Overview of the NTCIR-10 1CLICK-2 Task

Makoto P. Kato
Matthew Ekstrand-Abueg
Virgil Pavlu
Tetsuya Sakai
Takehiro Yamamoto
Mayu Iwata
Talk Outline

1. What is 1CLICK Task?
2. The 1CLICK-2 Task
3. Results
4. Summary and Future work
What is the 1CLICK Task?
Suppose that ...

- Finding answers for a question
  “what’s the difference between PDP and LCD?”
In the "ten-blue-link" paradigm

1. Enter query
2. Click SEARCH button
3. Scan ranked list of URLs
4. Click URL
5. Read URL contents
6. Get all desired information

More than one clicks needed before being satisfied
This is “One Click Access”

1CLICK ACCESS SYSTEM

differences between PDP and LCD

Search

LCD is better in terms of the weight, size and energy saving.

PDP shows a better black color, a faster response speed, and a wider view angle.
One Click Access

= Immediate + Direct Information Access
One Click Access

Enter query

Click SEARCH button

Get all desired information

The system outputs *X-string*

Task:
Given a search query, return a single textual output (X-string)

Go beyond the "ten-blue-link" paradigm, and tackle information retrieval rather than document retrieval.
Evaluation of 1CLICK Systems

- Manual/automatic matching between the X-string and nuggets

**Phone:** 046-223-3636  **Fax:** 046-223-3630  **Address:** 118-1 Nurumizu, Atsugi, 243-8551  **Email:** soumu@shonan-atsugi.jp

**X-string**

**Nuggets**

a sentence relevant to the information need for a query

- **Phone number:** 046-223-3636
- **Fax number:** 046-223-3630  
- **Address:** 118-1 Nurumizu, Atsugi

- Systems are required to present more important information earlier
Unlike nugget precision/recall, **S-measure** (position-aware weighted recall) says (a)<(b). **T-measure** (a kind of precision) says (b)>(c). **S#** (official evaluation metric) combines S and T.
1CLICK Challenges

- For participants
  - Multi-document summarization for a given query
  - Precise estimation of the nugget importance
    - Not binary but graded importance
  - Readability of X-strings

- For organizers
  - Efficient nugget construction
  - Flexible, feasible, and consistent nugget matching
  - Appropriate evaluation metrics
The 1CLICK-2 Task
1CLICK-2 Task Structure

Crawled

Documents -> Queries

Nuggets

Units for matching

iUnits (J) -> Vital strings (E)

Participants

Runs

Matched positions -> S# scores
Main tasks (English + Japanese)

- Given a search query, return a X-string (a single textual output)

Options

- Device types (the length limit for X-strings)
  - **DESKTOP**: 1,000/500 characters for English/Japanese
  - **MOBILE**: 280/140 characters for English/Japanese

- Source types (from which X-strings must be generated)
  - **MANDATORY**: only distributed documents
  - **ORACLE**: only distributed documents with an “ORACLE” list
  - **OPEN**: any resources
Given a search query, return the query type

For *componentized* evaluation

**Queries**

“michael jackson death”
“sylvester stallone”
“robert kennedy cuba”
“ichiro suzuki”
“atlanta airport”
“kyoto hot springs”
“parkinsons disease”
“why is the sky blue?”

**Query types**

ARTIST
ACTOR
POLITICIAN
ATHLETE
FACILITY
GEO
DEFINITION
QA
1CLICK-2 Task Structure

Crawled

Documents  Queries

Nuggets

Units for matching

iUnits (J)  Vital strings (E)

Participants

Runs

Matched positions  S# scores
8 query types used

(for which the user's information need can be satisfied by the search results page)

<table>
<thead>
<tr>
<th>1CLICK-1</th>
<th>1CLICK-2</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEBRITY</td>
<td>ARTIST</td>
<td>“michael jackson death”</td>
</tr>
<tr>
<td>FACILITY</td>
<td>ACTOR</td>
<td>“sylvester stallone”</td>
</tr>
<tr>
<td>DEFINITION</td>
<td>POLITICIAN</td>
<td>“robert kennedy cuba”</td>
</tr>
<tr>
<td>QA</td>
<td>ATHLETE</td>
<td>“ichiro suzuki”</td>
</tr>
<tr>
<td></td>
<td>FACILITY</td>
<td>“atlanta airport”</td>
</tr>
<tr>
<td></td>
<td>GEO</td>
<td>“kyoto hot springs”</td>
</tr>
<tr>
<td></td>
<td>DEFINITION</td>
<td>“parkinsons disease”</td>
</tr>
<tr>
<td></td>
<td>QA</td>
<td>“why is the sky blue?”</td>
</tr>
</tbody>
</table>

