Overview of NTCIR5 QAC3

Tsuneaki Kato	The University of Tokyo
Jun’ichi Fukumoto	Ritsumeikan University
Fumito Masui	Mie University

Introduction

- QAC is a series of challenges for evaluating QA technologies in Japanese
- QAC3 follows the same course as the previous two workshops
- The task limited to IAD task (QAC2 Subtask3)
- IAD Task assumes interactive use of QA systems, and evaluates the abilities needed under such circumstances
    - Proper interpretation of questions under a given dialogue context
    - Context processing abilities such as anaphora resolution and ellipsis handling

Talk Overview

- IAD task basic
- New trials in QAC3
- Process of QAC3
- Analysis and review
- Conclusion

Design: Individual Questions

- Factoid question that could be answered by names or values
  - Including common names such as names of species and body parts
  - Including titles of novels and movies
  - Excluding definition questions and why questions
- Extracting exact answers
  - Rather than the text snippets
- Returning one list of all and only correct answers

Design: Individual Questions

- Factoid question that could be answered by names or values
  - Including common names such as names of species and body parts
  - Including titles of novels and movies
  - Excluding definition questions and why questions
- Extracting exact answers
  - Rather than the text snippets
- Returning one list of all and only correct answers

Design: Setting

- Requesting systems to return answers to series of questions
  - This series of questions and the answers to those questions comprise a *simulated* information access dialogue
- A number of series are given in a BATCH manner
  - Boundary of series is given
  - It is NOT allowed to look ahead to the questions following the one currently being handled
- Two types of series: gathering and browsing
  - Type of the series is NOT given

Example of Series of Questions

- What genre does the “Harry Potter” series belong to?
- Who is the author?
- Who are the main characters in that series?
- When was the first volume published?
- What title does it have?
- How many volumes were published by 2001?
- How many languages has it been translated into?
- How many copies have been sold in Japan?

Series 02: Gathering Type

Design: Evaluation Measure

- Mean Modified F measure is employed, which takes account of both precision and recall
- An answer itself and an accompanying article that supports it are judged
- The correctness of an answer is determined according to the interpretation of a given question done by human assessors within the given context
  - The system’s answer to previous questions and its understanding of the context are irrelevant to the correctness

Design: Evaluation Measure

- Mean Modified F measure is employed, which takes account of both precision and recall
- An answer itself and an accompanying article that supports it are judged
- The correctness of an answer is determined according to the interpretation of a given question done by human assessors within the given context
  - The system’s answer to previous questions and its understanding of the context are irrelevant to the correctness

Constructing a Test Set

- Collecting questions
  - Questions to elicit information for a report on a given topic
  - Series of wh-type questions possibly with anaphoric expressions
- Choosing and rearranging collected series for gathering type series
- Using some questions as seeds of series and adding new questions to create flow to/from those questions for browsing type series

Constructing a Test Set

- Collecting questions
  - Questions to elicit information for a report on a given topic
  - Series of wh-type questions possibly with anaphoric expressions
- Choosing and rearranging collected series for gathering type series
- Using some questions as seeds of series and adding new questions to create flow to/from those questions for browsing type series

Reference Test Set

- For an isolated evaluation of context processing
- For examining degree of context dependency of questions
- Two accompanying test set for reference, which measure the ceiling and floor of the context processing
    - By manually resolving all anaphoric expressions including zero anaphora
    - By mechanically deleting anaphoric expressions

Examples of Questions in the Reference Test Set

- What genre does the “Harry Potter” series belong to?
- Who is the author of the “Harry Potter” series?
- Who are the main characters in the “Harry Potter” series?
- When was the first volume of the “Harry Potter” series published?
- What is the title of the first volume of the “Harry Potter” series?
- How many volumes of the “Harry Potter” series were published by 2001? …

Correspondents of Series 02

New Trials in QAC3

- Redefinition of the scope of answers and questions
  - Conventional and colloquial range expressions
  - Expressions of approximate or round numbers
  - Description of events and objects
- New evaluation measures
  - Multi-grade evaluation
  - Correct answer set (CAS)
- New method for test set construction
  - A WoZ method

New Evaluation Measure

- Two types of qualities of answers
  - Quality in expression   ex. full name / nickname
  - Quality of answer itself   ex. regular/reserve

⇒ Multi-graded evaluation

- More than one way to enumerate all correct answer
  - {“Three prefectures of Tokai region”}
  - {“Mie”, “Aichi”, Gifu”}

