Overview of the NTCIR-12 QA Lab-2 task

Hideyuki Shibuki (YNU), Kotaro Sakamoto (YNU, NII), Madoka Ishioroshi (NII), Akira Fujita (NII),
Yoshinobu Kano (Shizuoka Univ.), Teruko Mitamura, (CMU) Tatsunori Mori (YNU), Noriko Kando (NII, SOKENDAI)

Goal

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

Highlights

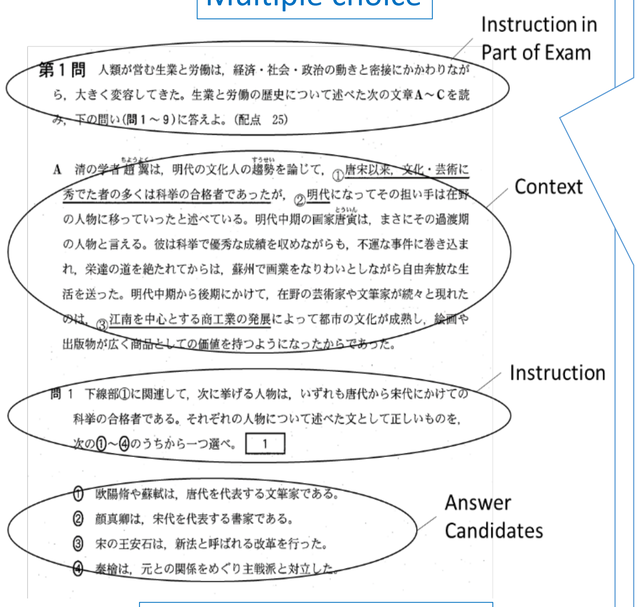
1. Multiple-choice and free-description type questions
2. Understanding of the surrounding context
3. Evaluation of free-description, especially, essay questions
4. Construction of a hierarchy of question formats (see the paper)
5. Competition with high-school students from all over Japan

Data Set

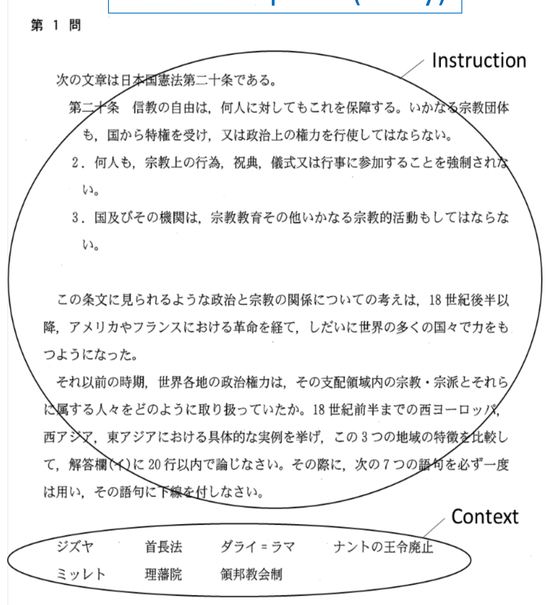
Exam name	Type	Training	Phase-1 (EN & JA)	Phase-2 (JA only)	Phase-3 (EN & JA)
Center Test	Multiple choice	97,01,03,05, 07,09	99	N/A	11
Secondary Exam	Free description	00,05,07,09	01,03,06,10	N/A	02,04,08,11
Yozemi (JA only)	Multiple choice	13b,13c	12,13a	N/A	13d,14a
Benesse (JA only)	Multiple choice	14Jun	14Nov	15Jun	14Sep
Sundai (JA only)	Free description	14Aug,14Nov	13Nov	15Aug	13Aug

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

Multiple choice



Free description (Essay)

