Text Summarization Challenge 3
Text Summarization Evaluation
at NTCIR Workshop

Takahiro FUKUSIMA
*Tsutomu HIRAO
Hidetsugu NANBA
Manabu OKUMURA

Tasks
- Two kinds of summarization -
  • Extraction
    – Extracting important sentences from document sets
    limitation: No of sentences
  • Abstraction
    – Producing summaries from document sets
    limitation: No of characters

The limitations were given by the organizers
The length: short, long

Data
• 30 document clusters
  – 30 event document clusters selected by organizers from both Mainichi and Yomiuri Newspaper articles
  – redundant source: almost 10 documents / topic
• Data given for participants
  – 30 document clusters
  – Titles of the clusters
  – (Questions about important parts concerning document clusters generated by human summarizers)*
  * The participants judges whether use this information or not

Participants
• 9 participants from university and governmental research org.
  – 10 systems for extraction
  – 9 systems for abstraction
  – University: 6, Gov.-Univ.:3

Evaluation Methods
• Extraction
  – Intrinsic evaluation
    • Precision, Coverage
  – Abstraction
  – Intrinsic evaluation
    • Content: Information Coverage
    • Readability: Quality Questions (We modified DUC's QQ for Japanese text)
    • Extrinsic evaluation
      • Pseudo-Question-Answering

Comparison among topics (short)

Evaluation Results (Readability)
• Q04: How many expressions which have same meanings but different term are there?
  – System average: 2.05 (short), 4.16 (long)
  – Human: 0.433, 1.133
• Q08: Does the summary have wrong chronological ordering (yesÆ-1, noÆ+1, otherÆ0)
  – System average: -0.21, -0.58
  – Human: 0.933, 0.800