Overview of the NTCIR-14 QA Lab-PoliInfo Task

Yasutomo Kimura[1,2], Hideyuki Shibuki[3], Hokuto Ototake[4], Yuzu Uchida[5], Keiichi Takamaru[6], Kotaro Sakamoto[3,7], Madoka Ishioroshi[7], Teruko Mitamura[8], Noriko Kando[7,9], Tatsunori Mori[3], Harumichi Yuasa[10], Satoshi Sekine[2], and Kentaro Inui[11,2]

[1] Otaru University of Commerce, Japan
[2] RIKEN, Japan
[3] Yokohama National University, Japan
[4] Fukuoka University, Japan
[5] Hokkai-Gakuen University, Japan
[6] Utsunomiya Kyowa University, Japan
[7] National Institute of Informatics, Japan
[8] Carnegie Mellon University, USA
[9] SOKENDAI, Japan
[10] Institute of Information Security, Japan
[11] Tohoku University, Japan

Abstract. The NTCIR-14 QA Lab-PoliInfo aims at real-world complex Question Answering (QA) technologies using Japanese political information such as local assembly minutes and newsletters. QA Lab-PoliInfo has three tasks, namely Segmentation, Summarization and Classification task. We describe the used data, formal run results, and comparison between human marks and automatic evaluation scores.

Keywords: NTCIR-14 · QA Lab · PoliInfo · question answering · political information · local assembly minutes · segmentation · summarization · classification.

1 Introduction

The QA Lab-PoliInfo (Question Answering Lab for Political Information) task at NTCIR 14 aims at complex real-world question answering (QA) technologies, to show summaries of the opinions of assembly members and the reasons and conditions for such opinions, from Japanese regional assembly minutes.

We reaffirm the importance of fact-checking because of the negative impact of fake news in the recent years. The International Fact-Checking Network of the Poynter Institute established that April 2 would be considered as International Fact-Checking Day from 2017. In addition, fact-checking is difficult for general Web search engines to deal with because of the filter bubble developed by Eli Pariser[1], which keeps users away from information that disagrees with their viewpoints. For fact-checking, we should confirm primary sources such as assembly minutes. The description of the Japanese assembly minutes is a transcript

	Fake News Challenge Stage 1	CLEF-2018 Fact Checking Lab	NTCIR QA Lab-PoliInfo
Dataset	News articles	Political debate	Assembly minutes and newsletter
Task	Stance Detection Classifying the stance of the body using both a headline and a body text. Output is as follows: 1. Agree 2. Disagree 3. Discussed 4. Unrelated	Task1: Check-worthiness Prediction which claim in a political debate should be prioritized for fact-checking. Task2: Factuality Checking the factuality of the identified worth-checking claims.	Task1: Segmentation Extracting of the range of primary information Task2: Summarization Summarizing of local assembly member's and governor's utterance Task3: Classification Classifying an utterance which includes fact-checkable statement and opinion for a political topic.
Number of training data	2,586 articles	1,400 sentences x 3 files	Segmentation : 298 set Summarization : 596 set Classification : 14 topic (includes 10,291 sentences)
Language	English	English and Arabic	Japanese

Fig. 1. Comparison with related shared tasks

of a speech, which is very long; therefore, understanding the contents, including the opinions of the members at a glance is difficult. New information access technologies to support user understanding are expected, which would protect us from fake news.

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

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

1) comprehensible summary of a topic;
2) beliefs and attitudes of assembly members;
3) mental spaces for other assembly members;
4) contexts, including reasons;
5) several topics in a speech; and
6) colloquial Japanese including dialect and slang.

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

2 Related Work

Fake News Challenge[12] and CLEF-2018 Fact Checking Lab[13] are shared tasks dealing with political information. Fake News Challenge conducted the Stance Detection task estimating the relative perspective (or stance) of two pieces of text relative to a topic, claim or issue. CLEF-2018 Fact Checking Lab conducted the Check-worthiness and Factuality tasks. Figure 1 shows a comparison with the related shared tasks.

[12] http://www.fakenewschallenge.org/
[13] http://alt.qcri.org/clef2018-factcheck/

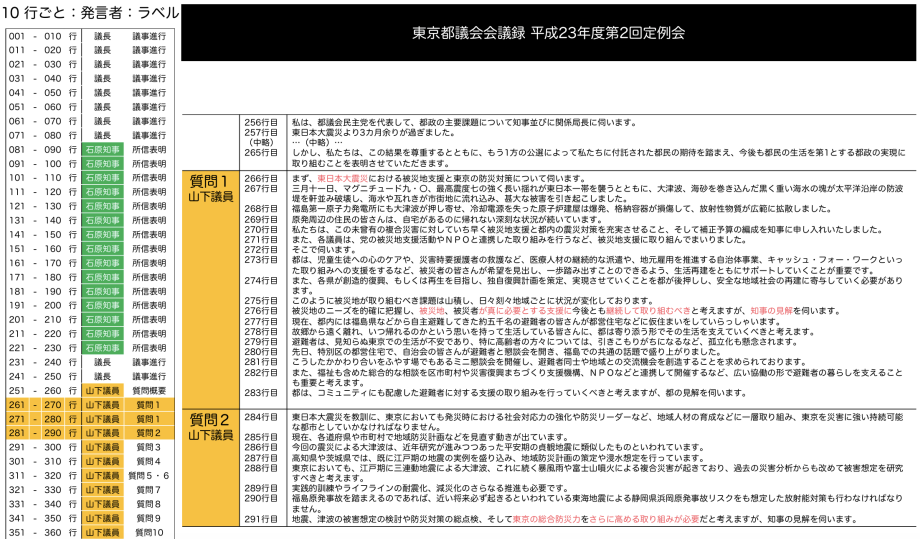

Fig. 2. Example of the plenary minutes of the Tokyo Metropolitan Assembly

3 Japanese Regional Assembly Minutes Corpus

Kimura et al.[4] constructed the Japanese Regional Assembly Minutes Corpus that collects minutes of plenary assemblies in 47 prefectures of Japan from April 2011 to March 2015. Figure 2 shows an example of the minutes of the Tokyo Metropolitan Assembly. Japanese minutes resemble a transcript. In the question-and-answer session, a member of assembly asks several questions at a time, and a prefectural governor or a superintendent answers the questions under his/her charge at a time. A speech is too long to understand the contents at a glance; therefore, information access technologies such as QA and automated summarization, will aid in understanding. For the QA Lab-PoliInfo task, we distributed a subset of the corpus, which is narrowed down to the Tokyo Metropolitan Assembly.

4 Task Description

We designed the segmentation, summarization and classification tasks. We put the tasks at the elemental technologies of information credibility or fact-checking for political information systems. Figure 3 shows a relation of the tasks. The segmentation task aims to find primary information corresponding to the given secondary information. The summarization task aims to generate brief texts considering argument structure such as questions and answers. The classification task aims to find pros and cons of a political topic and present their fact-checkable reasons. We preliminarily conducted the tasks at dry run. We discussed the

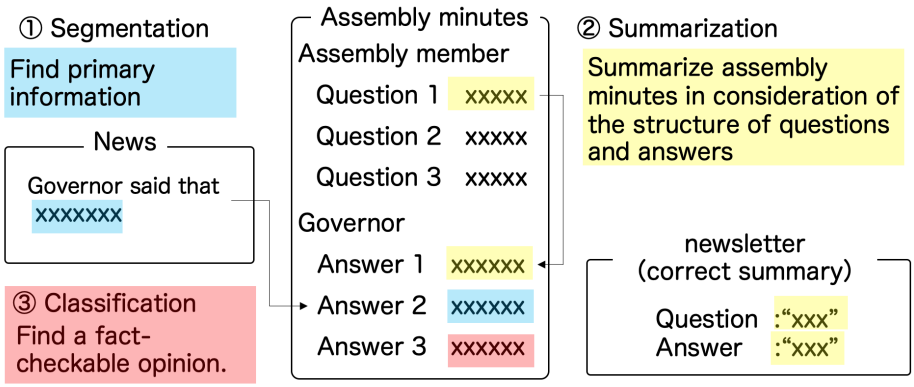

Fig. 3. Relation of the three tasks

results with participants via two round table meetings, and refined the tasks for Formal run. Only the Japanese task was conducted because we could not prepare minutes in other languages.

