Overview of the NTCIR-11 MobileClick Task

Makoto P. Kato (Kyoto University), Matthew Ekstrand-Abueg, Virgil Pavlu (Northeastern University), Tetsuya Sakai (Waseda University), Takehiro Yamamoto (Kyoto University) and Mayu Iwata (KDDI Corporation)
• You are finding answers for a question “what’s the difference between Organic EL and LCD?” in an electronics store
IR Systems in 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

Long way to get all desired information
MobileClick System

Task: Given a search query, return a two-layered textual output

Go beyond the "ten-blue-link" paradigm, and tackle information retrieval rather than document retrieval
• Present a concise summary (like "Knowledge Graph") to any kind of queries in a way that any users can easily reach their desired info.
Queries and Documents

• Queries
  – 50 English/Japanese queries
  – Most of which were highly specific (e.g. java vs python text processing)
    • Based on 1CLICK-2 task at NTCIR-10, the first paragraph of Wikipedia articles were good enough for named entity queries

  Examples

<table>
<thead>
<tr>
<th>hiphop clubs barcelona</th>
<th>Michoacan crafts materials</th>
<th>sears illinois</th>
</tr>
</thead>
<tbody>
<tr>
<td>why does turkey make you sleepy</td>
<td>ron paul tea party</td>
<td>aaron rodgers belt celebration</td>
</tr>
<tr>
<td>french landmarks</td>
<td>syrian civil war players</td>
<td>ukraine debt</td>
</tr>
</tbody>
</table>

• Documents
  – 300 - 400 commercial search engine results for each query
  – From which summaries are generated
Two Subtasks

Query

iUnit Retrieval Subtask

iUnits

iUnit Summarization Subtask

Two-layered Summary

<table>
<thead>
<tr>
<th>Score</th>
<th>iUnit</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>LCD is lighter than OLED</td>
</tr>
<tr>
<td>4</td>
<td>OLED shows a better black color</td>
</tr>
<tr>
<td>3</td>
<td>OLED has a wider view angle</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
iUnit Retrieval Subtask

- Generate a list of *iUnits* ranked according to their importance for a given query
  - iUnits: atomic information pieces relevant to a given query

**Input:** Query

OLED LCD difference

**Output:** List of iUnits

<table>
<thead>
<tr>
<th>Rank</th>
<th>iUnit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LCD is lighter than OLED</td>
</tr>
<tr>
<td>2</td>
<td>OLED shows a better black color</td>
</tr>
<tr>
<td>3</td>
<td>OLED has a wider view angle</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

…
iUnit Summarization Subtask

- For a given query and a list of *iUnits*, generate a *two-layered* textual output

**Input:** Query

**Input:** List of *iUnits*

<table>
<thead>
<tr>
<th>Rank</th>
<th><em>iUnit</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LCD is lighter than OLED</td>
</tr>
<tr>
<td>2</td>
<td>OLED shows a better black color</td>
</tr>
<tr>
<td>3</td>
<td>OLED has a wider view angle</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Output:**

- **Two-layered textual output**

  - **Advantage of OLED**
    - OLED shows a better black color, a faster response speed, and a wider view angle.
  - **Advantage of LCD**
    - LCD is better in terms of the weight, size and energy saving.
Example of iUnit Summarization Run

〈result qid="MC-SAMPLE-E-0001">
 〈firstlayer>
    LCD is better in terms of the weight, size and energy saving. OLED shows a better black color, a faster response speed, and a wider view angle.

 〈link id="1">Advantages of OLED</link>
 〈link id="2">Advantages of LCD</link>
</firstlayer>
 〈secondlayer id="1">
    OLED is ...
  </secondlayer>
 〈secondlayer id="2">
    LCD is ....
  </secondlayer>
</result>
Evaluation Methodology

• **Gold Standard iUnits**
  – For each query, assessors created gold standard iUnits (GiUnits) based on the document collection

• **iUnit Matching**
  – Assessors recorded the position at which GiUnits appear in the system output

• **Evaluation Metrics**
  – iUnit Retrieval: nDCG and Q-measure
  – iUnit Summarization: M-measure
Creating Gold Standard iUnits (GiUnits)

Atomic and relevant information pieces were recorded as GiUnits with their importance
For query "LCD OLED difference"

