification and information retrieval subtasks. For argument classification, the team achieved their most optimal results with a five-class BERT-based cascade model complemented with handcrafted rules. The rules were used to determine if the expression was monetary. Then each monetary expression was classified as a premise or conclusion in the first level of the cascade model. Finally, each premise was classified into one of the three premise classes, and each conclusion into one of the two conclusion classes. For information retrieval (i.e., relation ID detection or RID), their most optimal results were attained using a combination of a BERT-based binary classifier and the cosine similarity of tuples consisting of the monetary expression and budget BERT embeddings.

The SMLAB team's model for the Budget Argument Mining subtask has two inputs, three input sentences and the description of the target budget. In the subtask, participants have to search for the budgetID related to the input money expressions. Therefore, they opted to find the budgetID using cosine similarity based on BERT vectors.

The takelab's system for the Budget Argument Mining task utilizes a topic extraction method based on utterance classification.

Table 12: Scores of Question Answering subtask in formal run (ROUGE scores)

ID	Team	ROUGE (Recall)			ROUGE (F-measure)		
		N1	N2	R	N1	N2	R
Surface form							
310	nukl	0.4811	0.2601	0.4277	0.4629	0.2501	**0.4112**
313	nukl	0.4826	**0.2616**	**0.4288**	**0.4631**	**0.2505**	0.4110
311	nukl	0.4778	0.2499	0.4216	0.4571	0.2388	0.4030
288	ditlab	0.4826	0.2556	0.4256	0.4518	0.2364	0.3983
269	ditlab	0.4666	0.2443	0.4118	0.4485	0.2341	0.3962
271	ditlab	0.4651	0.2439	0.4106	0.4476	0.2340	0.3955
270	ditlab	0.4641	0.2424	0.4091	0.4461	0.2322	0.3936
273	ditlab	**0.4834**	0.2542	0.4252	0.4500	0.2335	0.3954
266	nukl	0.4508	0.2325	0.3988	0.4387	0.2240	0.3877
304	nukl	0.4554	0.2328	0.4012	0.4348	0.2209	0.3826
268	ditlab	0.4462	0.2235	0.3909	0.4263	0.2119	0.3740
258	nukl	0.4190	0.2068	0.3671	0.4225	0.2075	0.3708
291	ditlab	0.4560	0.2289	0.3994	0.4282	0.2112	0.3744
253	nukl	0.4411	0.2192	0.3885	0.4137	0.2016	0.3629
289	ditlab	0.4516	0.2200	0.3935	0.4187	0.2004	0.3648
225	ditlab	0.4339	0.2077	0.3793	0.4117	0.1978	0.3606
249	ditlab	0.4339	0.2077	0.3793	0.4117	0.1978	0.3606
Stem							
310	nukl	0.4883	0.2647	0.4323	0.4702	**0.2548**	**0.4158**
313	nukl	0.4897	**0.2659**	**0.4328**	**0.4703**	0.2548	0.4151
311	nukl	0.4858	0.2547	0.4270	0.4650	0.2438	0.4083
288	ditlab	0.4883	0.2604	0.4298	0.4574	0.2413	0.4025
269	ditlab	0.4732	0.2499	0.4167	0.4548	0.2395	0.4009
271	ditlab	0.4717	0.2494	0.4155	0.4539	0.2393	0.4002
270	ditlab	0.4705	0.2479	0.4139	0.4522	0.2374	0.3981
273	ditlab	**0.4903**	0.2600	0.4302	0.4569	0.2391	0.4004
266	nukl	0.4566	0.2371	0.4029	0.4447	0.2287	0.3920
304	nukl	0.4621	0.2381	0.4065	0.4412	0.2260	0.3877
268	ditlab	0.4526	0.2286	0.3957	0.4327	0.2167	0.3788
258	nukl	0.4243	0.2115	0.3712	0.4281	0.2124	0.3753
291	ditlab	0.4619	0.2324	0.4032	0.4336	0.2145	0.3782
253	nukl	0.4481	0.2236	0.3928	0.4207	0.2058	0.3671
289	ditlab	0.4567	0.2232	0.3968	0.4235	0.2036	0.3680
225	ditlab	0.4399	0.2119	0.3831	0.4172	0.2016	0.3641
249	ditlab	0.4399	0.2119	0.3831	0.4172	0.2016	0.3641
Content word							
310	nukl	0.3262	**0.1794**	0.3188	**0.3132**	**0.1711**	**0.3062**
313	nukl	**0.3270**	0.1794	**0.3196**	0.3129	0.1709	0.3061
311	nukl	0.3193	0.1756	0.3114	0.3051	0.1657	0.2976
288	ditlab	0.3212	0.1765	0.3139	0.3013	0.1639	0.2945
269	ditlab	0.3112	0.1678	0.3028	0.2992	0.1611	0.2909
271	ditlab	0.3104	0.1678	0.3020	0.2991	0.1614	0.2908
270	ditlab	0.3075	0.1662	0.2991	0.2961	0.1596	0.2878
273	ditlab	0.3127	0.1627	0.3045	0.2891	0.1486	0.2814
266	nukl	0.2931	0.1526	0.2845	0.2823	0.1456	0.2740
304	nukl	0.2947	0.1590	0.2878	0.2808	0.1502	0.2743
268	ditlab	0.2857	0.1530	0.2774	0.2732	0.1451	0.2650
258	nukl	0.2703	0.1409	0.2636	0.2713	0.1394	0.2645
291	ditlab	0.2808	0.1460	0.2726	0.2619	0.1351	0.2544
253	nukl	0.2776	0.1432	0.2711	0.2573	0.1300	0.2510
289	ditlab	0.2730	0.1433	0.2665	0.2547	0.1318	0.2484
225	ditlab	0.2616	0.1365	0.2543	0.2499	0.1304	0.2430
249	ditlab	0.2616	0.1365	0.2543	0.2499	0.1304	0.2430

