# Report from the NTCIR-10 1CLICK-2 Japanese Subtask: Baselines, Upperbounds and Evaluation Robustness

Makoto P. Kato
Kyoto University, Japan
kato@dl.kuis.kyoto-u.ac.jp

Tetsuya Sakai
Microsoft Research Asia, P.R.C.
tetsuyasakai@acm.org

Takehiro Yamamoto
Kyoto University, Japan
tyamamot@dl.kuis.kyoto-u.ac.jp

Mayu Iwata
Osaka University, Japan
iwata.mayu@ist.osaka-u.ac.jp

## ABSTRACT

The *One Click Access* Task (1CLICK) of NTCIR requires systems to return a concise multi-document summary of web pages in response to a query which is assumed to have been submitted in a mobile context. Systems are evaluated based on *information units* (or *iUnits*), and are required to present important pieces of information first and to minimise the amount of text the user has to read. Using the official Japanese results of the second round of the 1CLICK task from NTCIR-10, we discuss our task setting and evaluation framework. Our analyses show that: (1) Simple baseline methods that leverage search engine snippets or Wikipedia are effective for "lookup" type queries but not necessarily for other query types; (2) There is still a substantial gap between manual and automatic runs; and (3) Our evaluation metrics are relatively robust to the incompleteness of iUnits.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## General Terms

Experimentation

## Keywords

evaluation, information units, mobile environment, NTCIR, nuggets, summaries, test collections

## 1. INTRODUCTION

NTCIR (NII Testbeds and Community for Information access Research) is a sesquiannual evaluation forum that focuses primarily on Asian language information access. The *One Click Access* Task (1CLICK) of NTCIR requires systems to return a concise multi-document summary of web pages in response to a query which is assumed to have been submitted in a mobile context[1]. Systems are evaluated based on *information units* (or *iUnits*), and are required to present important pieces of information first and to minimise the amount of text the user has to read. Compared to nugget-based evaluations for summarisation and question answering, the task is novel in that the *position* of each information unit is utilised to evaluate the system output. Using the official Japanese results of the second round of the 1CLICK task from NTCIR-10, we discuss our task setting and evaluation framework. Our analyses show that: (1) Simple baseline methods that leverage search engine snippets or Wikipedia are effective for "lookup" type queries but not necessarily for other query types; (2) There is still a substantial gap between manual and automatic runs; and (3) Our evaluation metrics are relatively robust to the incompleteness of iUnits.

[1] http://research.microsoft.com/1CLICK/


*SIGIR'13,* July 28–August 1, 2013, Dublin, Ireland.


## 2. WHAT'S NEW AT 1CLICK-2

The 1CLICK task is defined as: *Given a query, return a textual summary which presents as much relevant information as possible within X characters, important pieces of information first, while minimising the amount of text the user has to read.* For *DESKTOP runs*, we let $X = 500$ (representing a few search engine snippets shown on a desktop PC), while for *MOBILE runs* we let $X = 140$ (representing a short text like a "tweet" shown on a mobile phone screen). For more details on the general task design of 1CLICK, we refer the reader to the Overview of *1CLICK-1* [7].

The main new features of 1CLICK-2 are as follows:

- Based on a study on mobile query logs [3], *1CLICK-1* considered four query types: CELEBRITY, LOCAL, DEFINITION and QA (question answering). However, after manually constructing iUnits for the 1CLICK-1 queries, we found that more fine-grained query types are necessary. For example, a person looking for information on a politician and a person looking for information on an actress probably want different types of information even though they are both celebrities. Therefore, at 1CLICK-2, we refined the CELEBRITY and LOCAL query types as follows: ARTIST, ACTOR, POLITICIAN, ATHLETE, FACILITY and GEO[2]. Thus we have a total of eight query types, which are shown below together with their information need we assumed:

  **ARTIST, ACTOR, POLITICIAN, ATHLETE (10 each)** user wants important facts about celebrities;

  **FACILITY (15)** user wants access and contact information for a particular landmark, facility etc.;

[2] For Example, "Tokyo Sky Tree" is a FACILITY query, while "Sushi bars near Tokyo Station" is a GEO query as it includes a geographical constraint.

