NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001326

Overview of the NTCIR-17 QA Lab-PoliInfo-4 Task

Yasuhiro Ogawa
Nagoya City University
Japan
ogawa@ds.nagoya-cu.ac.jp

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

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

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

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

Akio Kobayashi
National Agriculture and Food Research Organization
Japan
akio.kobayashi@naro.go.jp

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

ABSTRACT

The goal of the NTCIR-17 QA Lab-PoliInfo-4 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-4 consists of four subtasks: Question Answering-2, Answer Verification, Stance Classification-2, and Minutes-to-Budget Linking. In this paper, we present the data used and the results of QA Lab-PoliInfo-4's formal run.

TEAM NAME

Task Organizers

SUBTASKS

Overview

1 INTRODUCTION

NTCIR-17's Question Answering Lab for Political Information 4 (QA Lab-PoliInfo-4) task seeks to develop complex real-world question-answering (QA) techniques. In this task, the participants extract and summarize 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 2nd as International Fact-Checking Day. Fact-checking is difficult for general Web search engines because of the "filter bubble" coined by Pariser [25], 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 very long speech transcripts, making it challenging to understand the contents at a glance, such as members' opinions. New information access technologies to support user understanding are expected, which should protect us from fake news.

We use 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 evidence for each member's utterance;
5: Link to different language resources; and
6: Deal with colloquial Japanese, including dialect and slang.

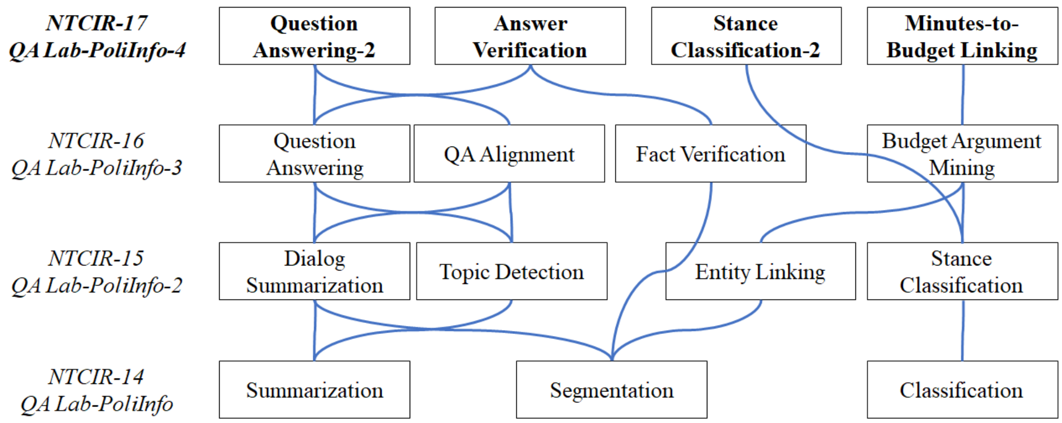

Figure 1: Relations between subtasks

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 designed several subtasks on political information in NTCIR-14, NTCIR-15, and NTCIR-16. NTCIR-17 QA Lab-PoliInfo-4 includes Question Answering-2, Answer Verification, Stance Classification-2, and Minutes-to-Budget Linking subtasks. The Question Answering-2 subtask is the same task as the Question Answering subtask in NTCIR-16 QA Lab-PoliInfo-3 [12] and its purpose is to return a concise answer to a question about the assembly minutes. The Answer Verification subtask is a combinational expansion of the Question Answering and Fact Verification subtasks in NTCIR-16 QA Lab-PoliInfo-3. The Answer Verification subtask aims to check the output of the Question Answering subtask. The Stance Classification-2 subtask is a successor of the Stance Classification in NTCIR-15 QA Lab-PoliInfo-2 [11] and aims to infer the speaker's stances on bills from the speeches of assembly members. The Minutes-to-Budget Linking subtask is a successor of the Budget Argument Mining in NTCIR-16 QA Lab-PoliInfo-3. The Minutes-to-Budget Linking seeks to identify argumentative components related to a budget item and then classify them based on 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 [13]. 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 [27]. SQuAD 2.0 combines the existing SQuAD with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones [26]. HotpotQA is a question answering dataset that 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 [22, 40].

2.3 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 [41]. Oshikawa et al. investigated the difference between fake news detection and other related tasks, and the importance of Natural Language Processing (NLP) solutions for fake news detection [24]. 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 [28]. Sharma et al. compiled a list of available datasets around fake news detection and summarized their characteristic features [29].

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

2.4 Fact-Checking

FEVER is a Fact Extraction and VERification Shared dataset that classifies whether human-written factoid claims could be supported or refuted using evidence retrieved from Wikipedia [35]. 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 [36]. 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 [2, 6].

2.5 Factual Error Detection

A factual error is a statement that contradicts the source documents. With the advent of large-scale language models (LLMs), automatic summarization has become as fluent as human summarization. However, LLMs suffer from a problem called "hallucination." In automatic summarization, hallucination causes the summary to include statements that are not actually found in the source documents; this problem is called factual error. Currently, in the field of summarization, much research is focused on detecting factual errors and creating corpora annotated with them [4, 8, 19, 31, 32, 42].

2.6 Stance Classification

Stance Classification (also known as Stance Detection) is a task that identifies the standpoint of the producer of a piece of text towards a given target [14]. The earliest competition on this task is SemEval-2016's shared task on Twitter stance detection [21]. Subsequently, various datasets have been created, mainly on social media [9, 33, 39]. On the other hand, stance classification on assemblies were not investigated for a long time, except for one earlier work [34], until our previous NTCIR-15 QA-Lab-PoliInfo-2 Stance Classification task.

2.7 Argument Mining

Research on argument mining has garnered considerable attention as a logic-based approach to NLP to capture the structure of arguments [7, 37]. Argument structure analysis is a typical task in argument mining that assigns labels (claim, premise) to discourse units of sentences and clauses [16]. Common processes in argument mining analysis include the identification of argumentative components, clause attributes, and relationships between clauses [17]. IBM Research AI presented "Project Debater," an autonomous debating system that can engage in a competitive debate with humans [30].

2.8 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 [5]. 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 [20]. 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 Question Answering-2, Answer Verification, Stance Classification-2, and Minutes-to-Budget Linking subtasks. We believe 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 tasks, which were published on the QA Lab-PoliInfo-4 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.

3.1 Question Answering-2

3.1.1 *Purpose.*

The purpose of the Question Answering-2 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, each question is not directly associated with its answer in the minutes, so participants must associate the input question with its answer in the minutes. This task is the same task as the QA Alignment subtask in NTCIR-16 QA Lab-PoliInfo-3.

3.1.2 *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, using 2020 data for the dry run and 2021 data for the formal run. For a complete description of the structure, see Appendix A.1.

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

Data size. See Table 1.

3.1.3 *Evaluation.*

For this task, we conducted automatic and human evaluation.

Automatic evaluation. We consider the answer summary in *Togikaidayori* the gold standard and calculated ROUGE scores [18]. 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*, from 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 ... ?"

[2]https://sites.google.com/view/poliinfo4/

Table 1: Statistics of data for the Question Answering subtask

Dataset		Sentences Question	Sentences Answer	Summaries	Date
Dry run	Train	150,194	72,128	7,627	September 2001 – December 2019
	Test	9,203	7,360	416	February 2020 – December 2020
Formal run	Train	159,397	79,488	8,046	September 2001 – December 2020
	Test	6,675	5,942	294	February 2021 – December 2021

Table 2: Statistics of data for the Answer Verification subtask

Dataset		Q&A pairs	facts	fakes
Train	NTCIR-16 QA	730	415	315
	GDADC	785	0	785
Test	Dry run	79	39	40
	Formal run	100	37	63

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.