ID	Team	ROUGE (Recall)			ROUGE (F-measure)		
		N1	N2	R	N1	N2	R
Surface form							
238	ditlab	0.4325	0.2093	0.3801	0.4083	0.1972	0.3588
222	ditlab	0.4415	0.2072	0.3832	0.4087	0.1925	0.3548
240	ditlab	0.4243	0.1999	0.3695	0.3979	0.1894	0.3472
246	ditlab	0.4259	0.1997	0.3722	0.3970	0.1853	0.3475
255	ditlab	0.4169	0.1907	0.3630	0.3956	0.1815	0.3453
244	ditlab	0.4229	0.1910	0.3629	0.4073	0.1862	0.3502
190	AKBL	0.4529	0.2096	0.3841	0.3850	0.1771	0.3267
228	ditlab	0.4448	0.1990	0.3819	0.4042	0.1813	0.3465
231	ditlab	0.4448	0.1990	0.3819	0.4042	0.1813	0.3465
256	ditlab	0.4207	0.1958	0.3679	0.3958	0.1832	0.3460
189	AKBL	0.3954	0.1843	0.3455	0.3808	0.1756	0.3325
223	ditlab	0.4224	0.1891	0.3616	0.4006	0.1798	0.3429
295	nukl	0.3824	0.1561	0.3301	0.3636	0.1472	0.3126
296	ditlab	0.3861	0.1519	0.3296	0.3466	0.1339	0.2948
293	ditlab	0.3623	0.1255	0.3036	0.3234	0.1106	0.2705
166	TO	0.2628	0.0482	0.2187	0.2422	0.0439	0.2001
Stem							
238	ditlab	0.4396	0.2140	0.3847	0.4148	0.2013	0.3628
222	ditlab	0.4483	0.2111	0.3875	0.4153	0.1960	0.3591
240	ditlab	0.4300	0.2038	0.3738	0.4031	0.1927	0.3510
246	ditlab	0.4323	0.2036	0.3770	0.4031	0.1888	0.3517
255	ditlab	0.4232	0.1952	0.3678	0.4015	0.1856	0.3497
244	ditlab	0.4303	0.1952	0.3664	0.4145	0.1902	0.3537
190	AKBL	0.4601	0.2140	0.3880	0.3910	0.1808	0.3301
228	ditlab	0.4517	0.2034	0.3867	0.4108	0.1854	0.3514
231	ditlab	0.4517	0.2034	0.3867	0.4108	0.1854	0.3514
256	ditlab	0.4292	0.2011	0.3735	0.4041	0.1880	0.3513
189	AKBL	0.4010	0.1872	0.3489	0.3867	0.1787	0.3361
223	ditlab	0.4277	0.1922	0.3655	0.4061	0.1827	0.3469
295	nukl	0.3890	0.1608	0.3352	0.3700	0.1516	0.3177
296	ditlab	0.3935	0.1553	0.3341	0.3532	0.1368	0.2988
293	ditlab	0.3697	0.1295	0.3098	0.3298	0.1138	0.2756
166	TO	0.2797	0.0558	0.2324	0.2580	0.0506	0.2126
Content word							
238	ditlab	0.2600	0.1337	0.2540	0.2460	0.1264	0.2402
222	ditlab	0.2580	0.1329	0.2509	0.2414	0.1239	0.2347
240	ditlab	0.2499	0.1304	0.2429	0.2359	0.1228	0.2293
246	ditlab	0.2504	0.1314	0.2462	0.2342	0.1226	0.2304
255	ditlab	0.2414	0.1238	0.2342	0.2308	0.1183	0.2241
244	ditlab	0.2360	0.1130	0.2284	0.2308	0.1110	0.2237
190	AKBL	0.2682	0.1404	0.2602	0.2306	0.1204	0.2236
228	ditlab	0.2476	0.1213	0.2388	0.2293	0.1114	0.2212
231	ditlab	0.2476	0.1213	0.2388	0.2293	0.1114	0.2212
256	ditlab	0.2413	0.1235	0.2346	0.2270	0.1162	0.2205
189	AKBL	0.2352	0.1222	0.2284	0.2248	0.1160	0.2183
223	ditlab	0.2314	0.1094	0.2235	0.2229	0.1042	0.2154
295	nukl	0.2047	0.1042	0.2004	0.1971	0.1007	0.1931
296	ditlab	0.1899	0.0919	0.1867	0.1715	0.0803	0.1681
293	ditlab	0.1605	0.0769	0.1570	0.1423	0.0674	0.1393
166	TO	0.0835	0.0357	0.0825	0.0767	0.0330	0.0756

Table 13: Scores of Question Answering subtask in formal run (human evaluation results)

ID	Team	Correspondence				Content				Well-formed				Overall			
		A	B	C	Score	A	B	C	Score	A	B	C	Score	A	B	C	Score
Gold		377	20	3	774	208	170	22	586	391	8	1	790	217	164	19	598
310	nukl	363	25	12	**751**	138	211	51	**487**	381	19	0	781	148	203	49	**499**
288	ditlab	348	33	19	729	138	200	62	476	379	17	4	775	142	200	58	484
269	ditlab	346	31	23	723	129	209	62	467	384	16	0	**784**	136	207	57	479
190	AKBL	320	42	38	682	104	196	100	404	381	6	13	768	103	203	94	409
166	TO	83	77	240	243	4	58	338	66	99	33	268	231	4	36	360	44

Table 14: Scores of Fact Verification subtask in formal run

ID	Team	F-Measure	Precision	Recall
232	AKBL	**0.8892**	**0.8951**	**0.9030**
307	AKBL	0.8874	0.8930	**0.9030**
306	AKBL	0.8866	0.8917	**0.9030**
203	AKBL	0.8608	0.8668	0.8718
202	AKBL	0.8506	0.8559	0.8610
292	Forst	0.8389	0.8466	0.8451
201	AKBL	0.8098	0.8139	0.8238
200	AKBL	0.8098	0.8139	0.8238
257	Forst	0.8040	0.8113	0.8110
272	10807010	0.7963	0.8014	0.8329
191	10807010	0.7876	0.7984	0.8199
192	10807010	0.7822	0.7899	0.8146
274	10807010	0.7734	0.7808	0.8098
229	AKBL	0.4853	0.4917	0.4866
165	TO	0.4488	0.4488	0.4488

9 CONCLUSION

We presented an overview of the NTCIR-16 QA Lab-PoliInfo-3 task. The goal of the task is to develop complex real-world question answering (QA) techniques and summarize the opinions of assembly members and their reasons and conditions, using minutes from Japanese assemblies. We conducted a dry run and a formal run, which included the QA Alignment, Question Answering, Fact Verification, and Budget Argument Mining subtasks. There were 143 submissions from 11 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 Verification of Political Claims. Task 1: Check-Worthiness. arXiv: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] Najah-Imane Bentabet, Rémi Juge, Ismail El Maarouf, Virginie Mouilleron, Dialekti Valsamou-Stanislawski, and Mahmoud El-Haj. 2020. The Financial Document Structure Extraction Shared task (FinToc 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*. COLING, Barcelona, Spain (Online), 13–22. https://aclanthology.org/2020.fnp-1.2

[4] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2020. Overview of the NTCIR-15 FinNum-2 Task: Numeral attachment in financial tweets. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo Japan*.