**GEO (15)** user wants access and contact information for entities with geographical constraints, e.g. sushi restaurants near Tokyo station;

**DEFINITION (15)** user wants to look up a phrase, etc.;

**QA (15)** user wants to know factual (but not necessarily factoid) answers to a natural language question.

The number of queries for each type is shown in parentheses: thus we have a total of 100 test queries. Furthermore, for the four CELEBRITY query types, we included *specialised* queries such as "jennifer gardner *alias*" as well as non-specialised ones such as "ichiro suzuki" as we wanted encourage participants to explore retrieval of highly focussed information rather than to summarise generic information from (say) Wikipedia.

- We tried to define an iUnit clearly. An iUnit is a factual statement, and usually satisfies the *relevance* (satisfies the information need behind the query partially or wholly), *atomicity* (cannot be further broken down into multiple iUnits), *credibility* (stated explicitly in at least one document) and *temporal validity* (holds true as of the date specified, e.g. July 4, 2012 for 1CLICK-2) criteria. Table 1 provides a few examples translated from one of our Japanese sample queries: here, iUnit I049 *entails* I050, but I049 is said to be atomic as it cannot be broken down into *multiple* iUnits[3]. The *vital string* column represents the *minimal* text required to convey the semantics of the iUnit to the user, which is used for computing our evaluation metrics. For example, without the string "160cm" it would be difficult for the system to tell the user that Keiko Kitagawa is 160 centimeters tall. Moreover, as shown in the "$w$" column, each iUnit was assigned a weight, which reflects the organisers' votes obtained on a five point scale.

- At 1CLICK-1, participating teams were allowed to use any resources available, which made fair comparisons across systems difficult. Therefore, at 1CLICK-2, we provided a set of baseline search results to participants, which consists of top-ranked Yahoo! API search results returned in response to each query[4], and asked them to produce textual output from these search results only. Runs that followed this rule are called *mandatory runs*. Following 1CLICK-1, we also allowed *oracle runs* (runs that assume that the documents from which the iUnits were extracted are known) and *open runs* (runs that use any additional resource besides the above mentioned data, e.g., Wikipedia, their own search results, etc.).

- We introduced a *length penalty* to our evaluation scheme. Figure 1 illustrates our ideas for evaluating the system output which we refer to as the *X-string* [6]. As our goal is to quickly satisfy the user's information need, our *S-measure*, a position-aware version of *weighted recall*, ranks System (b) above System (a). Let $I$ be the complete set of iUnits, $M(\subseteq I)$ be the set of iUnits found within an $X$-string, and $w(i)$ be the weight of $i \in I$. Then $S$ is defined as:

$$\text{S-measure} = \frac{\sum_{i \in M} w(i) \max(0, L - \text{offset}(i))}{\sum_{i \in I} w(i) \max(0, L - \text{offset}^*(i))}, \quad (1)$$

[3]However, in practice, we found it easier to construct iUnits by allowing them to entail multiple iUnits.

[4]The number of documents per query $k$ varied across queries as some documents could not be downloaded. The average of $k$ over the query set is 390.

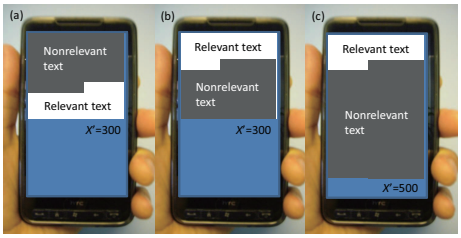


**Figure 1: Evaluating textual output at 1CLICK-2.**

**Table 1: Some translated examples of iUnits for an ACTOR query: "*Keiko Kitagawa*".**

| ID | entails | semantics | vital string | $w$ |
|---|---|---|---|---|
| I001 | | height: 160cm | 160cm | 11 |
| I004 | | born in 1986 | born 1986 | 18 |
| I049 | I050 | graduated from Meiji U. in 2009 | 2009 | 15 |
| I050 | | graduated from Meiji U. | Meiji U. graduate | 11 |

where $\text{offset}(i)$ is the offset position of $i$ within the $X$-string, while $\text{offset}^*(i)$ is the offset position of $i$ within the *Pseudo Minimal Output*, which is an "ideal" output used for normalisation [6]. Following a suggestion from previous work, we primarily use $L = 500$, and use $L = 250$ for additional analysis. These settings correspond to giving the user 60 and 30 seconds to gather information, respectively [5].

