NTCIR


# Overview of the NTCIR-12 QA Lab-2 Task

Hideyuki Shibuki*1, Kotaro Sakamoto*1,*2, Madoka Ishioroshi*2, Akira Fujita*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


# Introduction

- Goal

  Investigation of the **real-world complex Question Answering (QA)** technologies using Japanese **university entrance exams** and their English translation on the subject of “World history”.

NTCIR


# Highlights

- **Multiple-choice** and **free-description** type questions
- Many questions are not in a simple QA format, and require an **understanding of the surrounding context**
- Some questions require inference.
- **Evaluation of** free-description, especially, **essay questions**
- Construction of a **hierarchy of question formats**
- Comparison with high-school students from all over Japan

# University Entrance Exam Questions

- **Multiple choice** type questions
  - The National Center Test for University Admissions (EN,JA)
  - Benesse mock exams (JA)
  - Yozemi mock exams (JA)
- **Free description** type questions
  - Secondary exams from 5 Japanese universities (EN,JA)
    - The University of Tokyo
    - Kyoto University
    - Hokkaido University
    - Waseda University
    - Chuo University
  - Sundai mock exams (JA)

NTCIR

# 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。

```
<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>
```

NTCIR

# Question Types


Questions are classified into the following categories:

(A) Essay

- (A1) Complex essay: more than 200 characters (JA), 100 words (EN)
- (A2) Simple essay: less than 200 characters (JA), 100 words (EN)

(B) Term (NE)

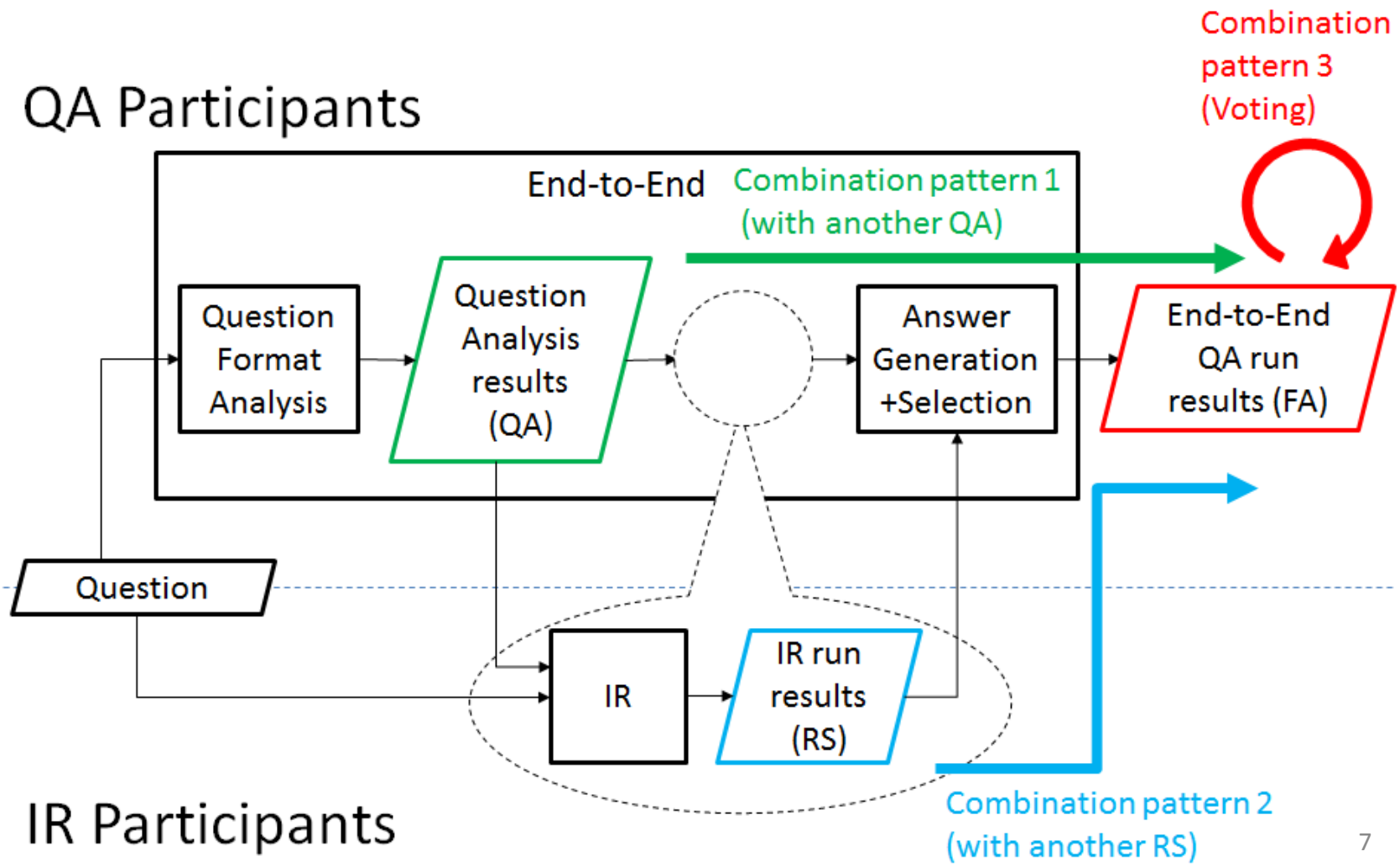
- (B1) Factoid
- (B2) Slot-Filling

(C) True-or-False

(D) Unique

- Including images, graphs, maps and/or tables
- Chronological reordering
- Others

# Combination patterns



NTCIR


# Tools

- Two Baseline Systems (Japanese and English)
  - UIMA module-based end-to-end QA systems
- One Passage Retrieval Module
  - to enhance the module-based collaboration.
- One Scorer and Format Checker for Center Test

NTCIR


# Resources

- Two Japanese high school textbooks
  - Available in Japanese
- World History Ontology
  - Available in Japanese
- A snapshot of Wikipedia
  - Available in English and Japanese

- Participants are free to use any resources

NTCIR


# 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

# Gold standard creation for free-description

- For Named Entity 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]

NTCIR


# Evaluation for free-description

- For Named Entity questions
  - Exact match
- For Essay questions
  - ROUGE-1 and -2 method
    - Morphology without stemming (JA)
    - Word without stemming (EN)
  - Pyramid method
    - Judgment by a human expert

NTCIR


# Schedule

- Phase 1 (EN & JA)
  - Aug 25, 2015: Formal run Topics release
  - Aug 25-31, 2015: Question Format Analysis
  - Sep 1-7, 2015: End-to-End QA and IR runs
  - Sep 8-14, 2015: Combination runs
- Phase 2 (JA only)
  - Oct 1, 2015: Formal run Topics release for Sundai (free-description)
  - Oct 1-8, 2015: End-to-End QA for Sundai
  - Oct 13, 2015: Formal run Topics release for Benesse (multiple-choice)
  - Oct 13-20, 2015: End-to-End QA for Benesse
- Phase 3 (EN & JA)
  - Dec 1, 2015: Formal run Topics release
  - Dec 1-7, 2015: Question Format Analysis
  - Dec 8-14, 2015: End-to-End QA and IR runs
  - Dec 15-21, 2015: Combination runs

