Overview of the NTCIR-11 QA-Lab Task

Hideyuki Shibuki[*1], Kotaro Sakamoto[*1, *2], Yoshionobu Kano[*3,*4, †], Teruko Mitamura[*5], Madoka Ishioroshi[*2], Kelly Y. Itakura[†], Di Wang[*5], Tatsunori Mori[*1], Noriko Kando[*2,*6]

*1: Yokohama National University, *2: National Institute of Informatics, *3: Shizuoka University, *4: PRESTO, *5: Language Technology Institute, Carnegie Mellon University, *6: The Graduate University for Advanced Studies (SOKENDAI) , †: formerly at *2

{shib|sakamoto|mori}@forest.eis.ynu.ac.jp, {ishioroshi|itakura|kando}@nii.ac.jp, kano@inf.shizuoka.ac.jp, {teruko+|diwang}@cs.cmu.edu

ABSTRACT

This paper describes an overview of the first QA Lab (Question Answering Lab for Entrance Exam) task at NTCIR 11. The goal of the QA lab is to provide a module-based platform for advanced question answering systems and comparative evaluation for solving real-world university entrance exam questions. In this task, "world history" questions are selected from The National Center Test for University Admissions and from the secondary exams at 5 universities in Japan. This paper also describes the used data, baseline systems and formal run results.

Categories and Subject Descriptors

H.3.4 [**INFORMATION STORAGE AND RETRIEVAL**]: Systems and Software – *Performance evaluation (efficiency and effectiveness), Question-answering (fact retrieval) systems.*

General Terms

Experimentation

Keywords

NTCIR 11, question answering, university entrance examination, module-based platform

1. INTRODUCTION

The goal of the QA lab is to provide a module-based platform for advanced question answering systems and comparative evaluation for solving real-world university entrance exam questions. In its first year at NTCIR 11, "world history" questions are selected from The National Center Test for University Admissions (Center Test) and from the secondary exams at 5 universities in Japan (Secondary Exam). Both Japanese and English translations of the topics (questions) are provided in an XML format. Although participants are welcome to use any resources except solutions to the exams, four sets of tagged corpus of the Japanese high school history textbooks published by 2 publishers and one referential version of a whole Wikipedia corpus were provided as knowledge sources. UIMA-based QA baseline systems were also provided to the participants both Japanese and English.

Some of the highlights are:

1. Solving real-world problems.
2. Many questions are not in a simple QA format, and require an understanding of the surrounding context.
3. Some questions require inference.

At NTCIR 11, most of the questions are True/False questions or factoid questions, with some questions involving short answers of 50 to 600 Japanese characters. In the next round at NTCIR 12, the plan is to include biology questions and increase the difficulty and the number of complex questions.

Table 1: Question types used in each phase

Phase	Center Test	Secondary Exam
Question Types	**multi-choice**	**various types including complex QA**
1	YES	N/A
2	YES	YES

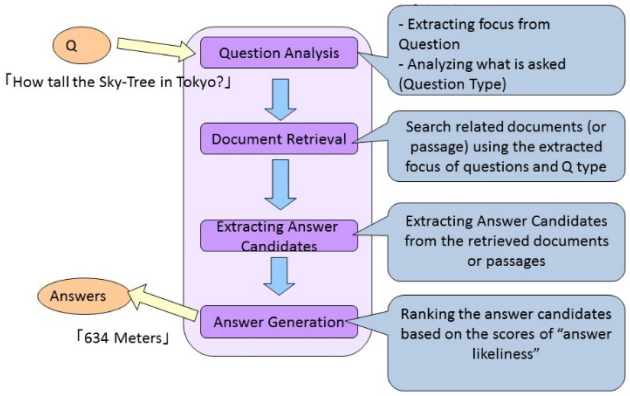


Figure 1: Module structure of the original QA platform

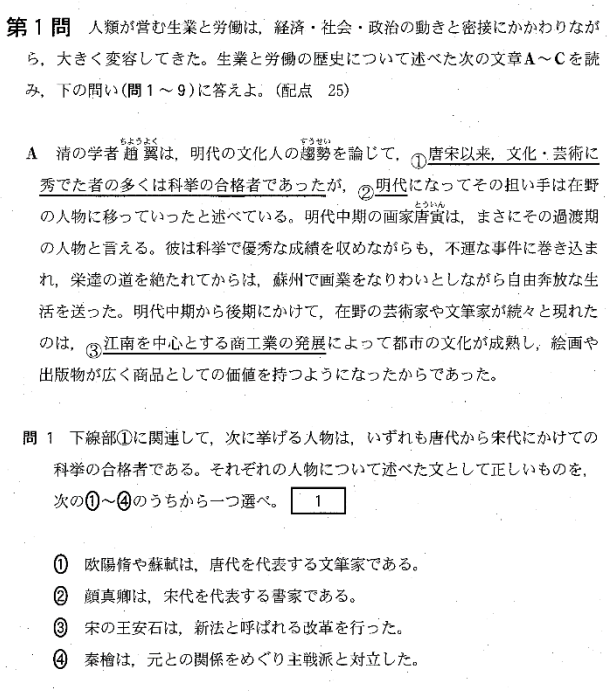
2. TASK DESCRIPTION

A single task is carried out in two separate phase. Each phase has a separate training set and test set with similar difficulty. The task is to return solutions that could be either True/False or short answers given a list of topics (questions) in an XML format.

Participants are free to participate any particular phase and either of Center Test or Secondary Exam. Table 1 shows question types used in each phase.

Figure 1 shows the module structure of the original QA platform. The module-based platform means participants can participate in any phase of QA system development. In Japanese baseline systems [1-2], the process is divided into 4 modules, question analysis, document retrieval, extraction of answer candidates, and answer generation, which are re-created based on the source-codes of MinerVA[3] and Javeline[4].

These baseline systems are originally usable for general-purpose question answering. To utilize them for multiple-choice type questions in Center Test, two modules were added – 1) the question format analysis module in which analyze the types of the questions (true or false, arrange the choices chronological order, blank-filling,



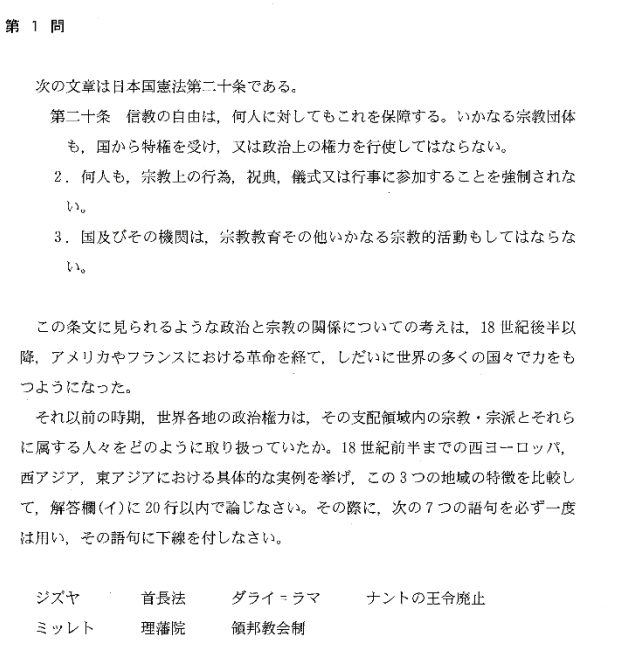
第１問　人類が営む生業と労働は，経済・社会・政治の動きと密接にかかわりながら，大きく変容してきた。生業と労働の歴史について述べた次の文章**A**～**C**を読み，下の問い(**問１**～**９**)に答えよ。(配点　25)