Based on [Li et al., SIGIR09]
1CLICK-2 Task Structure

- Documents
- Queries

- Nugget Extraction

- Nuggets
  - iUnits (J)
  - Vital strings (E)

- Participants

- Runs

- Matched positions

- S# scores
Nugget Extraction

- Nugget: a sentence relevant to a given query
  - e.g. For query “ichiro suzuki”,

  Ichiro is a professional baseball outfielder who is currently with the New York Yankees

- In Japanese 1CLICK-2, organizers worked very hard to collect all possible nuggets in advance

Very time-consuming process
Semi-automatic Nugget Extraction

- Was applied to English 1CLICK-2
- Mutual, iterative reinforcement between nuggets and documents

Refer to our SIGIR 2013 paper to see the performance:
Will have a DEMO on DAY-4 (6/21)
1CLICK-2 Task Structure

Crawled

Documents → Queries → Nuggets

Units for matching → iUnits (J) → Vital strings (E)

Participants

Runs

Matched positions → S# scores
Problems in 1CLICK-1 X-string-Nugget Matching

- Granularity problem

Phone: 046-223-3636. Fax: 046-223-3630.
Address: 243-8551.
Email: soumu@shonan-atsugi.jp.

X-string

What's the score?
- Very famous
- Less famous

Nuggets

- Her main work includes “I Will Always Love You”, "Lover for Life”, and “How Will I Know”.

Match?

Phone number: 046-223-3636
- Fax number: 046-223-3630
- Address: 118-1 Nurumizu, Atsugi, 243-8551

Importance problem
Breaking Nuggets into Finer-grained Units

Nuggets
“Murray tried to revive Jackson for ten minutes, at which point he realized he needed to call for help.”

Relevant, Atomic, and Dependent Information pieces

<table>
<thead>
<tr>
<th>ID</th>
<th>iUnits</th>
<th>vital string</th>
<th>w</th>
<th>dep</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Murray tried to revive Jackson</td>
<td>Murray tried to revive</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Murray tried to revive Jackson for ten minutes</td>
<td>ten minutes</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Murray realized he needed to call help</td>
<td>realized he needed to call help</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
1CLICK-2 Task Structure

- **Crawled**
  - Documents
  - Queries

- **Units for matching**
  - iUnits (J)
  - Vital strings (E)

- **Matching**
  - Matched positions

- **Runs**
  - Participants
  - S# scores
### Semi-automatic Matching

**Automatic matching**

After a year as a solo artist, Gaye ranked as the label's top-selling solo artist during the sixties. Due to solo hits including "How Sweet It Is (To Be Loved By You)", "Ain't That Peculiar", "I Heard It Through the Grapevine" and his duet singles with singers such as Mary Wells and Tammi Terrell, he was crowned "The Prince of Motown" and "The Prince of Soul". Notable for fighting the hit-making but restrictive Motown process in which performers and songwriters were kept separate, Gaye proved with albums like his 1971 *What's Going On* and his 1973 *Let's Get It On* that he was able to produce music without relying on the system. *What's Going On* and *I Want You* albums helped influence the quiet storm, urban adult contemporary and slow jam genres.

**Manual matching**

<table>
<thead>
<tr>
<th>Vital String</th>
<th>Context</th>
<th>Search:</th>
</tr>
</thead>
<tbody>
<tr>
<td>10. Let's Get It On</td>
<td>570-586</td>
<td><strong>Low</strong></td>
</tr>
<tr>
<td>11. inspiring fellow Motown artists</td>
<td>651-683</td>
<td><strong>Low</strong></td>
</tr>
<tr>
<td>12. Stevie Wonder and Michael Jackson</td>
<td>691-725</td>
<td><strong>Low</strong></td>
</tr>
<tr>
<td>13. mid-1970s work influenced quiet storm</td>
<td>745-846</td>
<td><strong>High</strong></td>
</tr>
<tr>
<td>14. mid-1970s work influenced slow jam genres</td>
<td>745-893</td>
<td><strong>High</strong></td>
</tr>
<tr>
<td>15. mid-1970s work influenced urban adult</td>
<td>745-875</td>
<td><strong>High</strong></td>
</tr>
</tbody>
</table>
1CLICK-2 Task Structure

- Documents
- Queries
- Nuggets
  - iUnits (J)
  - Vital strings (E)
- Participants
- Runs
- Matched positions
- S#-measure computation
- S# scores

Crawled
S, T, and S#-measure

**S-measure**

X-string

\[
S = \frac{1}{Z} \sum_{i \in M} w(i)d(i)
\]

w: weight, Z: normalization factor

d(i) = \max(0, L - \text{offset}(i))

discounts the iUnit (or VS) weight based on its offset

**T-measure**

\[T = \% \text{ of matched text}\]