3.1.4 Baseline system.
For the baseline system, we adopted the system proposed by the nukl team at PoliInfo-3, called 'Method 1,' and used T5 as a summarizer [23]. We concatenated the input question, its subtopic, and the answerer's entire utterance using a comma (,) as a separator. Then, we tokenized the concatenated text by SentencePiece [15] and input it into T5. However, an entire utterance can be long and sometimes exceed T5's input limit, so we selected a maximum number of last sentences from the utterance within the limit. We chose the last sentences because, in assembly, answerers often first touch on the topic of the question, then talk about the current situation, and finally talk about solutions or future measures.

3.2 Answer Verification

3.2.1 Purpose.
Short answers generated automatically in the Question Answering task of NTCIR-16 included many facts of the arguments, but there were mistakes with the answers. A short answer that contains a wrong part, even if the rest is correct, is regarded as a fake answer, so fact checking of generated answers is necessary. Because answers submitted in the NTCIR-16 Question Answering were evaluated for their truth by the participants and by using the Q&A pairs as training data, we could build a simple binary classifier that returns fact or fake when a question and its answer are given. However, the training data size was too small to build a robust and reliable classifier. In particular the lack of fake answers was a big problem. Therefore, we conducted the Answer Verification task to expand the training data set and improve the fact-checking classifier.

3.2.2 Fake answer generation.
Wallace et al. [38] proposed dynamic adversarial data collection (DADC), which is a continuous cycle of improving a prediction system based on adversarial data and collecting new adversarial cases against the improved prediction system. They showed that a robust system can be constructed by DADC. Although DADC is premised on crowdsourcing, there are problems such as the cost and incentives of annotators. Because gamification incentives were effective in encouraging annotators to act spontaneously and continuously [10], we applied DADC with gamification incentives (GDADC) to expand the data set for fake answers[3].

3.2.3 Data.

Input.

- Questions of Q&A sessions in the newsletters
- List of fact and fake answers to the question in a few lines
- The minutes of the Tokyo Metropolitan Assembly

Output.

- List of True or False (binary) (True indicates that the answer is not fake.)

Data size. See Table 2.

3.2.4 Evaluation. We used accuracy as the evaluation metric.

3.3 Stance Classification-2

3.3.1 Purpose.
The Stance Classification task aims at estimating a politician's position from her/his utterances. Taking a lesson from the last Stance Classification task evaluated at the NTCIR 15 QA Lab-PoliInfo-2, we took into account the following two aspects. First, we redesigned the classification task itself. In the last task, the information source of the classification was the assembly minutes as a whole. We found that members of an assembly tend to state their stance on a given topic explicitly at the beginning of their speech. While most of the participants successfully exploited that to achieve good performance, the use of such superficial expression does not match well with our purpose, i.e., estimating a politician's position from the contents of her/his utterances. Therefore, in the new Stance Classification-2 task, we focused on the classification of members' opinions about a given topic without any explicit statement on their stance. Second, we extended the target minutes to several local governments in Japan other than Tokyo Metropolitan Assembly.

In the Stance Classification-2 subtask, given an utterance of a politician associated with a topic (agenda), participant systems are

[3]https://sites.google.com/view/poliinfo4/game (in Japanese)

requested to classify it into two categories (agreement or disagreement).

3.3.2 Data.

We extracted politicians' utterances on the last day of a series of a regular meeting, in which they took a vote on a given topic; therefore they should have determined their position clearly. We only used a specific group of local governments in which topics were discussed one by one, so that we could assure that an extracted utterance was unambiguously associated with a topic. To ensure that the utterances have no explicit expression about speaker's stance, some pre-defined tokens were replaced with a special token `[STANCE]`. We chose '賛成' (agreement) and '反対' (disagreement) for such pre-defined tokens. At the same time, we utilized these tokens to assign a golden label to the utterances by using heuristic rules. Through our preliminary experiments, we found this method seldomly assigned incorrect labels.

We distributed two separate CSV files for training and test data, whose data fields are shown in A.3.1. In the test data, the 'stance' field is left blank and the participant systems are requested to be filled with either 'agreement' or 'disagreement.' For the dry run, we released 3,898 and 426 instances for the training and test data, respectively, which were constructed from 19 local governments in Aichi and Hokkaido Prefectures. For the training data of the formal run, we released 8,534 instances constructed from 26 local governments in Saitama Prefecture. For the test data of the formal run, we released 2,160 instances from 27 (the same 26 and one more) local governments in Saitama Prefecture and 80 instances from (hidden) one local government in Fukuoka Prefecture.

In addition to the regular training data above, we also released their UNMASKED version, in which the texts in the 'utterance' field are NOT masked, i.e., the pre-defined explicit tokens are not replaced with `[STANCE]` but are left unchanged, hoping the participants will use it for their system development.

3.3.3 Evaluation.

Our official evaluation metric is the accuracy of the predicted labels.

Table 3: Statistics of data for the MBLink

Formal run	Utterances	Table candidates	Fiscal year
Train	198	4,372	H29,H30,R1,R2,R4
Test	81	2,020	H28,R3

3.4 Minutes-to-Budget Linking

3.4.1 Purpose.

Minutes-to-Budget Linking (MBLink) aims to link minutes to budget tables when a sentence contained in the assembly minutes is given, and extracts the evidence for the discussion.

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.

We have been working on these issues since the NTCIR-16 Budget Argument Mining subtask. NTCIR-17 MBLink is the subtask that addresses the issues from Budget Argument Mining.

The characteristics of MBLink are as follows:

- The budget table under consideration is not a summarised version, but rather a comprehensive budget table encompassing detailed descriptions found in budget explanation documents.
- The unit of speech within the assembly minutes corresponds to individual sentences from the Otaru city mayor's statements, specifically those pertaining to budget elucidation.
- The evaluation ensures that no unnecessary tables are included, and that all necessary tables are present, thereby avoiding any superfluous information or deficiencies.

3.4.2 Data.

Input.

- Text of the mayor's utterances in the minutes (HTML format) : The minutes include the mayor's written remarks and are in HTML format, with one <p> tag for each statement.The utterance linked to the budget table is assigned the attribute data-mblink-sentence-id and is given a sentence ID.Training data are assigned the linked table ID as the data-mblink-table-ids attribute. If there is more than one table, the table IDs are given separated by single-byte spaces. Even in the case of test data, the data-mblink-sentence-id attribute is assigned only to the utterance linked to the budget table.
- Tables included in budget descriptions and other documents (HTML format) : Since the budget descriptions were published in PDF files, those PDF files were converted to HTML files. The <table> tag of each table is assigned the data-mblink-table-id attribute as the table ID.

Output. A file (JSON format) linking the budget table associated with the utterance.

Data size. See Table 3.

3.4.3 Dataset.

We used budget tables and minutes from Otaru City. Table 3 presents the number of utterances. The training data contained 198 utterances for the Otaru local assembly minutes. The test data contained 81 utterances for the Otaru local assembly minutes.