4.1 Segmentation Task

For the Segmentation task, the minutes of the Tokyo Metropolitan Assembly from April 2011 to March 2015 and a summary of a speech of a member of assembly described in *Togikai dayori*[14], a public relations paper of the Tokyo Metropolitan Assembly are given. The participants find the corresponding original speech from the minutes and answer positions of the first and last sentences of the found speech. As an evaluation measure, we used recall R_{seg}, precision P_{seg} and F-measure F_{seg} of concordance of the first and last sentences to the gold standard data. They were calculated using the following expressions:

$$R_{seg} = \frac{N_{cp}}{N_{gsp}} \tag{1}$$

$$P_{seg} = \frac{N_{cp}}{N_{sp}} \tag{2}$$

$$F_{seg} = \frac{2R_{seg}P_{seg}}{R_{seg} + P_{seg}} \tag{3}$$

where N_{cp} is the number of the first and last sentences of which the position is in concord with the gold standard position, N_{gsp} is the number of the gold standard positions; and N_{sp} is the number of sentence positions the participants submitted.

– *Dry run*

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

Table 1. Data fields used in the Segmentation task

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

Input: the minutes and a summary of a speech of member of assembly
Output: the first and last sentences of the original speech corresponding to the summary
Evaluation: recall, precision, and F-measure of the concordance rate of the first and the last sentences

In the round table meetings after the dry run, the participants reported that other sentences meant almost the same as the gold standard. We distinguished them, by refining the input of a single speech to a pair of question and answer speeches for the formal run.

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

4.2 Summarization Task

For the summarization task, a speech of a member of assembly and the limit length of summary are given. The participants generated a summary corresponding to the speech within the limit length. As an evaluation measure, we used the

Table 2. Data fields used in the summarization task

Field name	Explanation	Dry run	Formal run
ID	Identification code	○	○
Prefecture	Prefecture name	○	○
date	According to the Japanese calendar	○	○
Meeting	According to *Togikai dayori*	○	○
Speaker	Name of member of assembly	○	○
StartingLine	The number of first sentence	○	○
EndingLine	The number of last sentence	○	○
MainTopic	According to *Togikai dayori*	○	○
SubTopic	According to *Togikai dayori*	○	○
Summary	**Answer section**	○	○
Length	Limit length	○	○
Source	Speech of member of assembly	○	○

scores in the ROUGE[5] family and the scores of the quality questions by the participants. The ROUGE family means ROUGE-N1, -N2, -N3, -N4, -L, -SU4, and -W1.2. The quality questions were assessed by a three-grade evaluation (i.e., *A* to *C*) from viewpoints of content, formedness and total. However, for the content evaluation, we prepared an extra grade *X* because a summary that does not include contents of gold standard data may be acceptable. The quality question score $QQ(v)$ from viewpoint v was calculated using the following expressions:

$$QQ(v) = \frac{\sum_{s \in S} g(s, v)}{|S|} \tag{4}$$

$$g(s, v) = \begin{cases} 2\ (gradeA) \\ 1\ (gradeB) \\ 0\ (gradeC) \\ a\ (gradeX) \end{cases} \tag{5}$$

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

– *Dry run & formal run*
Input: a speech of a member of assembly in the minutes and a limit length of the summary
Output: a summary corresponding to the speech
Evaluation: ROUGE scores and participants assessment in terms of content, formedness and total.

Table 3. Data fields used in Classification task

Field name	Explanation	Dry run	Formal run
ID	Identification code	○	○
Topic	Political topic	○	○
Utterance	A sentence in the minutes	○	○
Relevance	**Answer section**	-	○
Fact-checkability	**Answer section**	-	○
Stance	**Answer section**	-	○
Class	**Answer section**	○	○

4.3 Classification Task

For the classification task, a political topic, such as "The Tsukiji Market should move to Toyosu." and a sentence in the minutes are given. The participants classify the sentence into the following three classes: support with fact-checkable reasons (S), against with fact-checkable reasons (A), and other (O). As evaluation measures, we used accuracy of all classes A. Then, recall $R_{cla}(c)$, precision $P_{cla}(c)$ and F-measure $F_{cla}(c)$ were used for each class c.

$$A = \frac{N_{cc}}{N_{ca}} \tag{6}$$

$$R_{cla}(c) = \frac{N_{cc}(c)}{N_{gsc}(c)} \tag{7}$$

$$P_{cla}(c) = \frac{N_{cc}(c)}{N_{sc}(c)} \tag{8}$$

$$F_{cla}(c) = \frac{2R_{cla}(c)P_{cla}(c)}{R_{cla}(c) + P_{cla}(c)} \tag{9}$$

where N_{acc} is the number of sentences of which the classified class is in concord with the gold standard class; N_{asc} is the number of all sentences, $N_{cc}(c)$ is the number of sentences, of which the gold standard class is c, that is classified into c, $N_{gsc}(c)$ is the number of sentences of which the gold standard class is c, and $N_{sc}(c)$ is the number of sentences classified into c.

– *Dry run*
Input: a political topic and a sentence in the minutes
Output: a class (support with fact-checkable reasons, against with fact-checkable reasons or other)
Evaluation: accuracy of all classes, recall of each class, precision of each class and F-measure of each class.

In the round table meetings after Dry run, we discussed basic factors of classification with participants, and agreed that the factors were relevance, fact-checkability and stance. The relevance means whether or not a given sentence

refer to a given topic. The fact-checkability means whether or not the sentence contains fact-checkable reasons. The stance means whether or not a speaker of the sentence agrees on the topic. However, we prepared the third stance, other (O), if a speaker stands neutral or has no relation to the topic. For Formal run, we refined the output to the factors besides class.

– *Formal run*
Input: a political topic and a sentence in the minutes
Output: a relevance (existance or absence), a fact-checkability (existance or absence), a stance (agree, disagree or other) and a class (support with fact-checkable reasons, against with fact-checkable reasons or other)
Evaluation: accuracy of all classes, recall of each class, precision of each class and F-measure of each class.

4.4 Schedule

The NTCIR-14 QA Lab-PoliInfo task has been run according to the following timeline:

February 20, 2018: QA Lab-PoliInfo Kickoff Meeting
March 20, 2018: NTCIR-14 Kickoff Event
April 19, 2018: 1st round table meeting
May 31, 2018: 2nd round table meeting
June 19, 2018: Dataset release

Dry Run
July 30, 2018: Task Registration Due for Dry Run
August 6 - 9, 2018: Dry Run (Segmentation & Classification Tasks)
August 13 - 16, 2018: Dry Run (Summarization Task)

August 30, 2018: 3rd round table meeting
October 29, 2018: 4th round table meeting

Formal Run
November 19, 2018: Task Registration Due for Formal Run (This is not required for Dry Run participants)
November 26 - 29, 2018: Formal Run (Segmentation & Classification Tasks)
December 3 - 6, 2018: Formal Run (Summarization Task)

NTCIR-14 CONFERENCE
February 1, 2019: Evaluation Result Release
February 1, 2019: Task overview paper release (draft)
March 15, 2019: Submission due of participant papers
May 1, 2019 Camera-ready participant paper due
June 10-13, 2019: NTCIR-14 Conference & EVIA 2019

Table 4. Active participating teams

Team ID	Organization
FU01*	Fukuoka University
FU02*	Fukuoka University
KitAi	Kyushu Institute of Technology
TTECH	Tokyo Institute of Technology
nami	Hitachi, Ltd.
nagoy	Nagoya University
akbl	Toyohashi University of Technology
ibrk	Ibaraki University
RICT	Ricoh Company, Ltd.
STARS	Hokkaido University
tmcit	Tokyo Metropolitan College of Industrial Technology
KSU	Kyoto Sangyo University
CUTKB	University of Tsukuba
LisLb	University of Tokyo
TO*	Task Organizers

*Task organizer(s) are in the team

5 Participation

Sixteen teams were registered, but only 15 teams (Table 4) participated.