A　清の学者趙翼は，明代の文化人の趨勢を論じて，①唐宋以来，文化・芸術に秀でた者の多くは科挙の合格者であったが，②明代になってその担い手は在野の人物に移っていったと述べている。明代中期の画家唐寅は，まさにその過渡期の人物と言える。彼は科挙で優秀な成績を収めながらも，不運な事件に巻き込まれ，栄達の道を絶たれてからは，蘇州で画業をなりわいとしながら自由奔放な生活を送った。明代中期から後期にかけて，在野の芸術家や文筆家が続々と現れたのは，③江南を中心とする商工業の発展によって都市の文化が成熟し，絵画や出版物が広く商品としての価値を持つようになったからであった。

問１　下線部①に関連して，次に挙げる人物は，いずれも唐代から宋代にかけての科挙の合格者である。それぞれの人物について述べた文として正しいものを，次の①～④のうちから一つ選べ。 1

① 欧陽脩や蘇軾は，唐代を代表する文筆家である。
② 顔真卿は，宋代を代表する書家である。
③ 宋の王安石は，新法と呼ばれる改革を行った。
④ 秦檜は，元との関係をめぐり主戦派と対立した。

Figure 2: Example of the Center Test



第 1 問

次の文章は日本国憲法第二十条である。

第二十条　信教の自由は，何人に対してもこれを保障する。いかなる宗教団体も，国から特権を受け，又は政治上の権力を行使してはならない。

2．何人も，宗教上の行為，祝典，儀式又は行事に参加することを強制されない。

3．国及びその機関は，宗教教育その他いかなる宗教的活動もしてはならない。

この条文に見られるような政治と宗教の関係についての考えは，18 世紀後半以降，アメリカやフランスにおける革命を経て，しだいに世界の多くの国々で力をもつようになった。

それ以前の時期，世界各地の政治権力は，その支配領域内の宗教・宗派とそれらに属する人々をどのように取り扱っていたか。18 世紀前半までの西ヨーロッパ，西アジア，東アジアにおける具体的な実例を挙げ，この 3 つの地域の特徴を比較して，解答欄(イ)に 20 行以内で論じなさい。その際に，次の 7 つの語句を必ず一度は用い，その語句に下線を付しなさい。

ジズヤ　　首長法　　ダライ＝ラマ　　ナントの王令廃止
ミッレト　　理藩院　　領邦教会制

Figure 3: Example of the Secondary Exam

etc. and answer the correct or wrong choice) comes first, and 2) selecting final answer (the choice) based on the answer candidates and their scores resulted from the QA-workflow and the question format analysis. English version started from the Javeline, but slightly different module structure [5]. In addition to the above a separate passage retrieval system which were constructed as an extension of [6] is provided for the participants for further use and comparison of the passage retrieval performance.

Originally the task was designed that the participants are free to use their own submission in the current module or baseline provided by the organizer to continue on to the next module, or stop altogether. This means it is possible to just participate in the document retrieval module of a single phase of QA lab. However, in the Phase 1, such combination runs were not possible and more coordination of the module structure is needed for further collaborations

2.1 Topics

Topics are provided in an XML format in both English and Japanese. The first part consists of true/false questions extracted from The National Center Test for University Admissions. The second part consists of factoid and complex, short answer questions extracted from secondary exams from 5 Japanese universities, including University of Tokyo. Figure 2 and 3 show examples of the Center Test and the Secondary Exam respectively. Table 2 shows tags used in the XML format, which originate from "torobo.dtd".

The formal run questions was made available at 0:00 (JST) of the first day of the Formal Runs at Phase 1 and Phase 2 at the Download Web site for each participating team.

2.1.1 Question Format

The questions were distributed in the following format.

For Center Test, one examination consists of several questions <question id=xxx>, and each question consists of instruction <instruction> which explaining the question, <data> text or several

Table 2: Tags of "torobo.dtd"

<exam> @source @subject @year	examination range source document subject year
<title>	title
<question> @id @minimal @answer_style @answer_type @knowledge_type @anscol	question range unique identification number "yes" if it has sub-questions. "no" otherwise question classification by answer styles question classification by answer types required knowledge ID list of answer columns
<instruction>	explanation of question
<data>	context and settings of question
<label>	number or symbol
<ansColumn> @id	answer column unique identification number
<choices> @anscol @comment	set of answer candidates ID list of answer columns comment
<choice> @comment	answer candidate Comment
<cNum>	number of the answer candidate
<uText> @id	text with underline unique identification number
<lText> @id	text with symbol (without underline) unique identification number
<blank> @id	blank position unique identification number
<ref> @target @comment	reference ID list of referred elements comment
 	new line position

passages describing the context and settings of the question, and several sub-questions. Each sub-question has an instruction and a set of multiple-choice answer candidates.

The <data> text usually just provides background information and/or context of the question and does not include the answer itself. The answers must be created using other knowledge resources.

For Secondary Exam, similar format to Center Text is used.

In the traditional question answering, a question is consisted of a single sentence with a rather standardized sentence structure like "Who is the prime minister of Japan?", "Who is Mr. Abe?" and so on. In the contrast to them, the questions appeared in the examinations are usually consist of multiple sentences, often require to understand the context from the instruction and data text. These are part of the challenging attributes that we have been tacking through this pilot task. And the technologies to solve such real world questions (understanding a question consists of multiple sentences and using context) shall be applicable for wide range of real world question answering applications in the future.

2.1.2 DTD

"torobo.dtd" is included in the "sample_questions_for_center_test.zip" file, which is available at the Download Website for each participating team. This will be included in the NTCIR-11 QA-Lab Test Collection.

2.1.3 Sample XML Format for the Question

A full sample XML file is included in the "sample_questions_for_center_test.zip" file, which is available at the Download Website for Each participating team and will be included in the NTCIR-11 QA-Lab Test Collection.

The below is an abridged sample. The tag structure is the same in English questions.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE exam PUBLIC "-//TOROBO//TOROBO ANNOTATION 1.0//EN" "torobo.dtd">
<?xml-stylesheet type="text/css" href="../torobo.css"?>