- OLED is better in contrast
- Less afterimage in OLED
- Less power consumption for LCD
- LCD is lighter
- The display of OLED consists of gas
- The display of LCD consists of liquid crystal
- OLED is self-luminous
- LCD uses backlight
- LCD has a narrow view angle
- OLED’s contrast in dark places is better
- OLED is excellent in expressive power of black
- OLED is easily scalable
- LCD shows smooth gradation
GiUnits in the system output were manually recorded with their position.
**Evaluation of iUnit Retrieval Runs**

Identify GiUnits from each participant iUnits (PiUnits)

<table>
<thead>
<tr>
<th>Run</th>
<th>GiUnit</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PiUnit1</td>
<td>GiUnit1</td>
<td>3</td>
</tr>
<tr>
<td>PiUnit2</td>
<td>GiUnit2</td>
<td>2</td>
</tr>
<tr>
<td>PiUnit3</td>
<td>GiUnit3</td>
<td>1</td>
</tr>
<tr>
<td>PiUnit4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Give scores to PiUnits based on the scores of identified GiUnits, where redundant PiUnits were ignored

<table>
<thead>
<tr>
<th>Run</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PiUnit1</td>
<td>0</td>
</tr>
<tr>
<td>PiUnit2</td>
<td>3</td>
</tr>
<tr>
<td>PiUnit3</td>
<td>1</td>
</tr>
<tr>
<td>PiUnit4</td>
<td>0</td>
</tr>
</tbody>
</table>

nDCG/Q = 0.6
1. The utility of a summary is measured by U-measure
   – Idea: "More important GiUnits earlier" is better

   \[ U = \sum_{pos=1}^{\text{ttl}} g(pos)D(pos) \]

   \[ D(pos) = 1 - \frac{pos}{L} \]

   \[ g(pos): \text{importance of GiUnit at pos} \]
   \[ D(pos): \text{decay function} \]
   \[ = \text{how much effort is required to reach pos} \]
   \[ L: \text{patience parameter (e.g. 280)} \]

   \[ U = g(80)D(80) = 2\left(1 - \frac{80}{280}\right) \]
Illustration of U-measure Computation

- Suppose that a user reads the first layer and the 2nd second layer of system output

\[ U = g(80)D(80) + g(150)D(150) = 2\left(1 - \frac{80}{280}\right) + 3\left(1 - \frac{150}{280}\right) \]
• 2. Users probabilistically read a summary
  – When they reach link $l_1$, they click on the link with probability $P(l_1)$

\[
P(\text{trailtext1}) = P(l_1)P(l_2) = 0.5 \times 0.2 = 0.1
\]

\[
P(\text{trailtext2}) = P(l_1)(1-P(l_2)) = 0.5 \times 0.8 = 0.4
\]

\[
P(\text{trailtext3}) = (1-P(l_1)) P(l_2) = 0.5 \times 0.2 = 0.1
\]
Evaluation Metric for iUnit Summarization

- **M-measure** = the expected utility of users who probabilistically click on links

\[ M = \sum_{tt \in T} P(tt)U(tt) \]

\[ T: \text{all possible trailtexts} \]

\[ M = P(tt_1)U(tt_1) + P(tt_2)U(tt_2) + \ldots = 0.1 \times 0.5 + 0.4 \times 0.3 + \ldots \]
Click Probability and Importance of iUnits

- Click probability
  - Estimated by assessors' voting
- Importance of iUnits
  - $g = 0$ if the iUnit is irrelevant to the link

1\textsuperscript{st} layer

- Some info.
- Link 1: $P(I_1) = 0.5$
- Link 2

2\textsuperscript{nd} layer

- Some info.
- $g = 3$ if iUnit is relevant to the anchor text of Link 2; otherwise 0
- $g = 3$
Challenges in MobileClick Task

• **iUnit Retrieval**
  – Estimating the importance of information pieces

• **iUnit Summarization**
  – Help users navigate so that they can efficiently reach their desired information
    • Clustering iUnits
    • Creating meaningful links for clustered iUnits
## Participating Teams

<table>
<thead>
<tr>
<th>Team name</th>
<th>Organization/Method</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KPNM</strong></td>
<td>Hunan University of Science and Technology, China</td>
</tr>
<tr>
<td></td>
<td>[Retrieval] Chain simple techniques based on statistical models and heuristic rules to extract significant text units</td>
</tr>
<tr>
<td><strong>IISR</strong></td>
<td>National Central University, Taiwan</td>
</tr>
<tr>
<td></td>
<td>[Retrieval] Classify each query into eight query types and set the weights of the extraction methods accordingly</td>
</tr>
<tr>
<td><strong>udel</strong></td>
<td>University of Delaware, USA</td>
</tr>
<tr>
<td></td>
<td>[Summarization] Simple re-ranking approach based on the cosine similarity between each iUnit and a dynamic &quot;model&quot; pseudo-document</td>
</tr>
<tr>
<td><strong>NTOUA</strong></td>
<td>National Taiwan Ocean University, Taiwan</td>
</tr>
<tr>
<td></td>
<td>[Summarization] Grouping by longest leading substring.</td>
</tr>
</tbody>
</table>
Baselines