6 Submissions

Table 5 shows the number of submitted runs (119 runs from 15 teams).

6.1 Dry Run

For Dry Run, 36 runs from 12 teams were submitted in total. For Segmentation task, 16 runs from 5 teams were submitted. For Summarization task, 6 runs from 5 teams were submitted. For Classification task, 14 runs from 8 teams were submitted.

6.2 Formal Run

For Dry run, 83 runs from 15 teams were submitted in total. For Segmentation task, 24 runs from 5 teams were submitted. For Summarization task, 14 runs from 7 teams were submitted. For Classification task, 45 runs from 11 teams were submitted.

Table 5. Number of submitted runs

Team ID	Dry run			Formal run		
	Segmentation	Summarization	Classification	Segmentation	Summarization	Classification
FU01	-	-	1	-	-	3
FU02	-	-	1	-	-	2
KitAi	-	-	-	-	2	-
TTECH	-	1	4	-	1	10
nami	11	-	-	11	-	-
nagoy	-	1	-	-	1	-
akbl	1	2	1	3	2	1
ibrk	-	-	1	-	-	2
RICT	1	-	1	5	-	7
STARS	-	-	4	-	-	4
tmcit	-	-	1	-	-	6
KSU	2	1	-	4	6	8
CUTKB	-	-	-	-	-	1
LisLb	-	-	-	-	1	1
TO	1	1	-	1	1	-
Sum	16	6	14	24	14	45

7 Result

7.1 Dry Run

Table 6 shows the results of Segmentation task. The best recall was 0.703 of akbl-1, the best precision was 0.859 of KSU-1, and the best F-measure was 0.570 of RICT-01.

Table 7 and 8 show the quality question scores and the ROUGE scores, respectively. Accordingly, akbl-01 achieved the best content scores and the best total score, regardless of the extra grade. The best formed score was 1.664 of akbl-02. For all ROUGE scores, akbl-01 acheived the best scores.

Table 9 shows the results of Classification task. The best accuracy (i.e. 0.823) was achieved by albl-01 and all STARS. For support, the best recall was 0.811 of FU01-01, the best precision was 0.400 of TTECH-03, and the best F-measure was 0.455 of TTECH-02. For against, the best recall was 0.708 of TTECH-02, the best precision was 0.375 of akbl-01, and the best F-measure was 0.314 of TTECH-03. For other, the best recall was 1.000 of ibrk-01 and all STARS, the best precision was 0.930 of TTECH-02, and the best F-measure was 0.903 of ibrk-01 and all STARS.

7.2 Formal Run

Table 10 shows the results of Segmentation task. The best recall was 1.000 of nami-11, the best precision was 0.940 of nami-01, and the best F-measure was 0.895 of RICT-01.

Table 11 and 12 show the quality question scores and the ROUGE scores, respectively. When extra grade was regarded as incorrect, the best content score was 0.886 of nagoy-01. When extra grade was regarded as correct, the best

Table 6. result of Segmentation task in Dry run

	R	P	F
nami-01	0.464 (311/670)	0.342 (311/909)	0.394
nami-02	0.458 (307/670)	0.339 (307/905)	0.390
nami-03	0.391 (262/670)	0.373 (262/702)	0.382
nami-04	0.479 (321/670)	0.304 (321/1,057)	0.372
nami-05	0.473 (317/670)	0.301 (317/1,053)	0.368
nami-06	0.396 (265/670)	0.354 (265/748)	0.374
nami-07	0.509 (341/670)	0.283 (341/1,203)	0.364
nami-08	0.503 (337/670)	0.281 (337/1,199)	0.361
nami-09	0.416 (279/670)	0.342 (279/815)	0.375
nami-10	0.370 (248/670)	0.420 (248/591)	0.393
nami-11	0.582 (390/670)	0.270 (390/1,444)	0.369
akbl-01	0.703 (471/670)	0.390 (471/1,207)	0.502
RICT-01	0.484 (324/670)	0.694 (324/467)	0.570
KSU-01	0.399 (267/670)	0.859 (267/311)	0.545
KSU-02	0.391 (262/670)	0.856 (262/306)	0.537
TO-01	0.267 (179/670)	0.056 (179/3,195)	0.093

Table 7. quality question scores of Summarization task in Dry run

	content		formed	total
	X=0	X=2		
TTECH-01	0.556	0.804	1.168	0.532
nagoy-01	0.156	0.204	0.856	0.168
akbl-01	0.644	1.036	1.656	0.784
akbl-02	0.608	0.968	1.664	0.744
KSU-01	0.000	0.000	0.064	0.000
TO-01	0.276	0.516	1.396	0.340
average	0.373	0.588	1.134	0.428

content score was 1.134 of KitAi-01. The best formed score was 1.955 of KSU-01, and the best total score was 0.912 of KitAi-01. For ROUGE scores, nagoy-01 achieved the best scores except some cases.

Table 13 shows the results of Classification task. The best accuracy was 0.942 of TTECH-07, -08 and -10. For support, the best recall was 0.731 of FU01-02, the best precision was 0.738 of KSU-03, -04, -07 and -08, and the best F-measure was 0.256 of TTECH-02. For against, the best recall was 1.000 of CUTKB-04, the best precision was 0.207 of TTECH-05, and the best F-measure was 0.216 of TTECH-05. For other, the best recall was 1.000 of TTECH-07, -08, -10, RICT-01, -05, -06 and STARS-01, the best precision was 0.947 of TTECH-02 and -05, and the best F-measure was 0.970 of TTHECH-07, -08 and -10.