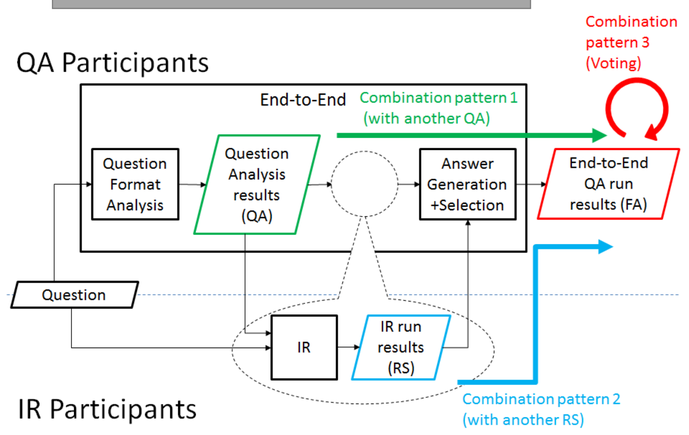


Question Data

XML Format

```
<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>
```

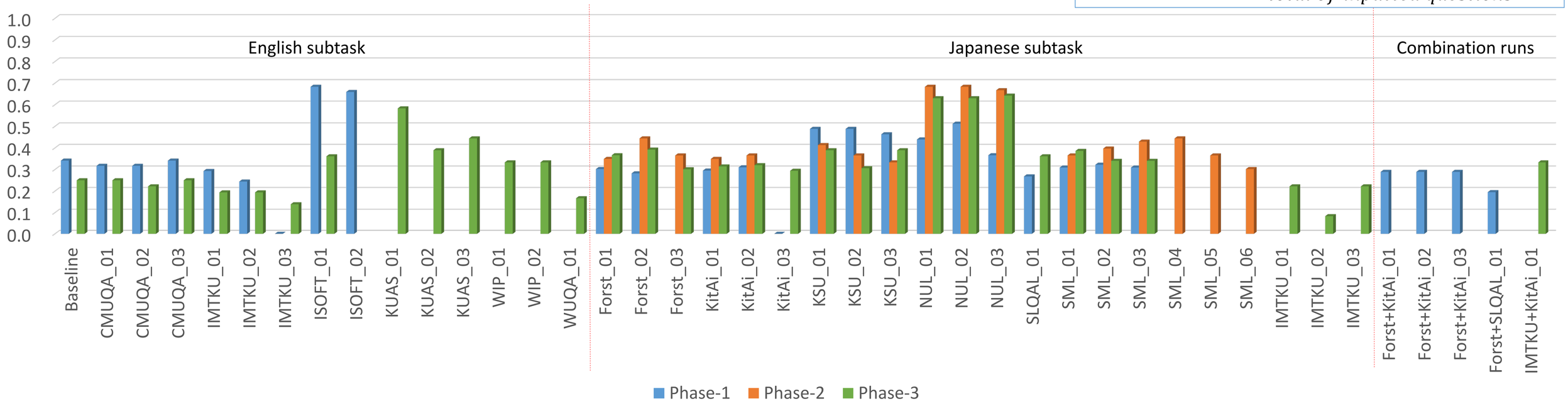
Combination Pattern



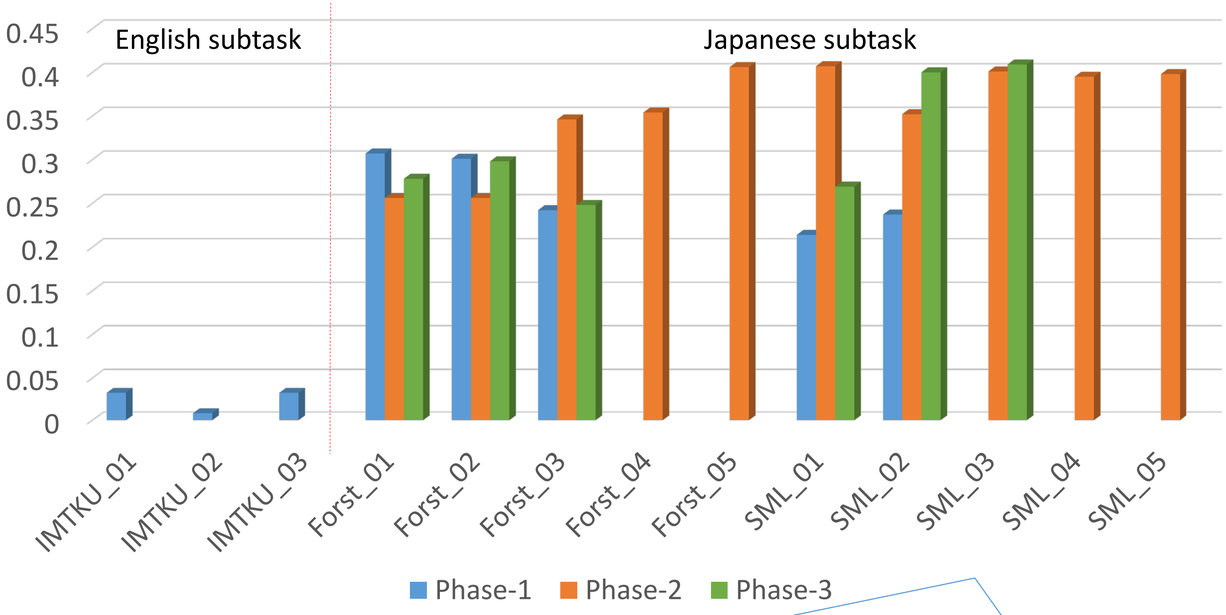
Results

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

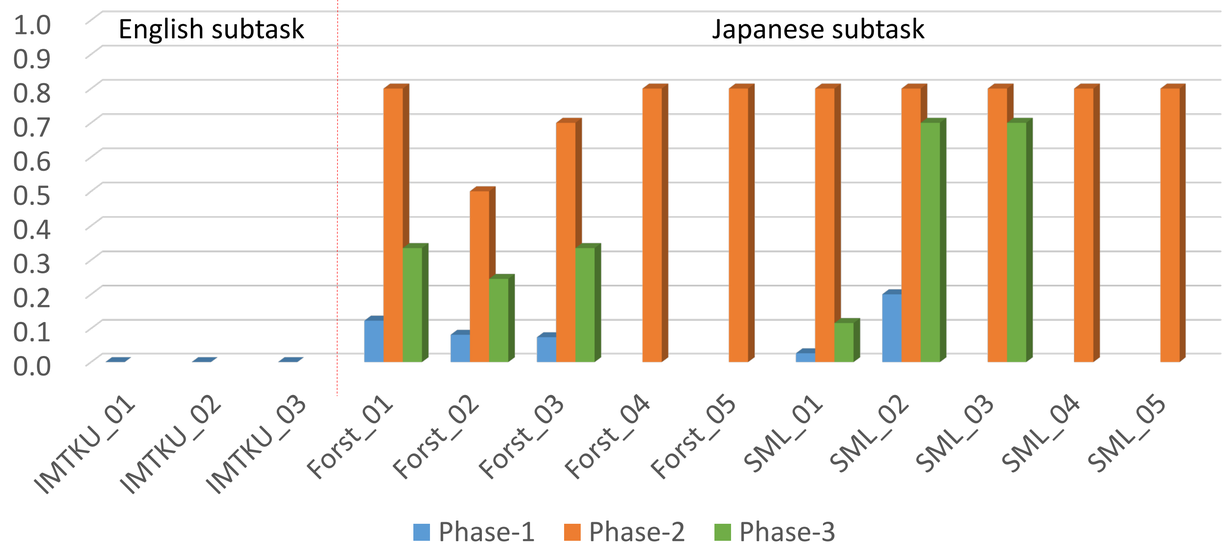
Correct rates of multiple-choice questions



ROUGE-1 scores of essay questions



Correct rates of free-description questions (except essay questions)

