NTCIR

Overview of the NTCIR-13 QA Lab-3 Task

Hideyuki Shibuki*1, Kotaro Sakamoto*1,*2, Madoka Ishioroshi*2, Yoshionobu Kano*3, Teruko Mitamura*4, Tatsunori Mori*1, Noriko Kando*2,*5

*1: Yokohama National University, *2: National Institute of Informatics, *3: Shizuoka University, *4: Carnegie Mellon University, *5: The Graduate University for Advanced Studies (SOKENDAI)

NTCIR

QA Lab

Introduction

- Goal
  - investigation of the **real-world complex Question Answering (QA)** technologies
  - as a joint effort of participants and appropriate evaluation metrics and methodologies for them
  - using Japanese **university entrance exams** and their English translation on the subject of “World history”

History of QA at NTCIR

NTCIR	Years	TSC Summarization)	QAC	CLQA	ACLIA	RITE	QA Lab	CLEF QA
					Module	Entailment		
2	2000-2001 feb	Single doc						
3	2001-2002 oct	Multi-doc	Factoid,List, Series					
4	2003-2004 june	Multi-doc	Factoid, Dialog(IAD)					
5	2004-2005 dec		Factoid, Dialog(IAD)	Factoid				
6	2006-2007 june	TREND Info	Complex	Factoid				
7	2007-2008 dec	TREND Info			Complex			
8	2009-2010 june				Complex			
9	2010-2011 dec					RITE		
10	2012-2013 june					RITE		Exam - Reading Comprehension
11	2013-2014 dec					RITEVAL	Exam	
12	2015-2016 june						Exam	
13	2016-2017dec						Exam	

Todai Robot
Apri 2011-2016

Overview of QA Lab-2 (1/2)

- Tasks 2 phases + mock exam
  - multiple-choice type, EN & JA (National Center Test)
  - free-description Question, EN & JA
  - mock exam
    - multiple-choice type, JA, Benesse ca. 430,000 students, 80%
    - multiple-choice type, JA, Yozemi ca.3500 students,
    - free-description type, JA, Sundai ca.10,000 Student Average+
- Multiple-choice type questions
  - 12 teams participated
  - Good results
- Free-description type questions
  - 3 teams participated

Compare with real human high school students

Unsatisfactory results Why?

Overview of QA Lab-2 (2/2)

- Free-description type questions
  - Named-Entity questions
    - Similar to factoid QA
    - Relatively good results
  - Essay questions
    - Complex and laborious task
    - Similar to query-biased multi-document summarization
    - Automated evaluation
    - There are many open problems

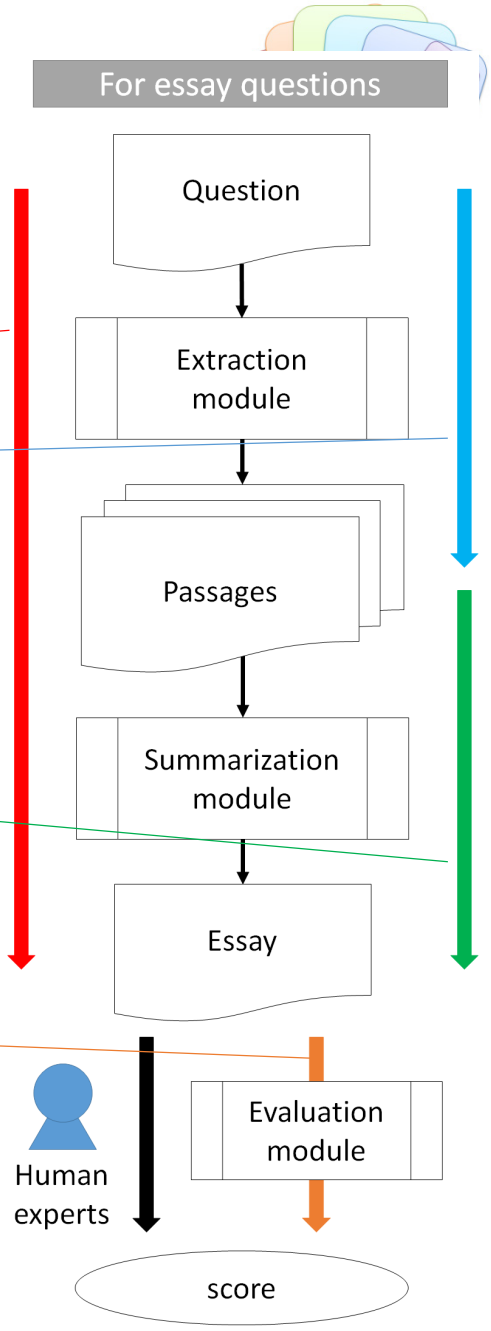
QA Lab-3 focused on Essay questions

What's new in the QA Lab-3

- Restructuring tasks based on question types
  - Multiple-choice question task
  - Term (Named-Entity) question task
  - Essay question task
- Breaking down Essay Question Task into manageable subtasks
  - Extraction subtask and Summarization subtask
  - Evaluation-method subtask for automated evaluation
- Research run for the progress so far

QA Lab 3 Breaking down essay questions

- End-to-End task
- Extraction subtask
  - First half of the end-to-end task
  - IN: Question
  - OUT: Passages including texts of Gold Standard Data
- Summarization subtask
  - Second half of the end-to-end task
  - IN: Question + Passages
  - OUT: Essay
- Evaluation-method subtask
  - For automated evaluation
  - IN: Question + Essays + Gold Standard Data
  - OUT: Ranking of Essays (with score)