8 Outline of the systems

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

Table 8. ROUGE scores of Summarization task in Dry run

		recall							F-measure						
		N1	N2	N3	N4	L	SU	W1.2	N1	N2	N3	N4	L	SU4	W1.2
surface form	TTECH-01	0.363	0.114	0.072	0.045	0.322	0.157	0.161	0.261	0.075	0.044	0.027	0.226	0.102	0.148
	nagoy-01	0.131	0.031	0.013	0.003	0.115	0.047	0.059	0.116	0.021	0.009	0.003	0.101	0.038	0.063
	akbl-01	0.388	0.145	0.092	0.060	0.351	0.182	0.173	0.317	0.114	0.071	0.046	0.283	0.141	0.180
	akbl-02	0.373	0.127	0.076	0.046	0.335	0.166	0.165	0.307	0.100	0.059	0.036	0.272	0.130	0.172
	KSU-01	0.106	0.007	0.001	0.000	0.095	0.024	0.048	0.135	0.009	0.001	0.000	0.121	0.031	0.070
	TO-01	0.207	0.051	0.022	0.013	0.186	0.070	0.093	0.189	0.046	0.021	0.013	0.167	0.062	0.105
stem	TTECH-01	0.391	0.131	0.085	0.055	0.342	0.177	0.172	0.281	0.087	0.052	0.033	0.239	0.115	0.159
	nagoy-01	0.136	0.036	0.019	0.006	0.121	0.052	0.063	0.119	0.024	0.011	0.005	0.103	0.040	0.066
	akbl-01	0.405	0.164	0.107	0.073	0.367	0.200	0.182	0.330	0.129	0.082	0.056	0.295	0.154	0.190
	akbl-02	0.392	0.146	0.091	0.058	0.353	0.185	0.175	0.321	0.115	0.071	0.045	0.286	0.143	0.182
	KSU-01	0.104	0.007	0.001	0.000	0.096	0.024	0.048	0.133	0.009	0.001	0.000	0.122	0.031	0.071
	TO-01	0.208	0.056	0.024	0.013	0.185	0.072	0.094	0.188	0.048	0.023	0.013	0.166	0.063	0.105
content word	TTECH-01	0.207	0.102	0.050	0.027	0.204	0.140	0.139	0.148	0.064	0.029	0.013	0.145	0.070	0.118
	nagoy-01	0.075	0.025	0.002	0.000	0.075	0.053	0.054	0.049	0.014	0.002	0.000	0.049	0.018	0.040
	akbl-01	0.263	0.124	0.072	0.049	0.255	0.162	0.171	0.204	0.094	0.053	0.028	0.196	0.100	0.156
	akbl-02	0.243	0.104	0.051	0.028	0.235	0.141	0.157	0.188	0.079	0.038	0.014	0.181	0.087	0.142
	KSU-01	0.009	0.000	0.000	0.000	0.009	0.001	0.006	0.010	0.000	0.000	0.000	0.010	0.001	0.007
	TO-01	0.088	0.017	0.010	0.000	0.087	0.034	0.059	0.074	0.017	0.010	0.000	0.074	0.026	0.058

The FU01 team tackled the classification task. For the Classification task, they recognized the relevance and the fact-checkability using words in the topic, and recognized the stance using the sentiment polarity dictionary.

The FU02 team tackled the classification task. For the Classification task, they recognized the relevance using words in the topic. The fact-chckablity was recognized by prepared clue expressions, and the stance was recognized by the fastText.

The KitAi team tackled the Summarization task. For the Summarization task, they took an approach of sentence extraction. They constructed pseudo training data based on five measures, and estimated the significance of sentences using SVR.

The TTECH team tackled the Summarization and the Classification tasks. For the Summarization task, they used sentence extraction based on redundancy-constrained knapsack problem. For the Classification task, they used SVM classifier taking account of a before- and an after-sentences as the context.

The nami team tackled the Segmentation task. For the Segmentation task, they proposed query reconstruction based on confidence and context information.

The nagoy team tackled the Summarization task. For the Summarization task, they used sentence extraction by random forest and sentence compression by heuristics and word frequency.

The akbl team tackled all tasks. For the Segmentation and the Summarization tasks, they used heuristics and TF-IDF values. For the Classification task, they used LSTM classifier and one-hot encoding.

Table 9. result of Classification task in Dry run

		support			against			other		
	A	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
FU01-01	0.326	0.811	0.130	0.224	0.292	0.292	0.292	0.265	0.833	0.402
FU02-01	0.410	0.351	0.105	0.105	0.292	0.103	0.152	0.428	0.807	0.559
TTECH-01	0.642	0.405	0.278	0.330	0.667	0.200	0.308	0.671	0.905	0.771
TTECH-02	0.494	0.541	0.392	0.455	0.708	0.113	0.195	0.470	0.930	0.624
TTECH-03	0.712	0.270	0.400	0.322	0.583	0.215	0.314	0.781	0.870	0.823
TTECH-04	0.497	0.514	0.373	0.432	0.583	0.103	0.175	0.488	0.879	0.628
akbl-01	0.762	0.216	0.205	0.210	0.125	0.375	0.188	0.887	0.845	0.865
ibrk-01	0.823	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.823	0.903
RICT-01	0.820	0.000	NaN	NaN	0.042	0.333	0.075	0.993	0.824	0.901
STARS-01	0.823	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.823	0.903
STARS-02	0.823	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.823	0.903
STARS-03	0.823	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.823	0.903
STARS-04	0.823	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.823	0.903
tmcit-01	0.727	0.000	0.000	NaN	0.167	0.087	0.114	0.869	0.834	0.851

The ibrk team tackled the Classification task. For the Classification task, they recognized the relevance and the fact-checkability using SVM classifier and one-hot encoding, and recognized the stance using the sentiment polarity dictionary.

The RICT team tackled the Segmentation and the Summarization tasks. For the Segmentation task, they segmented the whole minute in advance, and retrieved appropriate segments using the Elasticsearch. For the Classification task, they solved the relevance and the stance as anomaly detection problem, and the fact-checking was recognized by LSTM or SVM classifier.

The STARS team tackled the Classification task. For the Classification task, they recognized the relevance, the fact-checkability, and the stance using BiLSTM classifier and word-embedding. They also recognized the relevance using similarity between word vectors and so on.

The tmcit team tackled the Classification task. For the Classification task, they recognized the relevance based on cosine similarity. The fact-checkability was recognized by decision tree classifier, and the stance was recognized SVM classifier.

The KSU team tackled all tasks except the Classification task in the dry run. For the Segmentation task, they retrieved a relevant document, and segmented the document using heuristics and word frequency. For the Summarization task, they used LSTM encoder-decoder for abstractive summarization taking account of the subtopic. For the Classification task, they recognized the relevance using byte pair encoding, and recognized the fact-checking using LSTM classifier. The stance was recognized by SVM classifier using frequent words and N-gram.

The CUTKB team tackled the Classification task. For the Classification task, they used various classifiers such as LSTM, CNN, BERT, and their combinations. They also conducted a comparative study of them.

The LisLb team tackled the Summarization and the Classification tasks. For the Summarization task, they used rules according to Q&A pattern in the minutes. For the Classification task, they used an SVM classifier and word embedding.

Table 10. result of Segmentation task in Formal run

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

9 Conclusion

We described the overview of the NTCIR-14 QA Lab-PoliInfo task. The goal is realizing complex real-world question answering (QA) technologies, to show summaries of the opinions of assembly members and the reasons andconditions for such opinions, from Japanese regional assembly minutes. We conducted in a dry run and a formal run, which are including the segmentation, summarization, and classification tasks. Fifteen teams submitted 119 runs in total. We described the task description, the collection, the participation and the results.

References

1. Pariser, E.: The Filter Bubble: What the Internet Is Hiding from You. Penguin Group (2011)
2. Shibuki, H., Sakamoto, K., Kano, Y., Mitamura, T., Ishioroshi, M., Itakura, K. Y., Wang, D., Mori, T., Kando, N.: Overview of the NTCIR-11 QA-Lab Task. In: Proceedings of the 11th NTCIR Conference (2014)
3. Shibuki, H., Sakamoto, K., Ishioroshi, M., Fujita, A., Kano, Y., Mitamura, T., Mori, T., Kando, N.: Overview of the NTCIR-12 QA Lab-02 Task. Proceedings of the 12th NTCIR Conference (2016)
4. Kimura, Y., Takamaru, K., Tanaka, T., Kobayashi, A., Sakaji, H., Uchida, Y., Ototake, H., Masuyama, S.: Creating Japanese Political Corpus from Local Assembly Minutes of 47 Prefectures. In: Proceedings of Coling 2016 workshop, The 12th Workshop on Asian Language Resources, 78–85 (2016)

