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

Kotaro Sakamoto*1, *2, Madoka Ishioroshi*2, Yuta Fukuhara*1, Akihiro Iizuka*1, Hideyuki Shibuki*1, Tatsunori Mori*1, Noriko Kando*2, *3

Motivation

• QA is widely regarded as an advancement in IR.
• However, QA systems are not as popular as search engines in the real world.
• In order to apply QA systems to real-world problems
• We tackled the term question task and the essay question task including the evaluation-method subtask in Japanese
• 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 Sources

- 4 textbooks (Given in the task)
- World history event ontology (Given in the task)
- Glossary (6,081 words)
- Term Q & A collection (4,324 pairs)
- Essay Q & A collection (about 1,200 pairs from 6 books)
- Japanese thesaurus (about 300,000 entry words)
- English translation of the 4 textbooks by Google Translate
- English translation of the glossary by Google Translate
Term type answering (JA)

• The same pipeline as QA Lab-2’s Forst system

• Updates since QA Lab-2:
  • Using **keyword importance**
    • The later keywords appear in a question are more emphasized.
  • **Extending dictionary** for NE of world history
  • **Adding decision rules** for question types
  • Using **majority decision score** for answer selection
Essay type answering (JA)

• We developed 3 types of end-to-end system.
  1. QA Lab-2’s Forst system + MMR
     • The (minor) update is to add sentences from top in the MMR ranking when the answer is shorter than the length limitation
     • Using Okapi BM25 to extract sentences if there were no keywords; short essay questions.
     • (Released to the public) github.com/ktr-skmt/FelisCatusZero-multilingual
Essay type answering (JA)

- We developed 3 types of end-to-end system.
  1. QA Lab-2’s Forst system + MMR
  2. Use of `implicit keywords' that are question focuses but not stated positively
Essay type answering (JA)

- We developed 3 types of end-to-end system.
  1. QA Lab-2’s Forst system + MMR
  2. Use of `implicit keywords' that are question focuses but not stated positively
  3. Example based. Use of Q&A corpus (including test questions)
Essay evaluation method (JA)

• We developed two types of evaluation method systems
  1. Based on world history **terms**
     1. Simply counts the number of terms in essay
  2. Based on gold standard **nuggets**
     1. Segments an essay into sentences by punctuation
     2. Counts the number of nuggets that are matched with any one of sentences
        • If more than one term in a nugget are included in a sentence, the nugget is matched with the sentence.
## Evaluation Results – Term Question Task (JA)

<table>
<thead>
<tr>
<th></th>
<th>Correct rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase1</td>
<td>0.397</td>
</tr>
<tr>
<td>Phase2</td>
<td>0.273</td>
</tr>
</tbody>
</table>
Evaluation Results – Essay question’s end-to-end task

<table>
<thead>
<tr>
<th>Priority</th>
<th>Human expert (complex essay only)</th>
<th>ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.011</td>
<td>0.0523</td>
</tr>
<tr>
<td>2</td>
<td>0.0698</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(0.0887)</td>
<td></td>
</tr>
<tr>
<td>Phase 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0339</td>
<td>0.0385</td>
</tr>
<tr>
<td>2</td>
<td>0.0680</td>
<td></td>
</tr>
</tbody>
</table>
# Evaluation Results - Essay question’s evaluation method task

<table>
<thead>
<tr>
<th>Phase</th>
<th>Priority</th>
<th>Approach</th>
<th>Spearman’s Rho</th>
<th>Kendall’s Tau-b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase1</td>
<td>1</td>
<td>term</td>
<td>0.427</td>
<td>0.334</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>nugget</td>
<td>0.596</td>
<td>0.534</td>
</tr>
<tr>
<td>Phase2</td>
<td>1</td>
<td>term</td>
<td>-0.071</td>
<td>-0.049</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>nugget</td>
<td>0.404</td>
<td>0.360</td>
</tr>
</tbody>
</table>
Conclusion

• We participated in all phases of the term question task and the essay question task in Japanese
• Although the updates since the QA Lab-2 did not bring the major improvement
• Using `implicit keywords' extracted from question texts makes the results better
• The evaluation results of the evaluation method based on gold standard nuggets are moderate