Overview of the NTCIR-13 QA Lab-3 task

Hideyuki Shibuki (YNU), Kotaro Sakamoto (YNU, NII), Madoka Ishioroshi (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, Term and Essay type questions
2. Understanding of the surrounding context
3. Automatic evaluation of essays
4. Research run using the past QA Lab data

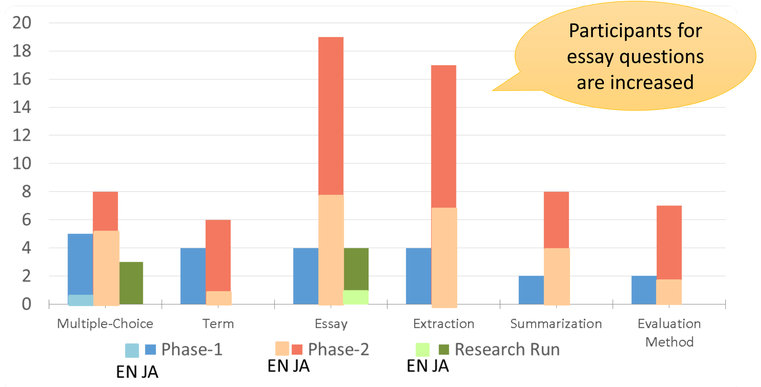
Data Set

Type	Training	Phase-1 (EN & JA)	Phase-2 (EN & JA)	Research run
Multiple choice	97,99,01,03,05, 07,09,11	12,13	14	07,11,13
Term & Essay	03,05,07,09,11	00,04,08,12,13	01,02,06,10,14	02,07,13

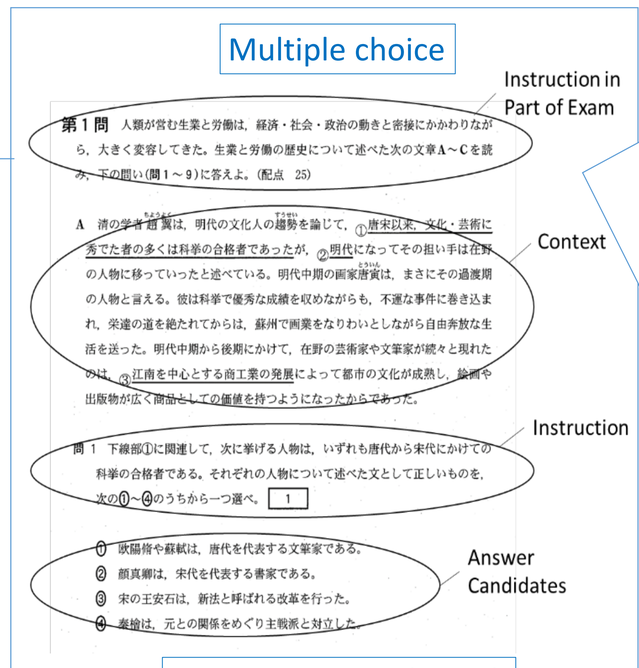
Participants are free to participate any particular phase and either of exams.

Submission

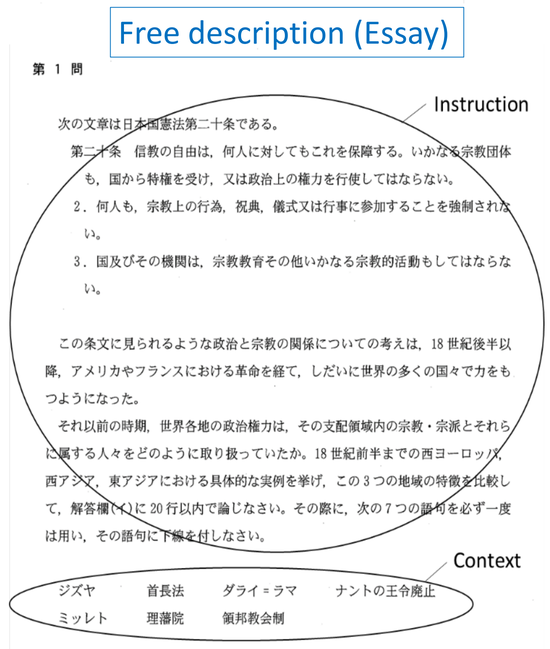
Phase-1:
24 runs from 6 teams
Phase-2
56 runs from 11 teams
Research run
6 runs from 4 teams