Research run

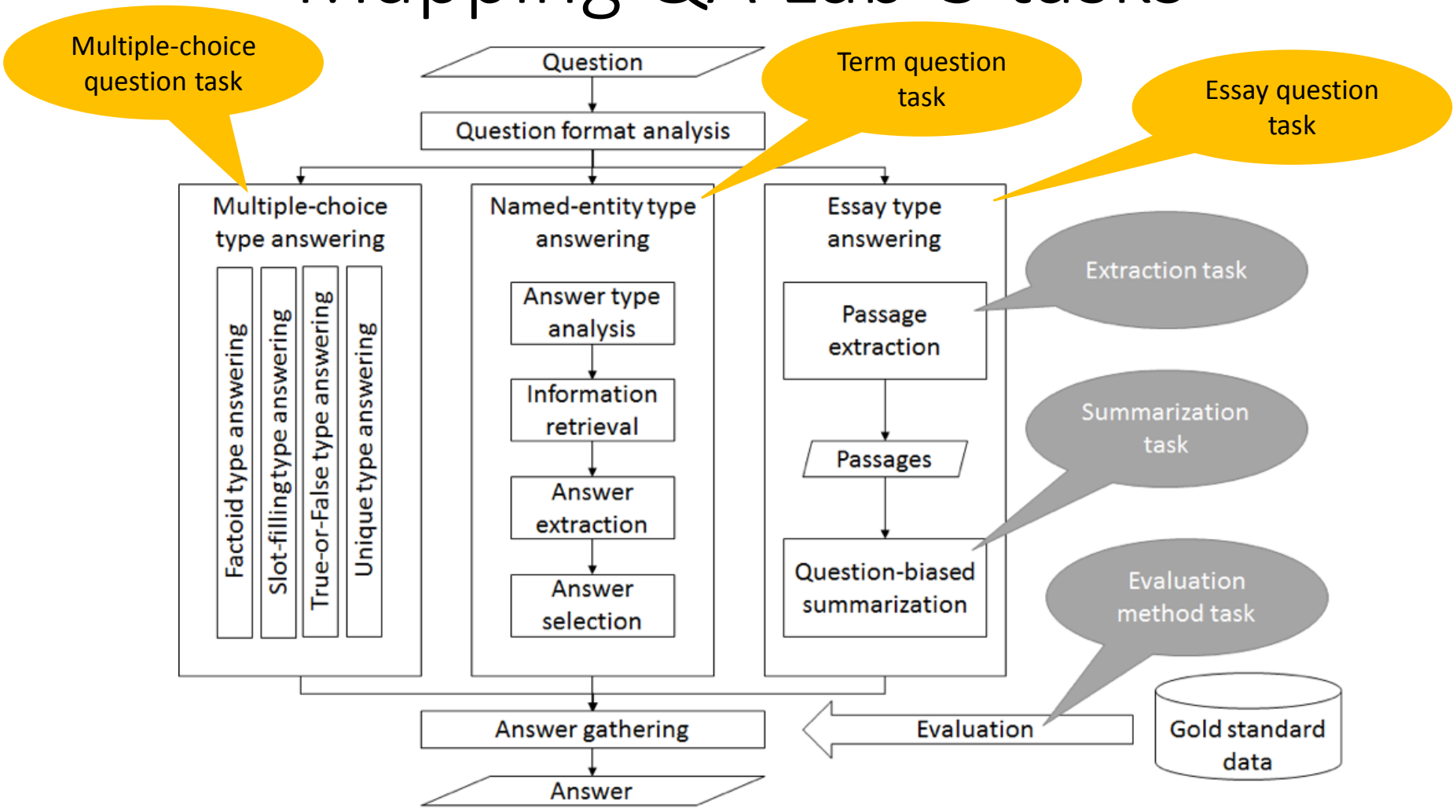
- How much did the QA technologies improved from QA Lab-1?
  - Using the same training/test sets as the past QA Lab runs, comparison with the past results
  - Only Multiple-choice and Essay end-to-end tasks

Task description

- Multiple-choice question task
- Term question task
- For Essay
  - End-to-End task
  - Extraction task
    - aiming to retrieve and extract texts that should be included in essay
  - Summarization task
    - aiming to generate an essay by summarizing the extracted texts
  - Evaluation-method task
    - aiming to automatically evaluate essays systems generated using gold standard essays

6 tasks in total

QA System Architecture Mapping QA Lab-3 tasks



Schedule

- Training data (EN & JA)
  - Jul 1, 2016: Released
- Phase 1 (EN & JA)
  - Feb 2-6, 2017: Term and Multiple-choice tasks
  - Feb 9-13, 2017: Essay End-to-End and Extraction tasks
  - Feb 16-20, 2017: Essay Summarization task
  - Feb 23 - Mar 1, 2017: Essay Evaluation-method task
- Phase 2 (EN & JA)
  - May 11-15, 2017: Term and Multiple-choice tasks
  - May 18-22, 2017: Essay End-to-End and Extraction tasks
  - May 25-29, 2017: Essay Summarization task
  - Jun 1-5, 2017: Essay Evaluation-method task
- Research Run (EN & JA)
  - Jul 6-10, 2017: Essay End-to-End and Multiple-choice tasks

Tasks in Each Phase

Question	Task	Phase-1	Phase-2	Research run
Multiple-choice	End-to-end	YES	YES	YES
Term	End-to-End	YES	YES	N/A
Essay	End-to-End	YES	YES	YES
	Extraction	YES	YES	N/A
	Summarization	YES	YES	N/A
	Evaluation-method	YES	YES	N/A

- Participants are free to participate any particular phase and either of exams.

NTCIR

QA Lab

Training and Test sets in Each Phase

Task	Formal run			Research run	
	Training	Phase-1	Phase-2	Training	Test
Multiple-choice	1997,1999,2001 2003,2005,2007 2009,2011	2012,2013	2014	1997,1999,2001 2003,2005,2007 2009,2011	2007,2011,2013
Term & Essay	2003,2005,2007 2009,2011	2000,2004,2008 2012,2013	2001,2002,2006 2010,2014	2000 to 2014	2002,2007,2013

- Multiple-choice questions
  - selected from the National Center Test for University Admissions
- Term and Essay questions
  - selected from secondary exams of the University of Tokyo

Question XML Format

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(ちょうよく)は，明代の文化人の趨勢(すうせい)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(とういん)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。

```
<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>
```

Question XML Format 1 (multiple choice)

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。

Questions (Multi-Sentence, Suggest Context)

Context

Sub-Questions

Multiple Choices

```
year= 2009 >
Center-2009--Main-SekaishiB<br/>
<title>
2009年度 本試験 世界史B<br/><br/>
</title>
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
<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/>
<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>
```

Ex. Complex Essay Type Question (U Tokyo 2005)

Q1. Throughout human history, while wars have brought about a great deal of anguish and tragedy, they have also given impetus to various kinds of initiatives in pursuit of peace and liberation beyond those calamities.

Broad Theme,

Even before World War II ended in 1945, the nations that comprised the Allies in that conflict had envisaged a variety of post-war scenarios with a view tocreation of a new framework for international order, including the concept of the United Nations. However, formation of the United Nations Organization (UN) was not enough to immediately bring about world peace. The confrontation *between the United States and the Soviet Union was linked to local nationalist movements,* giving rise to new conflicts. For instance, in China, a power struggle between the Nationalist Party and the Communist Party intensified during the course of the war against Japan, and became a factor in development of the so-called "Cold War" following the end of World War 2.

Background

How did events that occurred during World War II affect the post-war world up to the 1950s? Explain your conclusion. Make sure that you use each of the eight keywords listed below at least once, and underline them.

