Forst: Question Answering System for Term and Essay Questions at NTCIR-13 QA Lab-3 Task

K. Sakamoto (YNU, NII), M. Ishioroshi (NII), Y. Fukuhara (YNU), A. Iizuka (YNU), H. Shibuki (YNU), T. Mori (YNU), N. Kando (NII)

Motivation

- We tackled the term question task and the essay question task including the evaluation-method subtask.
- Our systems for the term question task and the essay question end-to-end subtask are successors of our systems at the QA Lab-2.

Knowledge Source

- 4 textbooks (given in the task)
- Q & A collection (4,324 pairs)
- Glossary (6,081 words)
- Japanese thesaurus (about 300,000 entry words)
- World history event ontology (given in the task)

Evaluation Method

- Write in the name of the head of this country at the time of inauguration of the Orient Express service.

Essay Type Answering

- **Question Analysis**
  - Keywords (essay-with-keywords question)
  - Character limit
  - Time range

- **Document Retrieval and Sentence Extraction**
  - Group sentences containing each keyword
  - Generate keywords (essay-without-keyword question)

- **Sentence Compression**
  - Select a sentence from top-5 sentences in the group

- **Sentence Grouping, Sentence Ranking, Answer Candidate Generation, Sentence Sorting, Answer Candidate Reduction, Answer Candidate Ranking, and Answer Selection**
  - Remove answer candidates which exceed the character limit
  - Rank the answer candidates based on the scores of "answer likeness"

Term Type Answering

- **Question Analysis**
  - Analyze what is asked for (semantic answer type)

- **Document Retrieval**
  - Search related documents by extracted lexical answer type

- **Answer Candidate Extraction and Answer Selection**
  - Extract answer candidates from the retrieved documents
  - Rank the answer candidates based on the scores of "answer likeness"

Changes since the QA Lab-2

- Using **keyword importance**
  - The later keywords appear in a question, the more emphasized.
- Extending dictionary for NE of world history
- Adding decision rules for question types
- Using **majority decision score** for answer selection

The concept is use of "implicit keywords" that are question focuses but not stated positively.

Evaluation Method

1. Based on **world history terms**
2. Based on a given set of **gold standard nuggets**
   - The system count the number of terms (nuggets) in an essay
   - The count is regarded as the essay score

Conclusion

- We participated in all phases of the term question task and the essay question task in Japanese.
- Although the changes since the QA Lab-2 did not bring the major improvement.
- Using 'implicit keywords' extracted from question texts makes the results better.
- The evaluation method using gold standard nuggets achieved the best results.

Results

<table>
<thead>
<tr>
<th>Term</th>
<th>Evaluation Method</th>
<th>Spearman's Rho</th>
<th>Kendall's Tau b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1</td>
<td>Priority</td>
<td>0.427</td>
<td>0.334</td>
</tr>
<tr>
<td>Phase 2</td>
<td>Priority</td>
<td>0.596</td>
<td>0.534</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Essay</th>
<th>Human (Complex essay only)</th>
<th>ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1</td>
<td>Priority</td>
<td>0.011</td>
</tr>
<tr>
<td>Phase 2</td>
<td>Priority</td>
<td>0.0698</td>
</tr>
</tbody>
</table>

The second system only answered **simple essay**