Multiple choice



Free description (Essay)

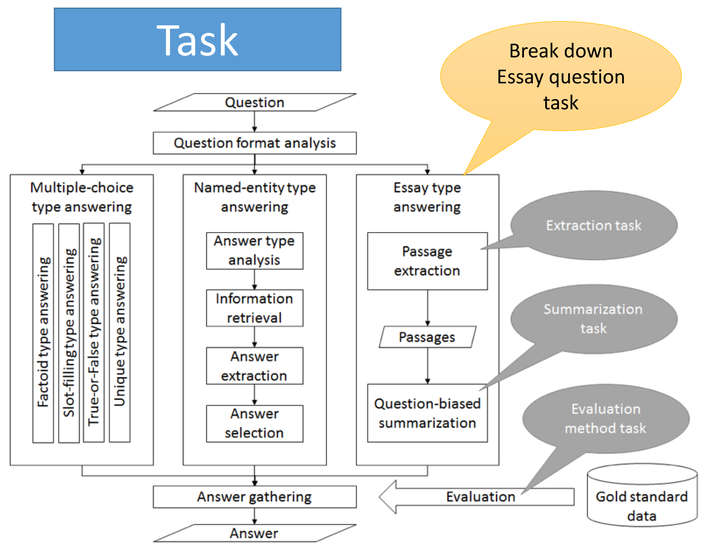


Question Data

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>
```

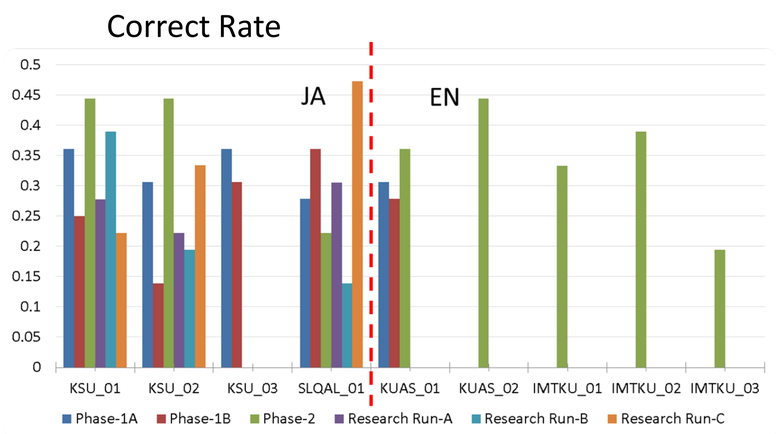
Task



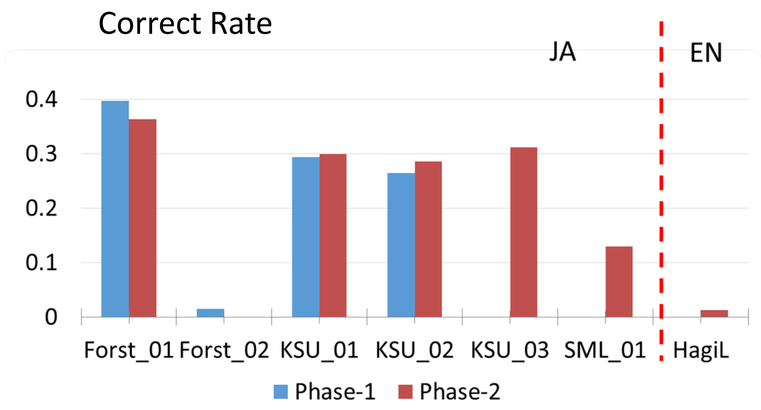
Results

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

Multiple-choice



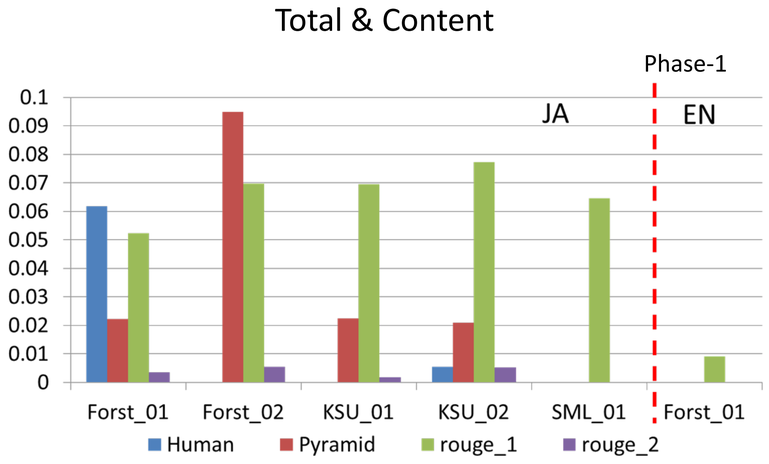
Term

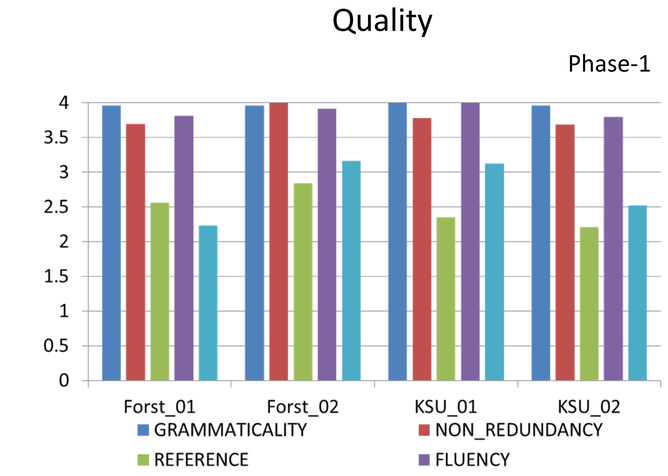


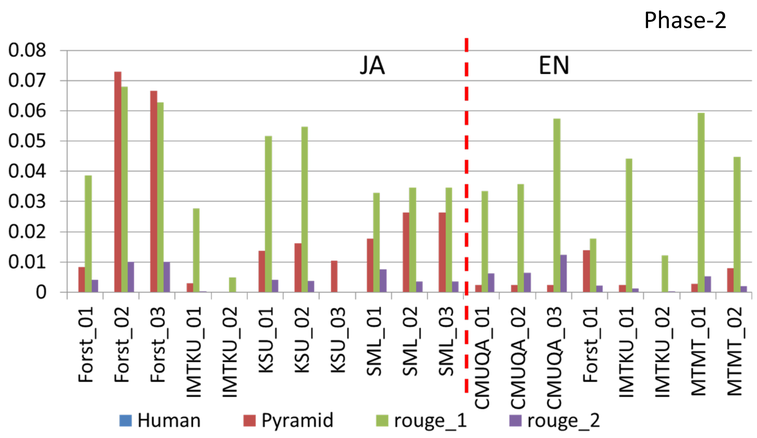
Essay

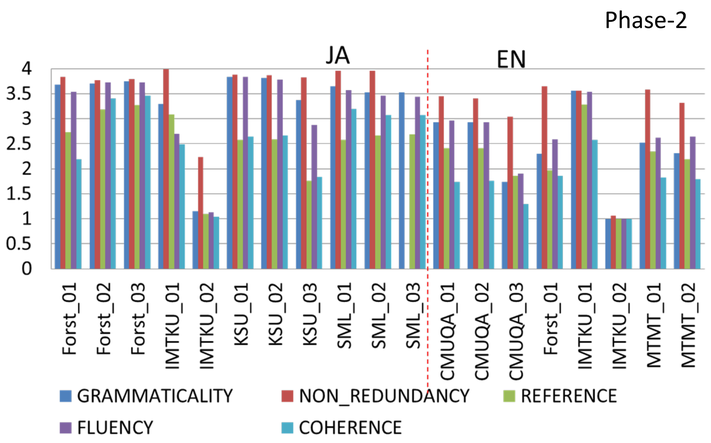
End-to-End

Evaluation:
Total: A human expert's mark Content: Pyramid, ROUGE-1, ROUGE-2
Quality: Grammaticality, Non redundancy, Reference clarity, Fluency, 'Coherence and content structure'









Most of contents are missed. The qualities of 'reference clarity' and 'coherence and content structure' are low. The improvement of them would enhance the total improvement

Extraction

Phase-1

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

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

Summarization

Phase-1

End-to-end Run

TeamID	Priority	Lang.	#of ques	#of N/A	content score Human	content score NUGGET	content score rouge_1	content score rouge_2	quality score GRAMMATICALITY	quality score NON_REDUNDANCY	quality score REFERENCE	quality score FLUENCY	quality