Question

Atlantic Charter, Constitution of Japan, Taiwan, Kim Il-Sung, East Germany, EEC (European Economic Community), Auschwitz, Palestinian refugees

KeyWords

NTCIR

QA Lab

Gold standard creation

- For Term questions
  - Several answers if there are different expressions
- For Essay questions
  - Reference complex essays written by three human experts
  - Reference simple essays written by a human expert
  - Nuggets extracted from references and assigned a weight [0(1)-3], and voted by three human experts [1-9]

Evaluation (1/2)

- For Term questions
  - Exact match
- For Essay questions (End-to-End)
  - Human expert's mark
  - Pyramid method
    - Judgment by participants
  - ROUGE-1 and -2 method
    - Morphology without stemming (JA)
    - Word without stemming (EN)
  - Quality questions
    - 4-level scale
    - Do not ascertain the truth of the essay
    - Grammaticality, Non-redundancy, Reference, Fluency, 'Coherence and content structure'

Evaluation (2/2)

- For Extraction task
  - Precision and recall of extracted texts including statements in Gold standard essay
- For Summarization task
  - Same as End-to-End task
- For Evaluation-method task
  - Rank correlation coefficient with human expert ranking

NTCIR

QA Lab

Collection

- 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.
  - Eight sets of National Center Tests (JA & EN)
  - Five sets of Second-stage Examinations (JA & EN)
  - Knowledge Sources (a snapshot of Wikipedia subset related to world history)
  - Right Answers

Knowledge Sources

- Two Japanese high school textbooks on world history, available in Japanese.
- A snapshot of Wikipedia, available in Japanese and in English. (Participants can also use the current up-to-date version).
  - Solr Instance with Indexed Wikipedia Subset (available in English) https://github.com/oaqa/ntcir-qalab-cmu-baseline/wiki/Solr-Instance-with-Indexed-Wikipedia-Subset
  - NTCIR-11 QA Lab Japanese subtask: Wikipedia Data Set http://warehouse.ntcir.nii.ac.jp/openaccess/qalab/11QALab-ja-wikipediadata.html
- World history ontology, available in Japanese. http://researchmap.jp/zoeai/event-ontology-EVT/

NTCIR

QA Lab

Right Answers

- Right answers for National Center Tests, available in English and Japanese.
- Right answers for Second-stage Examinations, available in English and Japanese.
- Reference essays and nuggets for Essays, available in Japanese.

NTCIR

QA Lab

Tools

- 1 baseline QA system for English, based on UIMA (CMU) https://github.com/oaqa/ntcir-qalab-cmu-baseline
- 1 baseline QA system for Japanese, based on YNU's MinerVA, CMU's Javelin and a question analysis module by Madoka Ishioroshi¥cite{Ishioroshi2014}, re-constructed and implemented as UIMA components by Yoshinobu Kano https://bitbucket.org/ntcirqalab/factoidqa-centerexam/
- Scorer and Format Checker for National Center Test https://bitbucket.org/ntcirqalab/qalabsimplescorer
- Passage Retrieval Engine passache https://code.google.com/p/passache/

For English Subtask

- The same content questions as Japanese ones
  - Translation from Japanese questions
  - Length limitation of essay was divided into a half by heuristics between Japanese characters and English words
    - Ex. 100 Japanese characters -> 50 English words
- Resources are different
  - No high school textbooks
  - No world history ontology
  - Larger size of Wikipedia

NTCIR

QA Lab

Active participating teams

- The following 11 teams

Team ID	Organization
KUAS	National Kaohsiung University of Applied Sciences
Forst	Yokohama National University
IMTKU	Tamkang University
SML	Nagoya University
KSU	Kyoto Sangyo University
SLQAL	Waseda University
CMUQA	Carnegie Mellon University
DGLab	DG Lab
tmkff	The National Center for University Entrance Examinations & Kyushu University
MTMT	Carnegie Mellon University
HagiL	Keio University

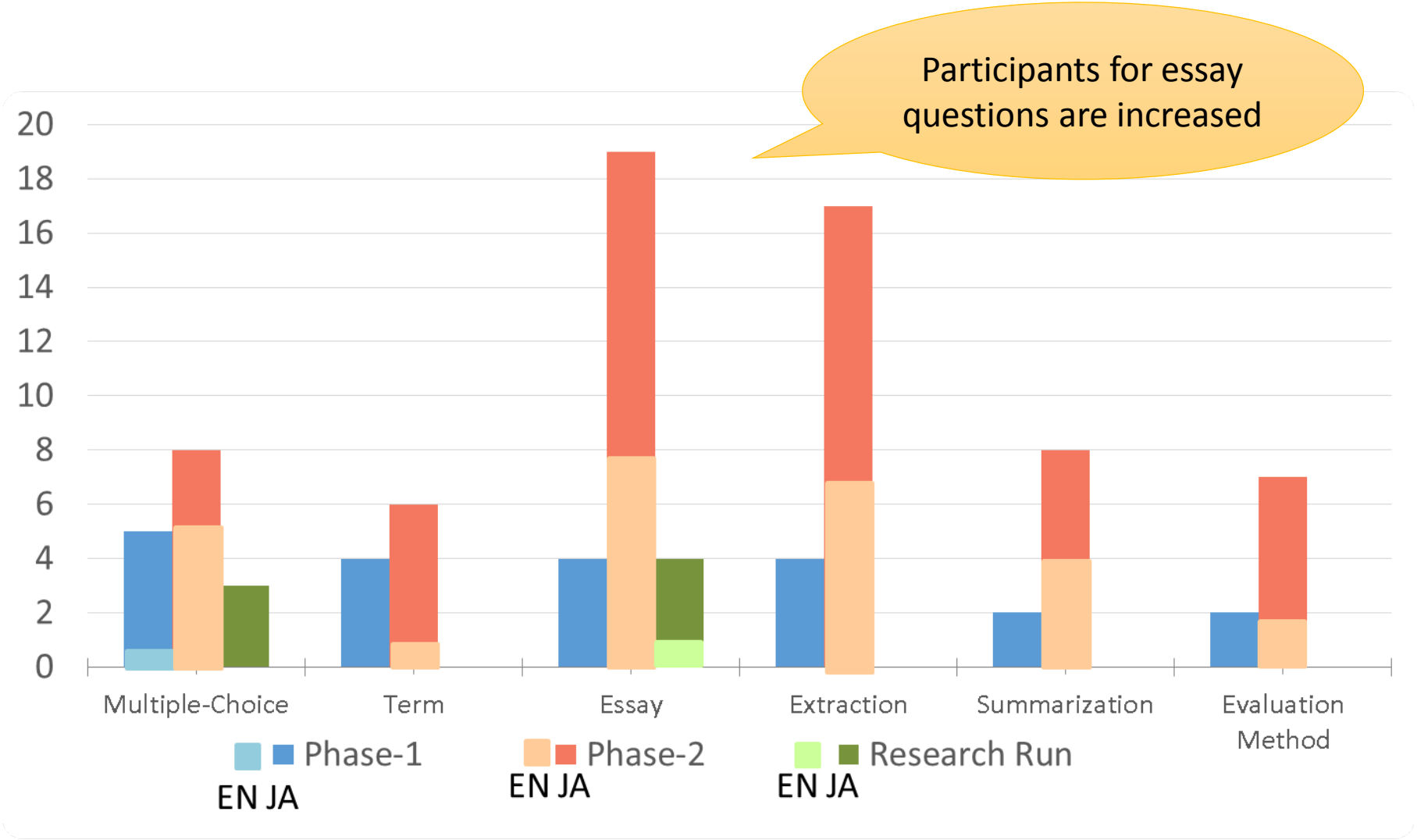
Submission number of each team

- 24 runs from 6 teams at Phase 1
- 56 runs from 11 teams at Phase 2
- 6 runs from 4 teams