In addition to S, we also used *T-measure* at 1CLICK-2, which is a precision-like metric that takes into account the fact that different pieces of information require different amount of space within the mobile phone screen: note that the vital strings vary in length in Table 1. Like precision, T ranks System (c) higher than System (b) in Figure 1. Let $X'$ be the length of an $X$-string (not the length limit), and $|v(i)|$ be the vital string length of iUnit $i \in M$. $T$ is defined as: $\text{T-measure} = \frac{\sum_{i \in M} |v(i)|}{X'}$. Furthermore, we use an F-measure-like combination of S and T called S♯ as the primary metric for ranking participating systems at 1CLICK-2[5] [5].

The general flow of the 1CLICK-2 task was as follows:

1. Organisers sampled queries from a mobile query log, and manually constructed iUnits for each query;
2. Task participants produced an $X$-string for each query automatically, and returned the results to the organisers;
3. Assessors used the *iUnit matching interface* [6] to identify iUnits within the $X$-string and to record their positions[6];
4. Organisers added some new iUnits to the gold standard data based on feedback from the assessors;
5. Weighted recall, S, T and S♯ were computed for each $X$-string and these were averaged across the query set for each submitted system.

We are currently exploring the possibility of semi-automating the iUnit extraction and matching processes by following the methodology used in the 1CLICK-2 *English* subtask [2]. Even though these attempts turned out to be successful to some extent, our current approach of manually extracting and matching iUnits is a necessary step to understand the problems of information access evaluation that goes beyond the evaluation of ranked document lists.

[5]$S\sharp = \frac{(1+\beta^2)TS}{\beta^2 T + S}$, $\beta = 10$

[6]Every $X$-string was assessed independently by two assessors so that two sets of matched iUnits were obtained. In this study, we define $M$ (See Eq. 1) as the *intersection* of these two sets. For each matched iUnit, we use the smaller of the two offset values obtained.

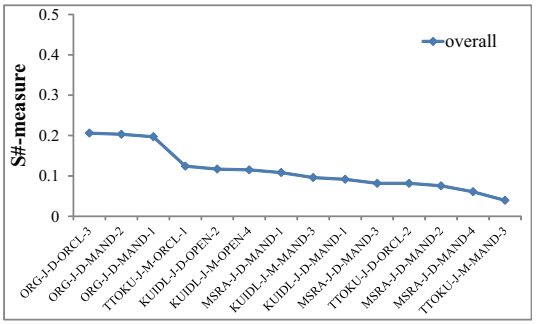


**Figure 2: Mean S♯-measure performances ($L = 500$) at 1CLICK-2. The $x$ axis represents runs sorted by Mean S♯ with the intersection iUnit match data.**

# 3. OFFICIAL RESULTS AND DISCUSSIONS

Figure 2 shows the official mean S♯-measure performances ($L = 500$) of runs submitted to the 1CLICK-2 Japanese subtask. The $x$ axis represents run names sorted by mean S♯. Figure 3 breaks down Figure 2 by showing similar results per query type. For the four CELEBRITY query types, we also show graphs for specialised and non-specialised queries separately. Hereafter, we shall discuss statistical significance based on a randomised version of *Tukey's Honestly Significant Differences (HSD) test* [1] at $\alpha = 0.05$.

## 3.1 Performances of Baselines

It can be observed that the three "ORG" runs are the overall top performers in Figure 2. These are actually simple baseline runs submitted by the organisers' team: ORG-J-D-MAND-1 is a DESKTOP mandatory run that outputs a concatenation of search engine snippets from the baseline search results; ORG-J-D-MAND-2 is a DESKTOP mandatory run that outputs the first sentences of a top-ranked Wikipedia article found in the baseline search results; ORG-J-D-ORCL-3 is similar to ORGs-J-D-MAND-1 but uses the sources of iUnits instead of the search results (an oracle run). These three runs significantly outperform the other runs, and are significantly indistinguishable from one another. Moreover, Figure 3 shows that these baseline runs outperform all participating runs with the four CELEBRITY query types as well as DEFINITION, while they are not as effective for FACILITY, GEO and QA. Furthermore, the graphs for CELEBRITY (where the runs have been sorted by the mean S♯ over *all* CELEBRITY queries) reveals that while the baseline runs are effective for the non-specialised queries, they are not necessarily so for the specialised ones. Recall that specialised CELEBRITY queries seek specific information about a celebrity such as "michael jackson *death*." Also, note that FACILITY, GEO and QA queries also seek specific information, e.g. contact information of restaurants, sentences that directly answer the natural language questions, and so on.