⇒ Introducing a concept of correct answer set

QAC3 Process and Schedule

- Declared in June 2004, at NTCIR4 Workshop meeting
- The formal run of QAC3 was conducted in April 2005
- The reference run was conducted in May 2005, without a strict deadline
- The final evaluation was delivered in August
- Seven teams (Sixteen systems) participated in the run

QAC3 Test Set

- 50 series (35 gathering + 15 browsing), 360 questions
- 5 to 10 questions in one series, the average is 7.2
- Average number of correct answers is 1.98, 204 questions have only one correct answer
- Multiple CASs is needed for 37 questions

- Sloppy definition of the gathering type series

Example of Series of Questions

- When did Asahi breweries Ltd. start selling their low-malt beer?
- What is the brand name?
- How much did it cost?
- What brands of low-malt beer were already on the market at that time?
- Which company had the largest share?
- How much low-malt beer was sold compared to regular beer?
- Which company made it originally?

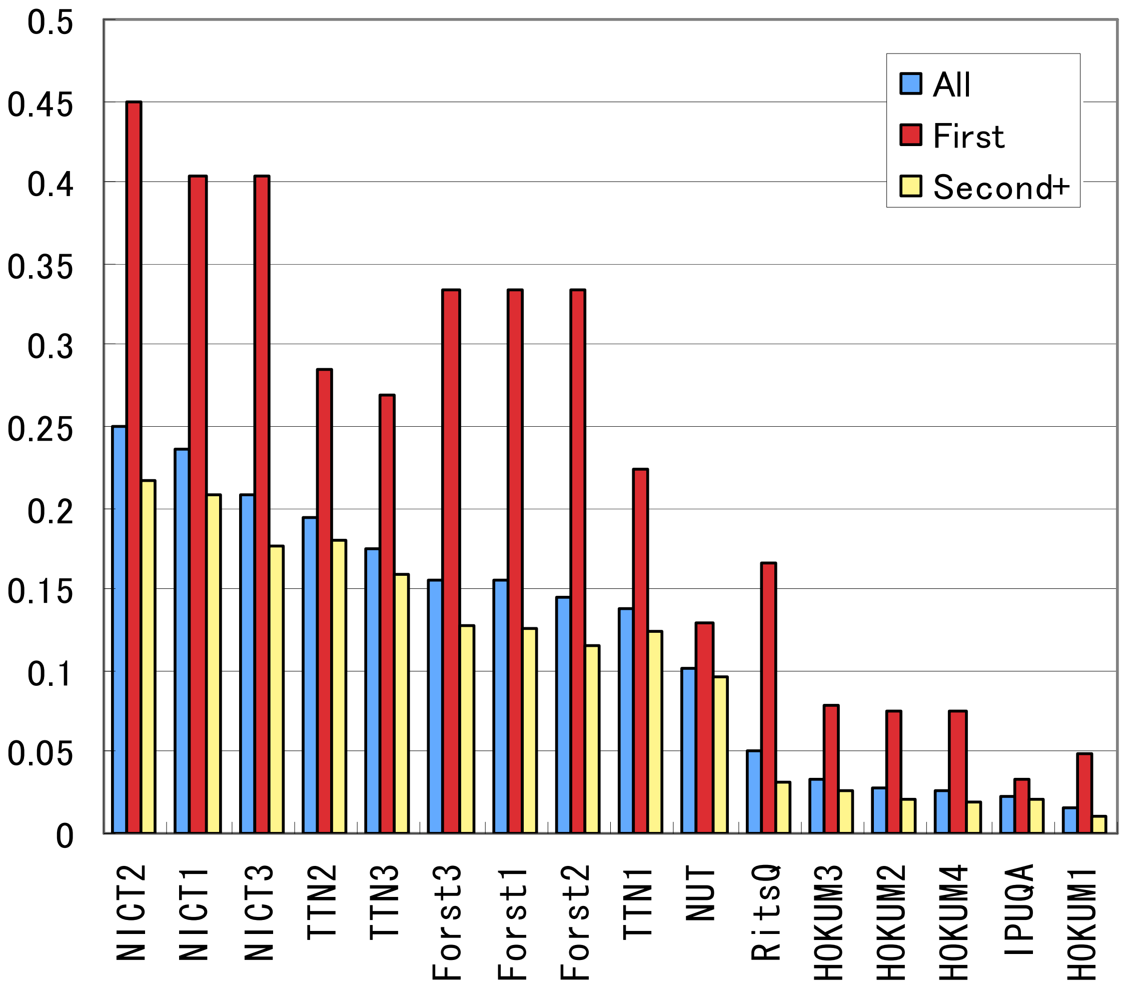
Series 04: Gathering Type

Example of Series of Questions

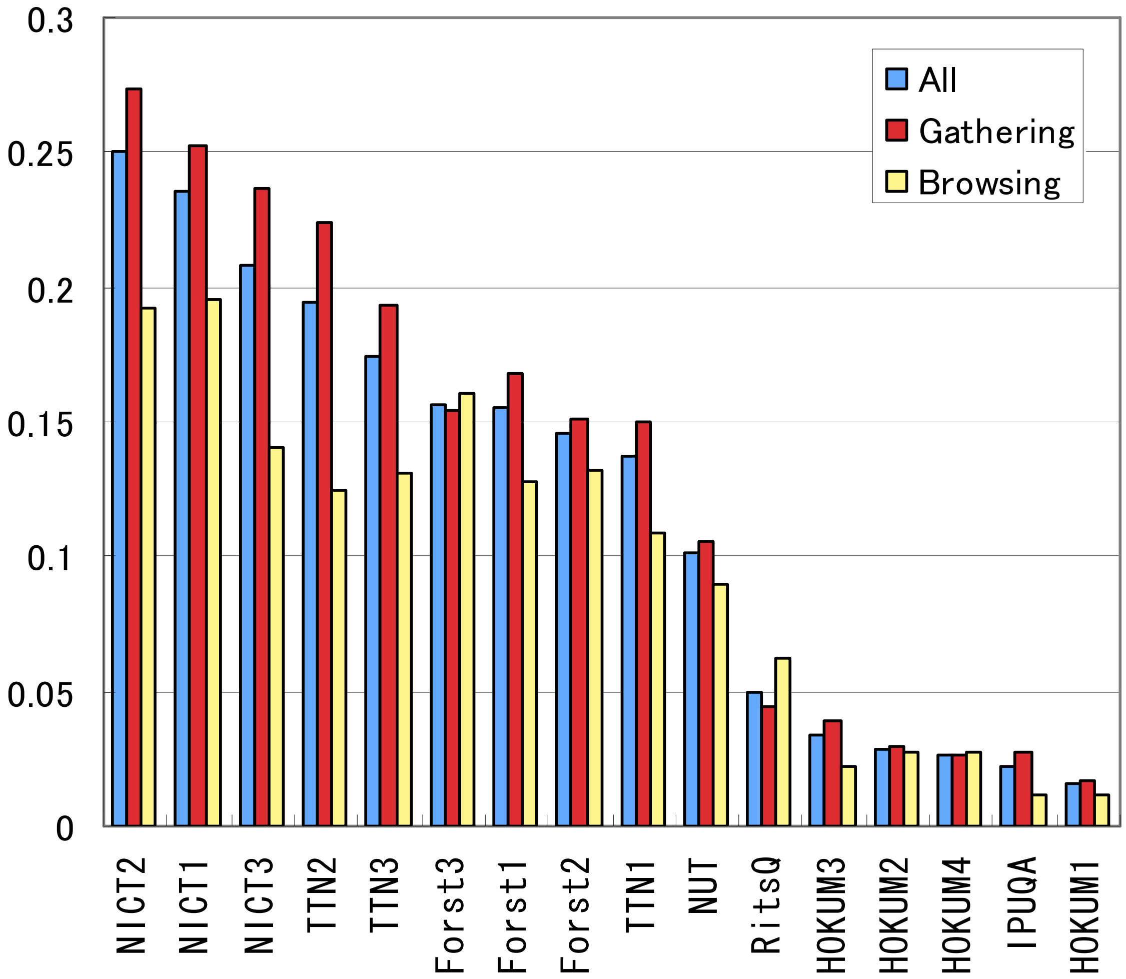
- Where was Universal Studio Japan constructed?
- Which train station is the nearest?
- Who is the actor who attended the ribbon-cutting ceremony on the opening day?
- What is the movie he was featured in that was released in the New Year season of 2001?
- What is the movie starring Kevin Costner released in the same season?
- What was the subject matter of that movie?
- What role did Costner play in that movie?