Three numbers separated by comma show submitted number at Phase 1, Phase 2 and Research run

Team ID	JA						EN					
	Choice	Term	Essay				Choice	Term	Essay			
			E2E	Ext	Sum	EvM			E2E	Ext	Sum	EvM
KUAS	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	1,2,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-
Forst	-,-,-	2,1,-	2,3,2	2,-,-	1,1,-	2,2,-	-,-,-	-,-,-	1,1,-	-,-,-	-,-,-	-,-,-
IMTKU	-,-,-	-,-,-	-,2,-	-,-,-	-,1,-	-,-,-	-,3,-	-,-,-	-,2,-	-,-,-	-,1,-	-,-,-
SML	-,-,-	-,1,-	1,3,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-
KSU	3,2,2	2,3,-	2,3,-	2,3,-	1,1,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-
SLQAL	1,1,1	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-
CMUQA	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,3,-	-,2,-	-,1,-	-,-,-
DGLab	-,-,-	-,-,-	-,-,1	-,-,-	-,2,-	-,2,-	-,-,-	-,-,-	-,-,1	-,-,-	-,2,-	-,2,-
tmkff	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,1,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-
MTMT	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,2,-	-,2,-	-,-,-	-,-,-
HagiL	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,-,-	-,1,-	-,-,-	-,-,-	-,-,-	-,-,-