**S#-measure**

The harmonic mean of S and T (official evaluation metric in 1CLICK-2)
1CLICK-2 Task Structure

- Documents
- Queries

Nuggets

iUnits (J)  Vital strings (E)

Participants

Runs

Matched positions

S# scores
# Participants

## English (5 teams, 28 runs)

<table>
<thead>
<tr>
<th>team name</th>
<th>MAIN</th>
<th>QC</th>
<th>organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>KUIDL</td>
<td>4</td>
<td>2</td>
<td>Kyoto University</td>
</tr>
<tr>
<td>NSTDB</td>
<td>6</td>
<td>0</td>
<td>Nara Institute of Science and Technology</td>
</tr>
<tr>
<td>NUIR</td>
<td>8</td>
<td>0</td>
<td>Northeastern University, USA</td>
</tr>
<tr>
<td>udem</td>
<td>4</td>
<td>0</td>
<td>University of Montreal</td>
</tr>
<tr>
<td>ut</td>
<td>2</td>
<td>2</td>
<td>University of Twente</td>
</tr>
</tbody>
</table>

## Japanese (5 teams, 22 runs)

<table>
<thead>
<tr>
<th>team name</th>
<th>MAIN</th>
<th>QC</th>
<th>organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUKB</td>
<td>0</td>
<td>2</td>
<td>Hokkaido University</td>
</tr>
<tr>
<td>KUIDL</td>
<td>4</td>
<td>2</td>
<td>Kyoto University</td>
</tr>
<tr>
<td>MSRA</td>
<td>4</td>
<td>1</td>
<td>Microsoft Research Asia</td>
</tr>
<tr>
<td>NUTKS</td>
<td>0</td>
<td>6</td>
<td>Nagaoka University of Technology</td>
</tr>
<tr>
<td>TTOKU</td>
<td>3</td>
<td>0</td>
<td>Tokyo Institute of Technology</td>
</tr>
</tbody>
</table>
Results
It can be observed that the three "ORG" runs are the overall top-ranked result in disagreement. The first Manual run (Manual-1) is similar to the Manual-2 run but uses the top-ranked Wikipedia article found in the baseline search results; it is the only mandatory run that outputs a concatenation of search engine snippets from the baseline search results; Manual-3 is the first Manual run that outputs a concatenation of search engine snippets only for certain facts, such as people (i.e., ACTOR, POLITICIAN, and ATHLETE) as well as DEFINITION, the score was computed based on the weighted recall performances per-top-ranked result (top-ranked string) shown in Figure 9: Japanese Subtask: Mean $S_I$-measure performances per-top-ranked string. Table 13 shows mean $S_I$-measure performances per-top-ranked string for all runs including the MANUAL runs, while they are not as effective for FACILITY, GEO and QA. Table 12 shows p-values between sets of $I$-measure and $S_I$-measure performances in disagreement. Moreover, Table 11 shows that these baseline runs outperform all submitted automatic runs. These three runs are statistically significant: Table 12 shows p-values for the significance of the difference between sets of $I$-measure performances. Another Wikipedia-based baseline...
TABLE 10 shows the official mean \( S^\# \)-measure performances per query and system for all runs, including the MANUAL runs. It can be observed that three of the four MANUAL runs far outperform the others. Figure 12 shows the mean \( S^\# \)-measure performances over the 73 queries for all runs including the MANUAL runs. These three runs are statistically significantly different from the other runs, and are significantly better than the other runs. Moreover, Table 11 shows that these baseline runs outperform all the other runs, and are significantly different from the other runs. The column \( S^\# \)-measure performances are computed based on the \( S^\# \)-measure data. Figure 13 shows the \( S^\# \)-measure performances over 100 Japanese queries for the MANUAL runs. The column \( S^\# \)-measure performances are computed based on the \( S^\# \)-measure data. Figure 14 shows the \( S^\# \)-measure performances over 100 Japanese queries for the MANUAL runs. The column \( S^\# \)-measure performances are computed based on the \( S^\# \)-measure data.
Recall that MANDATORY runs must be generated from a distributed document collection.
Too Strong Baselines?
Table 7: 1CLICK-2 runs submitted to English Main tasks. The run name is of a format "<team>-<lang>-<device>-<source>-<priority>".