In summary, the above results suggest that, while simple snippet-based and Wikipedia-based approaches are effective for "lookup" type queries (e.g. celebrity names and definitions), more sophisticated techniques are required to satisfy the user for other query types (e.g. queries that look for specific information).

## 3.2 Estimating Performance Upperbounds

We now address the following question: *Given the 1CLICK evaluation framework, what is the performance upperbound?*. To this end, we hired four subjects and asked them to manually create an

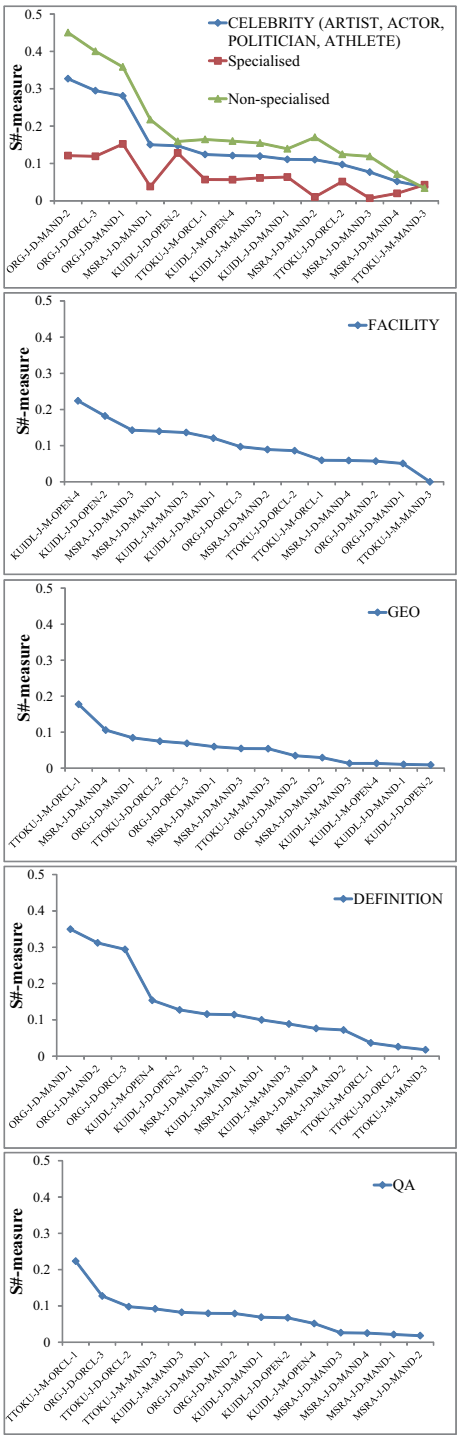


**Figure 3: Mean S♯-measure performances ($L = 500$) for each query type at 1CLICK-2. The $x$ axis represents runs sorted by Mean S♯ with the intersection iUnit match data.**

$X$-string for 73 queries from the official query set[7]. They were allowed to use any resources to formulate the $X$-strings: these four MANUAL runs are therefore *open* runs, and were manually assessed together with the submitted runs.

[7]We could not obtain manual $X$-string for the remaining 27 queries as the query set had not been finalised at that time.

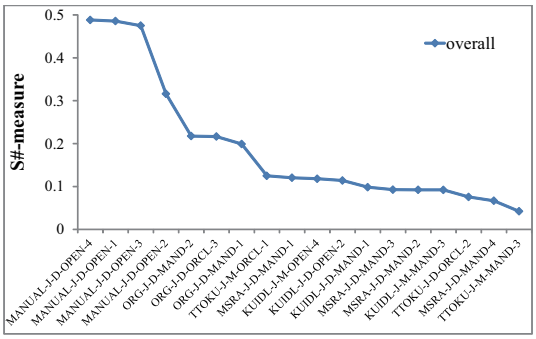


**Figure 4: Comparison between MANUAL and submitted runs in terms of mean S♯-measure ($L = 500$) over 73 queries. The $x$ axis represents runs sorted by Mean S♯ with the intersection iUnit match data.**

Figure 4 shows the mean S♯-measure ($L = 500$) over the 73 queries for all runs including the MANUAL ones. It can be observed that three of the four MANUAL runs far outperform the submitted automatic runs: these three runs are statistically significantly better than the other runs, and are statistically indistinguishable from one another. These results suggest that there are a lot of challenges for advancing the state-of-the-art of 1CLICK systems: a highly effective 1CLICK system needs to (a) find the right documents; (b) extract the right pieces from information from the documents; and (c) synthesise the extracted information to form an understandable text. It should also be noted that S♯ does not directly take into account the *readability* of text: in fact, all of the 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.