3.4.4 Evaluation.

We designed the MBLink score to avoid no superfluous or missing tables. The score is the macro-average of the F1 score of the linked table estimation results for each statement. Let S be the set of utterances in the test data and s_i the i-th utterance. The score is defined as follows using $\mathrm{Precision}_i$, Recall_i, and $\mathrm{F1}_i$ scores.

$$\mathrm{Precision}_i = \frac{\text{Number of table IDs output correctly}}{\text{Number of table IDs output}}$$

$$\mathrm{Recall}_i = \frac{\text{Number of table IDs output correctly}}{\text{Number of } \textit{data-mblink-table-ids} \text{ values}}$$

$$\mathrm{F1}_i = \frac{2 \cdot \mathrm{Precision}_i \cdot \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}$$

Table 4: Active participating teams

Team	Organization
AKBL*	Toyohashi University of Technology
ditlab	Denso IT Laboratory
fuys*	Fukuoka University
HUKB	Hokkaido University
IKM23	National Cheng Kung University CSIE IKM Lab
ISLab	National Kaohsiung University of Science and Technology
KIS	Shizuoka University
omuokdlb	Osaka Metropolitan University
Forst*	Yokohama National University (late submission only)
TO*	task organizers

*Task organizer(s) are in team

$$\text{Score} = \frac{1}{|S|}\sum_{i=1}^{|S|} \text{F1}_i$$

3.4.5 Baseline system.

We did not provide a baseline system. Instead, we submitted the results of a method that outputs 1 to 5 random table IDs for each statement.

4 SCHEDULE

The NTCIR-17 QA Lab-PoliInfo-4 task is running following this timeline:

September 28, 2022: NTCIR-17 kickoff meeting
November 12, 2022: QA Lab-PoliInfo-4 first round table meeting
April 17–21, 2022: QA Lab-PoliInfo-4 second round table meeting
June 17, 2023: QA Lab-PoliInfo-4 third round table meeting
March 8, 2023: Dataset release

Dry Run

March 6– July 3, 2023: Dry run

Formal Run

July 4, 2023: Update of dataset for formal run
July 4–15, 2023: Formal run
July 28 – August 16, 2023: Evaluation by participants
August 17, 2023: Evaluation by organizers
August 18, 2023: Evaluation Result Release

NTCIR-17 CONFERENCE

August 1, 2023: Task overview paper release (draft)
September 1, 2023: Submission due for participant papers
November 1, 2023: Camera-ready participant paper due
December 12–15, 2023: NTCIR-17 Conference

5 PARTICIPATION

Eleven teams registered for the task, but only eight teams participated actively, i.e., submitted results for the formal run. Table 4 shows the active participating teams.

Table 5: Number of submissions in dry run

Team	QA2	AV	SC2	MBLink	Total
AKBL	1	22	2	-	25
fuys	-	-	-	1	1
IKM23	6	-	2	-	8
ISLab	-	-	6	-	6
KIS	-	-	10	-	10
Subtotal	7	22	20	1	50
TO	1	1	1	-	3
Total	8	23	21	1	53

Table 6: Number of submissions in formal run

Team	QA2	AV	SC2	MBLink	Total
AKBL	1	6	1	7	15
ditlab	27	-	-	-	27
fuys	-	-	-	9	9
HUKB	5	-	-	-	5
IKM23	5	-	19	-	24
ISLab	-	-	6	-	6
KIS	-	-	20	-	20
omuokdlb	13	7	-	-	20
Subtotal	51	13	46	16	126
TO	1	2	-	1	4
Total	52	15	46	17	130

6 SUBMISSIONS

Tables 5 and 6 show the number of submissions for the dry run and the formal run, respectively. In the dry run, there were seven submissions from two teams for Question Answering-2, 22 submissions from one team for Answer Verification, 20 submissions from four teams for Stance Classification-2, and one submission from one team for Minutes-to-Budget Linking. In the formal run, there were 51 submissions from five teams for Question Answering-2, 13 submissions from two teams for Answer Verification, 46 submissions from four teams for Stance Classification-2, and 16 submissions from two teams for Minutes-to-Budget Linking. In total, there were 126 submissions from eight teams.

7 RESULTS

Tables 7, 9, 10, and 11 show the automatic evaluation results of Question Answering2, Answer Verification, Stance Classification-2, and Minutes-to-Budget Linking in the formal run, respectively.

Table 8 shows the human evaluation results of Question Answering-2. See Appendix for the results of Dry Run and the Late Submissions.

8 OVERVIEW OF PARTICIPANT SYSTEMS

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

The AKBL team participated in all four subtasks. For the Question Answering-2 subtask, a given question and its relevant answer segment extracted from the minutes are fed to their summarization model, for which they employ T5, a pre-trained language

Table 7: Scores of Question Answering-2 subtask in formal run (ROUGE scores)