[5] Katsuki Chousa, Masaaki Nagata, and Masaaki Nishino. 2020. SpanAlign: Sentence Alignment Method based on Cross-Language Span Prediction and ILP. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 4750–4761. https://doi.org/10.18653/v1/2020.coling-main.418

[6] 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.

[7] James B. Freeman. 2011. *Dialectics and the macrostructure of arguments : a theory of argument structure*. Technical Report.

[8] Katsumi Kanasaki, Jiawei Yong, Shintaro Kawamura, Shoichi Naitoh, and Kiyohiko Shinomiya. 2019. Cue-Phrase-Based Text Segmentation and Optimal Segment Concatenation for the NTCIR-14 QA Lab-PoliInfo Task. In *NII Conference on Testbeds and Community for Information Access Research*. Springer, 85–96.

[9] 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

[10] Tatsuki Kuribayashi, Hiroki Ouchi, Naoya Inoue, Paul Reisert, Toshinori Miyoshi, Jun Suzuki, and Kentaro Inui. 2019. An Empirical Study of Span Representations in Argumentation Structure Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4691–4698. https://doi.org/10.18653/v1/P19-1464

[11] John Lawrence and Chris Reed. 2019. Argument Mining: A Survey. *Computational Linguistics* 45, 4 (Dec. 2019), 765–818. https://doi.org/10.1162/coli_a_00364

[12] 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

[13] Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stéphane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2020. Financial Document Causality Detection Shared Task (FinCausal 2020). arXiv:2012.02505

[14] Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*. 1–10.

[15] Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Answering while Summarizing: Multi-task Learning for Multi-hop QA with Evidence Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2335–2345. https://doi.org/10.18653/v1/P19-1225

[16] Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A Survey on Natural Language Processing for Fake News Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 6086–6093. https://aclanthology.org/2020.lrec-1.747

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

[18] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/P18-2124

[19] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264

Table 15: Scores of Budget Argument Mining subtask in formal run

ID	Team	Score	AC	RID
299	JRIRD	**0.51064**	**0.58269**	0.61702
302	JRIRD	0.48936	0.56538	0.61702
300	OUC	0.44681	0.57115	**0.65957**
309	OUC	0.42553	0.53846	**0.65957**
263	OUC	0.42553	0.48269	**0.65957**
308	OUC	0.40426	0.56154	**0.65957**
305	OUC	0.40426	0.55769	**0.65957**
303	JRIRD	0.40426	0.54423	0.61702
251	OUC	0.40426	0.49038	**0.65957**
252	OUC	0.40426	0.48846	**0.65957**
250	OUC	0.40426	0.49038	0.57447
301	OUC	0.38298	0.47500	**0.65957**
277	OUC	0.38298	0.42308	**0.65957**
248	OUC	0.38298	0.49038	0.55319
278	OUC	0.29787	0.46154	**0.65957**
224	fuys	0.23404	0.56923	0.34043
239	rVRAIN	0.17021	0.47885	0.21277
287	rVRAIN	0.14894	0.47885	0.25532
230	OUC	0.14894	0.49038	0.17021
177	OUC	0.12766	0.37308	0.21277
254	rVRAIN	0.10638	0.47885	0.17021
234	OUC	0.08511	0.47500	0.17021
233	OUC	0.08511	0.42308	0.17021
219	OUC	0.08511	0.37308	0.17021
211	OUC	0.08511	0.37308	0.17021
212	OUC	0.08511	0.37308	0.14894
176	rVRAIN	0.06383	0.48462	0.21277
187	rVRAIN	0.06383	0.42692	0.21277
298	rVRAIN	0.04255	0.48462	0.25532
312	takelab	0.04255	0.39423	0.06383
280	takelab	0.04255	0.39423	0.04255
174	rVRAIN	0.00000	0.48462	0.17021
183	OUC	0.00000	0.37308	0.12766
209	takelab	0.00000	0.39423	0.00000
208	takelab	0.00000	0.39423	0.00000
279	takelab	0.00000	0.39423	0.00000
276	SMLAB	0.00000	0.38269	0.00000
265	SMLAB	0.00000	0.38269	0.00000
275	SMLAB	0.00000	0.38269	0.00000
175	SMLAB	0.00000	0.38269	0.00000
217	OUC	0.00000	0.37308	0.00000
172	SMLAB	0.00000	0.35962	0.00000
193	SMLAB	0.00000	0.32885	0.00000
164	TO	0.00000	0.13462	0.00000

[20] 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.

[21] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 3, Article 21 (Apr 2019), 42 pages. https://doi.org/10.1145/3305260

[22] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, and IBM Research AI. 2021. An autonomous debating system. *Nature* 591, 7850 (Mar 2021), 379–384. https://doi.org/10.1038/s41586-021-03215-w

[23] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal (Eds.). 2018. *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER).* Association for Computational Linguistics, Brussels, Belgium. https://aclanthology.org/W18-5500

[24] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 Shared Task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER).* Association for Computational Linguistics, Hong Kong, China, 1–6. https://doi.org/10.18653/v1/D19-6601

[25] Stephen E. Toulmin. 1958. *The Uses of Argument.* Cambridge University Press.

[26] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2369–2380. https://doi.org/10.18653/v1/D18-1259

[27] Xinyi Zhou and Reza Zafarani. 2018. Fake News: A Survey of Research, Detection Methods, and Opportunities. arXiv:1812.00315

A DATA FIELDS AND EXAMPLES

A.1 QA Alignment

A.1.1 Data Fields. The QA Alignment dataset consists of one file. The minutes data is the same as that of Question Answering with the following three items added.

QorA Speaker type (Questioner, Answerer, or Other)
QuestionerID Identifier of the questioner
QAID Identifier for questions and answers. The value of QAID is the same for the corresponding question and answer. In the test data, the value is -1, and participants will be asked for this value.

A.1.2 Examples.