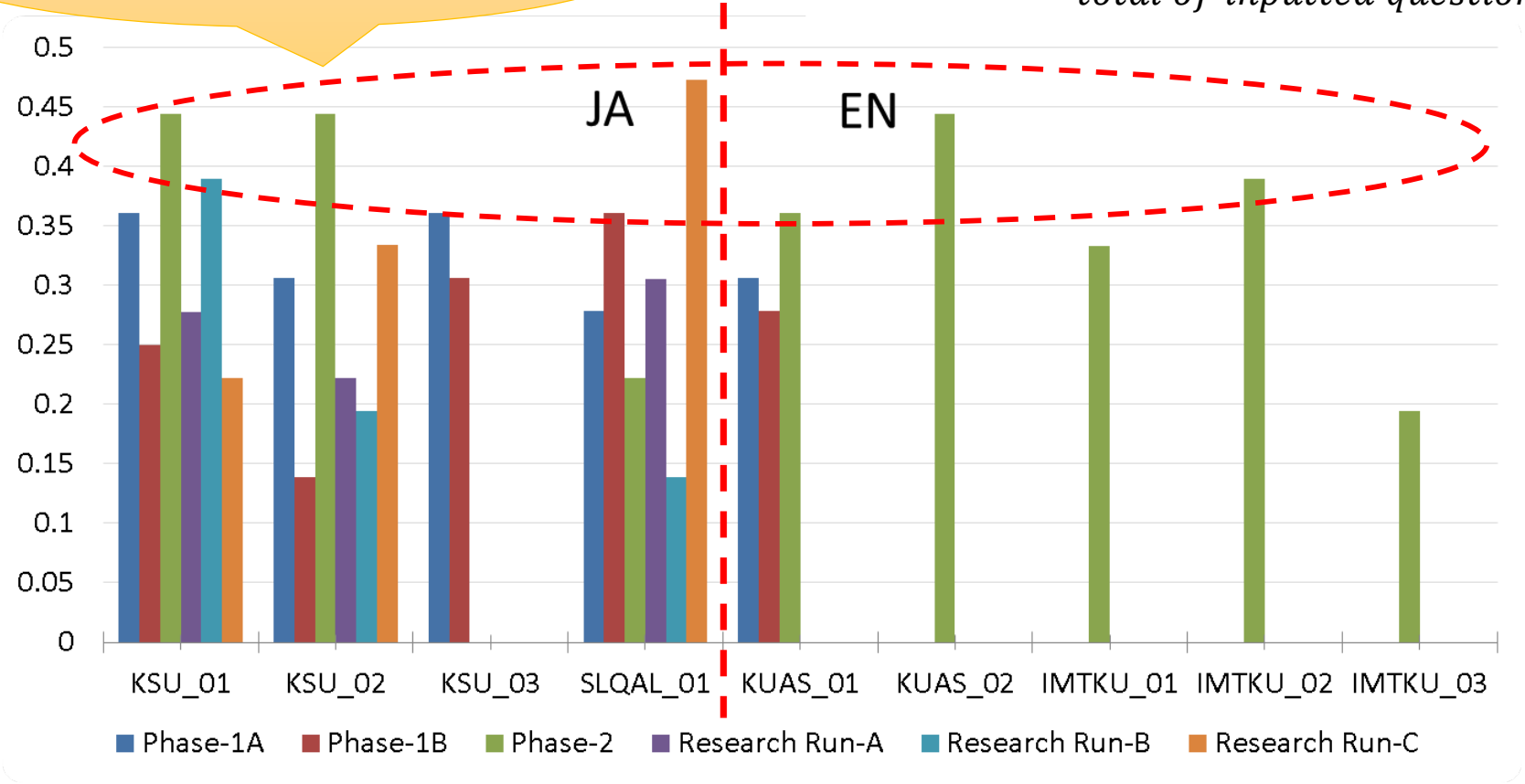
Total number of submissions



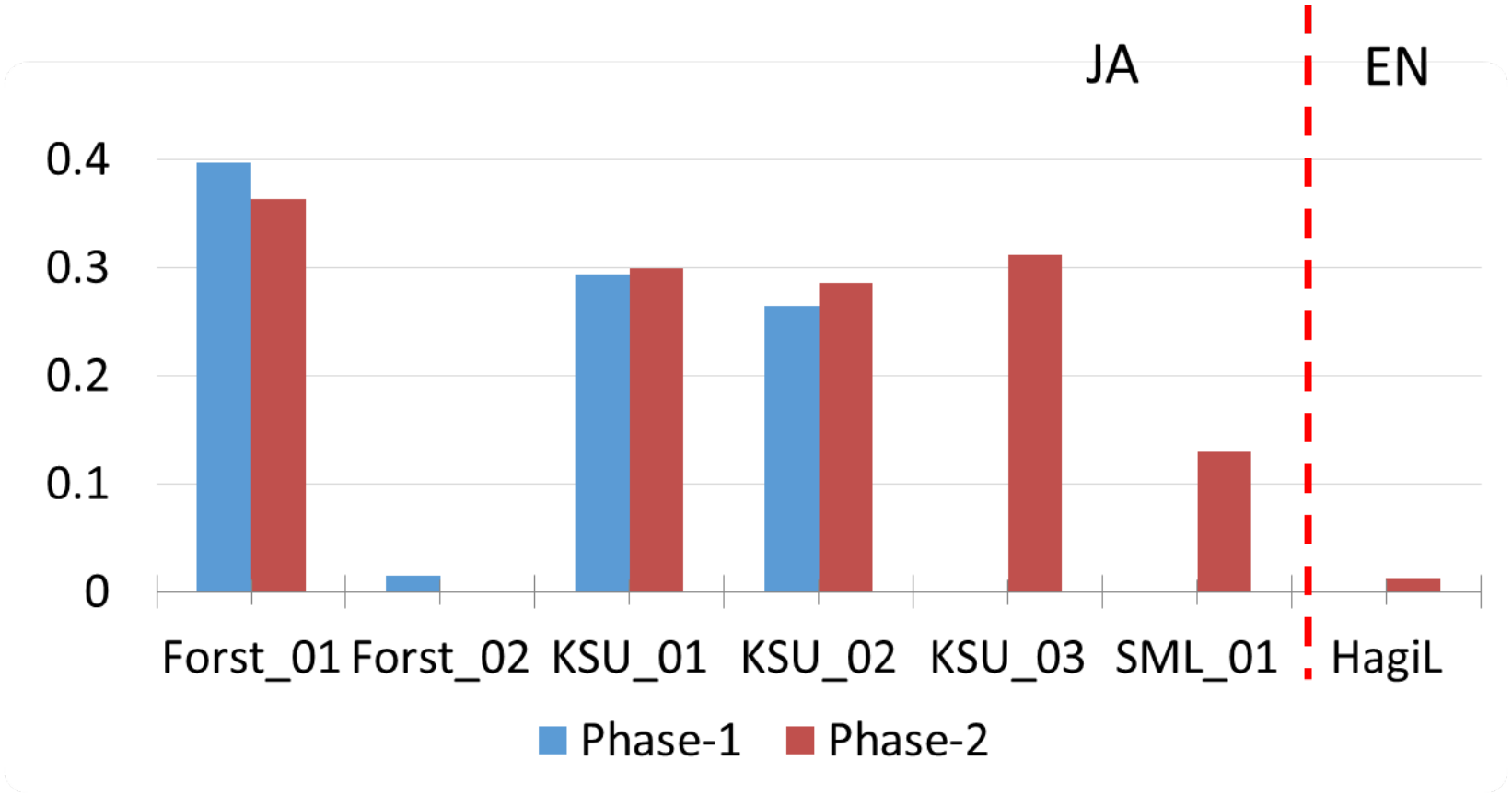
Correct rates in Multiple-choice task

The difference among the results was a little

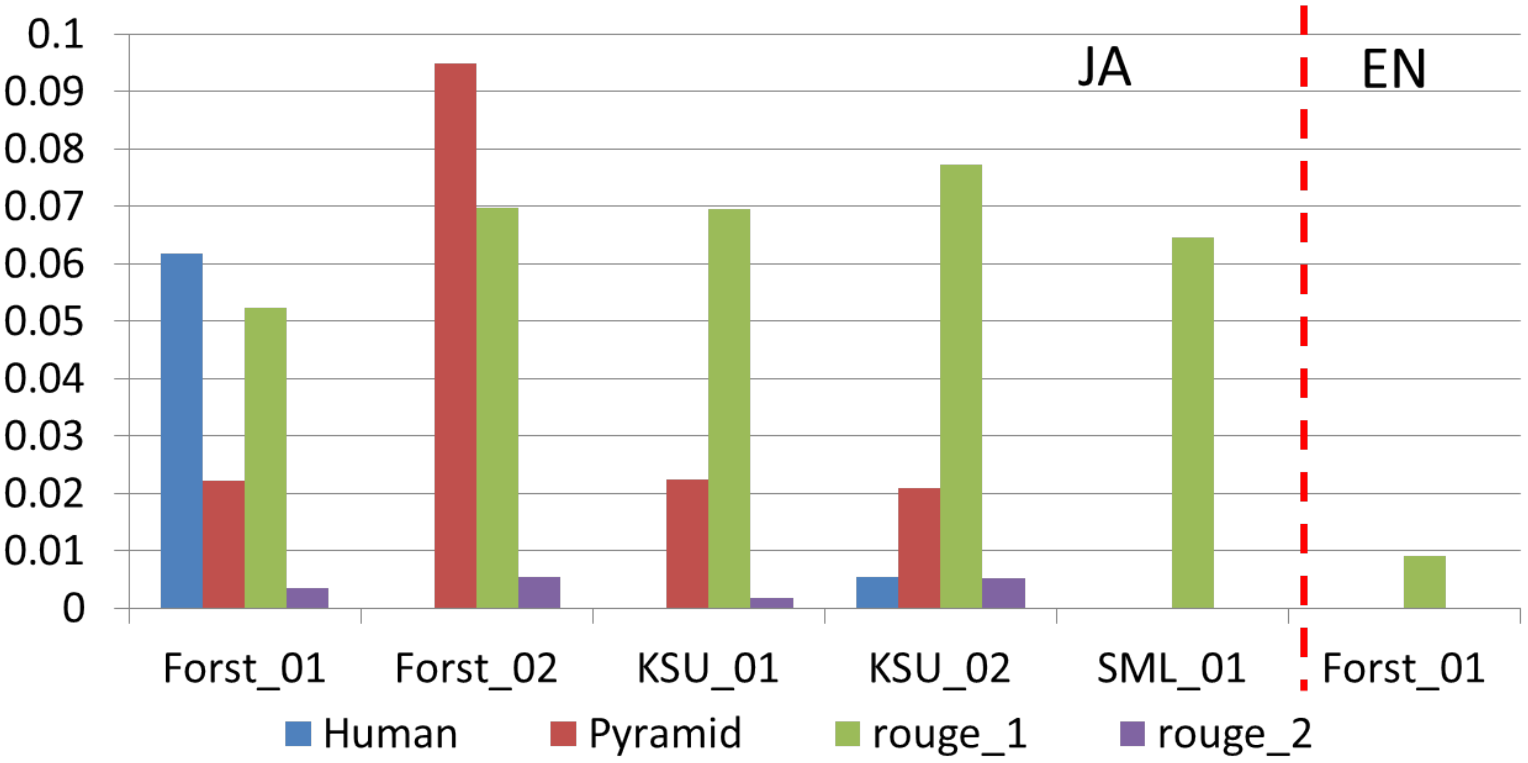
$$\text{Correct rate} = \frac{number\ of\ correct\ answers}{total\ of\ inputted\ questions}$$