# Phase-1 submission

- 58 submissions from 9 teams

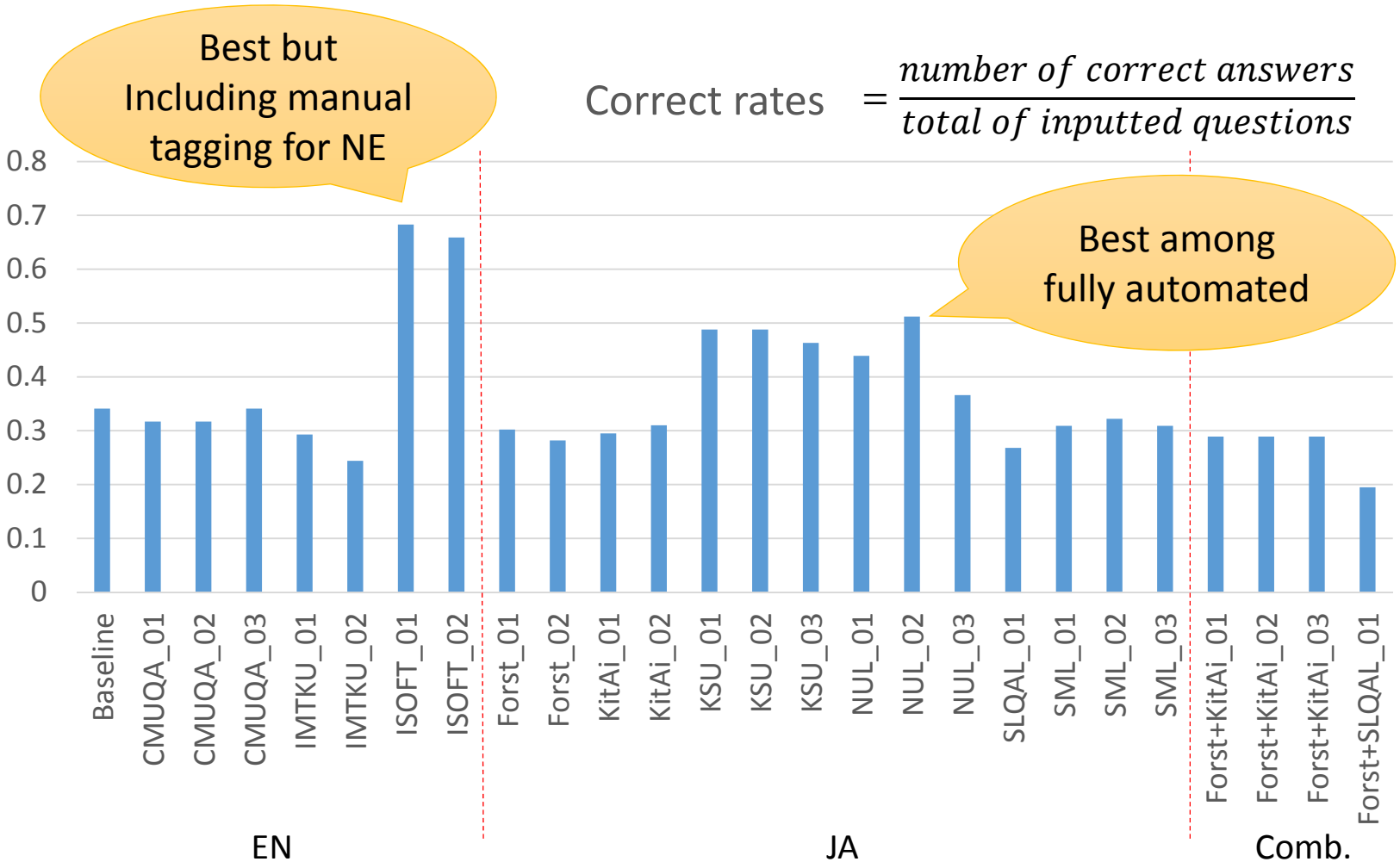
The number for free-description was small

| | EN | | JA | | | | |
|---|---|---|---|---|---|---|---|
| | CT (99) | SE (03) | CT (99) | Benesse (14Nov) | Yozemi (12,13a) | SE (03) | Sundai (13Nov) |
| CMUQA | 3 | - | - | - | - | - | - |
| IMTKU | 3 | 3 | | | | - | - |
| ISOFT | 2 | - | - | | - | - | - |
| Forst | - | - | 2(4) | 2(3) | 2(3) | 3 | 2 |
| KitAi | - | - | 2 | 2 | 2 | - | - |
| KSU | - | - | 3 | - | - | - | - |
| NUL | - | - | 3 | - | - | - | - |
| SLQAL | - | - | 1 | - | - | - | - |
| SML | - | - | 3 | 3 | 3 | 2 | 2 |