Listing 1: Answer sheet for the QA Alignment subtask

```
[
  ...
  {"Volume": "2011-2", "Number": "3", "Date": "2011-06-24", "Title": "平成二十三年東京都議会会議録第九号", "SpeakerPosition": "三十一番", "SpeakerName": "中村ひろし", "QuestionSpeaker": "中村ひろし（民主党）", "Speaker": "中村ひろし（民主党）", "Utterance": "都の財政運営についての考えを伺います。", "QorA": "Q", "QuestionerID": "2011_02_g01", "QAID": 1},
  {"Volume": "2011-2", "Number": "3", "Date": "2011-06-24", "Title": "平成二十三年東京都議会会議録第九号", "SpeakerPosition": "三十一番", "SpeakerName": "中村ひろし", "QuestionSpeaker": "中村ひろし（民主党）", "Speaker": "中村ひろし（民主党）", "Utterance": "次に、都政の課題と「二〇二〇年の東京」について伺います。", "QorA": "Q", "QuestionerID": "2011_02_g01", "QAID": 2},
  ...
  {"Volume": "2011-2", "Number": "3", "Date": "2011-06-24", "Title": "平成二十三年東京都議会会議録第九号", "SpeakerPosition": null, "SpeakerName": null, "QuestionSpeaker": null, "Speaker": null, "Utterance": "〔知事本局長秋山俊行君登壇〕", "QorA": "O", "QuestionerID": "", "QAID": 0},
  {"Volume": "2011-2", "Number": "3", "Date": "2011-06-24", "Title": "平成二十三年東京都議会会議録第九号", "SpeakerPosition": "知事本局長", "SpeakerName": "秋山俊行", "QuestionSpeaker": null, "Speaker": null, "Utterance": "「十年後の東京」計画の改定についてでありますが、これまで都は、「十年後の東京」計画のもと、環境、安全、福祉、産業などさまざまな分野で先進的な施策を推進してきましたが、計画期間が半ばを迎えたことに加えまして、東日本大震災によりまして新たな課題も明らかになってまいりました。", "QorA": "A", "QuestionerID": "2011_02_g01", "QAID": 2},
  ...
]
```

A.2 Question Answering

A.2.1 Data Fields. The Question Answering dataset consists of two types of files. The question summary data contains the following items.

ID Identifier of the utterance
Meeting Name of the minutes
Date Date (yyyy-mm-dd)
Headlines Summary of the questioner's entire utterances. Two sentences for each questioner regardless of the number of questions.
SubTopic Subtopic
QuestionSpeaker Questioner's name
QuestionSummary Summary of the question
AnswerSpeaker Answerer's name and position
AnswerSummary Summary of the answer (empty in the test file)

The minutes data contains the following items.

Date Date (yyyy-mm-dd)
Title Name of the minutes
SpeakerPosition Speaker's position or seat number
SpeakerName Speaker's name
QuestionSpeaker Questioner's name and position
Speaker Speaker's name and position
Utterance Utterance

A.2.2 Examples.

Listing 2: Answer sheet for the Question Answering subtask

```
[
  {"ID": "PoliInfo3-QA-v20210613-331-03-1-001",
    "Meeting": "平成 30年第 4回定例会",
    "Date": "2018-12-11",
    "Headlines": ["中小企業･小規模企業の支援を", "幼児教育無償化への都の対応は"],
    "SubTopic": "産業振興",
    "QuestionSpeaker": "小山くにひこ（都ファースト）",
    "QuestionSummary": "中小企業･小規模企業振興条例の理念に基づき、活力ある地域社会をつくり雇用の創出を。",
    "AnswerSpeaker": "知事",
    "AnswerSummary": "地域経済の持続的発展と雇用創出の実現のため効果の高い振興策を展開。"},
  {"ID": "PoliInfo3-QA-v20210613-331-03-1-002",
    "Meeting": "平成 30年第 4回定例会",
    "Date": "2018-12-11",
    "Headlines": ["中小企業･小規模企業の支援を", "幼児教育無償化への都の対応は"],
    "SubTopic": "産業振興",
    "QuestionSpeaker": "小山くにひこ（都ファースト）",
    "QuestionSummary": "農業は東京の持続的成長に必要不可欠。農業振興への今後の展開は。",
    "AnswerSpeaker": "知事",
    "AnswerSummary": "都市農地の保全、担い手の確保と育成･定着の体制整備、先進技術活用等、様々な施策を展開。"},
  {"ID": "PoliInfo3-QA-v20210613-331-03-1-003",
    "Meeting": "平成 30年第 4回定例会",
    "Date": "2018-12-11",
    "Headlines": ["中小企業･小規模企業の支援を", "幼児教育無償化への都の対応は"],
    "SubTopic": "ダイバーシティ･東京",
    "QuestionSpeaker": "小山くにひこ（都ファースト）",
    "QuestionSummary": "国の幼児教育無償化案では負担の軽減は十分とは言えず、また認可と認可外で格差が生じる。対応は。",
    "AnswerSpeaker": "知事",
    "AnswerSummary": "待機児童対策協議会で国と意見交換。国の動きを踏まえ適切に対応。"},
...
]
```

Listing 3: Minutes for the Question Answering subtask


```
[
  {"Volume": "2018-4", "Number": "2", "Date": "2018-12-11", "Title": "平成三十年東京都議会会議録第十六号", "SpeakerPosition": "百十五番", "SpeakerName": "小山くにひこ",
   "QuestionSpeaker": "小山くにひこ（都ファースト）", "Speaker": "小山くにひこ（都ファースト）",
   "Utterance": "東京都議会第四回定例会に当たり、都民ファーストの会東京都議団を代表して、小池知事及び教育長、関係局長に質問いたします。"},
...,
  {"Volume": "2018-4", "Number": "2", "Date": "2018-12-11", "Title": "平成三十年東京都議会会議録第十六号", "SpeakerPosition": "知事", "SpeakerName": "小池百合子",
   "QuestionSpeaker": "小山くにひこ（都ファースト）", "Speaker": "知事",
   "Utterance": "個々のライフスタイルに応じました柔軟な働き方を大会のレガシーとして浸透させて、時差ビズを新たな常識として定着させてまいります。"},
  {"Volume": "2018-4", "Number": "2", "Date": "2018-12-11", "Title": "平成三十年東京都議会会議録第十六号", "SpeakerPosition": "知事", "SpeakerName": "小池百合子",
   "QuestionSpeaker": "小山くにひこ（都ファースト）", "Speaker": "知事",
   "Utterance": "次に、幼児教育、保育の無償化についてでございます。"},
...,
]
```


A.3 Fact Verification

A.3.1 Data Fields. The Fact Verification dataset consists of two types of files. The input claim contains the following items.