<exam source="National Center For University Entrance Examination" subject="SekaishiB(main exam)" year="2009">
Center-2009--Main-SekaishiB<br/>
<title>
2009 年度 本試験 世界史Ｂ<br/><br/>
</title>
<question id="Q1" minimal="no">
<label>【１】</label>
<instruction>
<br/><br/> 人類が営む生業と労働は，経済・社会・政治の動きと密接にかかわりながら，大きく変容してきた。生業と労働の歴史について述べた次の文章Ａ～Ｃを読み，以下の問い(問１～９)に答えよ。<br/> (配点 25)<br/>
</instruction>
<data id="D0" type="text">
<label>Ａ</label><br/> 清の学者趙翼は，明代の文化人の趨勢を論じて，<uText id="U1"><label>(1)</label>唐宋以来，文化・芸術に秀でた者の多くは科挙の合格者であった</uText>が，<uText id="U2"><label>(2)</label>明代</uText>になってその担い手は在野の人物に移っていったと述べている。明代中期の画家唐寅は，まさにその過渡期の人物と言える。彼は科挙で優秀な成績を収めながらも，不運な事件に巻き込まれ，栄達の道を絶たれてからは，蘇州で画業をなりわいとしながら自由奔放な生活を送った。明代中期から後期にかけて，在野の芸術家や文筆家が続々と現れたのは，<uText id="U3"><label>(3)</label>江南を中心とする商工業の発展</uText>によって都市の文化が成熟し，絵画や出版物が広く商品としての価値を持つようになったからであった。<br/><br/>
</data>
<question anscol="A1" answer_style="multipleChoice" answer_type="sentence" id="Q2" knowledge_type="KS" minimal="yes">
<label>問 1 </label>
<instruction>
下線部<ref comment="" target="U1">(1)</ref>に関連して，次に挙げる人物は，いずれも唐代から宋代にかけての科挙の合格者である。それぞれの人物について述べた文として正しいものを，次の①～④のうちから一つ選べ。
</instruction>
<ansColumn id="A1"> 1 </ansColumn><br/>
<choices anscol="A1" comment="">
<choice ansnum="1">
<cNum>①</cNum> 欧陽脩や蘇軾は，唐代を代表する文筆家である。</choice>
<choice ansnum="2">
<cNum>②</cNum> 顔真卿は，宋代を代表する書家である。</choice>
<choice ansnum="3">
<cNum>③</cNum> 宋の王安石は，新法と呼ばれる改革を行った。</choice>
<choice ansnum="4">
<cNum>④</cNum> 秦檜は，元との関係をめぐり主戦派と対立した。</choice><br/></choices>
</question>
.....
</exam>
```

2.2 Gold Standard （Right Answers）

The Gold Standard and the scores for each question for Center Test were provided by the National Center Test for University Admissions. Table 3 shows tags used in the Gold Standard, which originate from "answerTable.dtd".

2.2.1 Gold Standard Format (DTD)

"answerTable.dtd" is available from the NTCIR-11 QA Lab test collection.

Table 3: Tags of "answerTable.dtd"

<answerTable> @filename	set of answers question XML filename
<data>	answer range
<section>	label of question group
<question>	label of question
<answer_column>	label of answer column
<answer>	number of the right answer candidate
<score>	points of the question
<answer_type>	@answer_type of <question>
<answer_style>	@answer_style of <question>
<knowledge_type>	@knowledge_type of <question>
<question_ID>	@id of <question>
<anscolumn_ID>	@id of <ansColumn>

Sample XML Format is as follows. A full XML files were available in the training data, "answer-0625training-center_test (1997,2001,2005,2009).zip" file, which is included in the NTCIR-11 QA-Lab test collection. The submission format for the final results is the same.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE answerTable SYSTEM "http://21robot.org/answerTable.dtd">
<answerTable filename="Center-2009--Main-SekaishiB">
<data>
<section>第 1 問</section>
<question>1</question>
<answer_column>1</answer_column>
<answer>3</answer>
<score>3</score>
<answer_type>sentence</answer_type>
<answer_style>multipleChoice</answer_style>
<knowledge_type>KS</knowledge_type>
<question_ID>Q3</question_ID>
<anscolumn_ID>A2</anscolumn_ID>
</data>
.....
</answerTable>
```

2.2.2 Scorer

To assess and evaluate the submitted run results, the scorer was prepared. For the center examination, the score for each question was defined by the National Center Test for University Exam and we followed the scores. One exam include about 30-40 sub-questions to be answered. The total score for one exam is in the range of 0 to 100.

2.3 Collections

Participants are free to use any resources available with the exception of the answer sets (readily available online in Japanese). In addition, the following resources are provided, but are not required to be used.

A) Japanese high school textbooks on world history, available in Japanese

B) A snapshot of Wikipedia, available in Japanese and in English. (Participants can also use the current up-to-date version)

C) World history ontology.

D) An ontology-annotated textbook.

2.4 Baseline Systems

UIMA-based Module-base QA System for module-based participation.

A) 1 baseline system in Japanese (Javelin from CMU and MinerVa from YNU)

B) 1 baseline system in English (CMU)

C) A wide-range of language annotation tools through Kachaco

D) NTCIR RITE resources and tools

2.4.1 Japanese Baseline Systems

KJP provided a toolkit as a baseline system for Japanese QA. This baseline system is implemented in a modular way allowing partial reuse. Each module is compliant with UIMA [7], and the modules are ready-to-run based on the Kachako [8] platform's automation feature.

UIMA is an open framework for interoperability of unstructured information software modules. UIMA provides a range of metadata schemes and processing engines, that are available as an Apache UIMA open source project [9]. Kachako is an integrated NLP platform and compatible components that are compliant with UIMA. Kachako aims to provide a couple of automation features for easy creation and evaluation of components in a modular way [10].

The QA modules in our baseline system is originally developed as Minerva [3] and Javelin [4], which are Japanese factoid QA system. Minerva's original Perl code was re-implemented in Java. Javelin was originally developed in Java. Both system was re-organized to be UIMA and Kachako compliant components where modules' I/O data type specifications are compatible with Kachako's data types.

We provided an end-to-end baseline solver for the History subjects of the Japanese Center Exam using the QA modules described above. This baseline solver [1] is originally developed in Perl, then we re-implemented everything into Java, UIMA and Kachako compliant. This baseline solver converts given problems into factoid-style questions, then the QA modules are called, finally a most confident choice is selected using our scoring methods.

2.4.2 English Baseline Systems

The English Baseline system provided by CMU is an UIMA-based modular question answering (QA) pipeline [5] that automatically answers multiple-choice questions for the entrance exams on world history. The pipeline consists of a XML collection reader, a question and answer choice analyzer, a document retrieval based evidence collector, a rule-based answer selection, and evaluation CAS consumers. This baseline system can correctly answer up to 53 of all 153 questions from the provided four years (1997, 2001, 2005, and 2009) training datasets.

Given a topic, contextual information (a brief excerpt on the topic), and specific question instructions, the question analysis component generates verifiable assertions for each answer choice. The evidencing component then turns these assertions into search queries, and validates them by running the queries against an indexed collection such as Wikipedia. The most plausible answer choice is selected based on retrieval scores of found documents.

To facilitate future collaborative efforts for designing and implementing question answering systems for the world history exams, this baseline's type system, collection reader, QA phases, and evaluator can also be served as a modular software platform for evaluating component performance. We summarize the pipeline phases as follows:

Collection Reader parses the information from the input XML, and stores them as annotations in UIMA CASs that can be processed by UIMA pipelines easily.

Question and Answer Choice Analyzer synthesize the question, the answer choice, along with instructions and contexts, and then generates assertions which can be validated in subsequent modules. An assertion is a self-sufficient statement sentence composed by resolving references and appending contexts. Each analyzed answer choice will be associated with one or more assertions and whether each assertion is expected to be correct or incorrect.

Evidencer evaluates the soundness of assertions. The baseline system's evidencer creates a retrieval query from the assertion, runs the query over a Solr indexed Wikipedia corpus, obtains the scores

of one or more highly ranked documents, and generates an evidence score for the assertion.

Answer Selection makes decision and picks the final answer based on the evidence scores for each answer choice's assertions from all evidencing components. In this baseline system, all evidence scores of an answer choice will be summed up to be combined as final evidence score. If the question asks for which answer choice is correct, the answer choice with the highest final evidence score will be selected. Otherwise the lowest final evidence scored answer choice will be chosen. The baseline's simple answer selection module also includes evaluation functionalities that print out the final accuracy and final scores of selected answer choice.

2.5 Schedule

The NTCIR-11 QA-Lab Pilot task has been run according to the following timeline:

July 14, 2014: Training data release (1) Knowledge resource (Textbooks, (J) wikipedia, (EJ) Center Test (multiple Choice, Years 2009,2005,2001,1997, EJ)
Aug 1, 2014: Baseline System (QA Base system), Release)
Aug 7, 2014: Hands - on - Tutorials on UIMA (at NII, Rm 1904) Hiroshi Kanayama (IBM), Yoshinobu Kano (PREST/NII)
Aug, 2014: Training data release (2) Secondary Exam (Complex Questions, (2009, 2005, EJ)

PHASE 1

< Center Test >
Sept 23, 2014: Formal run topics release (from the download website for each team)
Sept 23 - 30, 2014: End-to-End Question-answering and IR runs
Oct 1 - Oct 6, 2014: Combination Runs
Oct 14: Draft paper submission to the Task organizers

PHASE 2

< Center Test >
Oct 28, 2014: Formal run Topics release
Oct 28 – Nov. 3, 2014: End-to-End QA and IR runs
Nov 4 – 10, 2014: Combination runs

< Secondary Exam >
Nov 11, 2014: Test Topics release
Nov 11 – 17, 2014: End-to-End QA and IR runs
Nov 17 – 24, 2014: Combination runs
Nov 30: Paper Submission for the Proceedings, which will be available online at the Conference.

