Two-layered Summaries for Mobile Search: Does the Evaluation Measure Reflect User Preferences?

Makoto P. Kato (Kyoto U.), Tetsuya Sakai (Waseda U.), Takehiro Yamamoto (Kyoto U.), Virgil Pavlu (Northeastern U.), and Hajime Morita (Kyoto U.)
MOTIVATION AND TASK
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

Enter query → Click SEARCH button

System output

OLED LCD difference

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.

Advantage of OLED
Advantage of LCD

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.
iUnit Summarization Subtask at NTCIR-12

- Given a query, a set of iUnits, and a set of intents, generate a **two-layered summary**

**Input:** Query

- NTCIR

**Input:** iUnit set

- iUnit
  - A series of evaluation workshops
  - Designed to enhance IA research
  - ...

**Input:** Intents

- Intents

**Output:** Two-layered summary

The NTCIR Workshop is a series of evaluation workshops designed to enhance research in information access technologies including information retrieval, summarization, extraction, question answering, etc.

**Challenge**

Lay out iUnits so that any types of users can be immediately satisfied
Two-layered Summary in Action


Profile
Award
Film
Does the Evaluation Measure Reflect User Preferences?

**Research Question Addressed in This Work**

Does the Evaluation Measure Reflect User Preferences?

**M-measure**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.5</td>
<td>0.4</td>
</tr>
</tbody>
</table>

**User preference**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>10</td>
<td>4</td>
</tr>
</tbody>
</table>

**Which is higher?**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.5</td>
<td>0.4</td>
</tr>
</tbody>
</table>

**Which is better?**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>

**Same?**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>
DATA
Overview of Data

Queries
- napoleon

Web search

Documents

Extraction

iUnits
- Born on the island of Corsica
- Defeated at the Battle of Waterloo
- Established legal equality and religious toleration an innovator

Intentents
- Achievement
- Skill
- Career

Clustering

iUnit summarization
Queries and Documents

• **Queries**
  – 100 English/Japanese queries
  – Most of which were ambiguous/underspecified
  – **Selected from five categories:**
    celebrity, location, definition, and QA (similar to NTCIR 1CLICK-2)

Examples

<table>
<thead>
<tr>
<th>CELEBRITY</th>
<th>LOCATION</th>
<th>DEFINITION</th>
<th>QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>hulk hogan</td>
<td>bank adelanto</td>
<td>bitcoin</td>
<td>what is mirror made of</td>
</tr>
<tr>
<td>bruno mars</td>
<td>cafe killeen</td>
<td>divers disease</td>
<td>how to cook coleslaw</td>
</tr>
<tr>
<td>sharon stone</td>
<td>cincinnati art museum</td>
<td>windows 7</td>
<td>role of animal tail</td>
</tr>
</tbody>
</table>

• **Documents**
  – 500 commercial search engine results for each query
  from which iUnits were extracted
iUnits

• Definition
  – Atomic information pieces relevant to a given query

• The number of iUnits
  – 2,317 (23.8 iUnits per query) for English
  – 4,169 (41.7 iUnits per query) for Japanese

Examples of iUnits for query “Napoleon”

<table>
<thead>
<tr>
<th>Born on the island of Corsica</th>
<th>General of the Army of Italy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defeated at the Battle of Waterloo</td>
<td>One of the most controversial political figures won at the Battle of Wagram</td>
</tr>
<tr>
<td>Established legal equality and religious toleration an innovator</td>
<td>Baptised as a Catholic</td>
</tr>
<tr>
<td>Absent during Peninsular War</td>
<td>Cut off European trade with Britain</td>
</tr>
</tbody>
</table>
• An intent can be defined as
  – A specific interpretation of an ambiguous query
    ("Mac OS" and "car brand" for "jaguar"), or
  – An aspect of a faceted query
    ("windows 8" and "windows 10" for "windows")

• Obtained by clustering iUnits

<table>
<thead>
<tr>
<th>iUnits</th>
<th>Intents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Born on the island of Corsica</td>
<td></td>
</tr>
<tr>
<td>Defeated at the Battle of Waterloo</td>
<td></td>
</tr>
<tr>
<td>Established legal equality and religious</td>
<td></td>
</tr>
<tr>
<td>toleration an innovator</td>
<td></td>
</tr>
<tr>
<td>Absent during Peninsular War</td>
<td></td>
</tr>
</tbody>
</table>

Clustering

<table>
<thead>
<tr>
<th></th>
<th>Achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Skill</td>
</tr>
<tr>
<td></td>
<td>Career</td>
</tr>
</tbody>
</table>
EVALUATION
Per-intent iUnit Importance and Intent Probability

• Importance of iUnits in terms of an intent

  In terms of intent “Definition”

<table>
<thead>
<tr>
<th>iUnit</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A series of evaluation workshops</td>
<td>5</td>
</tr>
<tr>
<td>Task Registration Due 20/Jun./2016</td>
<td>3</td>
</tr>
</tbody>
</table>