Table 11. quality question scores in Formal run (max is 2)

	all-topic				single-topic				multi-topic			
	content		formed	total	content		formed	total	content		formed	total
	X=0	X=2			X=0	X=2			X=0	X=2		
KitAi-01	0.856	1.134	1.732	0.912	0.953	1.170	1.660	0.995	0.745	1.092	1.815	0.815
KitAi-02	0.788	1.035	1.308	0.667	0.849	1.028	1.340	0.722	0.717	1.043	1.272	0.603
TTECH-01	0.290	0.644	1.783	0.402	0.274	0.575	1.755	0.401	0.310	0.723	1.815	0.402
nagoy-01	0.886	1.104	1.619	0.899	0.953	1.179	1.642	1.028	0.810	1.016	1.592	0.750
akbl-01	0.722	1.005	1.833	0.826	0.708	1.009	1.844	0.849	0.739	1.000	1.821	0.799
akbl-02*	0.707	1.000	1.837	0.793	—	—	—	—	0.707	1.000	1.837	0.793
KSU-01	0.043	0.043	1.955	0.048	0.052	0.052	1.934	0.057	0.033	0.033	1.978	0.038
KSU-02	0.076	0.121	1.745	0.071	0.080	0.156	1.722	0.104	0.071	0.082	1.772	0.033
KSU-03	0.091	0.157	1.715	0.104	0.104	0.179	1.731	0.156	0.076	0.130	1.696	0.043
KSU-04	0.111	0.167	1.419	0.093	0.118	0.193	1.420	0.132	0.103	0.136	1.418	0.049
KSU-05	0.048	0.078	1.692	0.048	0.057	0.085	1.726	0.057	0.038	0.071	1.652	0.038
KSU-06	0.078	0.169	1.535	0.091	0.085	0.151	1.542	0.094	0.071	0.190	1.527	0.087
LisLb-01	0.720	0.942	1.237	0.591	0.722	0.920	1.349	0.684	0.717	0.967	1.109	0.484
TO-01	0.504	0.846	1.763	0.551	0.464	0.794	1.778	0.521	0.550	0.905	1.746	0.586
average	0.423	0.603	1.655	0.435	0.387	0.535	1.532	0.414	0.406	0.599	1.646	0.394

*akbl-02 did not submit single-type.

5. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In: Proceedings of the ACL-04 workshop 8 (2004)

Table 12. ROUGE scores in Formal run (all-topic)

		recall							F-measure						
		N1	N2	N3	N4	L	SU4	W1.2	N1	N2	N3	N4	L	SU4	W1.2
Surface Form	KitAi-01	0.440	0.185	0.121	0.085	0.375	0.217	0.179	0.357	0.147	0.096	0.067	0.299	0.168	0.188
	KitAi-02	0.390	0.174	0.113	0.078	0.320	0.200	0.154	0.343	0.154	0.101	0.069	0.281	0.173	0.176
	TTECH-01	0.278	0.060	0.035	0.020	0.216	0.092	0.096	0.240	0.055	0.031	0.018	0.187	0.079	0.111
	nagoy-01	0.459	0.200	0.131	0.089	0.394	0.229	0.186	0.361	0.151	0.097	0.064	0.305	0.169	0.192
	akbl-01	0.400	0.173	0.113	0.076	0.345	0.189	0.157	0.361	0.156	0.102	0.068	0.310	0.167	0.185
	akbl-02	0.326	0.124	0.080	0.057	0.269	0.147	0.112	0.320	0.119	0.077	0.055	0.262	0.141	0.144
	KSU-01	0.158	0.028	0.009	0.002	0.147	0.043	0.071	0.210	0.039	0.013	0.004	0.196	0.059	0.107
	KSU-02	0.185	0.043	0.021	0.014	0.167	0.063	0.080	0.230	0.056	0.027	0.017	0.209	0.080	0.116
	KSU-03	0.172	0.036	0.008	0.002	0.157	0.050	0.075	0.211	0.043	0.011	0.003	0.192	0.062	0.106
	KSU-04	0.171	0.044	0.013	0.002	0.153	0.055	0.072	0.219	0.056	0.017	0.003	0.195	0.072	0.106
	KSU-05	0.227	0.029	0.010	0.002	0.195	0.064	0.089	0.231	0.029	0.010	0.003	0.196	0.065	0.110
	KSU-06	0.221	0.038	0.013	0.004	0.187	0.065	0.086	0.230	0.038	0.012	0.004	0.192	0.067	0.108
	LisLb-01	0.251	0.120	0.079	0.058	0.211	0.132	0.103	0.226	0.107	0.071	0.051	0.188	0.115	0.118
	TO-01	0.267	0.093	0.061	0.045	0.230	0.117	0.105	0.272	0.086	0.052	0.036	0.233	0.110	0.133
Stem	KitAi-01	0.458	0.199	0.134	0.096	0.389	0.234	0.188	0.373	0.159	0.106	0.075	0.311	0.182	0.199
	KitAi-02	0.399	0.179	0.118	0.082	0.326	0.208	0.158	0.351	0.160	0.106	0.074	0.286	0.180	0.181
	TTECH-01	0.289	0.064	0.037	0.022	0.222	0.097	0.099	0.251	0.058	0.033	0.019	0.193	0.084	0.114
	nagoy-01	0.479	0.217	0.145	0.101	0.412	0.247	0.197	0.377	0.165	0.108	0.074	0.319	0.184	0.205
	akbl-01	0.415	0.184	0.122	0.083	0.357	0.203	0.164	0.375	0.165	0.110	0.074	0.322	0.179	0.195
	akbl-02	0.339	0.135	0.089	0.064	0.279	0.158	0.119	0.333	0.129	0.085	0.063	0.272	0.152	0.153
	KSU-01	0.161	0.028	0.010	0.002	0.148	0.044	0.071	0.214	0.040	0.013	0.004	0.197	0.061	0.108
	KSU-02	0.187	0.044	0.021	0.014	0.170	0.064	0.081	0.233	0.057	0.027	0.017	0.212	0.082	0.117
	KSU-03	0.175	0.036	0.008	0.002	0.159	0.052	0.075	0.217	0.044	0.011	0.003	0.196	0.065	0.108
	KSU-04	0.174	0.045	0.014	0.002	0.155	0.056	0.073	0.222	0.058	0.018	0.003	0.197	0.073	0.107
	KSU-05	0.230	0.029	0.010	0.002	0.199	0.066	0.090	0.236	0.030	0.010	0.003	0.201	0.067	0.112
	KSU-06	0.226	0.040	0.013	0.004	0.189	0.066	0.087	0.235	0.039	0.012	0.004	0.195	0.069	0.109
	LisLb-01	0.261	0.125	0.084	0.061	0.218	0.139	0.106	0.235	0.112	0.075	0.055	0.195	0.121	0.122
	TO-01	0.273	0.097	0.065	0.048	0.233	0.121	0.107	0.277	0.089	0.056	0.039	0.236	0.114	0.136
Content Word	KitAi-01	0.285	0.145	0.090	0.050	0.278	0.154	0.180	0.224	0.115	0.071	0.042	0.217	0.107	0.170
	KitAi-02	0.254	0.126	0.083	0.053	0.247	0.131	0.156	0.214	0.109	0.069	0.046	0.208	0.106	0.159
	TTECH-01	0.088	0.028	0.015	0.007	0.082	0.033	0.050	0.076	0.024	0.012	0.006	0.071	0.027	0.054
	nagoy-01	0.326	0.164	0.094	0.046	0.315	0.168	0.201	0.249	0.123	0.067	0.036	0.239	0.110	0.187
	akbl-01	0.256	0.113	0.065	0.034	0.247	0.124	0.148	0.224	0.098	0.056	0.031	0.216	0.100	0.158
	akbl-02	0.200	0.094	0.051	0.032	0.189	0.095	0.109	0.188	0.089	0.049	0.031	0.178	0.087	0.127
	KSU-01	0.048	0.001	0.000	0.000	0.047	0.007	0.032	0.059	0.001	0.000	0.000	0.058	0.009	0.043
	KSU-02	0.069	0.014	0.000	0.000	0.067	0.019	0.043	0.083	0.015	0.000	0.000	0.081	0.022	0.059
	KSU-03	0.041	0.002	0.000	0.000	0.041	0.007	0.027	0.050	0.002	0.000	0.000	0.050	0.008	0.036
	KSU-04	0.050	0.002	0.000	0.000	0.048	0.008	0.031	0.064	0.003	0.000	0.000	0.061	0.011	0.044
	KSU-05	0.067	0.002	0.000	0.000	0.062	0.013	0.041	0.063	0.003	0.000	0.000	0.057	0.011	0.043
	KSU-06	0.053	0.003	0.000	0.000	0.051	0.008	0.034	0.051	0.003	0.000	0.000	0.049	0.009	0.037
	LisLb-01	0.171	0.083	0.044	0.026	0.160	0.088	0.106	0.140	0.068	0.036	0.023	0.130	0.065	0.102
	TO-01	0.116	0.055	0.035	0.012	0.111	0.056	0.070	0.106	0.042	0.023	0.011	0.101	0.042	0.076