NTCIR-11 CONFERENCE

Dec 9, 2014 (AM): Round-table meeting
Dec 9 - 12, 2014: NTCIR-11 Conference

3. PARTICIPATION

Fifteen groups from nine countries were registered to the task. Nine groups from seven countries are the active research groups QA-Lab Pilot Task as shown in Table 4.

4. SUBMISSIONS

4.1 Format

End-to-end question answering results were submitted in the format described in Section 2.2.1. Each team could submit up to three runs for each condition with the priority to be evaluated and analyzed.

Table 4. Active participating groups

Team ID	Organization
sJanta	Begum Rokeya University, Rangpur & The Graduate University of Advanced Studies
CMUQA	Carnegie Mellon University
DCUMT	Dublin City University
nnlp	Hokkaido University (All Hokkaido)
Forst	Yokohama National University
FRDC_QA	Fujitsu R&D Center Co., Ltd
NUL	Nihon Unisys, Ltd.
KJP	Shizuoka University / PRESTO / NII
FLL	Fujitsu Laboratories Ltd. & Fujitsu R&D Center Co., Ltd

Each run is associated with a RunID which is an identity for each run and the RunID is used as a filename of the run result to be submitted. Based on NTCIR CLIR and ACLIA format as a base, the RunID is defined as follows;

[Topic File Name]_[TeamID]_[Language]_[RunType]_[Priority].[FileType]

The topic file name is used without the extension (.xml).

Two character language codes are as follows;
- EN (English)
- JA (Japanese)

Two character RunType codes are as follows;
- QA (Question analysis module output) (XMI and/or XML)
- RS (Information Retrieval result) (XMI and/or XML)
- IX (Information Extraction from retrieved documents) (XMI only)
- FA (Final Answer of the End-to-End QA and Combination Run output) (XML only)
- SD (System Description Form) (Text file)

Priority Parameter;
The "Priority" is two digits used to represent the priority of the run, taking 01 as the highest. It is used as a parameter for pooling and priority to be evaluated and analyzed the results.

File Type:
Please indicate the file type using "xmi" or "xml".
For System Description Form, please use "txt"

For example, suppose TEAM1 submitted an Information Retrieval result with a RunID:

Center-2009--Main-WorldHistoryB_TEAM1_EN_RS_01.xml

4.2 .Submission Result

For the Phase 1 Formal run, 13 runs from 7 teams were submitted in total. For the Phase 2 Formal run, 18 runs from 9 teams were submitted in total. The detail is shown in Table 5.

Table 5. The number of submitted run for Phase 2

Team ID	Phase 1			Phase 2			
	Center Test			Center Test			Secondary Exam
	End-to-End		Combination	End-to-End		Combination	End-to-End
	Japanese	English		Japanese	English		Japanese
CMUQA	-	3	2	-	3	2	-
DCUMT	1	-	-	1	-	-	1
FLL	3	-	-	3	-	-	-
Forst	1	-	-	1	-	-	1
FRDC_QA	-	1**	-	-	1	-	-
KJP	1	-	-	1	-	-	-
nnlp	1	-	-	1	-	-	-
NUL	-	-	-	2	-	-	-
sJanta	-	-	-	-	1	-	-

Table 6: Total scores for the submitted runs (Phase 1)

End-to-End Run								
TeamID	Lang.	Priority	Without A14 (perfect score = 97)			With A14 (perfect score = 100)		
			Total Score	# of Correct	Rate of Correct	Total Score	# of Correct	Rate of Correct
DCUMT***	JA	01	74	27	0.77	77	28	0.78
KJP*	JA	01	57	20	0.57	57	20	0.55
CMUQA*	EN	01	48	17	0.49	48	17	0.47
Forst*	JA	01	46	16	0.46	49	17	0.47
CMUQA*	EN	02	45	16	0.46	45	16	0.44
CMUQA*	EN	03	43	15	0.43	43	15	0.42
FLL	JA	01	41	14	0.40	41	14	0.39
FRDC_QA**	EN	01	37	13	0.37	37	13	0.36
FLL	JA	02	34	12	0.34	34	12	0.33
Baseline	EN	01	33	12	0.34	33	12	0.33
nnlp*	JA	01	31	11	0.31	31	11	0.31
FLL	JA	03	23	8	0.23	23	8	0.22
Baseline	JA	01	22	8	0.23	22	8	0.22
Combination Run								
TeamID	Method		Without A14 (perfect score = 97)			With A14 (perfect score = 100)		
			Total Score	# of Correct	# of Incorrect	Total Score	# of Correct	# of Incorrect
CMUQA	CMUQA_only01		55	19	0.54	55	19	0.53
CMUQA	CMUQA_all		52	18	0.51	52	18	0.50

*: task organizer(s) are in the team **: late submission ***: The draft paper was submitted on December 10 that was the first day of the conference. Therefore, we had no time to peruse the paper. Because the system description of this draft paper was not clear, we asked the authors to revise their paper. The revised paper has not been received yet. This is still pending.

5. RESULTS

5.1 Phase 1

5.1.1 End-to-End Run (Center Test)

For Center Test questions, we evaluate a system by total score according to the allotment of points indicated at the Center Test. We also use the rate of correct answers CR by the following expression:

$$CR = \frac{number\ of\ correct\ answers}{total\ of\ questions}.$$

Because every question is allotted 2 or 3 points, the difference between them has little impact on evaluation.

Table 6 shows the total scores and the rate of correct answers in Phase 1. Because of the error in the tag for the Question 14 in Japanese version, we have discarded the Answer 14 from the evaluation. DCUMT used deep learning and obtained the highest

Table 7: Total scores for the submitted runs (Phase 2)

End-to-End Run (Center Test)					
TeamID	**Lang.**	**Priority**	**Total Score**	**# of Correct**	**Rate of Correct**
DCUMT***	JA	01	72	28	0.68
KJP*	JA	01	53	21	0.51
Forst*	JA	01	49	19	0.46
FLL	JA	01	48	19	0.46
FLL	JA	03	43	17	0.41
FLL	JA	02	41	17	0.41
NUL	JA	02	40	16	0.39
CMUQA*	EN	02	34	14	0.34
NUL	JA	01	33	13	0.32
CMUQA*	EN	01	32	13	0.32
CMUQA*	EN	03	30	12	0.29
Baseline	EN	01	29	12	0.29
Baseline	JA	01	23	10	0.24
sJanta	EN	01	21	9	0.22
FRDC_QA	EN	01	21	8	0.20
nnlp*	JA	01	18	7	0.17
Combination Run (Center Test)					
TeamID	**Method**		**Total Score**	**# of Correct**	**# of Incorrect**
CMUQA*	LogisticRegression		71	28	0.68
CMUQA*	BiasedVoting		65	25	0.61
End-to-End Run (Secondary Exam)					
TeamID	**Lang.**	**Priority**		**# of Correct**	**# of Incorrect**
			ROUGE-1 Score	**ROUGE-2 Score**	**ROUGE-L Score**
DCUMT***	JA	01			
			0.072	0.034	0.072
Forst*	JA	01			
			0.125	0.062	0.097

*: task organizer(s) are in the team ***: The draft paper was submitted on December 10 that was the first day of the conference. Therefore, we had no time to peruse the paper. Because the system description of this draft paper was not clear, we asked the authors to revise their paper. The revised paper has not been received yet. This is still pending.

scores 74 (among 97). For the detailed scores, please see Appendix 1.

5.1.2 Combination Run (Center Test)

In Phase 1, CMU experimented combination runs of submitted end-to-end results from all teams from Japanese and English. Without prior knowledge about other team's system, un-weighted voting is used to combine all team's final answer selections. The answer choice with majority of votes was chosen as combination run's output. The smaller answer choice number is preferred to break ties. Voting over all the submissions is called CMUQA_all. Voting over only first submissions of each team is called CMUQA_only01.