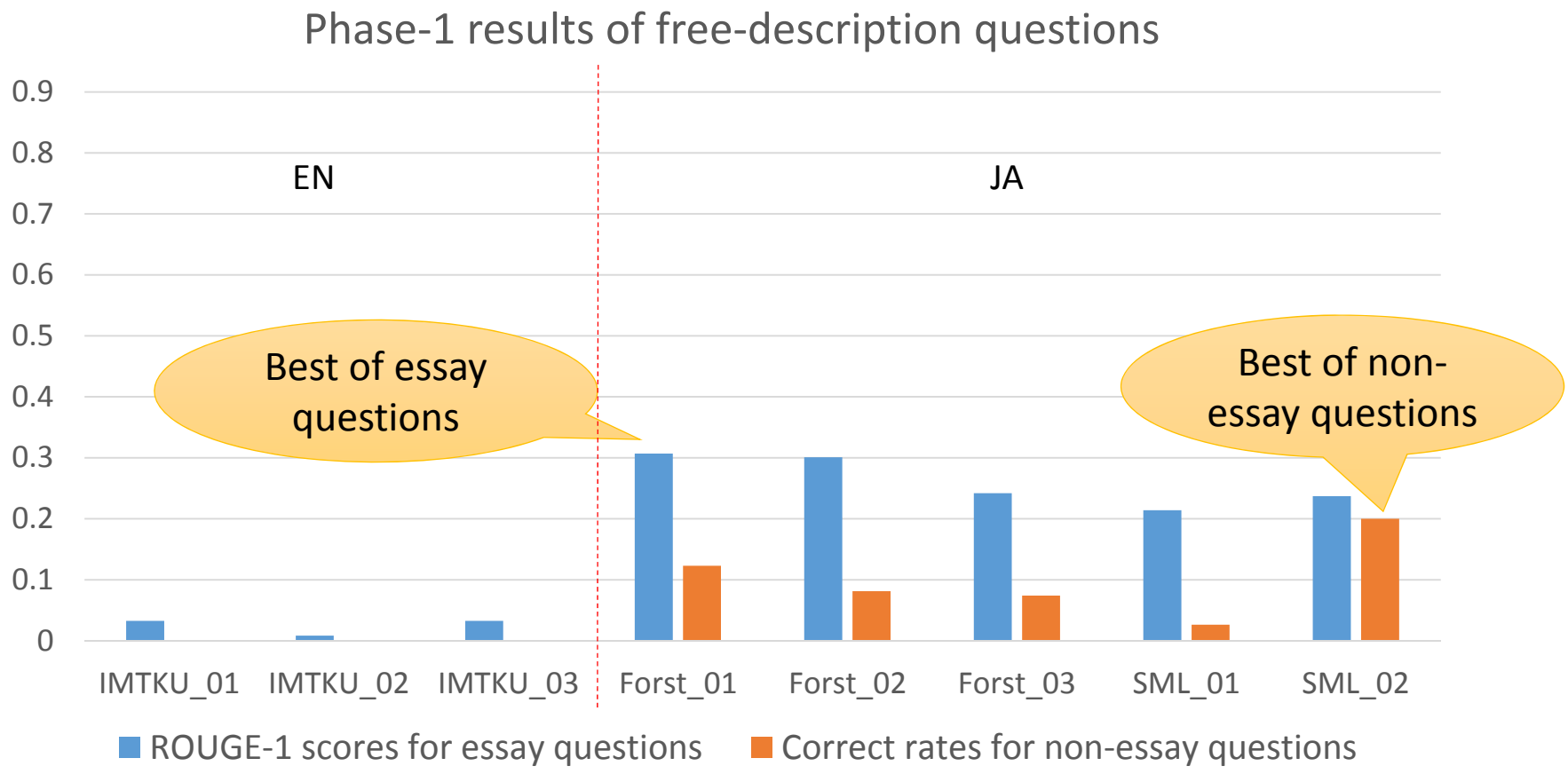
Combination with IR results

Bracketed numbers were the submission for combination runs

# Phase-1 multiple-choice results



# Phase-1 free-description results



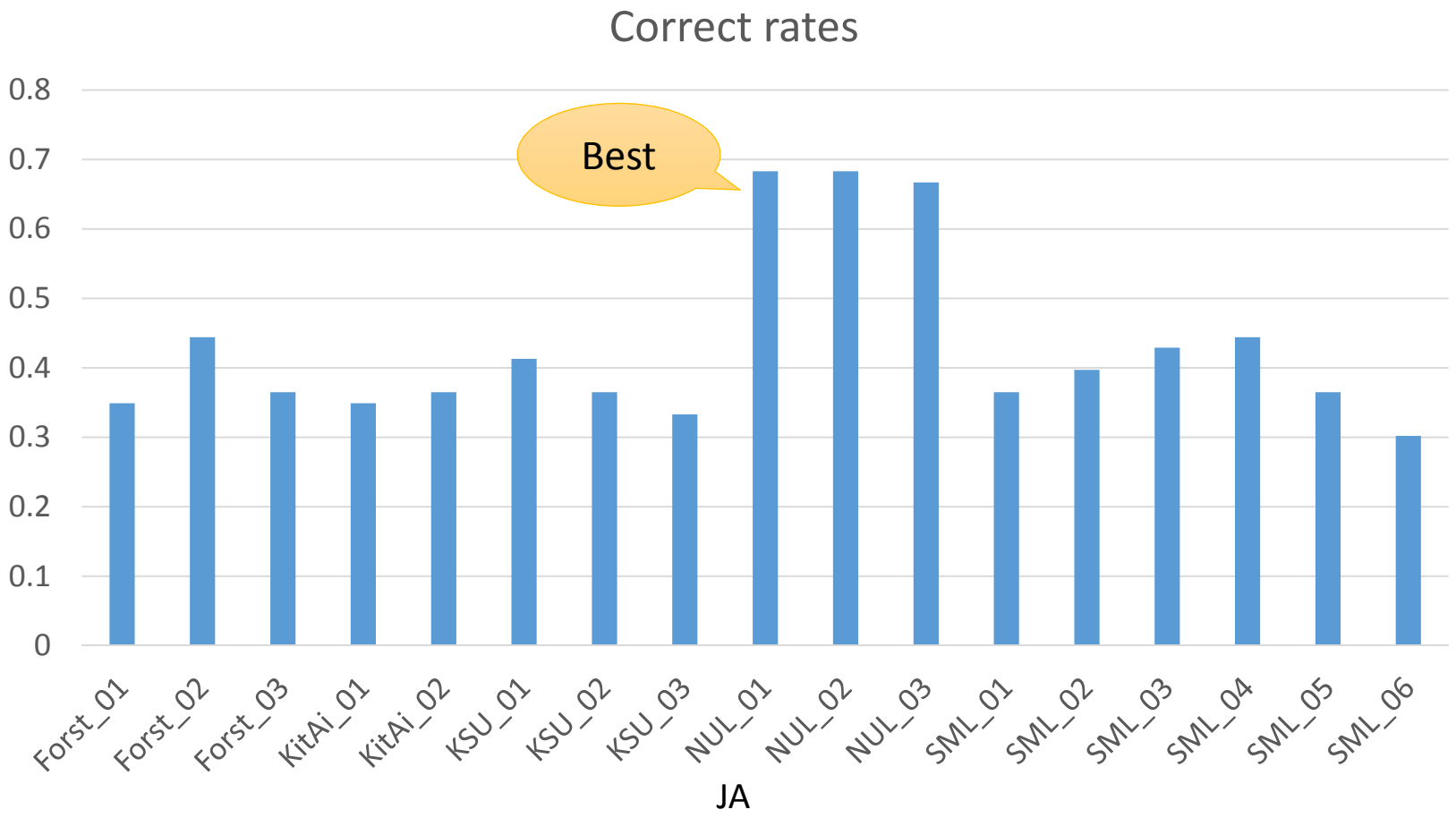
NTCIR


# Phase-2 submission

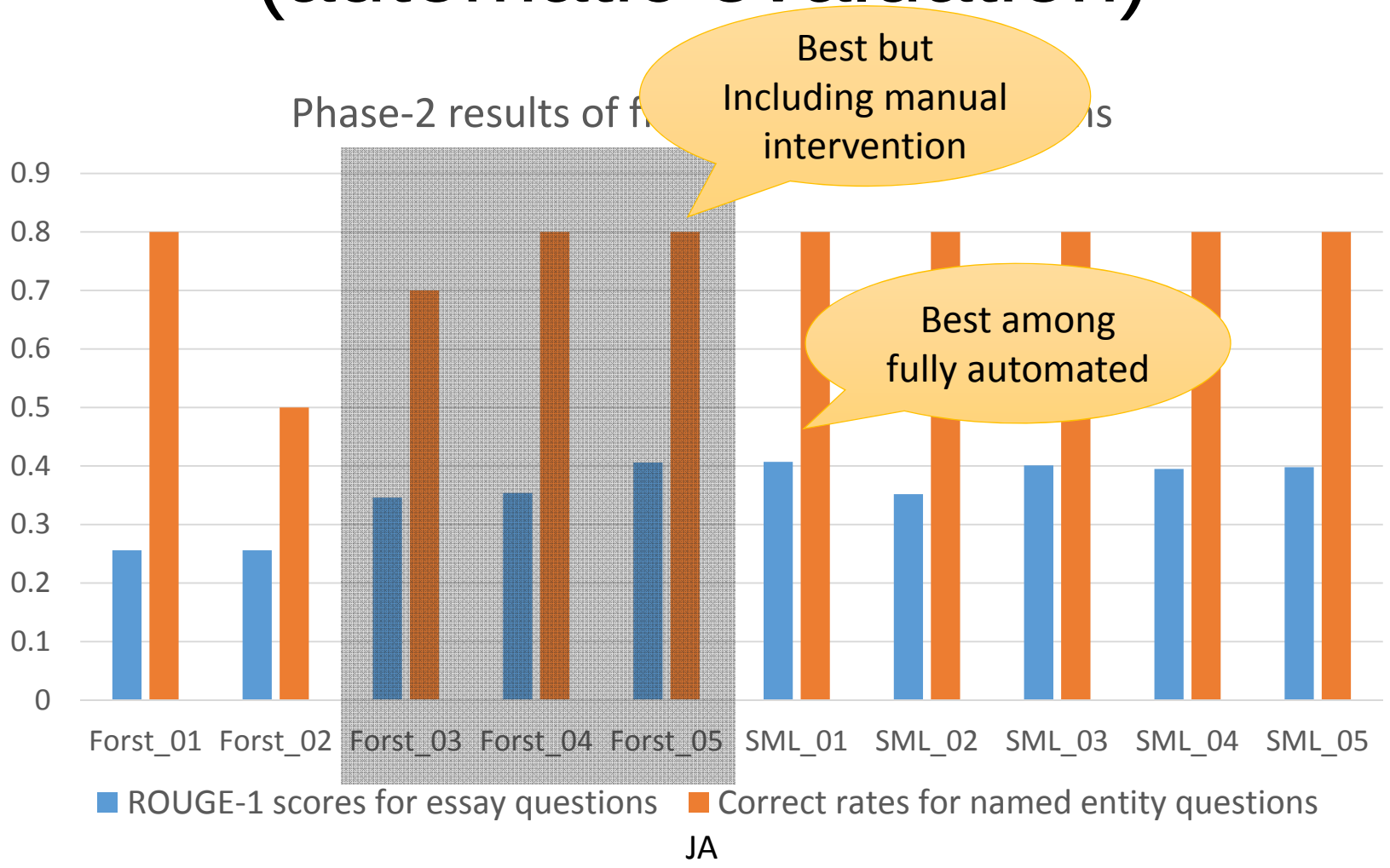
- Collaboration with Todai Robot Project
- Japanese subtask only
- 27 submissions from 5 teams