Table 13. result of Classification task in Formal run (Class salutariness)

		support			against			other		
	A	R	P	F	R	P	F	R	P	F
FU01-01	0.624	0.417	0.057	0.100	0.076	0.041	0.053	0.648	0.938	0.766
FU01-02	0.373	0.731	0.057	0.106	0.183	0.045	0.072	0.362	0.943	0.523
FU01-03	0.909	0.089	0.164	0.115	0.008	0.020	0.011	0.970	0.936	0.953
FU02-01	0.842	0.027	0.040	0.032	0.095	0.033	0.049	0.899	0.933	0.916
FU02-02	0.840	0.073	0.063	0.068	0.069	0.030	0.042	0.895	0.933	0.914
TTECH-01	0.923	0.046	0.163	0.072	0.015	0.133	0.027	0.987	0.935	0.960
TTECH-02	0.896	0.260	0.252	0.256	0.221	0.199	0.209	0.943	0.947	0.945
TTECH-03	0.919	0.116	0.254	0.159	0.069	0.200	0.103	0.978	0.938	0.958
TTECH-04	0.921	0.043	0.134	0.065	0.015	0.133	0.027	0.985	0.934	0.959
TTECH-05	0.897	0.251	0.251	0.251	0.225	0.207	0.216	0.944	0.947	0.945
TTECH-06	0.918	0.132	0.269	0.177	0.080	0.206	0.115	0.976	0.939	0.957
TTECH-07	0.942	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.942	0.970
TTECH-08	0.942	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.942	0.970
TTECH-09	0.926	0.000	0.000	NaN	0.000	NaN	NaN	0.982	0.941	0.961
TTECH-10	0.942	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.942	0.970
akbl-01	0.923	0.118	0.344	0.176	0.034	0.097	0.050	0.983	0.939	0.960
ibrk-01	0.731	0.178	0.063	0.093	0.202	0.045	0.074	0.770	0.934	0.844
ibrk-02	0.731	0.178	0.063	0.093	0.202	0.045	0.074	0.770	0.934	0.844
RICT-01	0.933	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.933	0.965
RICT-02	0.932	0.002	0.091	0.004	0.004	0.111	0.008	0.998	0.933	0.964
RICT-03	0.893	0.118	0.145	0.130	0.111	0.117	0.114	0.949	0.940	0.944
RICT-04	0.894	0.114	0.143	0.127	0.111	0.117	0.114	0.950	0.939	0.944
RICT-05	0.933	0.000	NaN	NaN	0.000	0.000	NaN	1.000	0.933	0.965
RICT-06	0.933	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.933	0.965
RICT-07	0.932	0.084	0.440	0.141	0.042	0.407	0.076	0.994	0.937	0.965
STARS-01	0.933	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.933	0.965
STARS-02	0.889	0.002	0.002	0.002	0.000	NaN	NaN	0.953	0.933	0.943
STARS-03	0.889	0.002	0.002	0.002	0.000	NaN	NaN	0.953	0.933	0.943
STARS-04	0.889	0.002	0.002	0.002	0.000	NaN	NaN	0.953	0.933	0.943
tmcit-01	0.875	0.282	0.139	0.186	0.000	NaN	NaN	0.925	0.943	0.934
tmcit-02	0.893	0.239	0.160	0.192	0.000	NaN	NaN	0.946	0.942	0.944
tmcit-03	0.873	0.296	0.142	0.192	0.000	NaN	NaN	0.922	0.943	0.932
tmcit-04	0.879	0.319	0.161	0.214	0.000	NaN	NaN	0.928	0.944	0.936
tmcit-05	0.898	0.267	0.189	0.221	0.000	NaN	NaN	0.950	0.942	0.946
tmcit-06	0.878	0.292	0.148	0.196	0.000	NaN	NaN	0.927	0.943	0.935
KSU-01	0.932	0.075	0.579	0.133	0.008	0.056	0.014	0.995	0.937	0.965
KSU-02	0.932	0.071	0.689	0.129	0.008	0.042	0.013	0.995	0.937	0.965
KSU-03	0.934	0.071	0.738	0.130	0.008	0.083	0.015	0.998	0.937	0.967
KSU-04	0.934	0.071	0.738	0.130	0.008	0.083	0.015	0.998	0.937	0.967
KSU-05	0.932	0.075	0.579	0.133	0.019	0.111	0.032	0.995	0.937	0.965
KSU-06	0.932	0.071	0.689	0.129	0.019	0.088	0.031	0.995	0.937	0.965
KSU-07	0.934	0.071	0.738	0.130	0.011	0.100	0.020	0.997	0.937	0.966
KSU-08	0.934	0.071	0.738	0.130	0.011	0.100	0.020	0.997	0.937	0.966
CUTKB-04	0.025	0.000	NaN	NaN	1.000	0.025	0.049	0.000	NaN	NaN
LisLb-01	0.914	0.021	0.065	0.032	0.037	0.080	0.051	0.976	0.935	0.955