<table>
<thead>
<tr>
<th>Run Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>KUIDL-E-D-MAND-5</td>
<td>IE based on HTML structure &amp; Web search snippet summarization</td>
</tr>
<tr>
<td>KUIDL-E-D-OPEN-6</td>
<td>Web-search-based query classification &amp; IE based on HTML structure &amp; Web search snippet summarization</td>
</tr>
<tr>
<td>KUIDL-E-M-MAND-7</td>
<td>IE based on HTML structure &amp; Web search snippet summarization</td>
</tr>
<tr>
<td>KUIDL-E-M-OPEN-8</td>
<td>Web-search-based query classification &amp; IE based on HTML structure &amp; Web search snippet summarization</td>
</tr>
<tr>
<td>NSTDB-E-D-MAND-1</td>
<td>EF: extracting non-overlapped elements in a document by element-score-order (INEX's Focused task in Ad hoc track)</td>
</tr>
<tr>
<td>NSTDB-E-D-MAND-2</td>
<td>ER: extracting non-overlapped elements with grouped by document (INEX's Relevant in context task in Ad hoc track)</td>
</tr>
<tr>
<td>NSTDB-E-D-MAND-3</td>
<td>DR: extracting non-overlapped elements with grouped by document and ordered by documents' score (INEX's Relevant in context task in Ad hoc track)</td>
</tr>
<tr>
<td>NSTDB-E-D-MAND-4</td>
<td>EB: extracting only one element per document by element-score-order (INEX's Best in Context task in Ad hoc track)</td>
</tr>
<tr>
<td>NSTDB-E-D-MAND-5</td>
<td>DB: extracting only one element per document by document-score-order (INEX's Best in Context task in Ad hoc track)</td>
</tr>
<tr>
<td>NSTDB-E-M-MAND-6</td>
<td>mobileEF: extracting non-overlapped elements in a document by element-score-order for mobile (INEX's Focused task in Ad hoc track)</td>
</tr>
<tr>
<td>NUIR-E-D-MAND-1</td>
<td>Concatenates snippets from top documents.</td>
</tr>
<tr>
<td>NUIR-E-M-MAND-2</td>
<td>Concatenates snippets from top documents.</td>
</tr>
<tr>
<td>NUIR-E-D-MAND-3</td>
<td>Takes first string of text from first Wikipedia document.</td>
</tr>
<tr>
<td>NUIR-E-M-MAND-4</td>
<td>Takes first string of text from first Wikipedia document.</td>
</tr>
<tr>
<td>NUIR-E-D-MAND-5</td>
<td>Takes wikipedia text with query terms.</td>
</tr>
<tr>
<td>NUIR-E-M-MAND-6</td>
<td>Takes wikipedia text with query terms.</td>
</tr>
<tr>
<td>udem-E-D-MAND-1</td>
<td>ILP Hunter-Gatherer – desktop</td>
</tr>
<tr>
<td>udem-E-M-MAND-2</td>
<td>ILP Hunter-Gatherer – mobile</td>
</tr>
<tr>
<td>udem-E-D-MAND-3</td>
<td>“wiki based pattern extraction + learning nugget weight + ILP” – desktop</td>
</tr>
<tr>
<td>udem-E-D-MAND-4</td>
<td>Basic Hunter-Gatherer – desktop</td>
</tr>
<tr>
<td>ut-E-D-OPEN-1</td>
<td>API-based Information Extraction System without partial matching</td>
</tr>
<tr>
<td>ut-E-D-OPEN-2</td>
<td>API-based Information Extraction System with partial matching</td>
</tr>
</tbody>
</table>

![Figure 6: English Subtask: S♯-score for all systems averaged across each query Meta-category.](image-url)

**Per-query-type Analysis: 1CLICK-2**

We want to make the evaluation a matter of system quality, and not of any other factors. Unfortunately, large other factors contributed to noise in the evaluation results, most prominently assessor quality.

Some systems are better at CELEBRITY queries, while other systems are better at non-CELEBRITY queries. For example, the ut-E-D-OPEN runs clearly do better at CELEBRITY queries, possibly because they are somewhat based on entities. Our baseline systems do better at overall coverage (non-CELEBRITY) because they are based on Wikipedia paragraphs.

**Text Readability:**

Text Readability is certainly an issue for the assessors, independent of the actual information on the summaries. Especially with many systems and queries, it is indeed quite possible that correct information in summaries was not matched with the appropriate nugget. Here is such example of a not-so-readable summary:

> "Events and tenants": "Preceded byNone"], "Pacific": "American Airlines Center Â° u HP Pavilion Â° u Honda Center Â° u Jobing.com Arena Â° u Staples Center"], "Broke ground": "April 29, 1993"], "Culture": "Celtics parquet floor Celtic ..."
Per-query-type Analysis in Japanese 1CLICK-2

KUIDL is the best for FACILITY

TTOKU is the best for GEO

Baselines are the best for DEF.