| | JA | |
|---|---|---|
| | **Benesse (15Jun)** | **Sundai (15Aug)** |
| Forst | **3** | **5** |
| KitAi | **2** | **-** |
| KSU | **3** | **-** |
| NUL | **3** | **-** |
| SML | **6** | **5** |

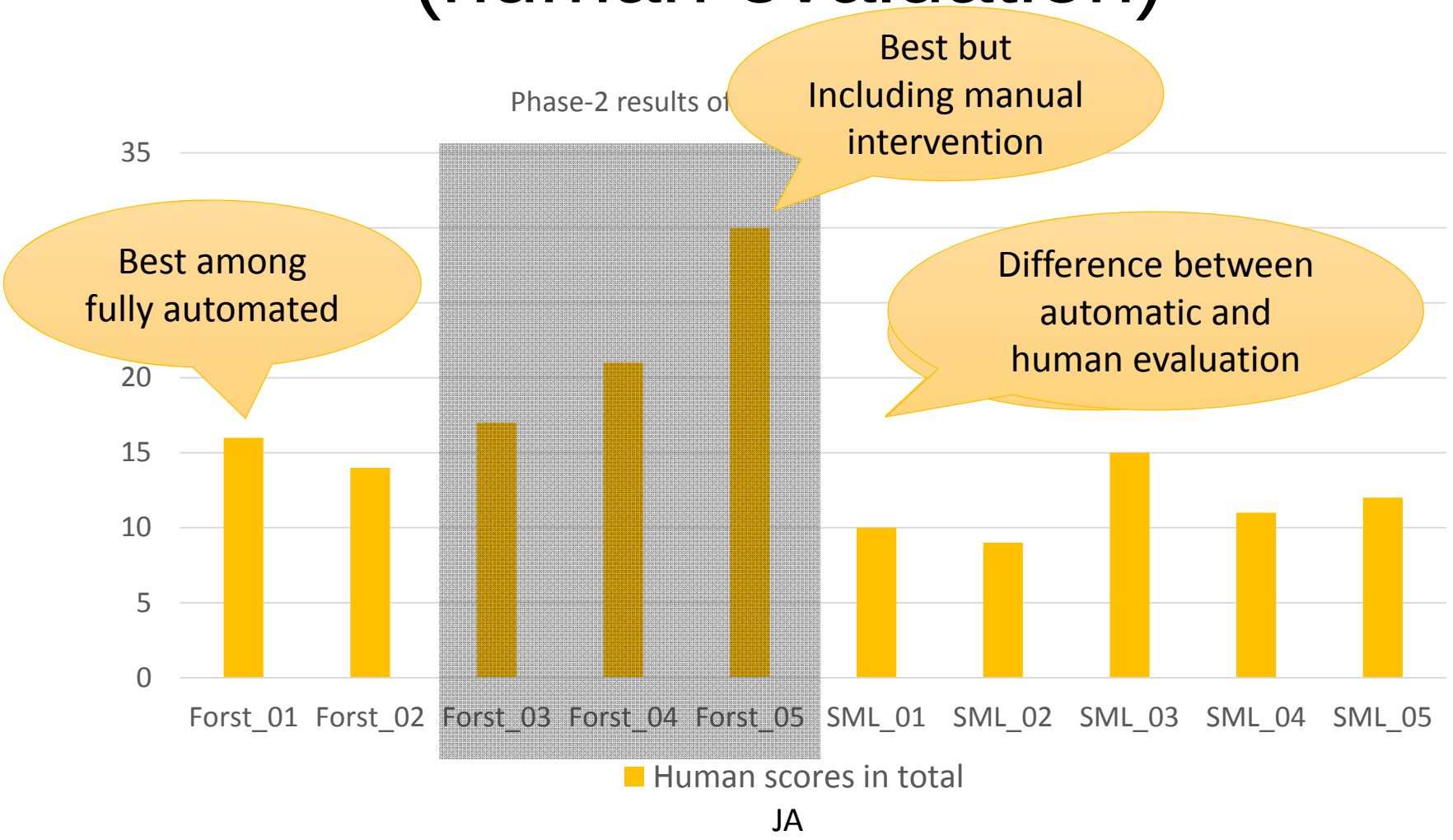
# Phase-2 multiple-choice results



# Phase-2 free-description results (automatic evaluation)



# Phase-2 free-description results (human evaluation)



# Phase-3 submission

- 63 submissions from 12 teams

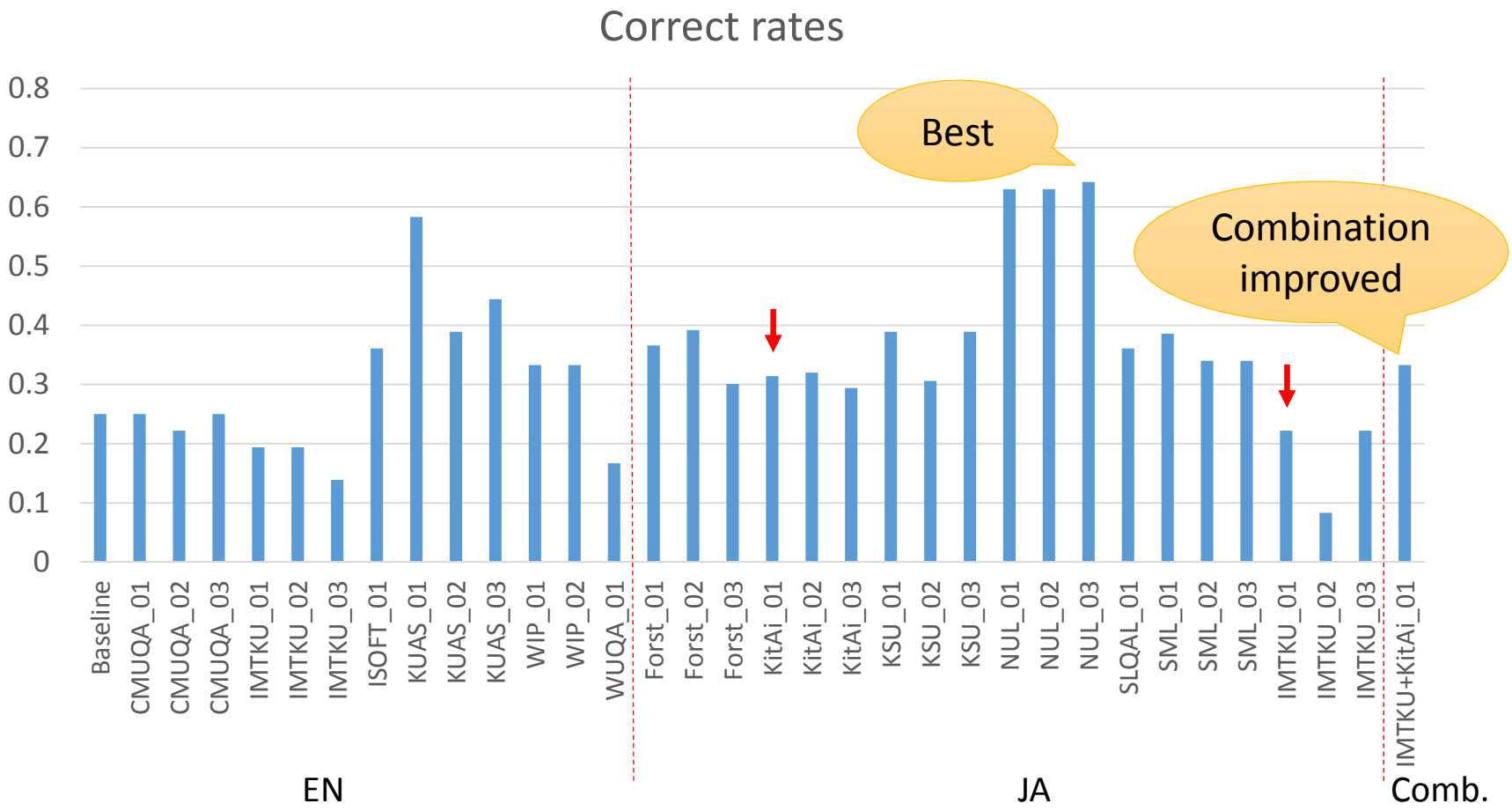
| | EN | | JA | | | | |
|---|---|---|---|---|---|---|---|
| | CT (11) | SE (11) | CT (11) | Benesse (14Sep) | Yozemi (13d,14a) | SE (11) | Sundai (13Aug) |
| CMUQA | 3 | - | - | - | - | - | - |
| KUAS | 3 | - | - | - | - | - | - |
| IMTKU | 3(1) | - | 3 | - | - | - | - |
| ISOFT | 1 | - | - | - | - | - | - |
| WIP | 2 | - | - | - | - | - | - |
| WUQA | 1 | - | - | - | - | - | - |
| Forst | - | - | 3 | 3 | 3 | 3 | 3 |
| KitAi | - | - | 3 | 3 | 3 | - | - |
| KSU | - | - | 3 | - | - | - | - |
| NUL | - | - | 3 | - | - | - | - |
| SLQAL | - | - | 1 | - | - | - | - |
| SML | - | - | 3 | 1 | 3 | 3 | 3 |