Series 24: Browsing Type

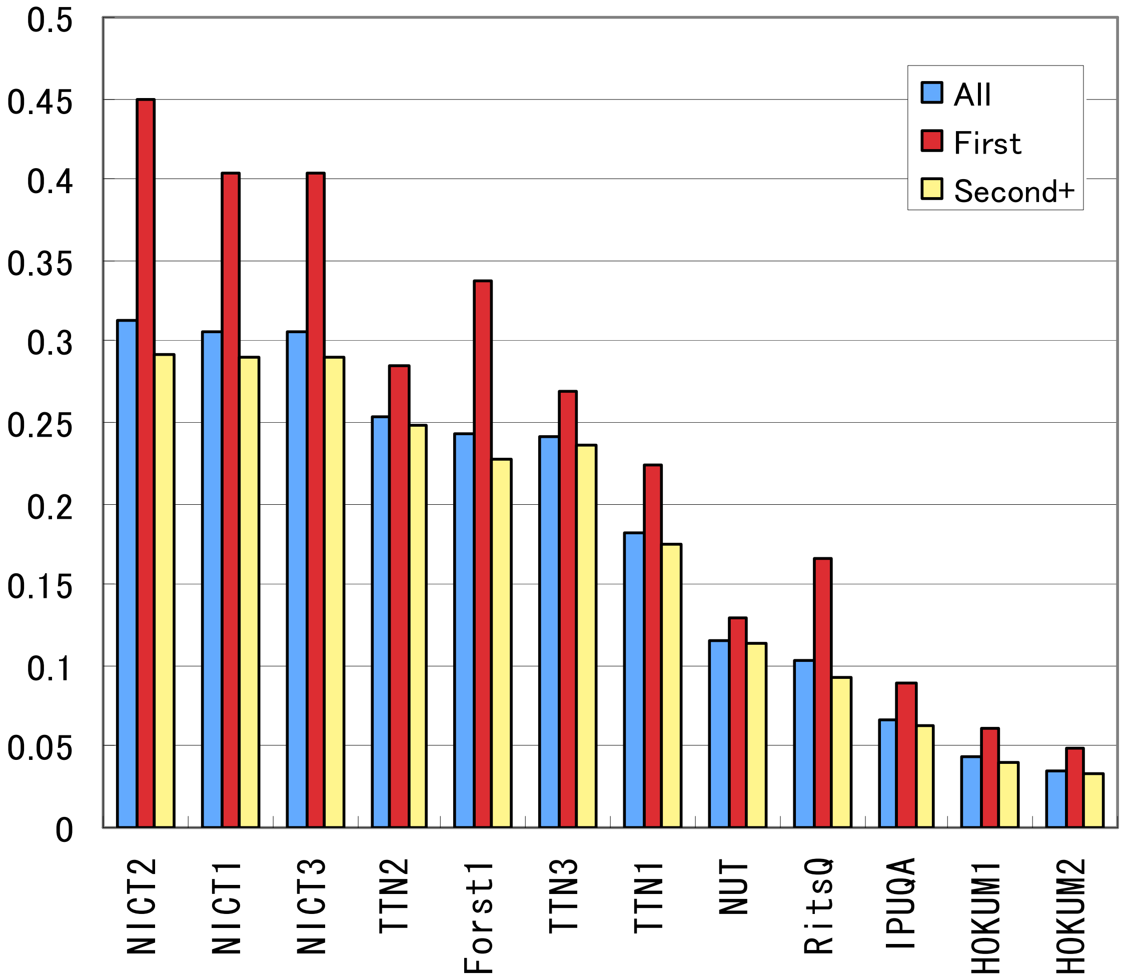
Evaluation by MMF



Differences on Series Type



Evaluation of Reference Run



Remaining Problems

- Insufficient for the purpose of constructing reusable test set
- Problems related task definition and test set construction remain unsolved
- Needs to discuss the direction of QA technologies

Conclusion

- Great success, although we face some problems for future work
- IAD task has become more sophisticated
- Several new methods1 were tried and evaluated

Information Access Dialogues

- Gathering Type
  - A concrete objective such as writing report or summary on a specific topic
  - All questions are concerning the common global topic
  - Each consecutive question shares a local context
- Browsing Type
  - No fixed topic of interest; the topic of interest varies as the dialogue progresses
  - No global topic covers a whole dialogues
  - Each consecutive question shares a local context