In summary, the comparison with the MANUAL runs shows that our task setting is challenging, and that there is a lot of room for improvement for the automatic 1CLICK systems. We hope that the future rounds of the 1CLICK task will help close the performance gap.

## 3.3 Robustness of Evaluation Metrics

Evaluating 1CLICK systems requires much manpower: for the 1CLICK-2 task, the organisers manually extracted over 6,000 iUnits for the 100 queries in advance, and further added some new iUnits based on assessors' feedback. In this section we address the following questions: *Do we need to try to find iUnits for a query exhaustively? What happens to the system ranking if the sets of iUnits were substantially incomplete?* To this end, we follow a practice from document retrieval evaluation for examining the effect of incomplete relevance assessments (e.g. [4]): we randomly downsample from the official sets of iUnits and examine the changes in the system ranking in terms of Kendall's tau rank correlation.

Figure 5 shows the effect of downsampling the iUnits on the system ranking for three of our evaluation metrics. It can be observed, for example, that even if we only have 50% samples of the official iUnit sets, the Kendall's tau between the original system ranking and the new ranking is around 0.90 (about seven pairs of runs swapped) or higher for both S♯ and weighted recall. Even with 10% samples, the tau is above 0.80. These results suggest that our evaluation framework is fairly robust to the incompleteness of iUnits.

Since even randomly downsampled iUnits yield relatively reliable evaluation results, exploring (semi)automatic approaches to iUnit extraction seems worthwhile. As we have mentioned earlier,

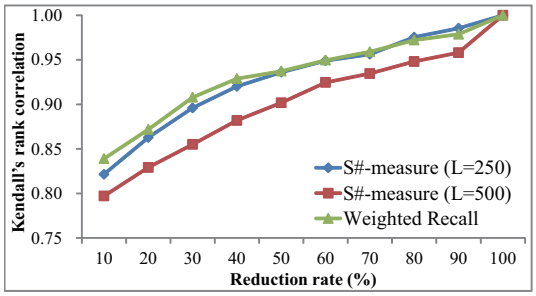


**Figure 5: Reduction rate ($x$ axis) vs. Kendall's rank correlation to the ranking based on a full iUnit set ($y$ axis).**

we are actually investigating how iUnit extraction and matching can be semi-automated.

# 4. FUTURE DIRECTIONS

One Click Access is an ambitious task that aims to satisfy the user immediately after he clicks the search button. Our analyses with the official NTCIR-10 1CLICK-2 results showed that: (1) Simple baseline methods that leverage search engine snippets or Wikipedia are effective for "lookup" type queries but not necessarily for other query types; (2) There is still a substantial gap between manual and automatic runs; and (3) Our evaluation metrics are relatively robust to the incompleteness of iUnits.

Our future work includes (a) Investigating the effect of removing the "additional" iUnits (ones that were added after the run submissions) and of "dependent" iUnits (those that are entailed by other iUnits); (b) Semi-automating the iUnit extraction and matching processes while ensuring their reliability; (c) User studies 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).

# 5. 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.

This work has been supported in part by the project: Grants-in-Aid for Scientific Research (No. 24240013) from MEXT of Japan.

# 6. REFERENCES

[1] B. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS*, 30(1):4:1–4:34, 2012.

[2] M. Ekstrand-Abueg, V. Pavlu, M. P. Kato, T. Sakai, T. Yamamoto, and M. Iwata. Exploring semi-automatic nugget extraction for japanese one click access evaluation. In *Proc. of ACM SIGIR 2013*, to appear.

[3] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and PC internet search. In *Proc. of ACM SIGIR 2009*, pages 43–50, 2009.

[4] T. Sakai and N. Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, 11(5):447–470, 2008.

[5] T. Sakai and M. P. Kato. One click one revisited: Enhancing evaluation based on information units. In *Proc. of AIRS 2012 (LNCS 7675)*, pages 39–51, 2012.

[6] T. Sakai, M. P. Kato, and Y.-I. Song. Click the search button and be happy: Evaluating direct and immediate information access. In *Proc. of ACM CIKM 2011*, pages 621–630, 2011.

[7] T. Sakai, M. P. Kato, and Y.-I. Song. Overview of NTCIR-9 1CLICK. In *Proc. of NTCIR-9*, pages 180–201, 2011.