The CMUQA_only01 approach obtained the better results than the CMUQA_all.

5.2 Phase 2

5.2.1 End-to-End Run (Center Test)

For the Phase 2 Center Test end-to-end run, 14 runs submitted from 9 teams including 2 new entries. Table 7 shows the total scores and the rate of correct answers in Phase 2. The DCUMT system obtained the highest scores 72 (among 100). Ranking of top teams made little difference between Phase 1 and Phase 2. For the detailed scores, please see Appendix 2.

5.2.2 Combination Run (Center Test)

The CMU team used two different methods for the combination runs.

Biased Voting: Since the DCUMT system's performance was considerably better than other teams in phase 1, CMU simply gives the DCUMT system more voting power in a phase 2 combination run which called biased voting. Specifically, the CMU team uses the same setup as CMUQA_only01 combination run in Phase 1, except that CMU gives DCUMT three votes on its choice instead of 1. When applying biased voting on phase 1 submissions, CMU found that giving DCUMT 3 votes outperform 2 votes, but assigning more than 3 votes will result in identical output as the DCUMT submission.

Logistic Regression: Additionally, CMU attempts to learn the voting weights with logistic regression based on all system's final answer choices and gold standard data from phase 1. To that end, CMU reduces the multiple choice task to a classification problem of individual answer choice, and selecting the answer choice with the highest classification confidence. Therefore, each answer choice becomes one classification instance. The binary value of whether a system selecting this answer choice is one dimension of features. The classification label is whether current answer choice matches the gold standard. CMU uses logistic regression as the classifier, trains it with phase 1 data, uses the learned weights to predict an answer with phase 2 end-to-end submissions, and submits another combination run called logistic regression.

The logistic regression approach obtained the better results than the biased voting.

5.2.3 End-to-End Run (Second Exam)

The DCUMT and Forst teams submitted for the Phase 2 Secondary Exam run. Because Secondary Exam includes short answer questions, we evaluated parts of short answer questions using ROUGE measures. We used the past exam question collection Akahon[11] as a reference of ROUGE evaluation. The Forst system obtained the highest ROUGE scores while DCUMT obtained the most number of correct answers except for short answer questions. For the detailed scores except short answer questions, please see Appendix 3.

6. OUTLINE OF THE SYSTEMS

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

Table 8: The number of questions and the average of correct answer

Question Format	Phase 1		Phase 2			
	Center Test		Center Test		Secondary Exam	
	# of Questions	Ave. of Correct Ans.	# of Questions	Ave. of Correct	# of Questions	Ave. of Correct
Factoid	5 (0.139)	1.5 (0.293)	1 (0.024)	0.4 (0.429)	152 (0.658)	
Blank	5 (0.139)	2.3 (0.467)	4 (0.098)	1.5 (0.375)	43 (0.186)	
True/False	23 (0.639)	11.2 (0.487)	27 (0.659)	11.0 (0.407)	15 (0.065)	
True/False-Combo	1 (0.028)	0.1 (0.067)	3 (0.073)	0.7 (0.238)	0 (0.000)	-
Time	1 (0.028)	0.1 (0.133)	1 (0.024)	0.3 (0.285)	1 (0.004)	
Graph	1 (0.028)	0.0 (0.000)	5 (0.122)	1.3 (0.257)	1 (0.004)	
ShortAnswer-over100					3 (0.013)	
ShortAnswer-within100					16 (0.069)	
Total	36 (1.000)	15.2 (0.422)	41 (1.000)	15.2 (0.371)	231 (1.000)	

Table 9: Detailed scores per question format for the Center Test end-to-end runs

Team ID	Lang.	Priority	Factoid		Blank		True/False		True/False-Combo		Time		Graph	
			Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2
DCUMT	JA	01	5/5	1/1	4/5	3/4	17/23	21/27	1/1	1/3	1/1	1/1	0/1	1/5
KJP	JA	01	2/5	1/1	4/5	2/4	13/23	15/27	0/1	1/3	1/1	0/1	0/1	2/5
Forst	JA	01	2/5	1/1	4/5	2/4	11/23	15/27	0/1	0/3	0/1	0/1	0/1	1/5
FLL	JA	01	2/5	0/1	2/5	1/4	10/23	14/27	0/1	2/3	0/1	1/1	0/1	1/5
		02	2/5	0/1	2/5	1/4	8/23	10/27	0/1	2/3	0/1	1/1	0/1	3/5
		03	0/5	0/1	0/5	2/4	8/23	13/27	0/1	0/3	0/1	1/1	0/1	1/5
CMUQA	EN	01	1/5	0/1	3/5	2/4	13/23	11/27	0/1	0/3	0/1	0/1	0/1	0/5
		02	0/5	1/1	3/5	2/4	12/23	8/27	0/1	1/3	0/1	0/1	0/1	2/5
		03	1/5	1/1	3/5	2/4	12/23	8/27	0/1	1/3	0/1	0/1	0/1	0/5
nnlp	JA	01	0/5	0/1	1/5	1/4	10/23	5/27	0/1	1/3	0/1	0/1	0/1	0/5
FRDC_QA	EN	01	2/5	0/1	1/5	0/4	10/23	8/27	0/1	0/3	0/1	0/1	0/1	0/5
NUL	JA	01	-	1/1	-	1/4	-	8/27	-	0/3	-	0/1	-	3/5
		02	-	0/1	-	2/4	-	10/27	-	1/3	-	0/1	-	3/5
sJanta	EN	01	-	0/1	-	0/4	-	8/27	-	0/3	-	0/1	-	1/5

The DCUMT [1] system used case-frame graphs as semantic representation. The graphs were acquired from textbooks parsed by their semantic parser using commonsense knowledge. The system also handled implicit arguments/relations, causality analysis, time analysis, and temporal order analysis by heuristics. Thereby, the system obtained the highest scores except for Graph and short answer questions.

The KJP system is specialized to True/False questions. The team focuses on the way to handle keyword distribution as a simple, important, domain-independent and language-independent fundamentals. The system achieved higher performance by taking account of a penalty score to scattered keywords around various texts. The team also provided the Japanese baseline system described in Section 2.4.1.

The Forst team aims to apply QA systems to real-world problems. The system consists of dedicated modules for each question format and common modules called by the dedicated modules as necessary. The system obtained the highest ROUGE scores for short answer questions. For short answer questions, the team pointed out that answering by text as high-compress-ratio informative summarization is needed.

The FLL system took an approach changing combination patterns of three solvers according to question format. The first solver is based on different search engines using multiple knowledge sources. The second solver is trained with virtual examples that are

[1] The draft paper was submitted on December 10 that was the first day of the conference. Therefore, we had no time to peruse the paper. Because the system description of this draft paper was not clear, we asked the authors to revise their paper. The revised paper has not been received yet. This is still pending.

textbook sentences randomly replaced words in the sentences. The third solver is for answering chronological order of historical events.

The CMUQA system is specialized to multiple choice questions. The system answered by aggregating evidencing scores of short historical descriptions included in questions and answer candidates. For standardization of historical events, the team pointed out that co-reference resolution, disambiguation and implicit temporal reference resolution are needed. The team also conducted the combination runs using several voting methods. Besides, the team provided the English baseline system described in Section 2.4.2.

The nnlp system used a date identification method to check for temporal overlaps between time periods in questions and their answer candidates. The team pointed out that unknown question types may appear and that taking measure to unknown question types is needed.

The FRDC_QA system used different features and classification models to create a ranking system. The system has expansibility by adding more features and using more models. The team pointed out that more external resources are needed.

The NUL system used their entailment recognizer to solve entailment problems converted from questions by matching the type of question answer pairs. The team pointed out that identifying question types and recognizing relationships between entities are necessary.

The sJanta system is an open domain system based on online Wikipedia. The system is simple and does not need any training corpus.