  In terms of intent “Schedule”

<table>
<thead>
<tr>
<th>iUnit</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A series of evaluation workshops</td>
<td>2</td>
</tr>
<tr>
<td>Task Registration Due 20/Jun./2016</td>
<td>5</td>
</tr>
</tbody>
</table>

• Intent probability \( P(i|q) \)

  – Probability of having intent i for a given query q

<table>
<thead>
<tr>
<th>Intent</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>0.4</td>
</tr>
<tr>
<td>Schedule</td>
<td>0.3</td>
</tr>
<tr>
<td>Tasks</td>
<td>0.3</td>
</tr>
</tbody>
</table>

For details, see our MobileClick-2 overview paper
• Consider single-layered summary evaluation

• **U-measure** [Sakai and Dou. SIGIR2013]
  – Higher if more important iUnits appear earlier

Create a list of iUnits by assuming that users read text from left to right, from top to bottom.

\[
U = \sum_{r=1} G(u_r) \left( 1 - \frac{\text{pos}(u_r)}{L} \right)
\]

- \(u_r\): r-th iUnit
- \(G(u)\): importance of u
- \(\text{pos}(u)\): offset of u from the beginning
- \(L\): patience parameter
M-measure

- **M-measure**
  - Expectation of U-measure over multiple *trailtexts*

\[ M = \sum_{t} P(t)U(t) \]

- *P(t)*: probability of trailtext *t*
- *U(t)*: U-measure of trailtext *t*

1. **Generate trailtexts by assuming that**
   - Users read a summary from the top of the first layer
   - Users click on an intent if they are interested in it

First-layer

- **u₁**
- **u₂**
- **u₃**
- **l₁ (intent 1)**

Second layer

- **u₄**

Trailtext

- **User interested in Intent 1** \( (P(i₁|q)) \)
- **User interested in Intent 2** \( (P(i₂|q)) \)
2. Compute the expectation of U-measure

First layer

\[ u_1 \quad u_2 \]
\[ u_3 \]

Second layer

\[ l_1 \text{ (intent 1) } \]
\[ u_4 \quad u_5 \]

Second layer

\[ l_2 \text{ (intent 2) } \]
\[ u_6 \]

Trailtext (t) (reading path)

\[ u_1 \quad u_2 \quad u_3 \]
\[ u_4 \quad u_5 \]
\[ u_1 \quad u_2 \quad u_3 \]
\[ u_6 \]

\[ P(t_1) = P(i_1 | q) = 0.75 \]

\[ P(t_2) = P(i_2 | q) = 0.25 \]

Because trailtext \( t_2 \) is read by users interested in \( i_2 \)

\[ M = \sum_t P(t)U(t) \]

\[ U = 0.44 \]

\[ M - \text{measure} = 0.36 \]

\[ P(t_1) = P(i_1 | q) = 0.75 \]

\[ P(t_2) = P(i_2 | q) = 0.25 \]
Pairwise Comparison

All possible pairs of 7 summaries for 25 queries were presented to about 14 users.
Instruction in Pairwise Comparison

• Users were asked to select either
  the left one is better, the right one is better, equally good, or equally bad

• Criteria:
  (1) How much useful information you can get from the summary, and
  (2) How quickly you can get useful information from the summary
Settings of M-measure

- **$L$ of U-measure in M-measure**
  
  \[ U = \sum_{r=1}^{\text{pos}(u_r)} G(u_r) \max \left( 0, 1 - \frac{\text{pos}(u_r)}{L} \right) \]

- $L$ is a patience parameter that controls how the gain of iUnits decreases as the user reads the text

- **Simple variants of M-measure**
  
  - Use only first layer
  - Use only second layer
  - Use a uniform distribution for $P(i|q)$
Interpretation of Results

Each dot represents a pair of systems (A, B) for a particular query.

Agreement = (#dots in Agree) / (#dots)

Diff. of M-measure (M(A) - M(B))
Experimental Results for Different Patient Parameters

- **LOW** agreement for **LOW** patience parameter (L=93.5)
  - Agreement: 36.4%

- **HIGH** agreement for **HIGH** patience parameter (L=24000)
  - Agreement: 74.2%

Agreement is high (70-74%) for both of the languages
Experimental Results for Simple Variants of M-measure

Use of the second layer and intent probability improves the agreement (but the first layer doesn’t)
Why did the only 2nd layer correlate to the user pref. well?

• **Possible explanations include**
  
  – The quality of the second layer correlates to the quality of the whole summary
  
  – Users decided the quality of the summary mainly based on the second layer
  
    • We asked the users to look at the second layer in the assessment
• **Conclusions**
  – **Proposed M-measure**
    • A special case of intent-aware U-measure for two-layered summarization
  – **Measured the agreement between M-measure and user preferences**
    • Agreement was high (70-74%)

• **Future work**
  – Error analysis
  – Address “why did the only second layer correlate to the user preferences well?”