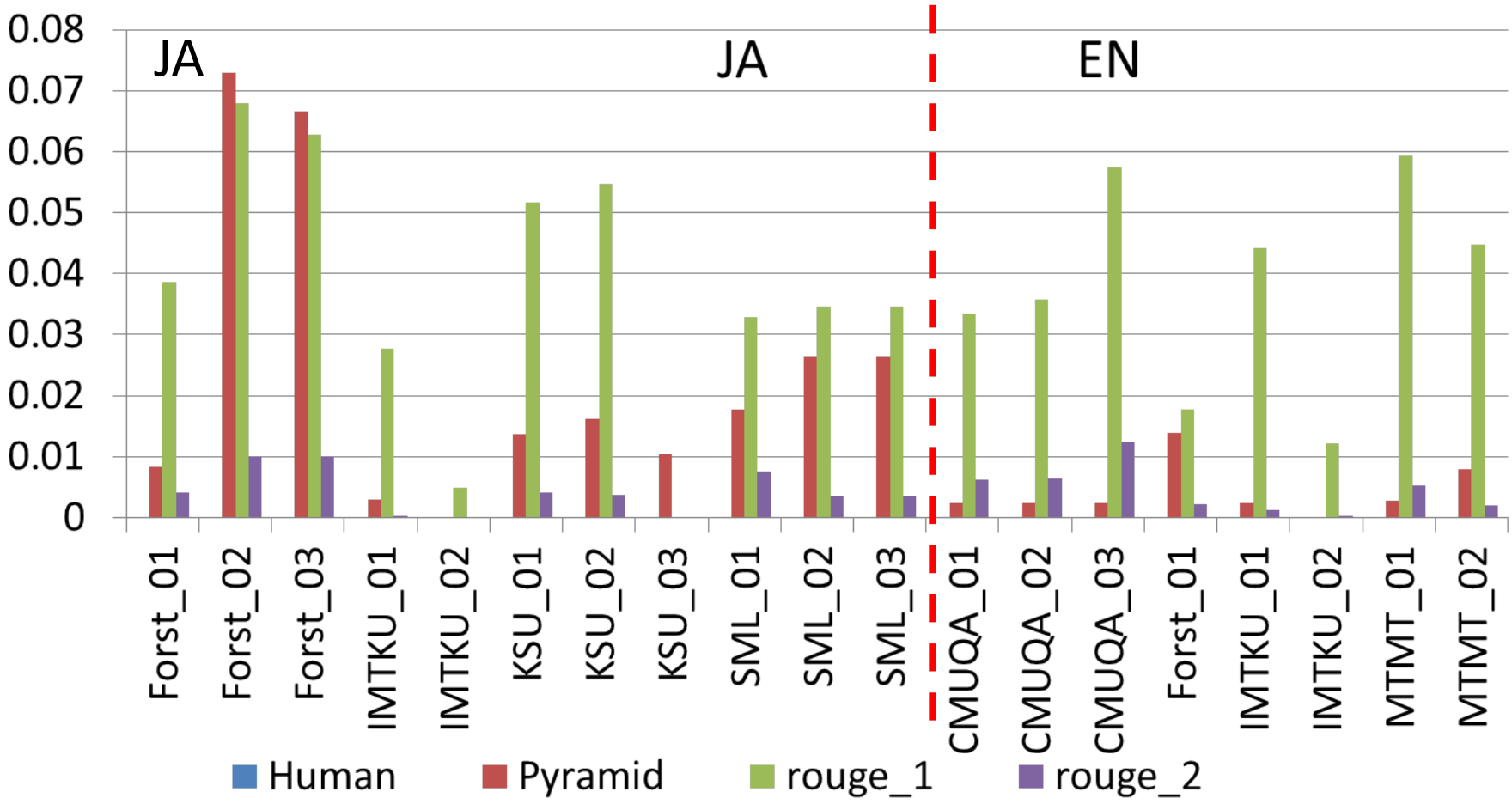
Correct rates in Term question task



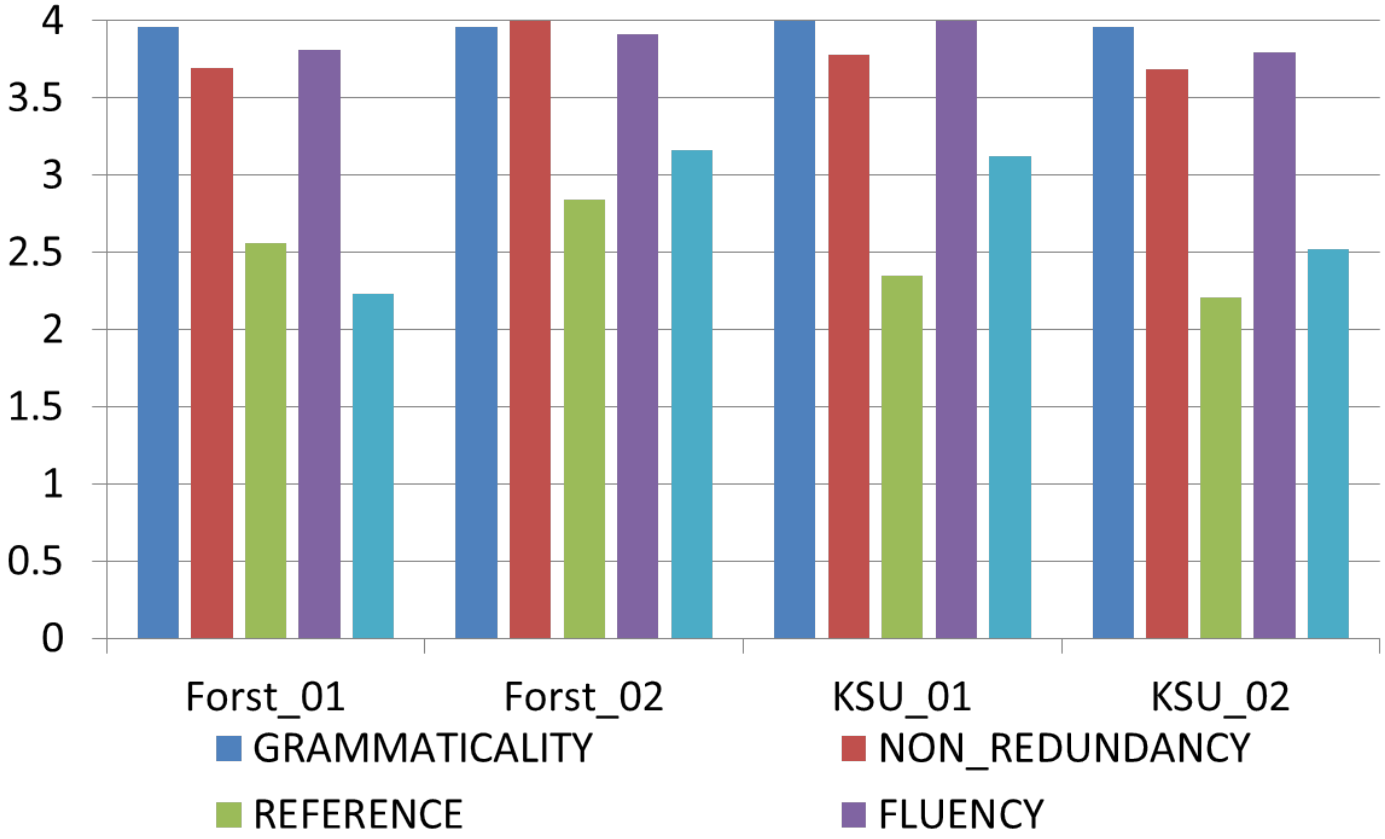
Human marks, Pyramid and ROUGE scores in Essay task at Phase 1