7. DISCUSSION

Table 8 shows the number of questions and the average of correct answer overall. Table 9 shows the detailed scores per question format for the Center Test end-to-end runs. True/False-Combo questions differ from True/False questions in expectation of judging the truth of each answer candidate absolutely. Note that True/False questions can be answered if an answer candidate is more correct/wrong than other candidates. True/False questions account for more than 60% of questions in Center Test. Overall, greater number of True/False questions a team could answer, higher place the team ranked in. The True/False and Blank questions could be answered relatively. On the other hand, the True/False-Combo, Time and Graph questions leave a lot of rooms for improvement in future work. For the secondary exams including more various types of question than the Center Test, only two teams could submit the result. The difficulty of applying QA systems to real world was emphasized. We expect a break through using advance summarization, argument structure analysis, and/or, text generation in the future.

We planned a *horizontal* module-based pipeline, where consisting of "question analysis", "document retrieval", "information extraction", "answer generation", and provided two open-source UIMA module-based end-to-end QA systems for Japanese and English, and one open-source passage retrieval to enhance the module-based collaboration. However, submitted modules or intermediate results are few because they used many different types of pipelines according to question format types, answer types, and knowledge sources used. The simple *horizontal* component-based integration was well-worked in the past NTCIR's QA tasks with fixed format at NTCIR, but not for the real-world highly complex questions like exams. Therefore, *vertical* module-based runs according to question format type, answer-type and/or knowledge will be needed rather than the *horizontal* integration.

8. CONCLUSION

This paper introduced the overview of the first QA Lab (Question Answering Lab for Entrance Exam) task at NTCIR 11. The goal of the QA lab is to provide a module-based platform for advanced question answering systems and comparative evaluation for solving real-world university entrance exam questions. In this task, "world history" questions are selected from The National Center Test for University Admissions and from the secondary exams at 5 universities in Japan. For the Phase 1 Formal run, 13 runs from 7 teams were submitted in total. For the Phase 2 Formal run, 18 runs from 9 teams were submitted in total. We described the characteristic aspects of the participating groups' systems and their contribution.

9. ACKNOWLEDGMENTS

Our thanks to participants, data distributers, and the answer creaters. Part of the task organization was supported by NII's Todai Robot Project[12]

10. REFERENCES

[1] Ishioroshi, M., Kano, Y., Kando, N. 2014. A study of multiple choice problem solver using question answering system. IPSJ NL-215 research report. (in Japanese)

[2] Kano, Y. 2014. Materials delivered at the Hands-on Tutorial for UIMA and the QA Lab baseline systems.

[3] Mori, T. 2005. Japanese question-answering system using A* search and its improvement. ACM Trans. Asian Lang. Inf. Process. 4(3): 280-304

[4] Shima, H., Lao, N., Nyberg, E., Mitamura, T. 2008. Complex Cross-lingual Question Answering as Sequential Classification and Multi-Document Summarization Task. In NTCIR-7 Workshop.

[5] https://github.com/oaqa/ntcir-qalab-cmu-baseline

[6] https://code.google.com/p/passache/

[7] Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E. W., Hampp, T., et al. (2006) Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report.

[8] http://kachako.org/

[9] http://uima.apache.org/

[10] Kano, Y. 2012. Kachako: a Hybrid-Cloud Unstructured Information Platform for Full Automation of Service Composition, Scalable Deployment and Evaluation. In the 1st International Workshop on Analytics Services on the Cloud (ASC), the 10th International Conference on Services Oriented Computing (ICSOC 2012).

[11] http://akahon.net/

[12] http://21robot.org/

(On 21st July 2015, we removed footnotes from Table 5, and put footnotes to Table 6, Table 7 and Section 6.)

Appendix 1: Scores for Each Question in the Submitted Runs (Phase 1 Center Test End-to-End Task)

TEAM_ID	LANG	PRIORITY	Toatl Score	A02	A36	A26	A03	A24	A33	A01	A17	A19	A23	A34	A04	A16	A27	A28	A35	A31	A20	A09	A21	A30
				True False	True False	True False	True False	Blank Combo	True False	True False	True False	Blank Combo	True False	Blank	True False	True False	True False	Factoid	True False	True False	Factoid	True False	True False	True False
DCUMT	JA	1	74	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	3	correct	correct	correct	correct	1	correct	correct	correct	2
KJP	JA	1	57	correct	correct	correct	correct	correct	2	correct	correct	correct	3	correct	correct	2	correct	1	correct	1	correct	1	correct	correct
CMUQA	CMUQA_only01		55	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	1	correct	2	correct	2
CMUQA	CMUQA_all		52	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	1	2	3	correct
CMUQA	EN	1	48	correct	correct	correct	correct	correct	correct	3	correct	correct	3	correct	correct	correct	correct	correct	correct	correct	1	correct	3	4
Forst	JA	1	46	correct	correct	2	correct	correct	1	correct	2	4	correct	correct	3	correct	correct	correct	2	2	1	2	correct	correct
CMUQA	EN	3	45	correct	correct	correct	correct	correct	correct	3	correct	correct	3	correct	correct	4	correct	correct	2	correct	1	correct	3	4
CMUQA	EN	2	43	correct	correct	correct	correct	correct	correct	3	correct	correct	2	correct	correct	correct	correct	3	correct	correct	1	2	3	4
FLL	JA	1	41	correct	correct	correct	correct	correct	correct	correct	2	1	correct	3	correct	4	1	correct	3	correct	correct	2	3	2
FRDCQA	EN	1	37	correct	correct	correct	correct	4	correct	correct	4	4	correct	correct	3	correct	1	4	2	correct	correct	2	3	4
FLL	JA	2	34	correct	correct	3	3	correct	correct	3	3	1	correct	3	correct	4	1	correct	3	1	correct	1	3	correct
CMUQA	EN	Baseline	33	correct	correct	correct	correct	correct	correct	correct	correct	1	correct	2	3	2	1	1	2	2	1	correct	1	4
nnlp	JA	1	31	correct	correct	correct	2	N/A	1	correct	correct	correct	correct	1	1	correct	1	1	correct	2	N/A	2	1	2
FLL	JA	3	23	correct	correct	correct	correct	1	correct	correct	correct	1	3	1	1	4	1	1	2	2	1	1	3	correct
KJP	JA	Baseline	22	4	4	correct	2	3	2	1	4	correct	3	3	3	4	1	3	correct	correct	3	correct	correct	4
# of runs producing the correct answer				14	14	13	12	11	11	10	10	9	9	9	8	8	8	8	8	7	6	5	5	5
correct rate (%)				93.3	93.3	86.7	80.0	73.3	73.3	66.7	66.7	60.0	60.0	60.0	53.3	53.3	53.3	53.3	53.3	46.7	40.0	33.3	33.3	33.3
score allocated to the answer				3	3	3	3	3	2	3	3	3	3	3	3	3	3	3	2	3	3	3	3	3