• **iUnit Retrieval**
  – Sentences in snippets of the search results
  – In order of appearance in the results

• **iUnit Summarization**
  – Output iUnits in descending order of the iUnit importance
  – "Headers" (e.g. h1 and h2) used as links
iUnit Retrieval Results

Systems implementing a ranking function on the organizer-provided iUnits and importance.

Best performer in retrieval

IISR achieved the highest performance
The baselines were strong, but there were systems that tended to be relatively strong, but naïve solutions to the task. The systems that re-ranked the organizer-provided iUnits. The last run of that team did not use the gold iUnits and the performance in their ranking methodology. The runs with an asterisk represent the ones between udel's runs.

Figure 4: system performance averaged over all queries. Runs marked with * are implementing a ranking function. Although udel performed best on average for four of its five runs, there are significant differences between all the pairs except udel1* vs udel2*. Randomized Tukey's HSD test [8] shows inherent differences between the categories. For instance, the udel categories by the style of answer expected, as stated in the corpus.

The evaluation metrics, nDCG and Q measure, are computed for a variety of cutoff thresholds. Table 4 shows submitted runs and descriptions of developed systems.

### Table 4: Submitted Runs and Descriptions

<table>
<thead>
<tr>
<th>Category</th>
<th>LOOKUPLIST</th>
<th>FACTFIND</th>
<th>COMPARISON</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>data mining course online</td>
<td>obamacare penalty</td>
<td>ivy bridge vs haswell</td>
</tr>
</tbody>
</table>

Performance for "Lookup list" and "Comparison" highly depend on the systems.
Table 4: Submitted runs and descriptions of developed systems.

<table>
<thead>
<tr>
<th>Run Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUM-NTOUA-E-MAND-1 Grouping by longest leading substring.</td>
</tr>
<tr>
<td>SUM-udel-E-MAND-1 Simple re-ranking approach based on the cosine similarity between each iUnit and a dynamic 'model' pseudo-document; At each step, Model doc is built using concatenation of iUnits that have been ranked so far, then select the doc least similar to model doc.</td>
</tr>
<tr>
<td>SUM-udel-E-MAND-4 Simple re-ranking approach based on the cosine similarity between each iUnit and a fixed 'model' pseudo-document; Model doc is constructed using the concatenation of top-10 docs for the query.</td>
</tr>
<tr>
<td>SUM-udel-E-MAND-5 Simple re-ranking approach based on the cosine similarity between each iUnit and a fixed 'model' pseudo-document; Model doc is built using all concatenated iUnits. These iUnits are constructed by ourselves by consecutive tokens from top-10 docs with a max of 70 characters long.</td>
</tr>
<tr>
<td>SUM-ORG-E-MAND-1 Organizers' Baseline: This method outputs gold standard iUnits in descending order of iUnit scores in the first layer, uses headers that appear at the same level in a HTML, and outputs iUnits similar to the text that follows the headers in the second layers.</td>
</tr>
</tbody>
</table>

7. CONCLUSIONS

This paper presents the overview of the MobileClick task at NTCIR-11. This task aims to develop a system that returns a concise summary of information relevant to a given query, and brings a structure into the summarization so that users can easily locate their desired information. Our task attracted four teams and received fourteen runs for the iUnit retrieval and summarization subtasks. In this paper, we mainly explained the task design, evaluation methodology, and analysis of the results. We have a plan to continue the MobileClick task at NTCIR-12, and look forward to an improvement in the performance at the next round.

8. ACKNOWLEDGMENTS

We thank the NTCIR-11 MobileClick participants for their effort in submitting the runs, and NTCIR-11 PC chairs for their great organization including pre-task discussion and suggestions regarding the task organization.

9. REFERENCES

[6] T. Sakai. On penalising late arrival of relevant documents in iUnit Summarization  Results
Table 4: Submitted runs and descriptions of developed systems.

