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

○Junichi Fukumoto	Ritsumeikan University
Tsuneaki Kato	University of Tokyo
Fumito Masui	Mie University
Tsunenori Mori	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

- We used four year (1998-2001) Mainichi Newspaper articles for target document set of QAC-4.
- 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

Aoyama Gakuin Univ. (HARAD)	1
Carnegie Mellon Univ. (LTI-J)	1
Hokkaido Univ. + Mie Univ. + Otaru Univ. of Commerce (HOMIO)	2
National Institute of Information and Communications Tech.(NICT)	2
NTT Communication Science Laboratories (NCQAW)	2
Ritsumeikan Univ. (RitsQ)	1
Toyohashi Univ. of Technology (TTH)	3
Yokohama National Univ. (Forst)	2

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

System ID	All answers	A	B	C	D	No answer
Forest1	591	45	104	34	408	0
Forest2	317	30	52	21	214	2
HOMIO1	100	5	4	7	84	0
HOMIO2	100	3	7	4	86	0
LTI-J	377	24	30	13	310	1
NCQAW1	330	37	15	6	272	32
NCQAW2	323	31	11	4	277	32
NICT1	345	25	65	14	241	0
NICT2	363	6	119	24	214	0
HARAD	204	21	7	7	169	38
RitsQ	286	31	6	14	235	15
TTH1	353	34	36	24	259	0
TTH2	394	22	42	24	306	0
TTH3	354	30	43	26	255	0
Sum	4236	344	541	222	3330	120
average	302.6	24.6	38.6	15.9	237.9	8.6

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