score 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 Human	content score NUGGET	content score rouge_1	content score rouge_2	quality score GRAMMATICALITY	quality score NON_REDUNDANCY	quality score REFERENCE	quality score FLUENCY	quality score COHERENCE
Forst		Exp	JA	5	0	0	0.00356	0.01	0.00118	4	3.6	2.5	4	2
Forst	1	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	1	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

Phase-2

End-to-end Run

TeamID	Priority	Lang.	#of ques	#of N/A	content score Human	content score NUGGET	content score rouge_1	content score rouge_2	quality score GRAMMATICALITY	quality score NON_REDUNDANCY	quality score REFERENCE	quality score FLUENCY	quality score COHERENCE
Forst	1	JA	27	0	0	0.00829	0.0385	0.0042	3.68	3.84	2.73	3.53	2.19
Forst	2	JA	21	6		0.073	0.068	0.0101	3.7	3.77	3.18	3.73	3.41
Forst	3	JA	21	6		0.0666	0.0627	0.0101	3.75	3.8	3.27	3.73	3.45
IMTKU	1	JA	21	6	0	0.00295	0.0277	0.000318	3.29	3.98	3.09	2.7	2.49
IMTKU	2	JA	19	8		0	0.00491	0	1.15	2.23	1.1	1.12	1.04
KSU	1	JA	26	1	0	0.0137	0.0517	0.00404	3.83	3.88	2.58	3.84	2.64
KSU	2	JA	26	1		0.0161	0.0548	0.00374	3.81	3.87	2.58	3.78	2.66
KSU	3	JA	27	0		0.0105	0	0	3.37	3.82	1.76	2.88	1.83
SML	1	JA	21	6		0.0178	0.0328	0.00751	3.64	3.95	2.57	3.57	3.19
SML	2	JA	22	5		0.0264	0.0346	0.00345	3.52	3.95	2.66	3.45	3.07