Human marks, Pyramid and ROUGE scores in Essay task at Phase 2

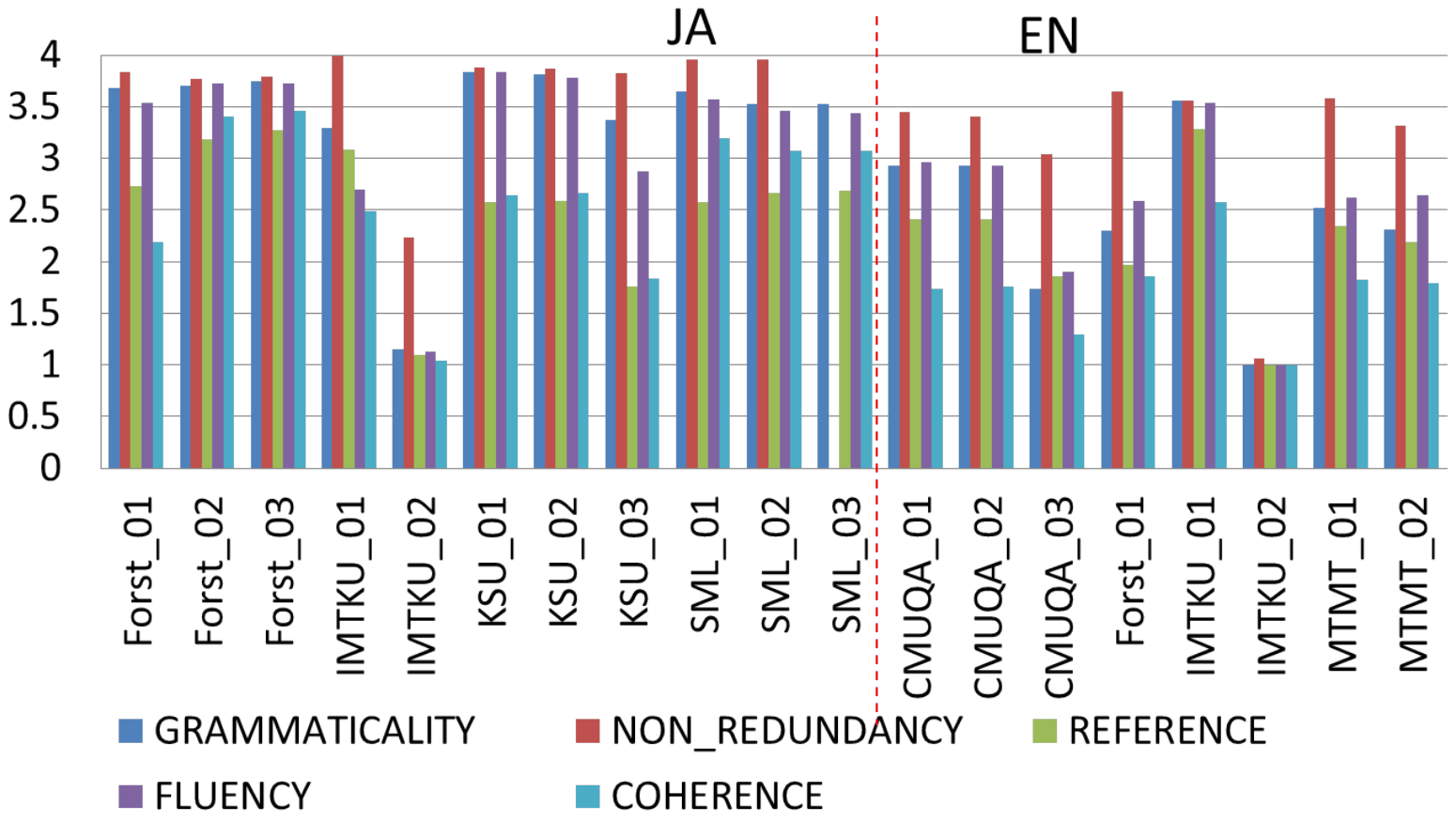


Quality question scores in Essay task at Phase 1



the qualities of ‘reference clarity’ and ‘coherence and content structure’ are low by and large. The improvement of the qualities may enhance the total improvement

Quality question scores in Essay task at Phase 2



the qualities of ‘reference clarity’ and ‘coherence and content structure’ are low by and large. The improvement of the qualities may enhance the total improvement

NTCIR

QA Lab

Results of Extraction task at Phase 1

Nugget recall is low

TeamID	Priority	Lang	Passage Precision	Nugget Recall	Ave. of tokens
Forst	1	JA	0.267	0.019	1037.6
KSU	1	JA	0.468	0.288	1147.5
KSU	2	JA	0.251	0.100	1483.5

$$\text{Passage precision} = \frac{number\ of\ passages\ including\ at\ least\ one\ gold\ standard\ nugget}{total\ of\ extracted\ passages}$$

$$\text{Nugget recall} = \frac{number\ of\ nugget\ included\ among\ the\ extracted\ passages}{total\ of\ gold\ standard\ nugget}$$

NTCIR

QA Lab

Results of Extraction task at Phase 2

TeamID	Priority	Lang	Passage Precision	Nugget Recall	Ave. of tokens
DGLab	1	JA	0.510	0.057	1875.6
DGLab	2	JA	0.479	0.044	1875.6
DGLab	3	JA	0.263	0.166	1459.2
Forst	1	JA	0.038	0.080	1578.0
Forst	2	JA	0.192	0.017	1324.4
IMTKU	1	JA	0.113	0.020	454.4
IMTKU	2	JA	0.000	0.000	336.25
KSU	1	JA	0.057	0.152	1591.8
KSU	2	JA	0.100	0.201	1592.6
KSU	3	JA	0.083	0.057	1597.6
CMUQA	1	EN	0.113	0.035	243.2
CMUQA	2	EN	0.088	0.026	274.2
DGLab	1	EN	0.087	0.035	770.4
DGLab	2	EN	0.117	0.035	770.4
IMTKU	1	EN	0.260	0.061	249.2
IMTKU	2	EN	0.234	0.058	249.2
MTMT	1	EN	0.009	0.032	797.2
MTMT	2	EN	0.014	0.019	782.4

Results of Summarization task at Phase 1

TeamID	Priority	source	Lang.	#of ques	#of N/A	content score				quality score				
						Human	NUGGET	rouge_1	rouge_2	GRAMMATICALITY	NON_REDUNDANCY	REFERENCE	FLUENCY	COHERENCE
Forst	1	Exp	JA	5	0	0	0.00356	0.01	0.00118	4	3.6	2.5	4	2
Forst		GSN+ExP	JA	5	0	0	0.00356	0	0	4	3.6	2.5	4	2
Forst		GSN	JA	5	0	0	0.00698	0	0	4	3.8	3.5	4	3
KSU	1	ExP	JA	4	1	0	0.00991	0.0223	0.00182	4	3.13	2	3.75	2
KSU		GSN+ExP	JA	4	1	0	0.00991	0.0223	0.00182	4	3.13	2	3.75	2
KSU		GSN	JA	5	0	0.0587	0.0527	0.0659	0.0279	4	3.8	2.8	4	3.5