Table 14. detail of Segmentation task in Formal run

	question						answer					
	R		P		F		R		P		F	
nami-01	0.794	(947/1,193)	0.949	(947/998)	0.864		0.856	(486/568)	0.922	(486/527)	0.888	
nami-02	0.841	(1,003/1,193)	0.819	(1,003/1,224)	0.830		0.912	(518/568)	0.918	(518/564)	0.915	
nami-03	0.977	(1,165/1,193)	0.497	(1,165/2,342)	0.659		1.000	(568/568)	0.501	(568/1,133)	0.668	
nami-04	0.614	(732/1,193)	0.829	(732/883)	0.705		0.692	(393/568)	0.763	(393/515)	0.726	
nami-05	0.515	(614/1,193)	0.936	(614/656)	0.664		0.632	(359/568)	0.923	(359/389)	0.750	
nami-06	0.637	(760/1,193)	0.631	(760/1,204)	0.634		0.692	(393/568)	0.715	(393/550)	0.703	
nami-07	0.791	(944/1,193)	0.949	(944/995)	0.863		0.810	(460/568)	0.902	(460/510)	0.853	
nami-08	0.820	(978/1,193)	0.938	(978/1,043)	0.875		0.856	(486/568)	0.922	(486/527)	0.888	
nami-09	0.858	(1,023/1,193)	0.810	(1,023/1,263)	0.833		0.912	(518/568)	0.918	(518/564)	0.915	
nami-10	0.990	(1,181/1,193)	0.448	(1,181/2,636)	0.617		1.000	(568/568)	0.501	(568/1,133)	0.668	
nami-11	1.000	(1,193/1,193)	0.087	(1,193/13,684)	0.160		1.000	(568/568)	0.273	(568/2,081)	0.429	
akbl-01	0.780	(931/1,193)	0.468	(931/1,990)	0.585		0.741	(421/568)	0.802	(421/525)	0.770	
akbl-02	0.878	(1047/1,193)	0.389	(1,047/2,694)	0.539		0.783	(445/568)	0.757	(445/588)	0.770	
akbl-03	0.638	(761/1,193)	0.439	(761/1,735)	0.520		0.694	(394/568)	0.801	(394/492)	0.743	
RICT-01	0.851	(1,015/1,193)	0.913	(1,015/1,112)	0.881		0.949	(539/568)	0.903	(539/597)	0.925	
RICT-02	0.811	(968/1,193)	0.871	(968/1,111)	0.840		0.949	(539/568)	0.923	(539/584)	0.936	
RICT-03	0.847	(1,010/1,193)	0.759	(1,010/1,331)	0.800		0.868	(493/568)	0.829	(493/595)	0.848	
RICT-04	0.828	(988/1,193)	0.715	(988/1,382)	0.767		0.680	(386/568)	0.839	(386/460)	0.751	
RICT-05	0.935	(1,116/1,193)	0.643	(1,116/1,735)	0.762		0.937	(532/568)	0.919	(532/579)	0.928	
KSU-01	0.835	(996/1,193)	0.192	(996/5,196)	0.312		0.662	(376/568)	0.841	(376/447)	0.741	
KSU-02	0.806	(962/1,193)	0.209	(962/4,603)	0.332		0.660	(375/568)	0.949	(375/395)	0.779	
KSU-03	0.899	(1,072/1,193)	0.612	(1,072/1,751)	0.728		0.655	(372/568)	0.857	(372/434)	0.743	
KSU-04	0.866	(1,033/1,193)	0.905	(1,033/1,141)	0.885		0.651	(370/568)	0.974	(370/380)	0.781	
TO-01	0.450	(537/1,193)	0.984	(537/546)	0.618		0.151	(86/568)	0.581	(86/148)	0.240	

Table 15. correlation coefficients between total score and ROUGE scores

	recall							F-measure						
	N1	N2	N3	N4	L	W1.2	SU4	N1	N2	N3	N4	L	W1.2	SU4
surface form	0.924	0.955	0.964	0.968	0.915	0.953	0.893	0.900	0.942	0.957	0.959	0.852	0.946	0.882
stem	0.928	0.959	0.968	0.972	0.918	0.956	0.900	0.912	0.950	0.965	0.968	0.866	0.954	0.894
content word	0.943	0.957	0.948	0.920	0.939	0.952	0.926	0.942	*0.963*	*0.953*	*0.924*	0.937	*0.956*	*0.935*

Table 16. detail of Classification task in Formal run (Rl and FC)

	relevance							fact-checkability						
		existence			absence				existence			absence		
	A	R	P	F	R	P	F	A	R	P	F	R	P	F
FU01-01	0.706	0.754	0.889	0.816	0.393	0.200	0.265	0.523	0.527	0.376	0.439	0.520	0.668	0.585
FU01-02	0.706	0.754	0.889	0.816	0.393	0.200	0.265	0.359	0.961	0.352	0.515	0.030	0.580	0.056
FU01-03	0.825	0.910	0.890	0.900	0.280	0.326	0.301	0.632	0.183	0.450	0.260	0.878	0.662	0.755
FU02-01	0.865	0.999	0.866	0.928	0.006	0.533	0.011	0.530	0.419	0.359	0.387	0.591	0.650	0.619
FU02-02	0.865	0.999	0.866	0.928	0.006	0.533	0.011	0.530	0.419	0.359	0.387	0.591	0.650	0.619
TTECH-01	0.853	0.930	0.903	0.917	0.361	0.445	0.399	0.708	0.422	0.631	0.506	0.865	0.732	0.793
TTECH-02	0.849	0.922	0.905	0.914	0.377	0.430	0.402	0.661	0.670	0.516	0.583	0.656	0.784	0.714
TTECH-03	0.835	0.903	0.906	0.904	0.397	0.389	0.393	0.710	0.480	0.617	0.540	0.837	0.746	0.789
TTECH-04	0.853	0.930	0.903	0.917	0.361	0.445	0.399	0.708	0.417	0.632	0.503	0.867	0.731	0.793
TTECH-05	0.849	0.924	0.904	0.914	0.368	0.430	0.397	0.659	0.650	0.515	0.574	0.664	0.776	0.716
TTECH-06	0.834	0.902	0.905	0.904	0.395	0.385	0.390	0.709	0.468	0.617	0.533	0.841	0.743	0.789
TTECH-07	0.988	1.000	0.988	0.994	0.000	NaN	NaN	0.709	0.335	0.475	0.393	0.855	0.767	0.809
TTECH-08	0.988	1.000	0.988	0.994	0.000	NaN	NaN	0.686	0.588	0.455	0.513	0.724	0.818	0.768
TTECH-09	0.988	1.000	0.988	0.994	0.000	NaN	NaN	0.702	0.235	0.444	0.308	0.885	0.748	0.811
TTECH-10	0.988	1.000	0.988	0.994	0.000	NaN	NaN	0.719	0.176	0.500	0.261	0.931	0.743	0.827
akbl-01	0.861	0.952	0.895	0.922	0.282	0.476	0.354	0.708	0.438	0.626	0.515	0.857	0.736	0.791
ibrk-01	0.865	1.000	0.865	0.928	0.000	NaN	NaN	0.646	0.000	NaN	NaN	1.000	0.646	0.785
ibrk-02	0.865	1.000	0.865	0.928	0.000	NaN	NaN	0.646	0.000	NaN	NaN	1.000	0.646	0.785
RICT-01	0.857	0.990	0.865	0.923	0.008	0.104	0.015	0.729	0.419	0.694	0.522	0.899	0.738	0.811
RICT-02	0.517	0.510	0.883	0.646	0.567	0.152	0.240	0.729	0.419	0.694	0.522	0.899	0.738	0.811
RICT-03	0.787	0.827	0.918	0.870	0.528	0.322	0.400	0.729	0.419	0.694	0.522	0.899	0.738	0.811
RICT-04	0.794	0.836	0.919	0.875	0.524	0.332	0.407	0.729	0.419	0.694	0.522	0.899	0.738	0.811
RICT-05	0.857	0.990	0.865	0.923	0.008	0.104	0.015	0.621	0.693	0.476	0.564	0.582	0.776	0.665
RICT-06	0.857	0.990	0.865	0.923	0.008	0.104	0.015	0.724	0.417	0.680	0.517	0.892	0.737	0.807
RICT-07	0.857	0.990	0.865	0.923	0.008	0.104	0.015	0.729	0.419	0.694	0.522	0.899	0.738	0.811
STARS-01	0.865	1.000	0.865	0.928	0.000	NaN	NaN	0.646	0.000	NaN	NaN	1.000	0.646	0.785
STARS-02	0.865	1.000	0.865	0.928	0.000	NaN	NaN	0.354	1.000	0.354	0.523	0.000	NaN	NaN
STARS-03	0.865	1.000	0.865	0.928	0.000	NaN	NaN	0.354	1.000	0.354	0.523	0.000	NaN	NaN
STARS-04	0.865	1.000	0.865	0.928	0.000	NaN	NaN	0.354	1.000	0.354	0.523	0.000	NaN	NaN
tmcit-01	0.767	0.814	0.907	0.858	0.463	0.279	0.348	0.652	0.630	0.507	0.562	0.665	0.766	0.712
tmcit-02	0.586	0.561	0.935	0.701	0.748	0.209	0.327	0.651	0.636	0.506	0.564	0.660	0.768	0.710
tmcit-03	0.767	0.814	0.907	0.858	0.463	0.279	0.348	0.650	0.636	0.504	0.562	0.658	0.767	0.708
tmcit-04	0.767	0.814	0.907	0.858	0.463	0.279	0.348	0.649	0.637	0.503	0.562	0.656	0.767	0.707
tmcit-05	0.586	0.561	0.935	0.701	0.748	0.209	0.327	0.649	0.636	0.504	0.562	0.657	0.767	0.708
tmcit-06	0.767	0.814	0.907	0.858	0.463	0.279	0.348	0.652	0.629	0.507	0.561	0.665	0.766	0.712
KSU-01	0.790	0.785	0.966	0.866	0.823	0.373	0.513	0.735	0.407	0.722	0.521	0.914	0.738	0.817
KSU-02	0.790	0.785	0.966	0.866	0.823	0.373	0.513	0.735	0.407	0.722	0.521	0.914	0.738	0.817
KSU-03	0.790	0.785	0.966	0.866	0.823	0.373	0.513	0.735	0.407	0.722	0.521	0.914	0.738	0.817
KSU-04	0.790	0.785	0.966	0.866	0.823	0.373	0.513	0.735	0.407	0.722	0.521	0.914	0.738	0.817
KSU-05	0.873	0.969	0.893	0.930	0.257	0.567	0.353	0.735	0.407	0.722	0.521	0.914	0.738	0.817
KSU-06	0.873	0.969	0.893	0.930	0.257	0.567	0.353	0.735	0.407	0.722	0.521	0.914	0.738	0.817
KSU-07	0.873	0.969	0.893	0.930	0.257	0.567	0.353	0.735	0.407	0.722	0.521	0.914	0.738	0.817
KSU-08	0.873	0.969	0.893	0.930	0.257	0.567	0.353	0.735	0.407	0.722	0.521	0.914	0.738	0.817
CUTKB-04	0.865	1.000	0.865	0.928	0.000	NaN	NaN	0.730	0.523	0.647	0.579	0.843	0.764	0.801
LISLab-01	0.727	0.829	0.843	0.836	0.200	0.183	0.191	0.544	0.313	0.366	0.337	0.680	0.627	0.652