ID Number that uniquely identifies the claim
Prefecture Meeting location
Date Date of the meeting
Meeting Title of the meeting
Speaker Speaker name of the utterance
UtteranceSummary Summary of the utterance
UtteranceType Type of utterance (question or answer)
ContextSummary Summary of entire dialog before and after the utterance
ContextWord Topic word related to the utterance
RelatedUtteranceSummary Another utterance related to the utterance. Example: When "UtteranceType" is "answer", "RelatedUtteranceSummary" is a summary of the question from which the answer was based.
StartingLine Target value of the task. "Line" of the predefined primary source corresponding to "UtteranceSummary". This field is -1 when "DocumentEntailment" is false.
EndingLine Target value of the task. "Line" of the predefined primary source corresponding to "UtteranceSummary". This field is -1 when "DocumentEntailment" is false.
DocumentEntailment Target value of the task (whether or not the claim is credible)

The predefined primary source contains the following items.

ID Identification code
Line Line number
Prefecture Prefecture name
Volume Volume
Number Number
Year Year
Month Month
Day Date
Title Title
Speaker Speaker
Utterance Utterance

A.3.2 Examples.

Listing 4: Answer sheet for the Fact Verification subtask


```
[
  {
    "ID": "00002",
    "Prefecture": "東京都",
    "Date": "23-9-28",
    "Meeting": "平成 23年_第３回定例会",
    "Speaker": "知事",
    "UtteranceSummary": "首都の知事として強い危機感に立ち、現場を踏まえて緊急になすべきことを建言した。日本再生に向けて速やかに行動して、都民･国民の不安を振り払ってもらいたい。",
    "UtteranceType": "answer",
    "ContextSummary": "放射能対策に一丸で取り組め。スポーツの力で復興の後押しを",
    "ContextWord": "新内閣への建言",
    "RelatedUtteranceSummary": "知事が込めた想いは。",
    "StartingLine": 8275,
    "EndingLine": 8283,
    "DocumentEntailment": false
  },
  ...
]
```


Listing 5: Minutes for the Fact Verification subtask


```
[
  ...
  {"ID":"130001_230928_828","Line":8274,"Prefecture":"東京都","Volume":"平成 23年_第３回","Number":"2","Year":23,"Month":9,"Day":28,"Title":"平成 23年_第３回定例会(第 12号)","Speaker":"石原慎太郎","Utterance":"鈴木あきまさ議員の代表質問にお答えいたします。"},
  {"ID":"130001_230928_829","Line":8275,"Prefecture":"東京都","Volume":"平成 23年_第３回","Number":"2","Year":23,"Month":9,"Day":28,"Title":"平成 23年_第３回定例会(第 12号)","Speaker":"石原慎太郎","Utterance":"まず、新内閣への建言に込めた思いについてでありますが、原発事故への対応など、国の政がかつてなく混迷し、国民は先が見えない不安を募らせております。"},
  {"ID":"130001_230928_830","Line":8276,"Prefecture":"東京都","Volume":"平成 23年_第３回","Number":"2","Year":23,"Month":9,"Day":28,"Title":"平成 23年_第３回定例会(第 12号)","Speaker":"石原慎太郎","Utterance":"それゆえ、新内閣は、まず何よりも速やかな東日本大震災からの復旧、復興、原発事故の収束と放射性物質対策に全力で取り組むことが求められております。"},
  ...
]
```

A.4 Budget Argument Mining

A.4.1 Data Fields. The Budget Argument Mining dataset consists of two types of files. The budget file contains the following items.

budgetId Identifier of the budget
budgetTitle Budget title
typesOfAccount Type of account (general or special)
department Name of competent ministry/department
url URL
budgetItem Budget item
categories Top Hierarchy, Category (Details – Overview)
budget Current year's budget
budgetLastYear Previous year's budget
description
budgetDifference Comparative increase or decrease

The meeting minutes file includes the following items.

date Date
localGovernmentCode Local government code (6 digits)
localGovernmentName Local government name
proceedingTitle Proceeding title
url URL
proceeding Speakers and their comments
- **speakerPosition** Position
- **speaker** Speaker
- **utterance** Speech
  - **moneyExpressions** Money expression included in the statement
    - **moneyExpression** Money expression
    - **relatedID** Related budget IDs, list type (can store multiple IDs)
    - **argumentClass** Argument labels

A.4.2 Examples.