TTOKU is the best for QA
Summary of Findings

- **Overall**
  - Baselines outperformed participants’ runs
  - ut is the second best in English 1CLICK-2

- **ARTIST, ACTOR, POLITICIAN, and ATHLETE**
  - Top performer: Baselines

- **FACILITY and LOCATION**
  - Top performer: KUIDL and TTOKU

- **DEFINITION and QA**
  - Top performer: TTOKU
Possible Problems

- **Readability problem**
  - Assessor matching mistakes are more probable on crabbed X-strings than readable ones (e.g. our simple baselines)

- **Wikipedia-is-enough problem**
  - For single-term queries, the first sentences from a Wikipedia article are effective enough
  - While specified queries such as “michael jackson death” require a summary from multiple documents
0.85+ accuracy achieved (by NUTKS&KUIDL)

- DIFFICULT: DEFINITION type
- EASY: CELEBRITY types (ARTIST, ACTOR, etc.)
Summary and Future work
Summary

- 1CLICK is immediate and direct information access that focuses more on information retrieval

- Several new features in 1CLICK-2
  - A new subtask
  - Semi-automatic nugget extraction
  - Finer-grained units for matching
  - Semi-automatic matching between X-strings and VVs

Results

- Opportunity for big improvement
- Some runs show good performances for some query types
- Readability problem for both participants and organizers
3.3 Robustness of Evaluation Metrics

Figure 4: Comparison between MANUAL and submitted runs

- The system ranking in terms of Kendall’s tau rank correlation is around 0.90 (about seven pairs of runs).
- Even if we only have 50% samples of the official sets of iUnits, the system ranking remains substantially incomplete.
- What happens to the system ranking if the sets of iUnits were exhaustively sampled?

- The effect of downsampling the iUnits on the system ranking is shown in Figure 5.
- Since even randomly downsampled iUnits yield relatively reliable evaluation results, exploring (semi)automatic approaches to iUnit extraction seems worthwhile.

- As we have mentioned earlier, evaluating 1CLICK systems requires much manpower. For the future rounds of the 1CLICK task, it will help to reduce the performance gap.
- The NEXT 1CLICK project has been supported in part by the project: Grants-in-Aid for Scientific Research.

Future work includes investigation of the effect of removing the user to read and understand the text, and the MANUAL runs were evaluated also in terms of readability (how easy it is for the user to read and understand the text), and the MANUAL runs outperformed the submitted runs in terms of this criterion as well.

The NEXT 1CLICK project aims to satisfy the challenges for advancing the state-of-the-art of 1CLICK systems:

- A highly effective 1CLICK system needs to find the right document from one another. These results suggest that there are a lot of gaps.
- The NEXT 1CLICK project investigates how our effectiveness metrics correlate with subjective assessments; and (d) Evaluation with more realistic mobile information needs that go beyond simple lookup (e.g. synthesising information from multiple sources).

6. REFERENCES

- T. Sakai and M. P. Kato. One click one revisited: Enhancing One Click Access is an ambitious task that aims to satisfy the human intelligence wall.
- J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and PC internet search.
- T. Sakai and N. Kando. On information retrieval metrics designed for investigating how our effectiveness metrics correlate with subjective assessments; and (d) Evaluation with more realistic mobile information needs that go beyond simple lookup (e.g. synthesising information from multiple sources).

ACKNOWLEDGEMENTS

- We thank the NTCIR-10 1CLICK-2 participants for their effort in producing the runs, and Lyn Zhu, Zeyong Xu, Yue Dai and Sudong Chung for helping us access to the mobile query log.

Future work includes:

- Investigating the effect of removing the user to read and understand the text, and the MANUAL runs outperformed the submitted runs in terms of this criterion as well.
- The NEXT 1CLICK project has been supported in part by the project: Grants-in-Aid for Scientific Research.

Future work includes:

- Investigating the effect of removing the user to read and understand the text, and the MANUAL runs outperformed the submitted runs in terms of this criterion as well.
- The NEXT 1CLICK project has been supported in part by the project: Grants-in-Aid for Scientific Research.
Welcome You to 1CLICK Session on DAY-4

- **TTOKU Summarization Based Systems at NTCIR-10 1CLICK-2 task**
  - Tokyo Institute of Technology team: impressive QA performance

- **MSRA at NTCIR-10 1CLICK-2**
  - Microsoft Research Asia team: top performer among Japanese MANDATORY runs

- **An API-based Search System for One Click Access to Information**
  - University of Twente team: top performer in English 1CLICK-2

- **Hunter Gatherer: UdeM at 1CLICK-2**
  - Université de Montréal team: top performer among English MANDATORY runs

- **XML Element Retrieval@1CLICK-2**
  - Nara Institute of Science and Technology team: unique approach to 1CLICK

