An Overview of the 4th Question Answering Challenge (QAC-4) at NTCIR Workshop 6

Junichi Fukumoto
Tsuneaki Kato
Fumito Masui
Tsunenori Mori

Ritsumeikan University
University of Tokyo
Mie University
Yokohama National University
Previous QACs

- Evaluation of open domain question answering
  - Main task (5 ranked answers)
  - List task (all answers)
  - Information Access Dialogue (IAD) task
- Factoid question in QAC-1,2,3
Purpose of QAC-4

- Evaluation of open domain question answering
- Beyond factoid type questions
  - Answer extraction of long answer questions
- Evaluation method for long answer questions
  - Open evaluation by participants
Task Description

- Question Answering Track
  - Question answering evaluation using non-factoid questions

- Evaluation Track
  - Open evaluation using QAC-4 evaluation results
Question Answering Track

- Question will be non-factoid type question such as why-type, definition, question which has answer consists of multiple noun phrases.
- There will be 100 questions which are natural ones, not generated using target documents.
- System returns a set of answers for a question.
- Participants have to return human made answers for questions.
Evaluation Track

- Participants can join evaluation of QA Track with their own evaluation method.
- Participants will evaluate correctness and appropriateness for given questions using their own evaluation method, human evaluation or automatic evaluation.
Question Set


- For Formal Run, we prepared 100 questions.
  - We select 100 questions from 120 original questions.
Question set development

- We prepared several topics extracted from target documents.
- We asked a person to make arbitrary questions toward these topics.
- Questions will be basically beyond factoid questions.
- We also asked the person to make answers for the questions.
  - Answer data which have different contents have different answer ID.
  - New answers are added from participants answer data.
## Task Participants

<table>
<thead>
<tr>
<th>Institution</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aoyama Gakuin Univ. (HARAD)</td>
<td>1</td>
</tr>
<tr>
<td>Carnegie Mellon Univ. (LTI-J)</td>
<td>1</td>
</tr>
<tr>
<td>Hokkaido Univ. + Mie Univ. + Otaru Univ. of Commerce (HOMIO)</td>
<td>2</td>
</tr>
<tr>
<td>National Institute of Information and Communications Tech. (NICT)</td>
<td>2</td>
</tr>
<tr>
<td>NTT Communication Science Laboratories (NCQAW)</td>
<td>2</td>
</tr>
<tr>
<td>Ritsumeikan Univ. (RitsQ)</td>
<td>1</td>
</tr>
<tr>
<td>Toyohashi Univ. of Technology (TTH)</td>
<td>3</td>
</tr>
<tr>
<td>Yokohama National Univ. (Forst)</td>
<td>2</td>
</tr>
</tbody>
</table>
Schedule

Apr. 15, 2006  Call For Participation
May 31, 2006  Deadline of task participation
Jun. 22, 2006  Sample question set delivery
Sep. 25, 2006  Question set delivery
Oct. 20, 2006  System results due
Nov. 1, 2006  Start of Evaluation
Feb. 9, 2007  Evaluation results release
Evaluation method

- Two assessors have made evaluation and there is no overlapping between them.
- Information of document ID was not considered because of limited resource.
- Correct answer set was made from prepared answer set plus some of participants’ human answer sets.
  - 1171 answers for 100 questions (final)
  - 4499 answers for 100 questions (original data)
Evaluation data

- 14 submissions from 8 teams.
- There were 14,050 answers for 100 questions.
- Human evaluation was done for limited answers.
  - Reduced answer set submitted by participants
  - Or extracted answers from top 4 answers.
Evaluation criterion

- Human evaluation measure
  - Level A: System answer has almost the same contents as one of the correct answers.
  - Level B: System answer includes the contents of one of the correct answers.
  - Level C: System answer includes some part (not all one) of the contents of the correct answers.
  - Level D: System answer includes no information of any of the contents of the correct answers.
## Evaluation results of system answers

<table>
<thead>
<tr>
<th>System ID</th>
<th>All answers</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>No answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forest1</td>
<td>591</td>
<td>45</td>
<td>104</td>
<td>34</td>
<td>408</td>
<td>0</td>
</tr>
<tr>
<td>Forest2</td>
<td>317</td>
<td>30</td>
<td>52</td>
<td>21</td>
<td>214</td>
<td>2</td>
</tr>
<tr>
<td>HOMIO1</td>
<td>100</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>84</td>
<td>0</td>
</tr>
<tr>
<td>HOMIO2</td>
<td>100</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>86</td>
<td>0</td>
</tr>
<tr>
<td>LTI-J</td>
<td>377</td>
<td>24</td>
<td>30</td>
<td>13</td>
<td>310</td>
<td>1</td>
</tr>
<tr>
<td>NCQA1</td>
<td>330</td>
<td>37</td>
<td>15</td>
<td>6</td>
<td>272</td>
<td>32</td>
</tr>
<tr>
<td>NCQA2</td>
<td>323</td>
<td>31</td>
<td>11</td>
<td>4</td>
<td>277</td>
<td>32</td>
</tr>
<tr>
<td>NICT1</td>
<td>345</td>
<td>25</td>
<td>65</td>
<td>14</td>
<td>241</td>
<td>0</td>
</tr>
<tr>
<td>NICT2</td>
<td>363</td>
<td>6</td>
<td>119</td>
<td>24</td>
<td>214</td>
<td>0</td>
</tr>
<tr>
<td>HARAD</td>
<td>204</td>
<td>21</td>
<td>7</td>
<td>7</td>
<td>169</td>
<td>38</td>
</tr>
<tr>
<td>RitsQ</td>
<td>286</td>
<td>31</td>
<td>6</td>
<td>14</td>
<td>235</td>
<td>15</td>
</tr>
<tr>
<td>TTH1</td>
<td>353</td>
<td>34</td>
<td>36</td>
<td>24</td>
<td>259</td>
<td>0</td>
</tr>
<tr>
<td>TTH2</td>
<td>394</td>
<td>22</td>
<td>42</td>
<td>24</td>
<td>306</td>
<td>0</td>
</tr>
<tr>
<td>TTH3</td>
<td>354</td>
<td>30</td>
<td>43</td>
<td>26</td>
<td>255</td>
<td>0</td>
</tr>
<tr>
<td>Sum</td>
<td>4236</td>
<td>344</td>
<td>541</td>
<td>222</td>
<td>3330</td>
<td>120</td>
</tr>
<tr>
<td>average</td>
<td>302.6</td>
<td>24.6</td>
<td>38.6</td>
<td>15.9</td>
<td>237.9</td>
<td>8.6</td>
</tr>
</tbody>
</table>
Discussions: question type

- Various type questions
  - Why-type
  - How-type
  - Definition-type
  - Question for process, opinion, effect, situation, mechanism, problems, and so on.

  -> difficult type questions
Discussions: human evaluation

- 4 kinds of evaluation criterion
  - A-type (correct)
  - B-type (including correct contents)
  - C-type (including a part of correct contents)
  - D-type (wrong answer)

- Difficult judgment
  - Main contents of an answer is not correct answer
  - Too long answer strings
  - Too many answers
Discussions: other issues

- One system answer includes two or more answer contents.
- Many ways to express the same contents of an answer
- Constraints on answer length
- There is only one participation for evaluation track.
Conclusions

- Evaluation on QA beyond factoid type questions
- Test set development
  - Question and answer set
  - A number of system answers and human answers from participants
  - Evaluation results by assessors

- Next step of question answering evaluation
  - QAC-5 (factoid + non-factoid)
  - Automatic evaluation