ID	Team	ROUGE (Recall)			ROUGE (F-measure)		
		N1	N2	R	N1	N2	R
		Surface form					
153	ditlab	**0.5742**	**0.3089**	**0.5021**	**0.4791**	**0.2579**	**0.4194**
174	ditlab	0.5722	0.3059	0.5011	0.4681	0.2503	0.4102
175	ditlab	0.5420	0.2839	0.4759	0.4683	0.2456	0.4116
176	ditlab	0.5513	0.2891	0.4832	0.4625	0.2424	0.4054
152	ditlab	0.5628	0.2884	0.4878	0.4694	0.2442	0.4082
134	omuokdlb	0.4948	0.2591	0.4384	0.4668	0.2447	0.4131
133	omuokdlb	0.4948	0.2591	0.4384	0.4668	0.2447	0.4131
130	ditlab	0.5649	0.2966	0.4914	0.4713	0.2473	0.4101
131	ditlab	0.5739	0.3030	0.4991	0.4697	0.2471	0.4080
123	ditlab	0.5545	0.2894	0.4826	0.4728	0.2474	0.4120
183	ditlab	0.5597	0.2917	0.4892	0.4556	0.2368	0.3984
122	ditlab	0.5365	0.2792	0.4659	0.4701	0.2455	0.4089
115	ditlab	0.5354	0.2704	0.4657	0.4645	0.2361	0.4043
146	ditlab	0.5478	0.2845	0.4768	0.4509	0.2333	0.3930
151	ditlab	0.5702	0.3040	0.4997	0.4616	0.2445	0.4039
204	HUKB	0.4509	0.2366	0.3996	0.4398	0.2310	0.3909
205	HUKB	0.4499	0.2362	0.3991	0.4396	0.2308	0.3908
201	HUKB	0.4499	0.2362	0.3991	0.4396	0.2308	0.3908
107	ditlab	0.5219	0.2633	0.4577	0.4511	0.2280	0.3952
108	ditlab	0.5412	0.2739	0.4716	0.4508	0.2280	0.3924
119	ditlab	0.5275	0.2681	0.4627	0.4561	0.2330	0.4006
154	omuokdlb	0.4968	0.2458	0.4334	0.4552	0.2269	0.3972
120	omuokdlb	0.4872	0.2463	0.4277	0.4515	0.2281	0.3967
158	HUKB	0.4413	0.2263	0.3914	0.4351	0.2238	0.3866
200	HUKB	0.4594	0.2391	0.4058	0.4392	0.2287	0.3888
177	omuokdlb	0.4740	0.2415	0.4151	0.4506	0.2297	0.3953
		Stem					
153	ditlab	0.5856	**0.3176**	**0.5108**	**0.4889**	**0.2656**	**0.4271**
174	ditlab	0.5851	0.3156	0.5092	0.4787	0.2582	0.4167
175	ditlab	0.5534	0.2926	0.4834	0.4782	0.2534	0.4182
176	ditlab	0.5627	0.2981	0.4906	0.4723	0.2501	0.4118
152	ditlab	0.5714	0.2964	0.4957	0.4762	0.2508	0.4144
134	omuokdlb	0.5011	0.2645	0.4437	0.4729	0.2502	0.4183
133	omuokdlb	0.5011	0.2645	0.4437	0.4729	0.2502	0.4183
130	ditlab	0.5761	0.3053	0.5000	0.4811	0.2548	0.4178
131	ditlab	**0.5863**	0.3125	0.5084	0.4802	0.2553	0.4162
123	ditlab	0.5652	0.2978	0.4905	0.4824	0.2549	0.4193
183	ditlab	0.5714	0.3007	0.4971	0.4653	0.2447	0.4052
122	ditlab	0.5458	0.2868	0.4730	0.4784	0.2524	0.4156
115	ditlab	0.5441	0.2771	0.4723	0.4719	0.2419	0.4098
146	ditlab	0.5578	0.2930	0.4847	0.4589	0.2402	0.3993
151	ditlab	0.5806	0.3127	0.5074	0.4701	0.2516	0.4101
204	HUKB	0.4607	0.2464	0.4082	0.4495	0.2407	0.3994
205	HUKB	0.4597	0.2460	0.4077	0.4493	0.2405	0.3994
201	HUKB	0.4597	0.2460	0.4077	0.4493	0.2405	0.3994
107	ditlab	0.5298	0.2697	0.4635	0.4579	0.2335	0.4002
108	ditlab	0.5497	0.2801	0.4782	0.4582	0.2331	0.3980
119	ditlab	0.5370	0.2739	0.4688	0.4643	0.2383	0.4062
154	omuokdlb	0.5043	0.2528	0.4403	0.4621	0.2339	0.4038
120	omuokdlb	0.4948	0.2515	0.4335	0.4584	0.2330	0.4020
158	HUKB	0.4511	0.2365	0.4005	0.4445	0.2343	0.3956
200	HUKB	0.4690	0.2480	0.4138	0.4484	0.2373	0.3965
177	omuokdlb	0.4840	0.2490	0.4229	0.4601	0.2370	0.4025
		Content word					
153	ditlab	0.3883	0.2108	0.3800	**0.3246**	**0.1765**	**0.3182**
174	ditlab	**0.4002**	**0.2138**	**0.3899**	**0.3246**	0.1733	0.3165
175	ditlab	0.3697	0.1971	0.3611	0.3197	0.1717	0.3125
176	ditlab	0.3751	0.1997	0.3653	0.3155	0.1678	0.3076
152	ditlab	0.3792	0.1924	0.3687	0.3147	0.1611	0.3063
134	omuokdlb	0.3340	0.1808	0.3300	0.3130	0.1698	0.3091
133	omuokdlb	0.3340	0.1808	0.3300	0.3130	0.1698	0.3091
130	ditlab	0.3753	0.2010	0.3670	0.3130	0.1680	0.3066
131	ditlab	0.3824	0.2050	0.3737	0.3125	0.1680	0.3061
123	ditlab	0.3643	0.1936	0.3560	0.3112	0.1668	0.3047
183	ditlab	0.3839	0.2011	0.3755	0.3110	0.1628	0.3042
122	ditlab	0.3488	0.1860	0.3398	0.3080	0.1659	0.3006
115	ditlab	0.3554	0.1837	0.3479	0.3068	0.1599	0.3002
146	ditlab	0.3726	0.1998	0.3635	0.3050	0.1636	0.2979
151	ditlab	0.3792	0.2075	0.3698	0.3039	0.1653	0.2965
204	HUKB	0.3107	0.1780	0.3051	0.3011	0.1742	0.2960
205	HUKB	0.3102	0.1786	0.3046	0.3008	0.1750	0.2956
201	HUKB	0.3102	0.1786	0.3046	0.3008	0.1750	0.2956
107	ditlab	0.3490	0.1820	0.3388	0.2996	0.1566	0.2909
108	ditlab	0.3618	0.1885	0.3511	0.2995	0.1556	0.2907
119	ditlab	0.3495	0.1808	0.3449	0.2995	0.1572	0.2955
154	omuokdlb	0.3261	0.1634	0.3195	0.2980	0.1516	0.2916
120	omuokdlb	0.3252	0.1719	0.3210	0.2975	0.1578	0.2935
158	HUKB	0.3037	0.1716	0.2981	0.2967	0.1696	0.2912
200	HUKB	0.3092	0.1747	0.3025	0.2949	0.1674	0.2887
177	omuokdlb	0.3108	0.1555	0.3034	0.2948	0.1517	0.2877