TEAM_ID	LANG	PRIORITY	Toatl Score	A07	A13	A32	A05	A06	A10	A22	A25	A11	A12	A14*	A15	A29	A18	A08
				Blank	True False	True False	True False	True False	Factoid	Factoid	True False	True False	True False	Blank Combo	Factoid	Time	True False Combo	Graph
DCUMT	JA	1	74	1	3	4	correct	correct	correct	correct	correct	correct	2	correct	correct	correct	correct	3
KJP	JA	1	57	correct	1	4	1	correct	correct	2	2	correct	2	1	1	correct	2	5
CMUQA	CMUQA_only01		55	1	1	correct	1	3	1	1	2	1	2	1	1	1	1	1
CMUQA	CMUQA_all		52	1	1	2	1	3	1	1	4	3	2	3	2	1	1	1
CMUQA	EN	1	48	1	correct	2	1	3	3	1	2	3	2	4	4	1	1	1
Forst	JA	1	46	correct	2	2	correct	3	1	correct	4	1	correct	correct	2	3	1	3
CMUQA	EN	3	45	1	correct	2	1	3	4	3	correct	3	2	4	4	1	1	1
CMUQA	EN	2	43	1	3	2	3	3	4	2	4	3	correct	4	2	1	4	1
FLL	JA	1	41	correct	2	correct	1	2	1	3	4	3	2	3	2	1	4	1
FRDCQA	EN	1	37	4	3	2	4	2	correct	3	correct	4	2	4	2	3	4	3
FLL	JA	2	34	correct	1	correct	4	correct	1	3	4	3	4	3	2	1	4	1
CMUQA	EN	Baseline	33	1	correct	2	4	4	4	correct	1	1	2	3	2	1	1	1
nnlp	JA	1	31	1	1	correct	correct	3	1	1	1	1	1	1	1	N/A	1	N/A
FLL	JA	3	23	2	1	3	1	2	1	1	4	3	1	3	2	1	1	1
KJP	JA	Baseline	22	4	correct	2	3	4	3	N/A	4	1	1	3	correct	1	1	N/A
# of runs producing the correct answer				4	4	4	3	3	3	3	3	2	2	2	2	2	1	0
correct rate (%)				26.7	26.7	26.7	20.0	20.0	20.0	20.0	20.0	13.3	13.3	13.3	13.3	13.3	6.7	0.0
score allocated to the answer				3	2	3	2	2	3	2	2	2	3	3	3	3	3	3

*: Because of the error in the tag for the Question 14 in Japanese version, we have discarded the Answer 14 from the evaluation.

Appendix 2: Scores for Each Question in the Submitted Runs (Phase 2 Center Test Task)

TEAM_ID	LANG	PRIORITY	Total Score	A20	A24	A32	A01	A11	A12	A13	A07	A09	A18	A41	A27	A29	A33	A38	A39	A10	A16	A22	A28	A35
				True False	Blank Combo	True False	True False	True False Combo	Blank	True False	Graph	True False	True False	True False	True False	True False	True False	True False	True False	True False	True False	True False	True False	True False	Factoid
DCUMT	JA	1	72	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	3	correct	correct	correct	correct
CMUQA	Logistic Regression		71	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct
CMUQA	BiasedVoting		65	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	correct	3	correct	correct	correct	correct
KJP	JA	1	53	correct	correct	correct	1	4	correct	correct	4	4	correct	3	1	2	correct	correct	correct	correct	correct	correct	correct	correct
Forst	JA	1	49	correct	correct	correct	2	4	correct	correct	4	correct	1	correct	2	correct	correct	1	2	correct	1	correct	4	correct
FLL	JA	1	48	correct	correct	correct	correct	correct	4	correct	correct	4	correct	1	correct	correct	correct	correct	correct	3	correct	1	correct	1
FLL	JA	3	43	correct	3	correct	correct	1	correct	correct	correct	4	correct	1	correct	correct	correct	1	correct	3	correct	1	correct	2
FLL	JA	2	41	correct	correct	3	correct	correct	4	correct	correct	4	2	2	1	correct	correct	correct	1	correct	correct	1	2	1
NUL	JA	2	40	2	correct	4	correct	correct	correct	1	correct	4	correct	correct	4	4	correct	correct	correct	2	2	correct	correct	1
CMUQA	EN	2	34	correct	correct	correct	3	correct	correct	2	correct	correct	2	1	correct	correct	4	1	4	3	2	1	4	correct
NUL	JA	1	33	correct	correct	3	correct	4	3	4	correct	4	correct	correct	4	1	4	1	correct	3	correct	3	correct	correct
CMUQA	EN	1	32	correct	correct	correct	3	4	correct	correct	4	correct	2	correct	correct	4	4	1	1	correct	2	1	2	1
CMUQA	EN	3	30	correct	correct	correct	3	correct	correct	2	4	correct	correct	2	correct	correct	4	4	4	1	2	1	2	correct
CMUQA	EN	Baseline	29	4	3	correct	3	correct	correct	1	2	correct	2	correct	4	2	4	1	1	correct	2	correct	2	4
KJP	JA	Baseline	23	4	3	4	correct	correct	1	4	4	4	correct	correct	4	4	4	correct	4	correct	1	correct	4	N/A
FRDCQA	EN	1	21	2	4	correct	correct	4	4	correct	2	correct	2	1	1	4	4	3	correct	3	1	1	1	2
sJanta	EN	1	21	1	N/A	4	correct	N/A	N/A	correct	correct	correct	N/A	correct	1	4	1	correct	2	correct	2	2	4	1
nnlp	JA	1	18	correct	N/A	2	2	correct	1	1	2	1	1	3	correct	1	1	1	1	1	2	3	1	1
# of runs producing the correct answer				13	12	12	11	11	11	11	10	10	10	10	9	9	9	9	9	8	8	8	8	8
correct rate (%)				72.2	66.7	66.7	61.1	61.1	61.1	61.1	55.6	55.6	55.6	55.6	50.0	50.0	50.0	50.0	50.0	44.4	44.4	44.4	44.4	44.4
score allocated to the answer				3	3	2	3	3	3	3	2	3	2	2	2	2	3	2	3	2	3	2	3	3

TEAM_ID	LANG	PRIORITY	Total Score	A26	A15	A23	A31	A37	A02	A03	A04	A14	A25	A36	A08	A21	A05	A06	A17	A19	A30	A34	A40
				True False	True False	True False	True False	Time	True False	Graph	True False	True False	True False	True False	Graph	True False	Blank Combo	True False	True False Combo	Graph	Blank	True False Combo	Graph
DCUMT	JA	1	72	correct	correct	4	4	correct	correct	1	3	4	correct	correct	4	4	correct	correct	1	1	1	4	3
CMUQA	Logistic Regression		71	correct	correct	4	4	correct	correct	1	3	4	correct	correct	4	4	2	correct	1	1	1	1	3
CMUQA	BiasedVoting		65	correct	correct	4	4	correct	1	1	3	4	3	correct	4	4	correct	1	1	1	1	4	3
KJP	JA	1	53	correct	correct	correct	1	4	3	2	3	correct	3	3	4	3	2	1	1	correct	1	correct	correct
Forst	JA	1	49	correct	2	1	correct	3	correct	2	1	3	correct	correct	3	correct	4	4	3	correct	4	3	3
FLL	JA	1	48	1	4	2	correct	correct	1	2	3	3	3	2	4	correct	2	1	correct	1	1	1	1
FLL	JA	3	43	1	4	2	correct	correct	1	2	3	3	3	1	4	correct	2	1	4	1	correct	4	1
FLL	JA	2	41	3	4	4	4	correct	4	correct	3	correct	correct	3	correct	3	2	2	correct	0	1	1	1
NUL	JA	2	40	correct	4	correct	4	N/A	1	correct	2	2	correct	2	correct	4	4	1	1	N/A	4	1	N/A
CMUQA	EN	2	34	1	1	correct	4	1	correct	correct	correct	4	3	4	4	1	2	1	1	1	4	1	1
NUL	JA	1	33	3	2	correct	4	N/A	1	correct	3	4	2	3	correct	4	2	1	4	N/A	1	4	N/A
CMUQA	EN	1	32	4	1	correct	correct	1	correct	2	correct	4	4	4	4	3	2	1	4	1	4	4	1
CMUQA	EN	3	30	1	1	correct	4	1	1	2	correct	4	3	4	4	1	2	1	1	1	4	1	1
CMUQA	EN	Baseline	29	4	1	4	correct	1	1	correct	1	correct	3	3	4	4	2	correct	1	1	correct	1	1
KJP	JA	Baseline	23	correct	4	2	4	1	4	1	1	correct	3	2	correct	1	3	4	1	1	3	1	1
FRDCQA	EN	1	21	4	4	4	correct	3	1	2	correct	2	3	4	4	correct	4	1	4	4	1	4	3
sJanta	EN	1	21	1	correct	2	4	N/A	1	1	correct	3	3	3	3	4	N/A	1	N/A	N/A	N/A	N/A	N/A
nnlp	JA	1	18	1	correct	1	1	N/A	1	1	2	correct	2	correct	3	1	correct	1	1	N/A	1	1	N/A
# of runs producing the correct answer				7	6	6	6	6	5	5	5	5	5	5	4	4	3	3	2	2	2	1	1
correct rate (%)				38.9	33.3	33.3	33.3	33.3	27.8	27.8	27.8	27.8	27.8	27.8	22.2	22.2	16.7	16.7	11.1	11.1	11.1	5.6	5.6
score allocated to the answer				3	2	2	3	2	2	2	2	2	2	3	2	2	3	3	2	3	2	2	2