Thank you!
Specific vs. Unspecific

queries=SPEC systems=ALL

S#-Measure

SPEC

NONSPECIFIC
Semi-automatic Nugget Extraction

- Iterative reinforcement between nuggets/documents
  - Document(s) selected based on matching high quality nuggets
  - Document(s) assessed for binary relevance Rel/NRel
  - New Nuggets introduced = sentences from Rel documents
  - Nuggets updated quality
    - Rel docs matching a nugget increase nugget quality
    - NRel docs matching a nugget decrease nugget quality
    - Non-matching docs don’t affect nugget quality
  - Judge can review/assess nuggets directly
  - Judge can extract manually a nugget (different than a sentence from Rel docs)
<table>
<thead>
<tr>
<th>run name</th>
<th>readability</th>
<th>trustworthiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>KUIDL-J-D-MAND-1.tsv</td>
<td>-1.3</td>
<td>-1.48</td>
</tr>
<tr>
<td>KUIDL-J-D-OPEN-2.tsv</td>
<td>-1.32</td>
<td>-1.27</td>
</tr>
<tr>
<td>KUIDL-J-M-MAND-3.tsv</td>
<td>-1.23</td>
<td>-2.42</td>
</tr>
<tr>
<td>KUIDL-J-M-OPEN-4.tsv</td>
<td>-1.67</td>
<td>-2.46</td>
</tr>
<tr>
<td>MANUAL-J-D-OPEN-1.tsv</td>
<td>0.92</td>
<td>1.08</td>
</tr>
<tr>
<td>MANUAL-J-D-OPEN-2.tsv</td>
<td>0.54</td>
<td>0.74</td>
</tr>
<tr>
<td>MANUAL-J-D-OPEN-3.tsv</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>MANUAL-J-D-OPEN-4.tsv</td>
<td>0.57</td>
<td>0.91</td>
</tr>
<tr>
<td>MSRA-J-D-MAND-1.tsv</td>
<td>-1.43</td>
<td>-1.47</td>
</tr>
<tr>
<td>MSRA-J-D-MAND-2.tsv</td>
<td>-1.8</td>
<td>-1.97</td>
</tr>
<tr>
<td>MSRA-J-D-MAND-3.tsv</td>
<td>-1.89</td>
<td>-2.08</td>
</tr>
<tr>
<td>MSRA-J-D-MAND-4.tsv</td>
<td>-2.14</td>
<td>-2.16</td>
</tr>
<tr>
<td>ORG-J-D-MAND-1.tsv</td>
<td>-0.26</td>
<td>-0.13</td>
</tr>
<tr>
<td>ORG-J-D-MAND-2.tsv</td>
<td>1.5</td>
<td>0.82</td>
</tr>
<tr>
<td>ORG-J-D-ORCL-3.tsv</td>
<td>-0.31</td>
<td>0.09</td>
</tr>
<tr>
<td>TTOKU-J-D-ORCL-2.tsv</td>
<td>-2.35</td>
<td>-2.27</td>
</tr>
<tr>
<td>TTOKU-J-M-MAND-3.tsv</td>
<td>-2.68</td>
<td>-3.23</td>
</tr>
<tr>
<td>TTOKU-J-M-ORCL-1.tsv</td>
<td>-1.55</td>
<td>-2.28</td>
</tr>
</tbody>
</table>
**Significant Test Result**

Table 12: Japanese Subtask: p-values of two-sided randomized Tukey’s HSD in terms of $S^2$-measure performances over 100 Japanese queries ($L = 500$). Bold font indicates p-values < $\alpha = 0.05$.

<table>
<thead>
<tr>
<th></th>
<th>KUIDL</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>KUIDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.997</td>
<td>1.000</td>
<td>0.998</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.952</td>
<td>1.000</td>
<td>0.378</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSRA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTOKU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Disagreement across Assessors

Figure 9: Japanese Subtask: Mean $S^\#$-measure performances over 100 queries ($L = 500$). The $x$ axis represents runs sorted by Mean $S^\#$ with the intersection iUnit match data.
Table 23: Japanese Subtask: Inter-rater agreement in terms of Cohen’s kappa coefficient, mean absolute error (MAE), and mean square error (MSE).