Both EN and JA

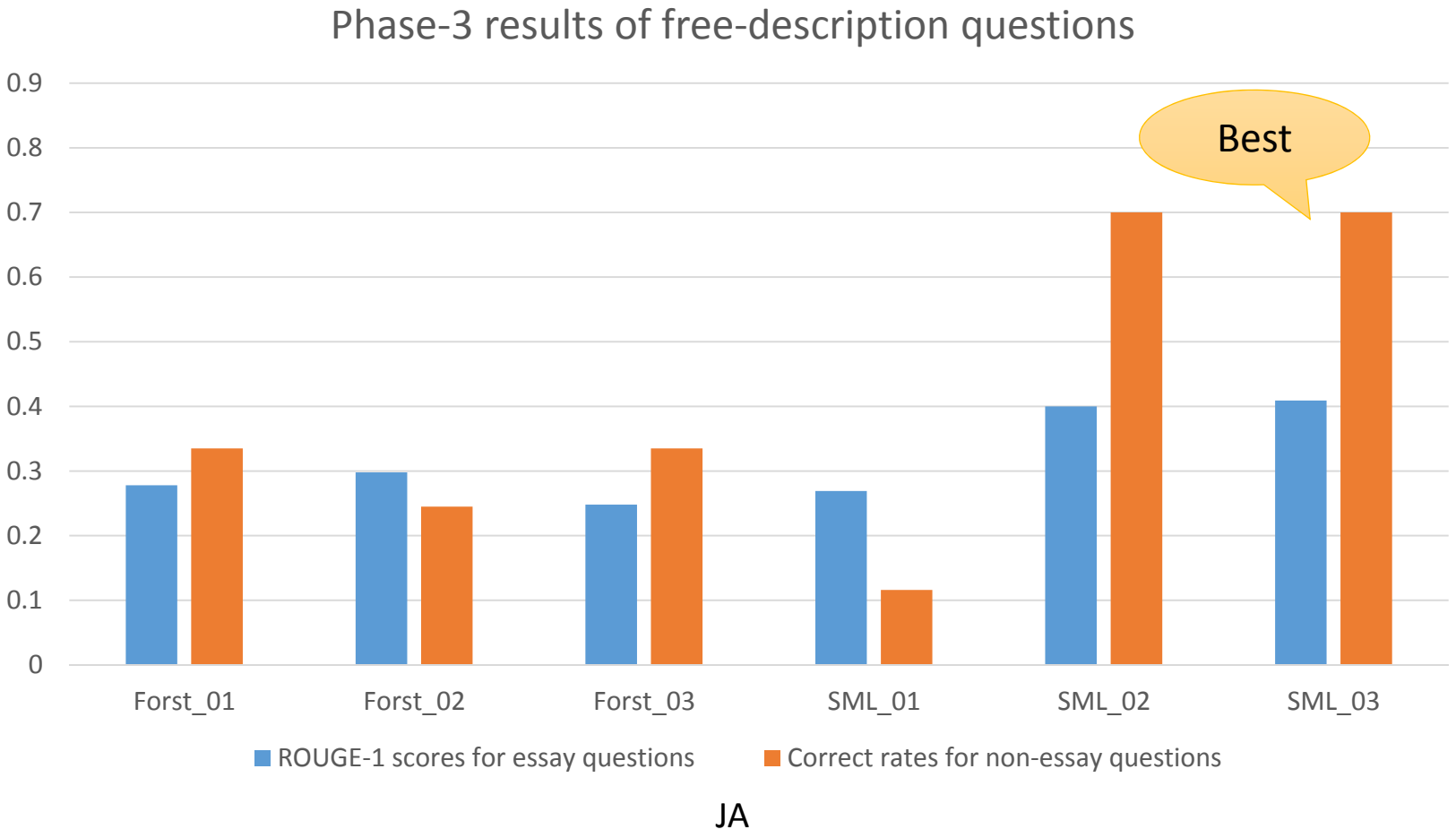
Combination with IR results

Bracketed numbers were the submission for combination runs

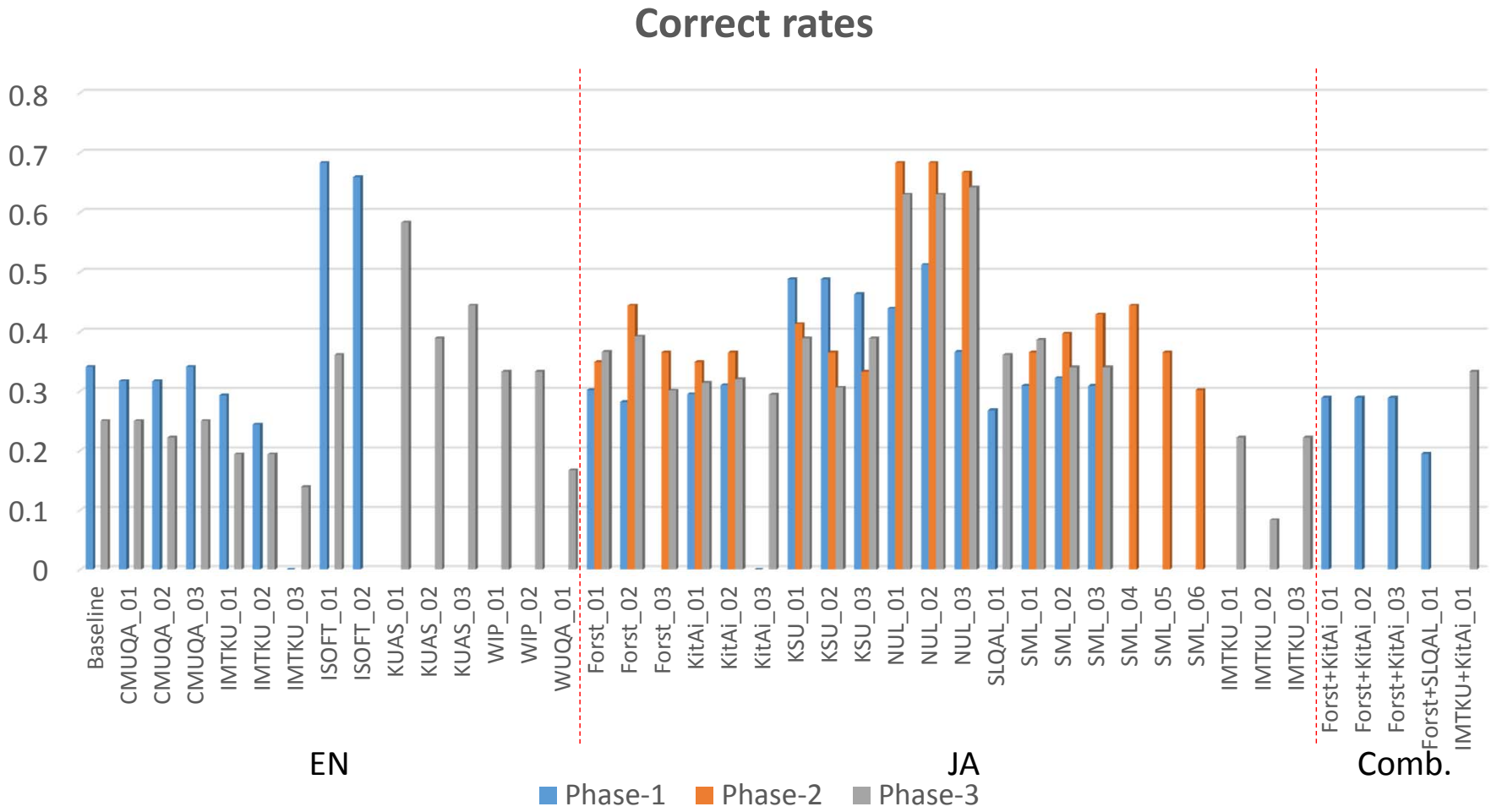
# Phase-3 multiple-choice results



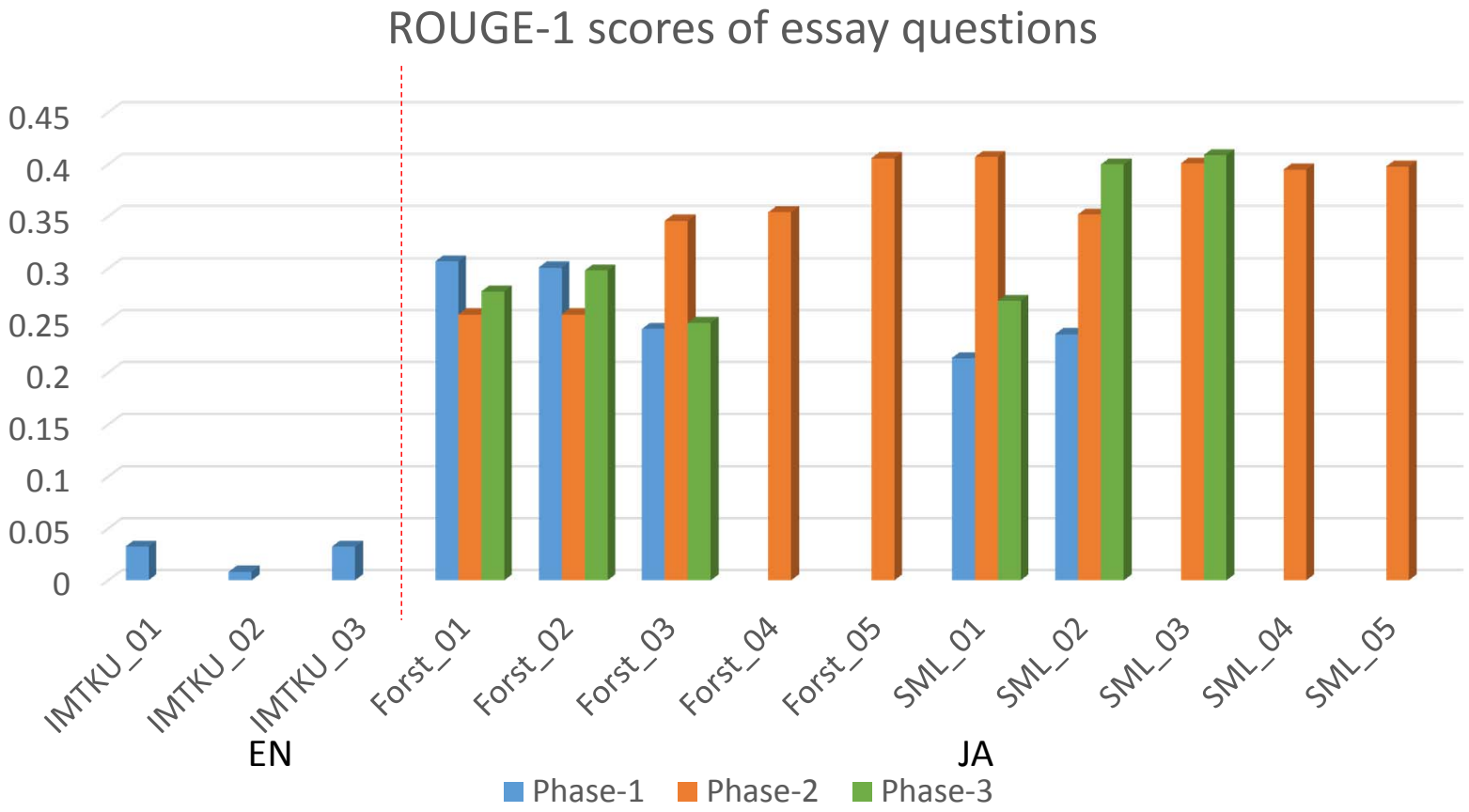
# Phase-3 free-description results



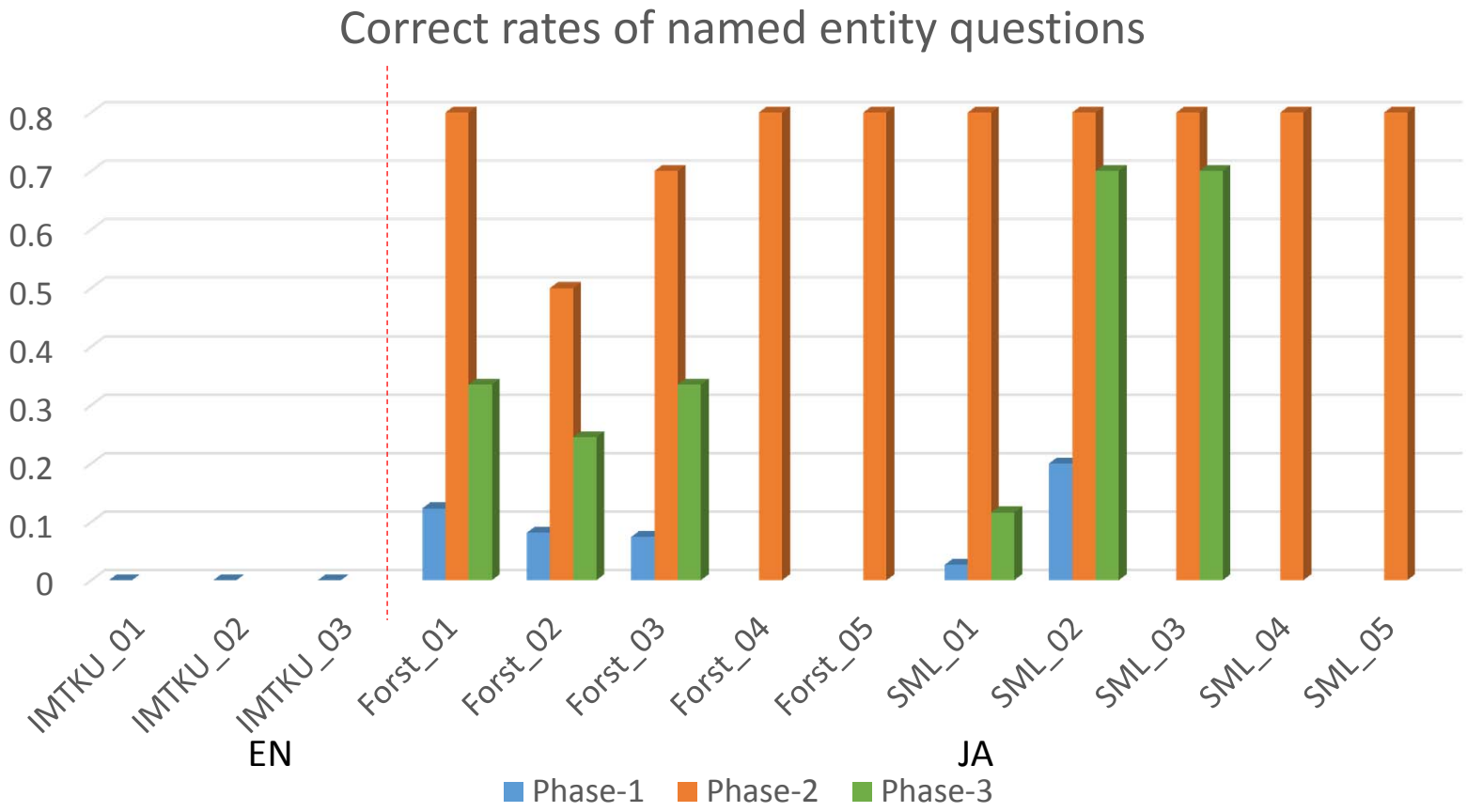