ID	Team	ROUGE (Recall)			ROUGE (F-measure)		
		N1	N2	R	N1	N2	R
		Surface form					
182	omuokdlb	0.4684	0.2436	0.4126	0.4493	0.2316	0.3953
79	ditlab	0.4894	0.2443	0.4283	0.4488	0.2231	0.3925
87	ditlab	0.4794	0.2439	0.4250	0.4416	0.2232	0.3908
106	ditlab	0.4686	0.2362	0.4120	0.4451	0.2226	0.3917
80	ditlab	0.4851	0.2460	0.4237	0.4457	0.2252	0.3898
140	omuokdlb	0.4878	0.2455	0.4263	0.4450	0.2223	0.3892
78	ditlab	0.4754	0.2401	0.4150	0.4419	0.2218	0.3849
90	ditlab	0.4753	0.2331	0.4153	0.4418	0.2190	0.3868
93	ditlab	0.4693	0.2315	0.4107	0.4354	0.2145	0.3812
95	ditlab	0.4662	0.2318	0.4114	0.4311	0.2133	0.3805
105	ditlab	0.4657	0.2294	0.4098	0.4315	0.2139	0.3798
101	TO	0.4711	0.2325	0.4132	0.4365	0.2136	0.3827
116	IKM23	0.4501	0.2212	0.3995	0.4326	0.2132	0.3839
139	omuokdlb	0.4578	0.2234	0.3980	0.4313	0.2100	0.3752
102	ditlab	0.4539	0.2244	0.3991	0.4267	0.2092	0.3746
159	omuokdlb	0.4628	0.2227	0.4042	0.4316	0.2072	0.3774
124	omuokdlb	0.4526	0.2157	0.3987	0.4265	0.2043	0.3761
132	omuokdlb	0.4452	0.2164	0.3912	0.4134	0.1982	0.3628
171	omuokdlb	0.4268	0.1935	0.3690	0.4207	0.1899	0.3636
94	ditlab	0.4234	0.1958	0.3703	0.4151	0.1920	0.3639
109	omuokdlb	0.4312	0.2020	0.3761	0.4048	0.1876	0.3535
84	IKM23	0.2522	0.0826	0.2226	0.2860	0.0927	0.2530
86	IKM23	0.2878	0.0973	0.2530	0.3024	0.1013	0.2669
85	IKM23	0.2812	0.0897	0.2449	0.2992	0.0927	0.2603
118	IKM23	0.3187	0.0990	0.2706	0.3128	0.0967	0.2653
192	AKBL	0.3314	0.1005	0.2840	0.3018	0.0928	0.2590
		Stem					
182	omuokdlb	0.4793	0.2512	0.4216	0.4593	0.2391	0.4036
79	ditlab	0.4961	0.2498	0.4341	0.4552	0.2283	0.3982
87	ditlab	0.4861	0.2477	0.4292	0.4479	0.2269	0.3950
106	ditlab	0.4757	0.2418	0.4184	0.4524	0.2281	0.3982
80	ditlab	0.4938	0.2524	0.4301	0.4534	0.2311	0.3956
140	omuokdlb	0.4950	0.2514	0.4322	0.4517	0.2277	0.3945
78	ditlab	0.4812	0.2446	0.4200	0.4472	0.2262	0.3896
90	ditlab	0.4846	0.2387	0.4218	0.4509	0.2244	0.3930
93	ditlab	0.4780	0.2371	0.4161	0.4432	0.2199	0.3863
95	ditlab	0.4726	0.2357	0.4168	0.4372	0.2172	0.3856
105	ditlab	0.4748	0.2347	0.4162	0.4402	0.2190	0.3860
101	TO	0.4784	0.2379	0.4194	0.4431	0.2189	0.3883
116	IKM23	0.4565	0.2272	0.4046	0.4383	0.2191	0.3886
139	omuokdlb	0.4655	0.2279	0.4038	0.4385	0.2143	0.3808
102	ditlab	0.4605	0.2288	0.4045	0.4330	0.2137	0.3798
159	omuokdlb	0.4690	0.2276	0.4094	0.4376	0.2121	0.3825
124	omuokdlb	0.4595	0.2213	0.4052	0.4329	0.2096	0.3820
132	omuokdlb	0.4553	0.2233	0.3985	0.4229	0.2048	0.3694
171	omuokdlb	0.4356	0.1984	0.3746	0.4293	0.1949	0.3691
94	ditlab	0.4294	0.2018	0.3756	0.4209	0.1983	0.3691
109	omuokdlb	0.4380	0.2073	0.3814	0.4112	0.1929	0.3588
84	IKM23	0.2558	0.0848	0.2254	0.2901	0.0951	0.2563
86	IKM23	0.2923	0.1004	0.2564	0.3069	0.1049	0.2701
85	IKM23	0.2861	0.0929	0.2477	0.3042	0.0964	0.2633
118	IKM23	0.3220	0.1009	0.2737	0.3164	0.0987	0.2688
192	AKBL	0.3386	0.1029	0.2878	0.3087	0.0948	0.2630
		Content word					
182	omuokdlb	0.3125	0.1620	0.3086	0.2941	0.1554	0.2904
79	ditlab	0.3223	0.1668	0.3143	0.2921	0.1515	0.2850
87	ditlab	0.3207	0.1701	0.3158	0.2878	0.1540	0.2832
106	ditlab	0.3040	0.1632	0.2968	0.2865	0.1537	0.2796
80	ditlab	0.3172	0.1602	0.3093	0.2855	0.1460	0.2788
140	omuokdlb	0.3192	0.1633	0.3136	0.2840	0.1442	0.2791
78	ditlab	0.3108	0.1634	0.3048	0.2826	0.1464	0.2771
90	ditlab	0.3039	0.1664	0.2994	0.2810	0.1543	0.2765
93	ditlab	0.3027	0.1577	0.2965	0.2781	0.1484	0.2723
95	ditlab	0.3037	0.1592	0.2976	0.2773	0.1448	0.2717
105	ditlab	0.3003	0.1609	0.2954	0.2760	0.1487	0.2715
101	TO	0.2994	0.1492	0.2933	0.2736	0.1385	0.2678
116	IKM23	0.2885	0.1446	0.2823	0.2724	0.1399	0.2669
139	omuokdlb	0.2934	0.1510	0.2866	0.2712	0.1420	0.2653
102	ditlab	0.2890	0.1452	0.2842	0.2678	0.1337	0.2632
159	omuokdlb	0.2880	0.1416	0.2838	0.2664	0.1326	0.2624
124	omuokdlb	0.2846	0.1360	0.2771	0.2638	0.1265	0.2567
132	omuokdlb	0.2848	0.1453	0.2798	0.2600	0.1332	0.2550
171	omuokdlb	0.2594	0.1159	0.2537	0.2549	0.1134	0.2491
94	ditlab	0.2594	0.1289	0.2545	0.2527	0.1281	0.2480
109	omuokdlb	0.2651	0.1312	0.2599	0.2453	0.1221	0.2405
84	IKM23	0.1214	0.0553	0.1188	0.1353	0.0624	0.1328
86	IKM23	0.1274	0.0624	0.1251	0.1332	0.0647	0.1314
85	IKM23	0.1208	0.0500	0.1172	0.1252	0.0493	0.1214
118	IKM23	0.1253	0.0484	0.1229	0.1208	0.0476	0.1185
192	AKBL	0.1239	0.0526	0.1214	0.1162	0.0502	0.1137

Table 8: Scores of Question Answering-2 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	93	6	1	192	47	47	6	141	96	3	1	195	69	28	3	166
153	ditlab	94	5	1	193	46	48	6	140	92	8	0	192	67	22	11	**156**
174	ditlab	94	6	0	**194**	47	48	5	**142**	84	12	4	180	65	25	10	155
116	IKM23	93	6	1	192	34	46	20	114	94	5	1	**193**	54	26	20	134
204	HUKB	85	10	5	180	23	61	16	107	87	9	4	183	46	39	15	131
134	omuokdlb	84	12	4	180	32	58	10	122	86	10	4	182	49	32	19	130
101	TO	86	8	6	180	35	49	16	119	89	6	5	184	48	29	23	125
192	AKBL	41	18	41	100	10	17	73	37	25	27	48	77	8	11	81	27

Table 9: Scores of Answer Verification subtask in formal run

ID	Team	Accuracy†	F-measure
70	AKBL	**0.88**	**0.8966**
77	AKBL	0.86	0.8814
69	AKBL	0.86	0.8772
138	AKBL	0.84	0.8621
68	AKBL	0.79	0.8073
161	AKBL	0.74	0.7451
129	omuokdlb	0.69	0.6804
191	omuokdlb	0.68	0.7922
160	omuokdlb	0.68	0.7922
194	omuokdlb	0.64	0.7778
169	omuokdlb	0.59	0.5176
128	omuokdlb	0.59	0.5176
104	TO	0.58	0.5227
103	TO	0.58	0.5227
168	omuokdlb	0.56	0.4762

† Since the number of the evaluated targets is 100, the accuracy has two significant digits.

model trained on Japanese text. For the Answer Verification subtask, their method first generates the pseudo-fake data automatically by round-trip translation. Then, it fine-tunes the pre-trained BERT with the training and pseudo-fake data. Their pseudo-fake data generation is based on three basic text operations: "Insertion and deletion of negation," "Conversion to antonyms," and "Subject-Object Exchange." For the Stance Classification-2 subtask, their best system combines the utterance and target as input and then classifies it using the binary classifier based on RoBERTa. For the Minutes-to-Budget Linking (MBLink) subtask, they employ Okapi BM25 to calculate the similarity between a given statement in the meeting minutes and the text extracted from each table in the budget table data.

The ditlab team participated in the Question Answering-2 subtask. First, they modified a QA Alignment system developed for the PoliInfo-3 QA Alignment subtask to compose paragraphs of related answer sentences. BM25 vectors were constructed for each paragraph of all answers, and the target answers were selected by the question summaries and subtopics based on the cosine similarity. Second, a T5 was used to summarize the associated answer. For creating fine-tuning data of T5, all data were used for ID153 and the data selection based on the Rouge scores was used for ID174.