Listing 6: Minutes for the Budget Argument Mining subtask


```
{
  "local": [
    {
      "date": "2019-02-15",
      "localGovernmentCode": "401307",
      "localGovernmentName": "福岡市",
      "proceedingTitle": ":平成 31年第 1 回定例会(第 1 日) 本文",
      "url": "",
      "proceeding": [
        ...
        {
          "speakerPosition": "市長",
          "speaker": "高島宗一郎",
          "utterance": "ただいま上程になりました議案 29 件について提案の趣旨を説明いたします。\
n まず、予算案について説明いたします。\
n 今回の補正規模は、一般会計 201 億 9,399万円の追加、特別会計 111億 609万円の追加、企業会計 33億 462万円の追加、合計 346億 470万円の追加となっております。
          その主な内訳は、国補正予算関連として 90億 7,012 万円の追加、そのうち、街路整備事業 37億 400万円の追加、公共下水道整備事業 29億 3,200万円の追加、このほか、年間執行見込みの増加に伴う教育、保育給付費 7 億 6,815万円の追加などとなっております。...そのほかの一般議案といたしまして、下水道施設の管理のかしに基づく損害賠償の額を決定するための議案、滞納学校給食費等の支払いを求めるため訴えの提起をするための議案、道路の新設、道路の組みかえ等に伴い、市道路線の認定及び変更を行うための議案 2 件を提出いたしております。\n 以上で説明を終わります。よろしく御審議をお願いします。",
          "moneyExpressions": [
            {
              "moneyExpression": "201億 9,399万円",
              "relatedID": null,
              "argumentClass": "Premise : 未来(現在以降)・見積"
            },
            {
              "moneyExpression": "111億 609万円",
              "relatedID": null,
              "argumentClass": "Premise : 未来(現在以降)・見積"
            },
            ...
            {
              "moneyExpression": "29億 3,200万円",
              "relatedID": [
                "ID-2019-401307-00-000099"
              ],
              "argumentClass": "Premise : 未来(現在以降)・見積"
            },
            {
              "moneyExpression": "7 億 6,815万円",
              "relatedID": [
                "ID-2019-401307-00-000031"
              ],
              "argumentClass": "Premise : 未来(現在以降)・見積"
            },
            ...
          ]
        },
        ...
      ]
    },
    ...
  ],
  "diet": [
    {
      "issueID": "120105261X02520200608",
      "imageKind": "会議録",
      "searchObject": 0,
      "session": 201,
      "nameOfHouse": "衆議院",
      "nameOfMeeting": "予算委員会",
      "issue": "第 25号",
      "date": "2020-06-08",
      "closing": null,
      "speechRecord": [
        {
          "speechID": "120105261X02520200608_002",
          "speechOrder": 2,
          "speaker": "麻生太郎",
          "speakerYomi": "あそうたろう",
          "speakerGroup": "自由民主党・無所属の会",
```

```
          "speakerPosition": "財務大臣・内閣府特命担当大臣（金融）",
          "speakerRole": null,
          "speech": "○麻生国務大臣　令和二年度第二次補正予算の大要につきましては、既に本会議において申し述べたところでありますが、予算委員会での御審議をお願いするに当たり、改めて御説明をさせていただきます。\r\n 最初に、一般会計予算の補正について申し上げます。\r\n 本補正予算につきましては、総額で三十一兆九千百三十四億円の歳出追加を行うこととしております。その内容としては、新型コロナウイルス感染症対策関係経費として、雇用調整助成金の拡充等に係る経費に四千五百十九億円、...\r\n 以上、令和二年度第二次補正予算の大要について御説明をさせていただきました。\r\n 何とぞ、御審議の上、速やかに御賛同いただきますようお願いを申し上げます。",
          "startPage": 1,
          "createTime": "2020-06-23 20:41:47",
          "updateTime": "2020-06-24 10:24:07",
          "speechURL": "https://kokkai.ndl.go.jp/#/detail?minId=120105261X02520200608&spkNum=2&single",
          "moneyExpressions": [
            {
              "moneyExpression": "三十一兆九千百三十四億円",
              "relatedID": null,
              "argumentClass": "Premise : 未来（現在以降）・見積"
            },
            {
              "moneyExpression": "四千五百十九億円",
              "relatedID": [
                "R2-MHLW-BUDGET-02-FIXED-000018"
              ],
              "argumentClass": "Premise : 未来（現在以降）・見積"
            },
            ...
          ]

        }
      ],
      "meetingURL": "https://kokkai.ndl.go.jp/#/detail?minId=120105261X02520200608",
      "pdfURL": "https://kokkai.ndl.go.jp/#/detailPDF?minId=120105261X02520200608"
    },
    ...
  ]
}
```


Listing 7: Budget information for the Budget Argument Mining subtask


```
{
  "local": {
    ...
    "401307": [
      ...
      {
        "budgetId": "ID-2019-401307-00-000031",
        "budgetTitle": "平成３１年度当初予算案",
        "url": "https://www.city.fukuoka.lg.jp/data/open/cnt/3/67165/1/05.H31juuyousesaku.pdf?20190308134622",
        "budgetItem": "安心して生み育てられる環境づくり",
        "budget": "97,173,470千円",
        "categories": [],
        "typesOfAccount": null,
        "department": "こども未来局",
        "budgetLastYear": null,
        "description": " 増加する保育ニーズに対応するため、保育所の新設や増改築の他、企業主導型保育事業や幼稚園における２歳児受け入れの促進など多様な手法により、3,000人分の保育の受け皿を確保する。また、様々な就労形態に対応するため、休日・夜間における保育や延長保育及び、子育ての負担感を軽減する一時預かり事業などを継続して実施するとともに、病気やその回復期にある乳幼児等を一時的に保育する病児・病後児デイケア事業の実施や、公立保育所における医療的ケア児のモデル的な受け入れの拡大により、多様な保育サービスの充実を図る。障がい児保育については、社会情勢等の変化による保育ニーズの高まりなどを踏まえ、総合的な制度の見直しを行う。 さらに、保育所等の増加に伴い必要な保育士等を確保するため、潜在保育士の再就職にあたって就職準備金の貸付等を行う事業を実施するとともに、正規雇用の保育士に対する家賃助成や新たに奨学金返済の支援を行い、市内保育所への就職促進や離職防止を図る。あわせて、保育業務のＩＣＴ化推進のためのシステムや事故防止のための機器の導入に要する費用の助成を行うとともに、保育事業に新たに参入する事業者を訪問し、助言等を行う巡回支援事業を引き続き実施するなど、保育の質の維持・向上を図る。 また、国が 2019年 10月１日の実施を目指している「幼児教育・保育の無償化」については、国の動向を踏まえ、適切に対応する。 母親と子どもの心と体の健康づくりの推進や乳幼児の虐待予防を強化するため、妊婦健康診査の公費助成や乳幼児健康診査、新生児の先天性代謝異常検査を継続して実施するとともに、先天性難聴を早期発見し、早期療育につなげるため、新生児聴覚検査に要した費用の助成を開始する。 妊娠期からの相談支援体制の強化を図るため、各区の子育て世代包括支援センターにおいて関係各課が連携して、妊娠期から子育て期までの切れ目のない支援を行う。産後早期の母親への支援の充実を図るため、宿泊や日帰りによる産後ケア事業や、産後ヘルパー派遣事業を実施するとともに、引き続き、助産師等の専門職による乳児がいる家庭への全戸訪問を実施するなど、母子保健事業を推進する。 また、子どもを望む夫婦に対する一般不妊治療費助成事業や特定不妊治療費助成事業を継続して実施する。さらに、不妊専門相談センターにおいて、不妊や不育に関する専門的な相談に応じるとともに、妊娠・出産に関する正しい知識の普及啓発に取り組む。 ひとり親家庭の生活の安定と向上のため、自立支援給付金事業を拡充するとともに、ひとり親家庭支援センターでの就業相談や自立支援プログラム策定事業を実施し、就業や自立に向けた支援に取り組む。 また、配偶者からの暴力被害者の相談・支援を行うＤＶ相談・支援推進事業については、ＤＶ被害者の相談・支援のほか、研修や広報・啓発に継続して取り組む。 子育てにかかる経済的負担の軽減を図るため、第３子優遇事業及び、児童手当・児童扶養手当の支給を継続するとともに、未婚の児童扶養手当受給者への臨時・特別給付金の支給を実施する。 少子化対策として、毎月 1～7日を“「い～な」ふくおか子ども週間 ”とし、子どもたちをバックアップする運動の普及・啓発に引き続き取り組む。",
        "budgetDifference": null
      },
      ...
      {
        "budgetId": "ID-2019-401307-00-000099",
        "budgetTitle": "平成３１年度当初予算案",
        "url": "https://www.city.fukuoka.lg.jp/data/open/cnt/3/67165/1/05.H31juuyousesaku.pdf?20190308134622",
        "budgetItem": "下水道整備",
        "budget": "23,362,000千円",
        "categories": [],
```

```
          "typesOfAccount": null,
          "department": "道路下水道局",
          "budgetLastYear": null,
          "description": " 下水道サービスを継続的に提供するため、管渠・ポンプ場・処理場における老朽施設の改築更新を最重点として、計画的に取り組む。また、重点地区を定めた「雨水整備Ｄｏプラン 2026」(案)により、引き続き雨水対策を進める。特に、天神周辺地区については、都心部の雨水対策を強化した「レインボープラン」により、従来の流下型施設の整備に加え、雨水流出抑制施設の導入も進める。さらに、地震被害を軽減するための既存施設の耐震化に取り組む。また、新たなまちづくりに併せた施設の整備、公共用水域の水質保全のための合流式下水道の改善(分流化)など、管渠・ポンプ場・処理場の整備を計画的に推進し、都市環境の向上に努める。加えて、資源の有効利用を図るため、下水処理水による再生水利用を推進するとともに、下水汚泥固形燃料化施設の導入をはじめとした再生可能エネルギーの活用に積極的に取り組む。",
          "budgetDifference": null
        },
        ...
      ]
    },
    "diet": [
      ...
      {
        "budgetId": "R2-MHLW-BUDGET-02-FIXED-000018",
        "budgetTitle": "令和２年度 厚生労働省第二次補正予算(案)の概要",
        "url": "https://www.mhlw.go.jp/wp/yosan/yosan/20hosei/dl/20hosei03.pdf",
        "budgetItem": "○ 雇用調整助成金の抜本的拡充",
        "budget": "７,７１７億円",
        "categories": [
          "(１)雇用を守るための支援",
          "第３ 雇用調整助成金の抜本的拡充をはじめとする生活支援"
        ],
        "typesOfAccount": null,
        "department": "厚生労働省",
        "budgetLastYear": null,
        "description": "新型コロナウイルス感染症の影響により休業する事業主を支援するため、４月１日以降に開始される賃金締切期間中の休業について、９月まで雇用調整助成金の日額上限を 8,330 円から 15,000 円まで特例的に引き上げる。同時に解雇等を行わない中小企業の助成率を 10/10に引き上げ、緊急対応期間を９月まで延長する。また、支給処理に係る人員体制の強化及び社会保険労務士との協力体制の構築等により、雇用調整助成金の支給の迅速化を図る。",
        "budgetDifference": null
      },
      ...
    ]
  }
```