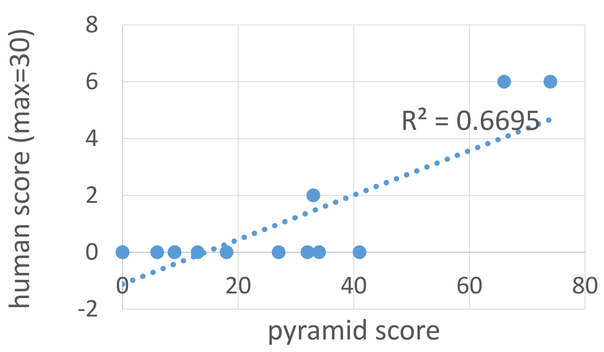


How much does ROUGE-1 score affect the evaluation of essay questions?

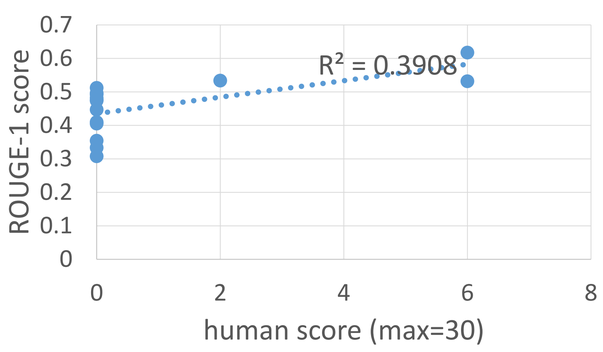
A human expert evaluated outputs with top priority.
Three human experts made nuggets for pyramid method.

A weak/moderate positive correlation was found, but ROUGE-1 is not enough to get logical consistency.

Correlation between human and pyramid scores



Correlation between human and ROUGE-1 scores



Correlation between pyramid and ROUGE-1 scores