Table 10: Scores of Stance Classification-2 subtask in formal run

ID	Team	Accuracy	ID	Team	Accuracy
198	KIS	**0.9728**	144	KIS	0.9536
181	KIS	0.9705	164	KIS	0.9469
197	KIS	0.9696	97	IKM23	0.9464
110	KIS	0.9688	82	IKM23	0.9455
111	KIS	0.9679	89	IKM23	0.9420
170	IKM23	0.9656	143	KIS	0.9415
186	KIS	0.9652	91	IKM23	0.9411
185	KIS	0.9652	142	KIS	0.9371
208	IKM23	0.9643	117	ISLab	0.9326
179	IKM23	0.9634	211	AKBL	0.9308
195	KIS	0.9629	162	KIS	0.9304
178	IKM23	0.9621	81	IKM23	0.9263
167	IKM23	0.9621	163	KIS	0.9254
199	KIS	0.9621	88	IKM23	0.9205
184	KIS	0.9621	75	IKM23	0.9161
196	KIS	0.9612	76	IKM23	0.9103
187	KIS	0.9612	98	ISLab	0.8946
145	KIS	0.9607	193	ISLab	0.8786
173	IKM23	0.9603	180	ISLab	0.8719
113	IKM23	0.9571	73	ISLab	0.8598
126	IKM23	0.9563	156	ISLab	0.8567
165	KIS	0.9563	83	IKM23	0.5433
114	IKM23	0.9536	74	IKM23	0.0844

Table 11: Scores of MBLink subtask in formal run

ID	Team	F-measure	ID	Team	F-measure
188	fuys	**0.3371**	150	AKBL	0.2289
112	fuys	0.2860	125	fuys	0.2157
137	AKBL	0.2577	121	fuys	0.2136
148	AKBL	0.2555	166	fuys	0.1877
147	AKBL	0.2554	172	fuys	0.0822
189	AKBL	0.2534	149	AKBL	0.0253
127	fuys	0.2419	92	fuys	0.0143
141	fuys	0.2351	67	TO	0.0031
190	AKBL	0.2348			

The fuys team participated in the MBLink subtask. They viewed this task as a binary classification problem that takes the text of sentences and tables as input and outputs, whether related or not, and used a fine-tuned BERT-based classification model.

The HUKB team proposed a system for the Question Answering-2 subtask. Their proposed system is divided into three steps. First, they found the sentence at the beginning of the same topic as the input question from the respondent's utterances and extracted the candidate sentences. Next, they found the sentences where the respondent seemed to answer the input question directly, using BERT. Finally, they entered the selected sentences with the input question into the T5-based summarizer and generated the answer summary.

The IKM23 team withdrew from this task, so the details of their system were not submitted.

The ISLab participated in the Stance Classification-2 subtask. They proposed two frameworks for determining the stance in utterances. The first framework involves concatenating the BERT model with the Bi-LSTM model to form a comprehensive decision-making model, while the second concatenates the Curie model with the ChatGPT model.

The KIS team participated in the Stance Classification-2 subtask. They additionally pretrained the Japanese pretrained LUKE model with a Masked Language Model on the Diet proceedings dataset in order to adapt it to the political domain. They also preprocessed the model using the head-tail method to truncate utterances longer than the maximum input length.

The omuokdlb team participated in the Question Answering-2 and Answer Verification subtasks. In Question Answering-2, they used BERT to match the question summary and the answer utterances. They then generated a summary of the answer to the question using a T5. In Answer Verification, they created binary classifiers using BERT to determine whether or not answers.

The Forst team tackled the Answer Verification subtask. They submitted their data as late submissions. The method they submitted is to input "QuestionSummary," "AnswerOriginal," and "AnswerSummary" items together and have ChatGPT classify them.

9 CONCLUSION

We presented an overview of the NTCIR-17 QA Lab-PoliInfo-4 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 Question Answering-2, Answer Verification, Stance Classification-2, and Minutes-to-Budget Linking subtasks. There were 183 submissions from 8 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] Hou Pong Chan, Qi Zeng, and Heng Ji. 2023. Interpretable Automatic Fine-grained Inconsistency Detection in Text Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 6433–6444. https://doi.org/10.18653/v1/2023.findings-acl.402

[5] 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*.

[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] Mingqi Gao, Xiaojun Wan, Jia Su, Zhefeng Wang, and Baoxing Huai. 2023. Reference Matters: Benchmarking Factual Error Correction for Dialogue Summarization with Fine-grained Evaluation Framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 13932–13959. https://doi.org/10.18653/v1/2023.acl-long.779

[9] Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance Detection in COVID-19 Tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 1596–1611. https://doi.org/10.18653/v1/2021.acl-long.127

[10] Jue Hou, Ilmari Kylliäinen, Anisia Katinskaia, Giacomo Furlan, and Roman Yangarber. 2022. Applying Gamification Incentives in the Revita Language-learning System. In *Proceedings of the 9th Workshop on Games and Natural Language Processing within the 13th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 7–16. https://aclanthology.org/2022.games-1.2

[11] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Teruko Mitamura, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Tatsunori Mori, Kenji Araki, Satoshi Sekine, and Noriko Kando. 2020. Overview of the NTCIR-15 QA Lab-PoliInfo Task. *Proceedings of The 15th NTCIR Conference* (12 2020).

[12] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Ken ichi Yokote, Kazuma Kadowaki, Tatsunori Mori, Kenji Araki, Teruko Mitamura, and Satoshi Sekine. 2022. Overview of the NTCIR-16 QA Lab-PoliInfo-3 Task. *Proceedings of The 16th NTCIR Conference*.

[13] 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

[14] Dilek Küçük and Fazli Can. 2020. Stance Detection: A Survey. *ACM Comput. Surv.* 53, 1, Article 12 (feb 2020), 37 pages. https://doi.org/10.1145/3369026

[15] Taku Kudo and John Richardson. 2018. Sentencepiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Brussels, Belgium, 66–71. https://doi.org/10.18653/v1/D18-2012

[16] 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

[17] 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

[18] 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

[19] Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023. On Improving Summarization Factual Consistency from Natural Language Feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 15144–15161. https://doi.org/10.18653/v1/2023.acl-long.844

[20] 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

[21] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 31–41.

[22] 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

[23] Yasuhiro Ogawa, Michiaki Satou, Takahiro Komamizu, and Katsuhiko Toyama. 2022. nagoy Team's Summarization System at the NTCIR-14 QA Lab-PoliInfo. (6 2022).

[24] 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

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

[26] 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

[27] 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

[28] 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.

[29] 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

[30] 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

[31] Amir Soleimani, Christof Monz, and Marcel Worring. 2023. NonFactS: Non-Factual Summary Generation for Factuality Evaluation in Document Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 6405–6419. https://doi.org/10.18653/v1/2023.findings-acl.400

[32] Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the Factual Consistency of Large Language Models Through News Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 5220–5255. https://doi.org/10.18653/v1/2023.findings-acl.322

[33] Mariona Taulé, M Antonia Martí, Francisco M Rangel, Paolo Rosso, Cristina Bosco, Viviana Patti, et al. 2017. Overview of the task on stance and gender detection in tweets on Catalan independence at IberEval 2017. In *CEUR Workshop Proceedings*, Vol. 1881. CEUR-WS, 157–177.

[34] Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *arXiv preprint cs/0607062* (2006).

[35] 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

[36] 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

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

[38] Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. 2022. Analyzing Dynamic Adversarial Training Data in the Limit. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 202–217. https://doi.org/10.18653/v1/2022.findings-acl.18

[39] Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In *Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24*. Springer, 907–916.

[40] 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

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

[42] Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. 2023. Annotating and Detecting Fine-grained Factual Errors for Dialogue Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 6825–6845. https://doi.org/10.18653/v1/2023.acl-long.377

A DATA FIELDS AND EXAMPLES

A.1 Question Answering

A.1.1 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.1.2 Examples.

Listing 1: 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 2: 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.2 Answer Verification

A.2.1 Data fields. The Answer Verification dataset consists of two types of files. The answer 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
- **`AnswerOriginal`** Text corresponding to the answer in the minutes

`PredictedClass` Fact (true) or Fake (false) (empty in the test file)

The minutes data is the same as that of the Question Answering dataset.

A.2.2 Examples.