<table>
<thead>
<tr>
<th>Run Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUM-NTOUA-E-MAND-1 Grouping by longest leading substring.</td>
<td></td>
</tr>
<tr>
<td>SUM-udel-E-MAND-1 Simple re-ranking approach based on the cosine similarity between each iUnit and a</td>
<td>dynamic 'model' pseudo-document; At each step, Model doc is built using</td>
</tr>
<tr>
<td></td>
<td>concatenation of iUnits that have been ranked so far, then select the</td>
</tr>
<tr>
<td></td>
<td>doc least similar to model doc.</td>
</tr>
<tr>
<td>SUM-udel-E-MAND-4 Simple re-ranking approach based on the cosine similarity between each iUnit and a</td>
<td>fixed 'model' pseudo-document; model doc is constructed using the</td>
</tr>
<tr>
<td></td>
<td>concatenation of top-10 docs for the query.</td>
</tr>
<tr>
<td>SUM-udel-E-MAND-5 Simple re-ranking approach based on the cosine similarity between each iUnit and a</td>
<td>fixed 'model' pseudo-document; Model doc is built using all concatenated</td>
</tr>
<tr>
<td></td>
<td>iUnits. These iUnits are constructed by oursleves by consecutive tokens</td>
</tr>
<tr>
<td></td>
<td>from top-10 docs with a max of 70 characters long.</td>
</tr>
<tr>
<td>SUM-ORG-E-MAND-1 Organizers’ Baseline: This method outputs gold standard iUnits in descending order</td>
<td>of iUnit scores in the first layer, uses headers that appear at the</td>
</tr>
<tr>
<td></td>
<td>same level in a HTML, and outputs iUnits similar to the text that</td>
</tr>
<tr>
<td></td>
<td>follows the headers in the second layers.</td>
</tr>
</tbody>
</table>

Furthermore, we drilled down the results by the category of queries. Figure 7 shows per-category M (L=840) of each run. The trend of each category in the iUnit summarization subtask seems different from that in the iUnit retrieval subtask: MobileClick systems performed well for LOOKUPLIST, while they could not achieve high performances for FACTFINDING.

7. CONCLUSIONS

This paper presents the overview of the MobileClick task at NTCIR-11. This task aims to develop a system that returns a concise summary of information relevant to a given query, and brings a structure into the summarization so that users can easily locate their desired information. Our task attracted four teams and received fourteen runs for the iUnit retrieval and summarization subtasks. In this paper, we mainly explained the task design, evaluation methodology, and analysis of the results. We have a plan to continue the MobileClick task at NTCIR-12, and look forward to an improvement in the performance at the next round.

8. ACKNOWLEDGMENTS

We thank the NTCIR-11 MobileClick participants for their effort in submitting the runs, and NTCIR-11 PC chairs for their great organization including pre-task discussion and suggestions regarding the task organization.

9. REFERENCES


**Summary**

- **MobileClick** task aims to achieve *information* retrieval systems that appropriately lay out information in two layers.

- Evaluation was based on **iUnits** and **M-measure** that approximates utility of users who read the two-layered summary.

- **IISR** achieved the highest performance in retrieval, while summarization was a difficult task to achieve high performance.
<table>
<thead>
<tr>
<th></th>
<th>Title</th>
<th>Organization/Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Improving iUnit Retrieval with Query Classification and Multi-Aspect iUnit Scoring: The IISR System at NTCIR-11 MobileClick Task</td>
<td>National Central University, Taiwan</td>
</tr>
<tr>
<td></td>
<td>[Retrieval] Classify each query into eight query types and set the weights of the extraction methods accordingly</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Description of the NTOU MobileClick System at NTCIR-11</td>
<td>National Taiwan Ocean University, Taiwan</td>
</tr>
<tr>
<td></td>
<td>[Summarization] Grouping by longest leading substring.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Udel @ NTCIR-11 MobileClick Track</td>
<td>University of Delaware, USA</td>
</tr>
<tr>
<td></td>
<td>[Summarization] Simple re-ranking approach based on the cosine similarity between each iUnit and a dynamic &quot;model&quot; pseudo-document</td>
<td></td>
</tr>
</tbody>
</table>
Round Table Session

• **DAY-3 (Dec. 11 (Thu))**

• **Location:**
  Seminar Room 1904, 19F

• **Time:**
  16:05 – 18:00
"Invisible Baseline" Problem
In MobileClick-2, we will keep the basic task design, but promise to help you continuously improve your systems.

Before the run submission deadline, our real time scoreboard will show the current performance of participant systems.