<table>
<thead>
<tr>
<th>assessor pairs</th>
<th>Kappa</th>
<th>MAE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_7$</td>
<td>$a_1$</td>
<td>0.782</td>
<td>10.4</td>
</tr>
<tr>
<td>$a_1$</td>
<td>$a_2$</td>
<td>0.738</td>
<td>15.5</td>
</tr>
<tr>
<td>$a_2$</td>
<td>$a_3$</td>
<td>0.717</td>
<td>5.32</td>
</tr>
<tr>
<td>$a_3$</td>
<td>$a_4$</td>
<td>0.755</td>
<td>3.87</td>
</tr>
<tr>
<td>$a_4$</td>
<td>$a_5$</td>
<td>0.743</td>
<td>6.01</td>
</tr>
<tr>
<td>$a_5$</td>
<td>$a_6$</td>
<td>0.751</td>
<td>4.77</td>
</tr>
<tr>
<td>$a_6$</td>
<td>$a_7$</td>
<td>0.688</td>
<td>4.18</td>
</tr>
<tr>
<td>$a_8$</td>
<td>$a_9$</td>
<td>0.818</td>
<td>4.07</td>
</tr>
<tr>
<td>$a_9$</td>
<td>$a_{10}$</td>
<td>0.770</td>
<td>5.46</td>
</tr>
<tr>
<td>$a_{10}$</td>
<td>$a_{11}$</td>
<td>0.704</td>
<td>6.41</td>
</tr>
<tr>
<td>$a_{11}$</td>
<td>$a_8$</td>
<td>0.753</td>
<td>4.61</td>
</tr>
<tr>
<td>average</td>
<td></td>
<td>0.747</td>
<td>6.41</td>
</tr>
</tbody>
</table>
### Correlation across Utility, ROUGE, and S# with Manual/Automatic Matching

<table>
<thead>
<tr>
<th>CATEG</th>
<th>Utility VS ROUGE</th>
<th>Utility VS Man</th>
<th>Utility VS Auto</th>
<th>ROUGE VS Man</th>
<th>ROUGE VS Auto</th>
<th>Man VS Auto</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTOR</td>
<td>0.68</td>
<td>0.58</td>
<td>0.56</td>
<td>0.35</td>
<td>0.34</td>
<td>0.93</td>
</tr>
<tr>
<td>ARTIS</td>
<td>0.59</td>
<td>0.53</td>
<td>0.53</td>
<td>0.24</td>
<td>0.32</td>
<td>0.82</td>
</tr>
<tr>
<td>ATHLE</td>
<td>0.49</td>
<td>0.49</td>
<td>0.52</td>
<td>0.20</td>
<td>0.17</td>
<td>0.85</td>
</tr>
<tr>
<td>POLIT</td>
<td>0.53</td>
<td>0.42</td>
<td>0.45</td>
<td>0.05</td>
<td>0.04</td>
<td>0.71</td>
</tr>
<tr>
<td>GEO</td>
<td>0.38</td>
<td>0.36</td>
<td>0.43</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.70</td>
</tr>
<tr>
<td>FACIL</td>
<td>0.54</td>
<td>0.42</td>
<td>0.43</td>
<td>0.34</td>
<td>0.33</td>
<td>0.77</td>
</tr>
<tr>
<td>DEFIN</td>
<td>0.48</td>
<td>0.63</td>
<td>0.58</td>
<td>0.37</td>
<td>0.29</td>
<td>0.91</td>
</tr>
<tr>
<td>QA</td>
<td>0.67</td>
<td>0.51</td>
<td>0.43</td>
<td>0.39</td>
<td>0.28</td>
<td>0.83</td>
</tr>
<tr>
<td>ALLQ</td>
<td>0.59</td>
<td>0.62</td>
<td>0.58</td>
<td>0.26</td>
<td>0.18</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Table 2: Pairwise comparisons of evaluation metrics used for this task against utility as perceived by the assessors. Values taken by averaging scores over each category and comparing induced rankings via Kendall’s Tau. Man and Auto are manual and automatic matches combined with the S# evaluation metric.
Correlation across Utility, Estimated Readability, and Readability

<table>
<thead>
<tr>
<th>CATEG</th>
<th>Utility VS Read-F-S#</th>
<th>Utility VS Read-S-S#</th>
<th>Readability VS Read-F</th>
<th>Readability VS Read-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTOR</td>
<td>0.56</td>
<td>0.58</td>
<td>0.18</td>
<td>0.22</td>
</tr>
<tr>
<td>ARTIS</td>
<td>0.56</td>
<td>0.57</td>
<td>0.03</td>
<td>0.13</td>
</tr>
<tr>
<td>ATHLE</td>
<td>0.49</td>
<td>0.49</td>
<td>0.41</td>
<td>0.42</td>
</tr>
<tr>
<td>POLIT</td>
<td>0.31</td>
<td>0.33</td>
<td>-0.09</td>
<td>-0.13</td>
</tr>
<tr>
<td>GEO</td>
<td>0.28</td>
<td>0.27</td>
<td>0.26</td>
<td>0.27</td>
</tr>
<tr>
<td>FACIL</td>
<td>0.50</td>
<td>0.50</td>
<td>0.35</td>
<td>0.41</td>
</tr>
<tr>
<td>DEFIN</td>
<td>0.56</td>
<td>0.55</td>
<td>0.14</td>
<td>0.09</td>
</tr>
<tr>
<td>QA</td>
<td>0.53</td>
<td>0.50</td>
<td>-0.30</td>
<td>-0.17</td>
</tr>
<tr>
<td>ALLQ</td>
<td>0.54</td>
<td>0.53</td>
<td>0.16</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 3: Kendall’s Tau pairwise comparisons of $S\#$ using readability metrics versus Utility, and of only the readability scores versus assessed Readability, all using rankings induced by average scores over query categories.
iUnits

Information pieces that satisfy the following properties

- **Relevant**: can satisfy the user’s information need
- **Atomic**: cannot be broken down into multiple iUnits
- **Dependent**: can depend on other iUnits

Example nugget:

“Murray tried to revive Jackson for ten minutes, at which point he realized he needed to call for help.”