# Overview of multiple-choice results



# Overview of essay results



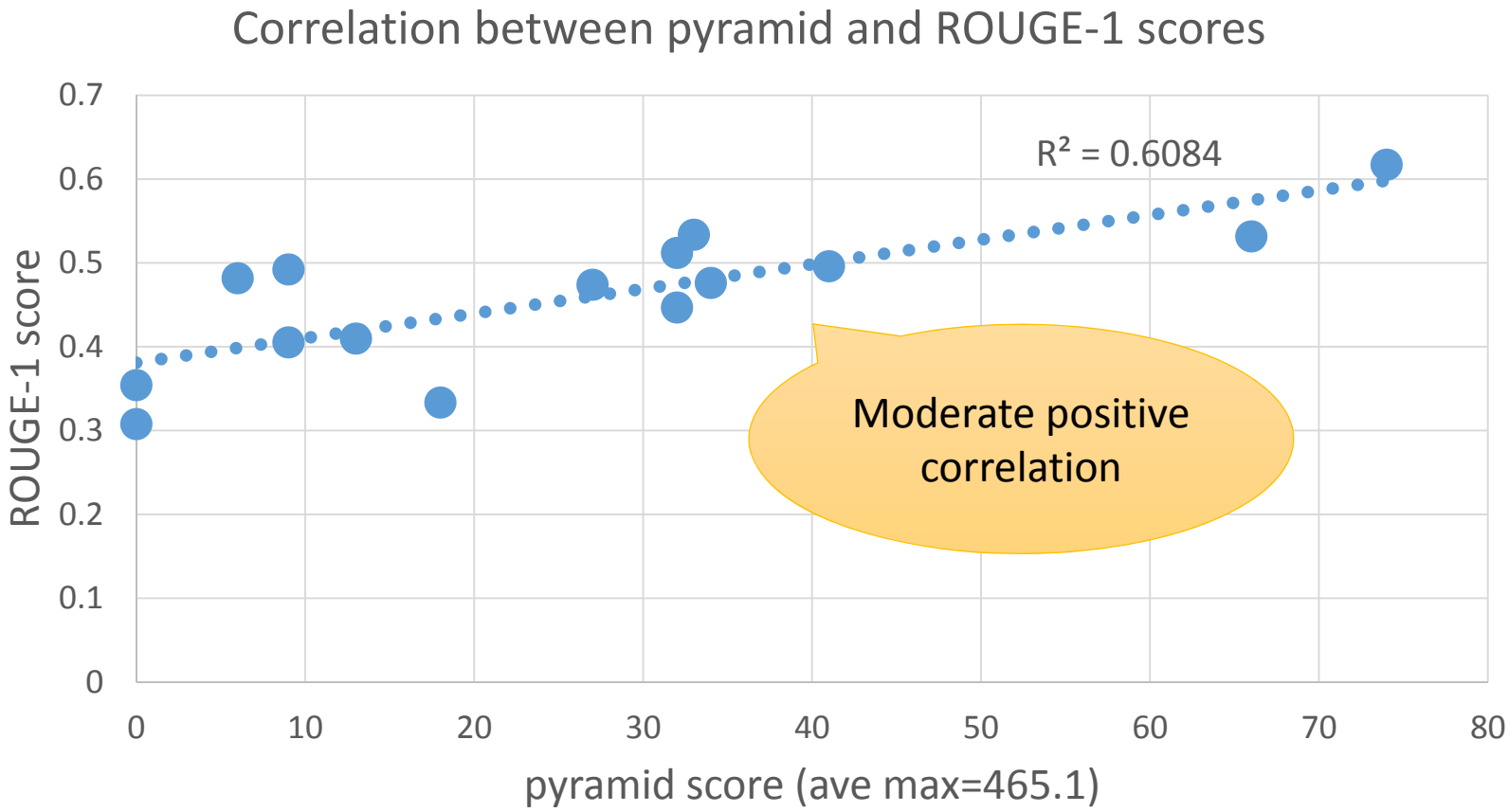
# Overview of named entity results



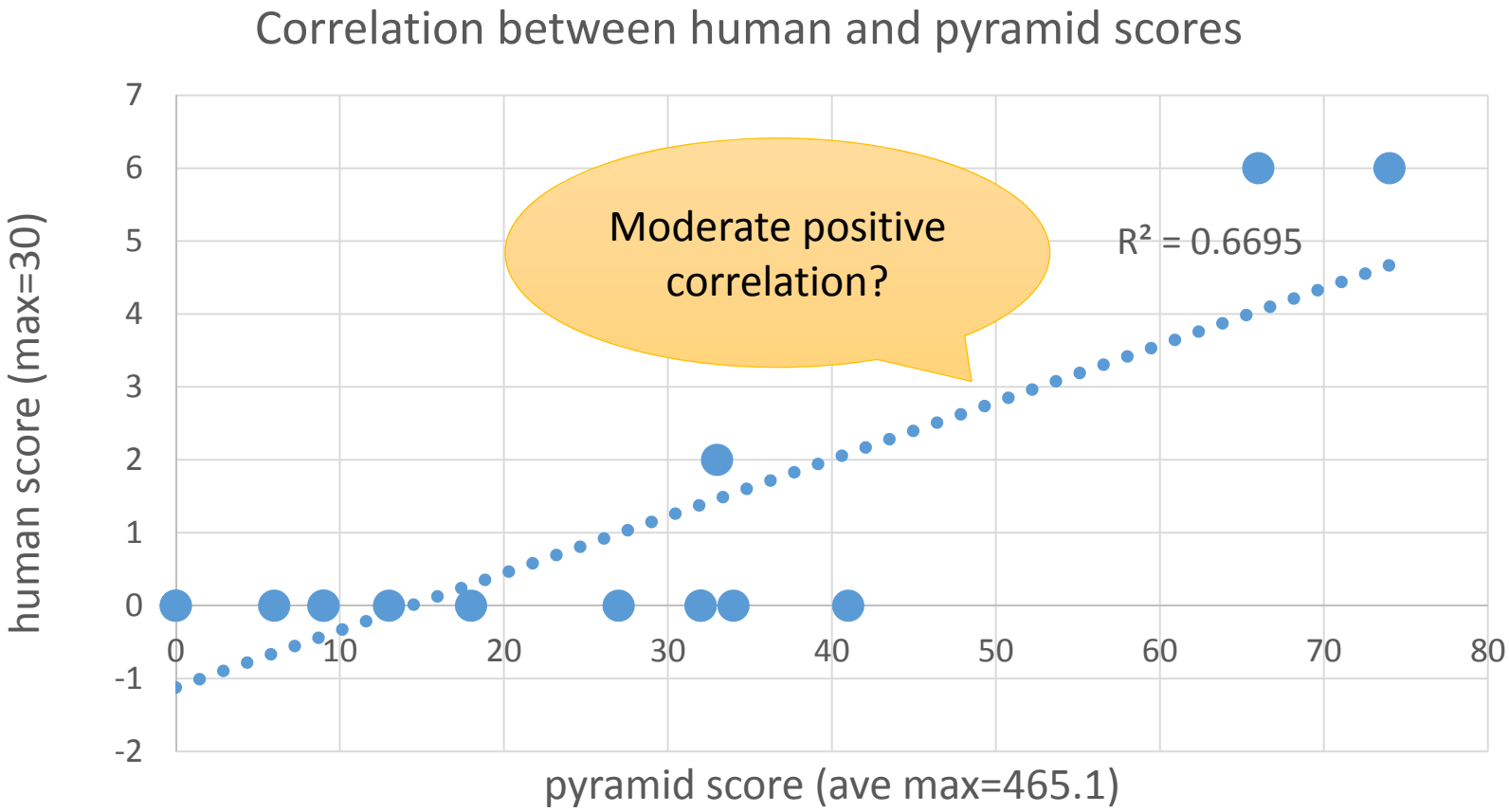
# Automatic vs. Human Evaluation

- ROUGE-1 method as automatic evaluation