Table 16: Scores of Question Answering subtask in dry run (ROUGE scores)

ID	Team	ROUGE (Recall)			ROUGE (F-measure)		
		N1	N2	R	N1	N2	R
Surface form							
105	nukl	0.4591	**0.2423**	**0.4048**	**0.4361**	**0.2261**	**0.3835**
100	AKBL	0.4475	0.2169	0.3838	0.3958	0.1898	0.3401
78	AKBL	**0.4669**	0.2215	0.3983	0.3878	0.1834	0.3313
95	nukl	0.4240	0.1910	0.3657	0.4002	0.1786	0.3452
76	AKBL	0.4645	0.2196	0.3960	0.3842	0.1812	0.3278
77	AKBL	0.4057	0.1897	0.3455	0.2772	0.1172	0.2328
42	TO	0.2668	0.0555	0.2224	0.2516	0.0518	0.2084
Stem							
105	nukl	0.4637	**0.2462**	**0.4091**	**0.4399**	**0.2295**	**0.3872**
100	AKBL	0.4554	0.2231	0.3906	0.4020	0.1951	0.3457
78	AKBL	**0.4741**	0.2282	0.4045	0.3939	0.1889	0.3366
95	nukl	0.4294	0.1967	0.3695	0.4051	0.1837	0.3486
76	AKBL	0.4726	0.2267	0.4031	0.3906	0.1869	0.3336
77	AKBL	0.4219	0.2057	0.3618	0.2886	0.1273	0.2436
42	TO	0.2811	0.0634	0.2351	0.2652	0.0591	0.2202
Content word							
105	nukl	**0.3063**	**0.1615**	**0.3003**	**0.2861**	**0.1479**	**0.2803**
100	AKBL	0.2759	0.1450	0.2682	0.2416	0.1263	0.2351
78	AKBL	0.2840	0.1559	0.2772	0.2336	0.1278	0.2280
95	nukl	0.2448	0.1179	0.2390	0.2311	0.1111	0.2253
76	AKBL	0.2809	0.1519	0.2750	0.2306	0.1248	0.2257
77	AKBL	0.2784	0.1580	0.2708	0.1797	0.0982	0.1746
42	TO	0.0916	0.0435	0.0899	0.0879	0.0415	0.0861

Table 17: Scores of QA Alignment subtask in dry run

ID	Team	F-Measure	Precision	Recall
89	AKBL	**0.8202**	0.8268	**0.8186**
117	AKBL	0.8193	**0.8269**	0.8165
59	AKBL	0.8076	0.8137	0.8074
156	ditlab	0.7614	0.7488	0.7815
151	ditlab	0.7597	0.7599	0.7667
160	ditlab	0.7579	0.7462	0.7793
159	ditlab	0.7576	0.7462	0.7782
155	ditlab	0.7516	0.7516	0.7586
150	ditlab	0.7452	0.7463	0.7512
127	ditlab	0.7332	0.7309	0.7427
149	ditlab	0.7112	0.7166	0.7136
34	TO	0.6652	0.6526	0.6857

B RESULTS OF DRY RUN

Tables 16, 17, 19, and 20 show the automatic evaluation results of Question Answering, QA Alignment, Fact Verification, and Budget Argument Mining subtasks in the dry run, respectively. Table 18 shows the human evaluation results of Question Answering in the dry run. Because the test data and/or metrics of Fact Verification and Budget Argument Mining were revised during the dry run, we separated the results by period.