Table 17. result of Classification task in Formal run (stance agreeing)

		agree			disagree			other		
	A	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
FU01-01	0.393	0.739	0.181	0.291	0.199	0.155	0.174	0.353	0.811	0.492
FU01-02	0.393	0.739	0.181	0.291	0.199	0.155	0.174	0.353	0.811	0.492
FU01-03	0.734	0.097	0.243	0.139	0.072	0.152	0.097	0.915	0.789	0.847
FU02-01	0.632	0.079	0.169	0.107	0.246	0.120	0.161	0.768	0.783	0.775
FU02-02	0.633	0.112	0.153	0.129	0.117	0.080	0.095	0.779	0.779	0.779
TTECH-01	0.782	0.249	0.466	0.325	0.188	0.578	0.284	0.938	0.814	0.872
TTECH-02	0.772	0.355	0.476	0.406	0.354	0.427	0.387	0.888	0.836	0.861
TTECH-03	0.774	0.295	0.490	0.368	0.299	0.419	0.349	0.908	0.826	0.865
TTECH-04	0.777	0.275	0.438	0.338	0.192	0.576	0.288	0.926	0.816	0.867
TTECH-05	0.769	0.379	0.474	0.421	0.357	0.420	0.386	0.881	0.838	0.859
TTECH-06	0.774	0.300	0.491	0.372	0.304	0.421	0.353	0.907	0.826	0.865
TTECH-07	0.747	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.747	0.855
TTECH-08	0.747	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.747	0.855
TTECH-09	0.666	0.405	0.342	0.371	0.000	NaN	NaN	0.774	0.778	0.776
TTECH-10	0.744	0.145	0.475	0.222	0.000	NaN	NaN	0.954	0.763	0.848
akbl-01	0.780	0.325	0.512	0.398	0.258	0.391	0.311	0.915	0.832	0.872
ibrk-01	0.686	0.183	0.218	0.199	0.218	0.221	0.220	0.823	0.800	0.811
ibrk-02	0.782	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.782	0.877
RICT-01	0.781	0.000	NaN	NaN	0.000	0.000	NaN	0.999	0.781	0.877
RICT-02	0.733	0.042	0.162	0.067	0.039	0.081	0.053	0.926	0.783	0.848
RICT-03	0.568	0.516	0.270	0.354	0.417	0.176	0.248	0.593	0.852	0.699
RICT-04	0.575	0.513	0.272	0.355	0.417	0.180	0.251	0.602	0.852	0.706
RICT-05	0.781	0.000	NaN	NaN	0.000	0.000	NaN	0.999	0.781	0.877
RICT-06	0.781	0.000	NaN	NaN	0.000	0.000	NaN	0.999	0.781	0.877
RICT-07	0.808	0.295	0.630	0.402	0.194	0.579	0.291	0.962	0.827	0.890
STARS-01	0.782	0.000	NaN	NaN	0.000	NaN	NaN	1.000	0.782	0.877
STARS-02	0.748	0.003	0.008	0.004	0.000	NaN	NaN	0.957	0.785	0.862
STARS-03	0.748	0.003	0.008	0.004	0.000	NaN	NaN	0.957	0.785	0.862
STARS-04	0.748	0.003	0.008	0.004	0.000	NaN	NaN	0.957	0.785	0.862
tmcit-01	0.735	0.428	0.371	0.397	0.194	0.373	0.256	0.846	0.826	0.836
tmcit-02	0.737	0.451	0.373	0.409	0.215	0.405	0.281	0.842	0.831	0.837
tmcit-03	0.739	0.441	0.377	0.406	0.165	0.384	0.231	0.852	0.827	0.839
tmcit-04	0.743	0.476	0.392	0.430	0.169	0.377	0.233	0.850	0.832	0.841
tmcit-05	0.737	0.480	0.382	0.426	0.179	0.372	0.242	0.840	0.832	0.836
tmcit-06	0.741	0.418	0.374	0.395	0.168	0.390	0.234	0.858	0.825	0.841
KSU-01	0.802	0.230	0.683	0.345	0.237	0.402	0.298	0.961	0.829	0.890
KSU-02	0.799	0.201	0.724	0.314	0.254	0.370	0.301	0.961	0.829	0.890
KSU-03	0.801	0.171	0.720	0.277	0.202	0.420	0.273	0.973	0.820	0.890
KSU-04	0.799	0.153	0.732	0.253	0.214	0.404	0.280	0.973	0.820	0.890
KSU-05	0.802	0.230	0.683	0.345	0.237	0.402	0.298	0.961	0.829	0.890
KSU-06	0.799	0.201	0.724	0.314	0.254	0.370	0.301	0.961	0.829	0.890
KSU-07	0.801	0.171	0.720	0.277	0.202	0.420	0.273	0.973	0.820	0.890
KSU-08	0.799	0.153	0.732	0.253	0.214	0.404	0.280	0.973	0.820	0.890
CUTKB-04	0.033	0.015	0.677	0.029	0.017	0.625	0.034	0.038	0.778	0.073
LISLab-01	0.706	0.069	0.124	0.089	0.126	0.157	0.140	0.869	0.795	0.830