SML	3	JA	22	5		0.0264	0.0346	0.00345	3.52	0	2.68	3.43	3.07
CMUQA	1	EN	27	0	0	0.00241	0.0334	0.0063	2.93	3.44	2.41	2.96	1.73
CMUQA	2	EN	27	0		0.00241	0.0358	0.00635	2.93	3.41	2.41	2.93	1.76
CMUQA	3	EN	27	0		0.00241	0.0573	0.0124	1.74	3.04	1.85	1.9	1.3
Forst	1	EN	17	10		0.014	0.0177	0.0021	2.29	3.65	1.97	2.59	1.85
IMTKU	1	EN	16	11	0	0.00234	0.0441	0.00124	3.56	3.56	3.29	3.53	2.58
IMTKU	2	EN	15	12	0		0.0121	7.86E-05	1	1.07	1	1	1
MTMT	1	EN	26	1		0.00282	0.0593	0.00523	2.52	3.58	2.34	2.62	1.83
MTMT	2	EN	24	3	0	0.00787	0.0448	0.0019	2.31	3.31	2.19	2.65	1.79

Summrization Run

TeamID	Priority	source	Lang.	#of ques	#of N/A	content score Human	content score NUGGET	content score rouge_1	content score rouge_2	quality score GRAMMATICALITY	quality score NON_REDUNDANCY	quality score REFERENCE	quality score FLUENCY	quality score 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

ExP: set of all passages submitted in Extraction task
GSN: set of gold standard nuggets
GSN+ExP: merged set of the above 2 sets

Evaluation:

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

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

Evaluation-method

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

Evaluation:
Rank correlation coefficient with human expert ranking