- Three sets of passages are provided as input
  - ExP: set of all passages submitted in Extraction task
  - GSN: set of gold standard nuggets
  - GSN+ExP: merged set of the above 2 sets

Results of Summarization task at Phase 2

TeamID	Priority	source	Lang.	#of ques	#of N/A	content score			quality score				
						Human	NUGGET	rouge_1	rouge_2	GRAMMATICALITY	NON_REDUNDANCY	REFERENCE	FLUENCY	COHERENCE
DGLab		Exp	JA	5	0	0	0.00641	0.0246	0.00169	4	2.87	3.5	3.7	2.47
DGLab	1	GSN+ExP	JA	5	0		0.0414	0.0603	0.0305	3.93	3.03	3.3	3.4	2.67
DGLab		GSN	JA	5	0		0.0464	0.0617	0.0317	4	3.03	3.3	3.4	2.67
DGLab		Exp	JA	5	0		0.0129	0.0229	0.000782	3.8	2.77	3.07	3.5	2.4
DGLab	2	GSN+ExP	JA	5	0		0.0468	0.0627	0.0299	3.93	3.03	3.1	3.4	2.57
DGLab		GSN	JA	5	0		0.0475	0.0627	0.0299	4	3.03	3.1	3.47	2.7
Forst		Exp	JA	5	0		0.00143	0.00797	0.000175	3.47	3.93	2.63	3.07	2.37
Forst	1	GSN+ExP	JA	5	0		0.00143	0	0	3.47	3.93	2.63	3.07	2.37
Forst		GSN	JA	5	0		0.00737	0	0	4	4	3.1	4	3.1
IMTKU	1	Exp	JA	5	0		0.00295	0	0	3.13	3.87	3.03	2.57	2.63
KSU		Exp	JA	5	0		0.0074	0.0252	0.00214	3.9	3.3	2.8	4	2.4
KSU	1	GSN+ExP	JA	5	0		0.00521	0.0264	0.00359	3.57	3.47	2.2	3.13	2.23
KSU		GSN	JA	3	2		0.0269	0.0682	0.0354	3.93	3.9	3.9	3.37	3.47
CMUQA	1	GSN	EN	5	0		0.0198	0.0708	0.0338	4	3.3	2.9	4	2.5
DGLab		Exp	EN	5	0	0	0.00335	0.0255	0.00249	2.1	2.5	2.6	1.5	1.5
DGLab	1	GSN+ExP	EN	5	0		0.0254	0.0635	0.0305	4	2.4	2.7	3.3	2.5
DGLab		GSN	EN	5	0		0.026	0.0636	0.0308	4	2.5	3	3.5	2.63
DGLab		Exp	EN	5	0		0.00321	0.026	0.00246	2.4	2.7	2.6	1.6	2
DGLab	2	GSN+ExP	EN	5	0		0.0288	0.066	0.0329	4	2.4	2.6	3.3	2.5
DGLab		GSN	EN	5	0		0.0292	0.0661	0.0329	4	2.8	3	3.4	2.5
IMTKU	1	Exp	EN	5	0		0.00262	0	0	3.8	3.4	3.2	3.5	2.4

Understandably GSN is better,
but the content scores are low

Comparison with End-to-End results (Phase 1)

End-to-end Run													
TeamID	Priority	Lang.	#of ques	#of N/A	content score				quality score				
					Human	NUGGET	rouge_1	rouge_2	GRAMMATICALITY	NON_REDUNDANCY	REFERENCE	FLUENCY	COHERENCE
Forst	1	JA	26	1	0.011	0.0221	0.0523	0.00351	3.96	3.69	2.56	3.81	2.23
Forst	2	JA	22	5		0.095	0.0698	0.00536	3.95	4	2.84	3.91	3.16
Forst	3	JA	24	3	0.0339	0.219	0.0887	0.00953	4	3.9	3.15	3.39	3.27
KSU	1	JA	16	11	0	0.0224	0.0695	0.00178	4	3.78	2.34	4	3.13
KSU	2	JA	24	3	0.00097	0.0209	0.0772	0.00533	3.96	3.69	2.21	3.79	2.52
SML	1	JA	22	5			0.0646	0					
Forst	1	EN	22	5			0.00921	0					

Summrization Run													
TeamID	Priority	source	Lang.	#of ques	#of N/A	content score				quality score				
						Human	NUGGET	rouge_1	rouge_2	GRAMMATICALITY	NON_REDUNDANCY	REFERENCE	FLUENCY	COHERENCE
Forst	1	Exp	JA	5	0	0	0.00356	0.01	0.00118	4	3.6	2.5	4	2
Forst		GSN+ExP	JA	5	0	0	0.00356	0	0	4	3.6	2.5	4	2
Forst		GSN	JA	5	0	0	0.00698	0	0	4	3.8	3.5	4	3
KSU		ExP	JA	4	1	0	0.00991	0.0223	0.00182	4	3.13	2	3.75	2
KSU		GSN+ExP	JA	4	1	0	0.00991	0.0223	0.00182	4	3.13	2	3.75	2
KSU		GSN	JA	5	0	0.0587	0.0527	0.0659	0.0279	4	3.8	2.8	4	3.5

Worse than End-to-End

Because of lacking unity?

NTCIR

QA Lab

Results of Evaluation-method task

TeamID	Priority	Lang	Spearman's Rho	Kendall's Tau-b
Phase 1				
Forst	1	JA	0.427	0.334
Forst	2	JA	0.596	0.534
Pyramid		JA	0.728	0.638
ROUGE-1		JA	0.677	0.568
ROUGE-2		JA	0.599	0.472
Phase 2				
Forst	1	JA	-0.071	-0.049
Forst	2	JA	0.404	0.360
tmkff	1	JA	0.193	0.212
DGLab	1	JA	0.200	0.167
DGLab	2	JA	0.341	0.303
DGLab	1	EN	0.333	0.286
DGLab	2	EN	-0.160	-0.067
Pyramid		JA	0.428	0.381
ROUGE-1		JA	0.620	0.588
ROUGE-2		JA	0.120	0.062
Pyramid		EN	0.086	0.073
ROUGE-1		EN	-0.263	-0.206
ROUGE-2		EN	-0.343	-0.273

Rank correlation coefficient with human expert ranking

No result could be better than Pyramid and ROUGE scores

Because DGLab graded by deducting marks, we calculated their correlation coefficients by inversing their sign

Future

- We are still struggling with failure analysis and analysis of effectiveness of each element
- Human assessment of essays need to review and re-analysis (too strict)

- NTCIR-14 Poli-Info
  - Balanced view
- Session & Breaout Thursday After noon
- Posters Friday Lunch

NTCIR

Thank you for your attention!