Listing 3: Answer sheet for the Answer Verification subtask

```
[
  { "ID": "PoliInfo3-QA-v20211120-338-04-8-003",
    "Meeting": "令和 2年第 2回定例会",
    "Date": "2020-06-03",
    "Headlines": [ "葛西臨海公園の魅力向上を", "ドクターヘリ導入を加速化せよ" ],
    "SubTopic": "ドクターヘリ",
    "QuestionSpeaker": "上野和彦(公明党) ",
    "QuestionSummary": "3年度からの運航を目指し取り組むべき。本格導入に向けた決意は。",
    "AnswerSpeaker": "知事",
    "AnswerSummary": "小型ドクターヘリとの併用により機能強化が進むよう、令和 3年度導入に向け着実に取り組む。",
    "AnswerOriginal": "次に、ドクターヘリについてのご質問であります。お話の小型ヘリを活用したドクターヘリですが、短時間での離陸など機動力が高く、救急医療の効率的な提供に寄与しております。都といたしまして、現在、東京消防庁と連携をして、遠距離運航や夜間飛行が可能な東京型ドクターヘリを多摩や島しょ地域において運用しております。そして、小型ドクターヘリと併用することによって、都の救急医療体制の機能強化が進みますように、令和三年度の導入に向けまして着実に取り組んでまいります。",
    "PredictedClass": true },
  { "ID": "PoliInfo3-QA-v20210613-332-05-2-001",
    "Meeting": "平成 31年第 1回定例会、第 1回臨時会",
    "Date": "2019-02-28",
    "Headlines": [ "違う一人一人が等しく尊重され", "一人一人の笑顔が輝く社会を" ],
    "SubTopic": "インクルーシブな公園整備",
    "QuestionSpeaker": "龍円あいり(都ファースト) ",
    "QuestionSummary": "取組は。",
    "AnswerSpeaker": "知事",
    "AnswerSummary": "障害者団体等にヒアリング。砧公園と府中の森公園を対象に 31年度末完成を目指す。",
    "AnswerOriginal": "まず、インクルーシブな公園整備についてのお尋ねでございます。誰もが自分らしく輝くことのできるダイバーシティーの実現に向けまして、都立公園において、障害の有無にかかわらず全ての子供たちが安全に楽しむことができる遊び場、これを整備することは重要でございます。都といたしまして、今年度、障害児の保護者、そして障害者団体、障害児保育の現場、ユニバーサルデザインの有識者など、さまざまな方々にヒアリングを行ってまいりました。その中で、体を支える力が弱い子供さんたちが揺れる感覚を楽しめるそんな遊具や直射日光を避けることのできる休憩場所の設置など、さまざまなご意見をいただいたところでございます。こうしたご意見を踏まえまして、現在、砧公園と府中の森公園を対象に、具体的な設計を行っておりまして、平成三十一年度末の完成を目指し整備を進めてまいります。今後とも、都立公園でこうした取り組みを進めていくことで、障害の有無にかかわらず、全ての子供たちがともに遊び、また、学ぶ機会を積極的に提供してまいります。",
    "PredictedClass": true },
  { "ID": "PoliInfo3-QA-v20211120-338-04-9-002",
    "Meeting": "令和 2年第 2回定例会",
    "Date": "2020-06-03",
    "Headlines": [ "新型コロナの患者数等の予測は", "東京アラート発動基準の根拠は" ],
    "SubTopic": "新型コロナ",
    "QuestionSpeaker": "川松真一朗(自民党) ",
    "QuestionSummary": "ロードマップの 3つの指標の根拠は。",
    "AnswerSpeaker": "福祉保健局長",
    "AnswerSummary": "感染状況、医療提供体制、モニタリングに関する指標のうち 3つを目安に設定。不明率と陽性者増加比は参考値とする。",
    "AnswerOriginal": "次に、モニタリング指標の考え方についてでございますが、都では、緊急事態措置に基づく自粛要請の緩和及び再要請を検討する際に、判断の目安として、感染状況、医療提供体制、モニタリングに関する七つの指標を定め、そのうち三つについて、感染拡大時の状況や国の対処方針の考え方も参考に目安となる数値を設定いたしました。新規陽性者数は感染拡大の兆候を把握するもので、第一波の感染拡大局面とした時期の水準を踏まえ、緩和の目安を一日当たり二十人未満と設定いたしました。接触歴等不明率は、市中感染の拡大状況を把握するものでございますが、新規陽性者のうち、接触歴不明者が一日当たり十人未満となるよう、五〇%未満を目安といたしました。陽性者増加比でございますが、一未満であれば新規感染者数は減少、それを超えれば増加傾向を示すため、緩和の目安を一未満としてございます。これらの指標の運用につきましては、国の動向や感染者の状況等に応じて柔軟に実施するほか、新規陽性者数の数値が十人以下となった場合には、接触歴等不明率と陽性者増加比は参考値とすることとしております。また、こうした感染状況の指標につきまして、一定期間の動向を見ながら、医療提供体制などその他の指標も勘案した上で、東京アラートの発動を判断することとしております。",
    "PredictedClass": false },
...
]
```

A.3 Stance Classification-2

A.3.1 Data fields. The Stance Classification-2 dataset consists of a CSV file. Its data fields are as follows.

- **id** Question ID (Japanese local government ID and serial number)
- **prefecture** Name of the prefecture
- **assembly** Name of the local government
- **meeting** Name and serial number of the regular meeting
- **date** Date of the meeting
- **speaker** Speaker name of the utterance
- **utterance** An utterance by a politician whose explicit tokens are replaced with `[STANCE]`
- **target** Topic of the utterance
- **stance** ‘Agreement’ or ‘Disagreement’

A.4 MBLink

A.4.1 Data fields. The MBLink dataset consists of two types of files.

The meeting minutes file includes the following items.

- **`data-mblink-sentence-id`** Sentence linked to the budget table

The budget table file contains the following items.

- **`data-mblink-table-ids`** A table ID is assigned to the <table> tag of each table

A.4.2 Examples.

Listing 4: Minutes for the MBLink subtask

```
＜html ＞
   ...
  <p class="annotate" data-mblink-sentence-id="otaru_h29
-01-sent21" data-mblink-table-ids="otaru_h29-tab16
otaru_h29-tab207">
   それでは、平成 29年度の予算編成についてですが、収入状況は、市税の伸びが期待できないことに加え、地方譲与税や交付金、さらには実質的な地方交付税の減少が見込まれ、引き続き大変厳しい状況にあります。
  </p>
     ...
</html>
```

Listing 5: Budget tables for the MBLink subtask

```
 <html>
 ...
  <table border="1" data-mblink-table-id="otaru_h29-tab0
">
   <tr>
    <td colspan="2" data-mblink-cell-id="otaru_h29-tab0-
r0c0" rowspan="2" style="vertical-align:middle;">
     <p>
      <span class="font27">
       会区分
      </span>
     </p>
     <p>
      <span class="font27">
       計
      </span>
     </p>
     <p>
      <span class="font27">
       別議決年月日
      </span>
     </p>
    </td>
    <td data-mblink-cell-id="otaru_h29-tab0-r0c1" style="
vertical-align:middle;">
     <p>
      <span class="font27">
       当初予算額
      </span>
     </p>
    </td>
...
</html>
```

Listing 6: Answer sheet for the MBLink subtask

```
[
    {
        "sentenceID": "otaru_h28-01-sent16",
        "tableIds": [
            "otaru_h28-tab663",
            "otaru_h28-tab128",
            "otaru_h28-tab802",
            "otaru_h28-tab304",
            "otaru_h28-tab273",
            "otaru_h28-tab4"
        ]
    },
    {
        "sentenceID": "otaru_h28-01-sent21",
        "tableIds": [
            "otaru_h28-tab380",
            "otaru_h28-tab344",
            "otaru_h28-tab498",
            "otaru_h28-tab320"
        ]
    },
    ...
]
```

B RESULTS OF DRY RUN

Tables 12, 13, 14, and 15 show the automatic evaluation results of Question Answering, QA Alignment, Fact Verification, and Budget Argument Mining subtasks in the dry run, respectively.