Table 18: Scores of Question Answering subtask in dry run (human evaluation results)

ID	Team	Correspondence				Content				Well-formed				Overall			
		A	B	C	Score	A	B	C	Score	A	B	C	Score	A	B	C	Score
Gold		139	9	2	287	71	74	5	216	143	6	1	292	85	62	3	232
105	nukl	113	24	13	**250**	42	70	38	**154**	137	9	4	**283**	51	64	35	**166**
100	AKBL	101	17	32	219	43	52	55	138	137	8	5	282	49	47	54	145
42	TO	50	28	72	128	2	37	111	41	32	28	90	92	7	28	115	42

Table 19: Scores of Fact Verification subtask in dry run

ID	Team	F-Measure	Precision	Recall
November 17, 2021				
161	10807010	**0.8209**	**0.8321**	**0.8417**
162	10807010	0.8067	0.8187	0.8256
157	AKBL	0.7912	0.7947	0.8012
158	AKBL	0.7909	0.7947	0.8006
August 8, 2021				
129	AKBL	**0.8809**	**0.8764**	**0.8854**
124	AKBL	**0.8809**	**0.8764**	**0.8854**
119	AKBL	0.8700	0.8648	0.8752
147	AKBL	0.8677	0.8612	0.8743
146	AKBL	0.8650	0.8615	0.8686
132	AKBL	0.8650	0.8615	0.8686
148	AKBL	0.8523	0.8462	0.8584
118	AKBL	0.8318	0.8258	0.8378
113	10807010	0.8221	0.8187	0.8256
144	10807010	0.8210	0.8187	0.8232
122	10807010	0.8210	0.8187	0.8232
114	10807010	0.8140	0.8038	0.8245
83	AKBL	0.8095	0.8053	0.8137
110	10807010	0.8094	0.8050	0.8139
145	10807010	0.8084	0.8129	0.8040
120	Forst	0.8046	0.8002	0.8091
84	AKBL	0.8035	0.7912	0.8162
94	Forst	0.8025	0.7951	0.8101
96	Forst	0.8003	0.7781	0.8238
107	10807010	0.7978	0.7825	0.8137
97	Forst	0.7926	0.7728	0.8134
88	Forst	0.7609	0.7370	0.7863
85	AKBL	0.7590	0.7511	0.7671
65	AKBL	0.7477	0.7506	0.7448
81	Forst	0.6789	0.6785	0.6793
106	10807010	0.4430	0.4430	0.4430
41	TO	0.4430	0.4430	0.4430
79	AKBL	0.4187	0.4166	0.4209
June 13, 2021				
23	AKBL	**0.8143**	**0.8141**	**0.8144**

Table 20: Scores of Budget Argument Mining subtask in dry run

ID	Team	Score	AC	RID
October 27, 2021				
154	fuys	**0.1277**	**0.5692**	0.1702
128	fuys	**0.1277**	0.5365	0.1702
111	OUC	**0.1277**	0.3731	**0.2128**
112	rVRAIN	0.0638	0.4846	**0.2128**
163	OUC	0.0000	0.3731	0.0000
131	takelab	0.0000	0.3346	0.0000
130	takelab	0.0000	0.3673	0.0000
126	rVRAIN	0.0000	0.4846	0.1702
125	takelab	0.0000	0.3404	0.0213
123	OUC	0.0000	0.3731	0.1277
116	fuys	0.0000	0.5365	0.0000
109	rVRAIN	0.0000	0.4846	0.0000
108	TO	0.0000	0.1346	0.0000
July 28, 2021				
98	fuys	**0.5058**	-	-
104	rVRAIN	0.4462	-	-
101	rVRAIN	0.4462	-	-
87	rVRAIN	0.3981	-	-
86	rVRAIN	0.3865	-	-
93	fuys	0.3596	-	-
90	OUC	0.3269	-	-
80	rVRAIN	0.2519	-	-
92	OUC	0.1308	-	-
29	TO	0.0981	-	-
28	TO	0.0250	-	-
91	OUC	0.0077	-	-
82	rVRAIN	0.0058	-	-

C RESULTS OF LATE SUBMISSIONS

Although the deadline was November 30, we accepted submissions until March 10 for the same dataset as that used in the formal run. They were treated as late submissions. Tables 21, 22, 23, and 24 show the automatic evaluation results of the late submissions of QA Alignment, Question Answering, Fact Verification, and Budget Argument Mining subtasks, respectively.

Table 21: Scores of late submissions in QA Alignment subtask

ID	Team	F-Measure	Precision	Recall
333	ditlab	**0.8348**	**0.8739**	0.8045

Table 22: Scores of late submissions in Question Answering subtask (ROUGE scores)

ID	Team	ROUGE (Recall)			ROUGE (F-measure)		
		N1	N2	R	N1	N2	R
Surface form							
316	nukl	0.4601	0.2290	0.4030	0.4378	0.2178	0.3826
Stem							
316	nukl	0.4689	0.2341	0.4090	0.4464	0.2228	0.3885
Content word							
316	nukl	0.2906	0.1607	0.2846	0.2787	0.1520	0.2726

Table 23: Scores of late submissions in Fact Verification subtask

ID	Team	F-Measure	Precision	Recall
341	Forst	0.8563	0.8591	0.8642
339	Forst	0.7980	0.7989	0.8065
340	Forst	0.7970	0.7964	0.8058
338	Forst	0.6857	0.6864	0.6925

Table 24: Scores of late submissions in Budget Argument Mining subtask

ID	Team	Score	AC	RID
320	OUC	0.3830	0.4346	**0.6596**
319	OUC	0.3617	0.4212	**0.6596**
318	OUC	0.3617	0.5058	**0.6596**
321	fuys	0.2340	0.5365	0.3404
337	Ibrk	0.0000	0.3731	0.0000
336	Ibrk	0.0000	0.3577	0.0000
335	Ibrk	0.0000	0.3577	0.0000
334	Ibrk	0.0000	0.3577	0.0000
332	Ibrk	0.0000	0.3462	0.0000
331	Ibrk	0.0000	0.3154	0.0000
330	Ibrk	0.0000	0.3385	0.0000
329	Ibrk	0.0000	0.3462	0.0000
328	Ibrk	0.0000	0.3654	0.0000
327	Ibrk	0.0000	0.0654	0.0000
326	Ibrk	0.0000	0.0654	0.0000
317	OUC	0.0000	0.5058	0.0000