- A human expert evaluated outputs with top priority
  - Giving a score [0-30]
  - Judging whether nuggets are included
- Comparison among ROUGE-1, pyramid and human scores

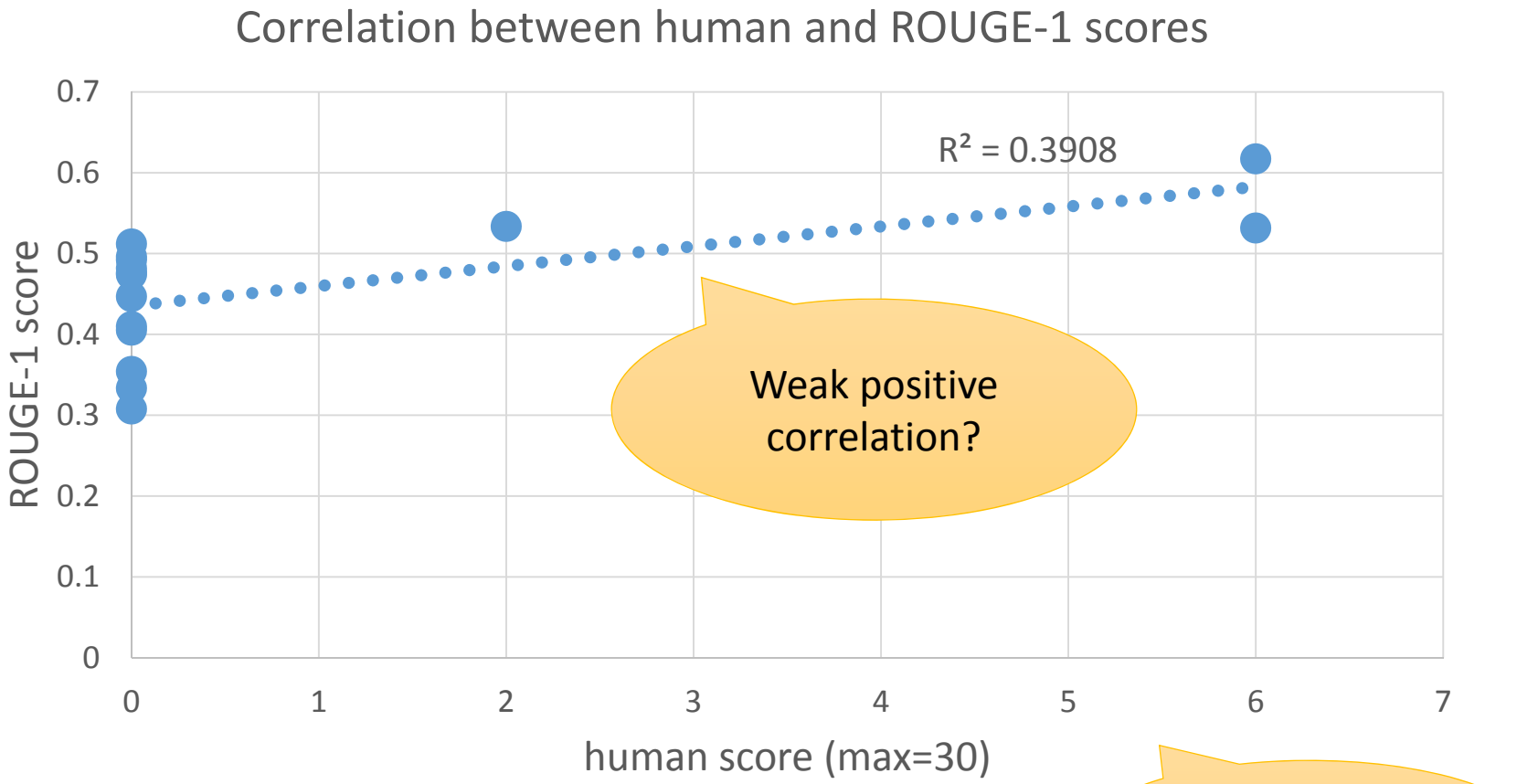
# Comparison between Pyramid and ROUGE-1 scores



# Comparison between Pyramid and Human scores



# Comparison between Human and ROUGE-1 scores



NTCIR


# Future research and subtasks for essay questions (NTCIR-13)

- End-to-End task
- Extraction subtask
- Summarization subtask
- Evaluation method subtask

- QA Lab Breakout Session:
  - June 9 (Th), 16:00-17:30
  - Conference Room 1

NTCIR


# Thank you for your attention!