Table 12: Scores of Question Answering-2 subtask in dry run (ROUGE scores)

ID	Team	ROUGE (Recall)			ROUGE (F-measure)		
		N1	N2	R	N1	N2	R
Surface form							
7	TO	**0.4605**	**0.2404**	**0.4077**	**0.4369**	**0.2255**	**0.3864**
44	IKM23	0.3430	0.1455	0.2972	<u>0.3670</u>	<u>0.1596</u>	<u>0.3200</u>
56	IKM23	0.3410	0.1342	0.2949	0.3595	0.1429	0.3125
31	IKM23	<u>0.4328</u>	<u>0.1525</u>	<u>0.3446</u>	0.3590	0.1220	0.2840
30	IKM23	0.3896	0.1372	0.3132	0.3551	0.1202	0.2849
57	IKM23	0.2990	0.1111	0.2648	0.3284	0.1232	0.2922
29	IKM23	0.3380	0.0984	0.2718	0.3236	0.0914	0.2605
58	AKBL	0.0903	0.0263	0.0794	0.1037	0.0288	0.0912
Stem							
7	TO	**0.4673**	**0.2455**	**0.4131**	**0.4435**	**0.2303**	**0.3919**
44	IKM23	0.3496	0.1498	0.3021	<u>0.3738</u>	<u>0.1639</u>	<u>0.3253</u>
56	IKM23	0.3469	0.1373	0.2991	0.3658	0.1460	0.3170
31	IKM23	<u>0.4433</u>	<u>0.1592</u>	<u>0.3528</u>	0.3685	0.1277	0.2912
30	IKM23	0.3987	0.1430	0.3189	0.3643	0.1253	0.2909
57	IKM23	0.3033	0.1144	0.2685	0.3331	0.1266	0.2962
29	IKM23	0.3475	0.1032	0.2785	0.3336	0.0958	0.2675
58	AKBL	0.0919	0.0269	0.0807	0.1056	0.0296	0.0927
Content word							
7	TO	**0.2993**	**0.1614**	**0.2937**	**0.2811**	**0.1527**	**0.2758**
44	IKM23	0.1911	<u>0.0933</u>	0.1858	<u>0.2044</u>	<u>0.1010</u>	<u>0.1989</u>
56	IKM23	0.1807	0.0847	0.1749	0.1893	0.0891	0.1837
31	IKM23	<u>0.2189</u>	0.0816	<u>0.1969</u>	0.1843	0.0648	0.1659
30	IKM23	0.1903	0.0752	0.1783	0.1733	0.0659	0.1622
57	IKM23	0.1508	0.0736	0.1487	0.1652	0.0815	0.1632
29	IKM23	0.1496	0.0509	0.1395	0.1454	0.0475	0.1352
58	AKBL	0.0364	0.0180	0.0359	0.0407	0.0191	0.0401

Table 13: Scores of Answer Verification task in dry run

ID	Team	Accuracy	F-measure
65	AKBL	**0.8608**	**0.8608**
64	AKBL	**0.8608**	**0.8608**
61	AKBL	**0.8608**	0.8533
60	AKBL	0.8481	0.8378
15	AKBL	0.8354	0.8116
23	AKBL	0.8228	0.8250
62	AKBL	0.8101	0.7887
34	AKBL	0.8101	0.7826
22	AKBL	0.8101	0.7761
16	AKBL	0.8101	0.8101
17	AKBL	0.7975	0.7647
37	AKBL	0.7848	0.7606
33	AKBL	0.7848	0.7463
32	AKBL	0.7595	0.7077
26	AKBL	0.7468	0.6667
24	AKBL	0.7468	0.7436
35	AKBL	0.7215	0.6452
41	AKBL	0.7089	0.6102
27	AKBL	0.7089	0.5965
3	TO	0.7089	0.6230
25	AKBL	0.6456	0.4615
28	AKBL	0.5823	0.3265
21	AKBL	0.5696	0.2609

Table 14: Scores of Stance Classification-2 subtask in dry run

ID	Team	Accuracy	ID	Team	Accuracy
45	KIS	**0.9624**	40	ISLab	0.9200
14	KIS	**0.9624**	59	ISLab	0.9106
46	KIS	0.9600	18	AKBL	0.9082
12	KIS	0.9506	9	KIS	0.9059
42	KIS	0.9482	53	ISLab	0.8988
13	KIS	0.9459	20	IKM23	0.8894
11	KIS	0.9294	47	KIS	0.8565
19	AKBL	0.9271	50	IKM23	0.8141
54	ISLab	0.9224	52	ISLab	0.5176
10	KIS	0.9224	5	TO	0.5082
63	ISLab	0.9200			

Table 15: Scores of MBLink subtask in dry run

ID	Team	F-measure
55	fuys	**0.0113**

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. These were treated as late submissions. Tables 16, 17, 18, and 19 show the automatic evaluation results of the late submissions of Question Answering-2, Answer Verification, Stance Classification-2, and Minutes-to-Budget Linking subtasks, respectively.

Table 16: Scores of late submissions in Question Answering-2 subtask (ROUGE scores)

ID	Team	ROUGE (Recall)			ROUGE (F-measure)		
		N1	N2	R	N1	N2	R
Surface form							
242	AKBL	**0.4665**	**0.2336**	**0.4106**	<u>0.4253</u>	**0.2133**	<u>0.3753</u>
240	omuokdlb	<u>0.4526</u>	<u>0.2157</u>	<u>0.3987</u>	**0.4265**	<u>0.2043</u>	**0.3761**
231	AKBL	0.3351	0.1068	0.2884	0.3047	0.0984	0.2632
Stem							
242	AKBL	**0.4709**	**0.2384**	**0.4144**	<u>0.4293</u>	**0.2176**	<u>0.3787</u>
240	omuokdlb	<u>0.4595</u>	<u>0.2213</u>	<u>0.4052</u>	**0.4329**	<u>0.2096</u>	**0.3820**
231	AKBL	0.3410	0.1105	0.2927	0.3100	0.1014	0.2667
Content word							
242	AKBL	**0.3026**	**0.1619**	**0.2979**	**0.2738**	**0.1489**	**0.2700**
240	omuokdlb	<u>0.2846</u>	<u>0.1360</u>	<u>0.2771</u>	<u>0.2638</u>	<u>0.1265</u>	<u>0.2567</u>
231	AKBL	0.1432	0.0658	0.1396	0.1309	0.0593	0.1277

Table 17: Scores of late submissions in Answer Verification task

ID	Team	Accuracy	F-measure
233	AKBL	**0.85**	**0.878**
232	AKBL	**0.85**	**0.878**
234	AKBL	0.75	0.7619
223	Forst	0.58	0.5116

Table 18: Scores of late submissions in Stance Classification-2 subtask

ID	Team	Accuracy	ID	Team	Accuracy
227	KIS	**0.9741**	225	KIS	0.9589
221	KIS	<u>0.9728</u>	224	ISLab	0.9326
228	KIS	0.9701	230	ISLab	0.9165
222	KIS	0.9701	216	ISLab	0.8786
220	KIS	0.9616	226	ISLab	0.8674
219	KIS	0.9616	229	ISLab	0.8638

Table 19: Scores of late submissions in MBLink subtask

ID	Team	F-measure
241	fuys	**0.3666**
239	fuys	<u>0.3177</u>
237	fuys	0.3177
236	OUC	0.2024
235	OUC	0.2024
238	fuys	0.1828