<table>
<thead>
<tr>
<th>ID</th>
<th>iUnits</th>
<th>weight</th>
<th>dep</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>Murray tried to revive Jackson</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>002</td>
<td>Murray tried to revive Jackson for ten minutes</td>
<td>4</td>
<td>001</td>
</tr>
<tr>
<td>003</td>
<td>Murray realized he needed to call help</td>
<td>1</td>
<td>001</td>
</tr>
</tbody>
</table>
Vital strings

- Minimally adequate natural language expression
  - Obtained from either nuggets or iUnits

- Example nugget:
  “Murray tried to revive Jackson for ten minutes, at which point he realized he needed to call for help.”

<table>
<thead>
<tr>
<th>ID</th>
<th>iUnits</th>
<th>weight</th>
<th>Dep</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>Murray tried to revive</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>002</td>
<td>ten minutes</td>
<td>4</td>
<td>001</td>
</tr>
<tr>
<td>003</td>
<td>realized he needed to call help</td>
<td>1</td>
<td>001</td>
</tr>
</tbody>
</table>
MOBILE (280 chars) runs in English 1CLICK-2

Search engine's top-ranked result snippets from the baseline search results; ORG-J-D-MAND-2 is a DESKTOP mandatory run that outputs a concatenation of search engine top-ranked result snippets only for those 73 queries. Table 14 and its graphs drawn in Figures 13 and 12 show the per-query performances over the 73 queries. It can be observed that three of the four MANUAL runs far outperform the others. Figure 11 shows the mean S♯-measure performances over 100 Japanese queries for the submitting automatic runs. These three runs are statistically significantly indistinguishable from one another: Table 12 shows p-values of 0.05, 0.15, 0.25, 0.2, and 0.3. Moreover, Table 11 shows that these baseline runs outperform all participating runs with the four celebrity query types (i.e. ARTIST, ACTOR, POLITICIAN, and ATHLETE) as well as DEFINITION, OF, and PLACE in both of the query types and Figures 10 and 11 visualize the performances shown in Table 10. These are actually simple baseline runs shown in the table.

The offset of iUnit matches is defined as the minimum offset of iUnit matches by two assessors in both of the runs. The score was computed based on the weighted recall performances over 100 Japanese queries for the submitted runs except MANUAL runs. The column S♯-measure performances (queries=ALL systems=E-M) shown in Figure 9: Japanese Subtask: Mean S♯-measure (queries=ALL systems=E-M). The x-axis represents runs sorted with the intersection iUnit match data. The y-axis shows the S♯-measure performances between sets of iUnit matches. The first set of iUnits consists of the union of the sources of iUnits instead of the search results (an oracle run). These runs are ranked by the mean S♯-measure performances over 100 Japanese queries (systems=E-D). Figure 5: English Subtask: Mean S♯-measure performances (queries=ALL systems=E-D). The x-axis represents runs sorted with the intersection iUnit match data. The y-axis shows the S♯-measure performances between sets of iUnit matches. The first set of iUnits consists of the union of the sources of iUnits instead of the search results (an oracle run). These runs are ranked by the mean S♯-measure performances over 100 Japanese queries (systems=E-D).
Evaluation Metric: S-measure

nDCG **discounts** documents based on ranks

S-measure **discounts** nuggets based on offsets (positions in X-string)

Ranked list of documents

```
document
document
document
document
```

X-string

```
nugget

nugget

nugget
```

Sakai, Kato, Song: Click the Search Button and Be Happy: Evaluating Direct and Immediate Information Access, ACM CIKM 2011
1CLICK-2 Task Structure

Crawled

Documents ➔ Queries

Nugget Extraction

Feedback

Units for matching ➔ Nuggets

Vital strings (E) ➔ iUnits (J)

Matching

Matched positions ➔ S#-measure computation ➔ S# scores

Participants

Runs

Part of test collection

Organizers’ work

S#-measure

Units for matching

Feedback

Crawled
Suppose that ...

- Searching for information about his highlights in a movie with a mobile device
Organizers worked hard, too

3,927 nuggets for 100 queries

Very time-consuming process