Appendix 3: Scores for Each Question except Short Answer in the Submitted Runs (Phase 2 Secondary Exam End-to-End Task)

The University of Tokyo		
	DCUMT	Forst
Q11	incorrect	incorrect
Q13	N/A	incorrect
Q14	correct	correct
Q15	incorrect	correct
Q16	correct	correct
Q17a	N/A	incorrect
Q17b	N/A	incorrect
Q18	N/A	incorrect
Q19a	N/A	incorrect
Q19b	N/A	incorrect
Q20	correct	incorrect
Q22	correct	incorrect
Q23	correct	incorrect
# of correct answers		
	5	3
# of incorrect answers		
	2	10
# of N/A		
	6	0

Kyoto University		
	DCUMT	Forst
Q2a	correct	N/A
Q2b	correct	incorrect
Q2c	correct	incorrect
Q2d	correct	incorrect
Q2e	correct	incorrect
Q2f	N/A	N/A
Q2g	correct	incorrect
Q2h	correct	incorrect
Q2i	incorrect	incorrect
Q2j	incorrect	incorrect
Q2k	correct	incorrect
Q2l	correct	incorrect
Q2m	incorrect	N/A
Q2n	correct	incorrect
Q2o	incorrect	N/A
Q4	correct	correct
Q5	correct	correct
Q6	correct	correct
Q7	incorrect	incorrect
Q9	incorrect	correct
Q10	correct	correct
Q11	N/A	correct
Q13	correct	incorrect
Q14	incorrect	incorrect
Q15	correct	incorrect
Q16	correct	incorrect
Q17	correct	incorrect
Q18	incorrect	incorrect
Q19	correct	incorrect
Q20	correct	incorrect
Q22a	correct	incorrect
Q22b	N/A	correct
Q24	correct	incorrect
Q25	incorrect	incorrect
Q28	incorrect	incorrect
Q29	incorrect	correct
Q31	correct	incorrect
Q32	incorrect	incorrect
Q35	correct	incorrect
Q36	correct	incorrect
Q37	incorrect	incorrect
Q38	correct	incorrect
Q39	correct	incorrect
Q40	incorrect	correct
Q41	incorrect	correct
Q42	correct	incorrect
Q44	incorrect	incorrect
Q47	incorrect	incorrect
Q48	correct	correct
Q49	N/A	incorrect
Q50	incorrect	incorrect
Q51	N/A	incorrect
Q52	correct	incorrect
# of correct answers		
	30	11
# of incorrect answers		
	18	38
# of N/A		
	5	4

Hokkaido University		
	DCUMT	Forst
Q2a	incorrect	N/A
Q2b	correct	N/A
Q2c	correct	N/A
Q2d	correct	incorrect
Q3a	correct	incorrect
Q3b	correct	incorrect
Q5	correct	correct
Q6	correct	incorrect
Q6	correct	incorrect
Q7a	incorrect	incorrect
Q7b	N/A	incorrect
Q8	correct	incorrect
Q11a	correct	incorrect
Q11b	correct	N/A
Q11c	correct	N/A
Q11d	N/A	incorrect
Q11e	N/A	incorrect
Q11f	incorrect	N/A
Q11g	correct	N/A
Q11h	correct	incorrect
Q11i	correct	incorrect
Q11j	correct	N/A
Q12	N/A	N/A
Q14a	incorrect	incorrect
Q14b	correct	correct
Q14c	N/A	incorrect
Q15	incorrect	incorrect
Q16	N/A	incorrect
Q18a	incorrect	incorrect
Q18b	incorrect	incorrect
Q18c	N/A	incorrect
Q18d	N/A	incorrect
Q18e	correct	incorrect
Q19	N/A	incorrect
Q22a	correct	N/A
Q22b	incorrect	N/A
Q22c	N/A	incorrect
Q22d	correct	N/A
Q23	correct	incorrect
Q27	incorrect	incorrect
Q28	correct	correct
# of correct answers		
	23	3
# of incorrect answers		
	9	27
# of N/A		
	10	12

Waseda University		
	DCUMT	Forst
Q2	N/A	N/A
Q3	correct	N/A
Q4	correct	N/A
Q5	b	N/A
Q6	b	N/A
Q7	c	N/A
Q8	correct	N/A
Q9	correct	N/A
Q10	b	N/A
Q11	d	correct
Q12	incorrect	correct
Q13	incorrect	correct
Q15	N/A	N/A
Q16	correct	correct
Q17	correct	N/A
Q18	correct	N/A
Q19	correct	N/A
Q20	a	N/A
Q21	correct	N/A
Q22	correct	N/A
Q23	correct	N/A
Q24	correct	N/A
Q25	correct	correct
Q26	N/A	correct
Q27	correct	correct
Q29	N/A	N/A
Q30	correct	correct
Q31	correct	N/A
Q32	correct	correct
Q33	b	N/A
Q34	b	N/A
Q35	correct	N/A
Q36	correct	N/A
Q37	correct	N/A
Q38	correct	correct
Q39	correct	correct
Q40	correct	correct
Q42a	N/A	correct
Q42b	N/A	N/A
Q42c	N/A	N/A
Q44	correct	N/A
Q45	correct	correct
Q46	b	N/A
Q47	b	correct
Q48	a	N/A
Q49	correct	correct
Q50	e	correct
Q51	a	N/A
Q52	c	b
Q53	correct	N/A
# of correct answers		
	27	17
# of incorrect answers		
	16	1
# of N/A		
	7	32

Chuo University		
	DCUMT	Forst
Q3	correct	incorrect
Q4	correct	correct
Q5	correct	incorrect
Q6	incorrect	correct
Q7a	incorrect	incorrect
Q7b	N/A	incorrect
Q8a	correct	correct
Q8b	correct	incorrect
Q9	correct	correct
Q10	incorrect	incorrect
Q11	correct	correct
Q12a	N/A	incorrect
Q12b	N/A	incorrect
Q13	correct	incorrect
Q14	correct	incorrect
Q15	correct	correct
Q16	incorrect	correct
Q17	correct	incorrect
Q18	correct	incorrect
Q19	correct	incorrect
Q20	incorrect	correct
Q21	correct	incorrect
Q22	correct	correct
Q23	N/A	incorrect
Q24	N/A	correct
Q25	correct	correct
Q26	correct	incorrect
Q27a	correct	incorrect
Q27b	N/A	incorrect
Q30	correct	incorrect
Q31	N/A	incorrect
Q32	incorrect	incorrect
Q33	correct	incorrect
Q34	correct	incorrect
Q35	correct	incorrect
Q36	N/A	incorrect
Q37	correct	incorrect
Q38	incorrect	incorrect
Q39	correct	incorrect
Q40	N/A	incorrect
Q41	correct	incorrect
Q42	correct	incorrect
Q43	correct	incorrect
Q44	correct	incorrect
Q45a	N/A	incorrect
Q45b	N/A	incorrect
Q45c	N/A	incorrect
Q45d	N/A	incorrect
Q45e	N/A	N/A
Q47	correct	incorrect
Q48	correct	incorrect
Q49	correct	incorrect
Q50	correct	correct
Q51	incorrect	incorrect
# of correct answers		
	32	12
# of incorrect answers		
	8	41
# of N/A		
	14	1